2018

A High-Throughput Approach to Uncover Novel Roles of APOBEC2, a Functional Orphan of the AID/APOBEC Family

Linda Molla

A HIGH-THROUGHPUT APPROACH TO UNCOVER NOVEL ROLES OF APOBEC2, A FUNCTIONAL ORPHAN OF THE AID/APOBEC FAMILY

A Thesis Presented to the Faculty of

The Rockefeller University

in Partial Fulfillment of the Requirements for

the degree of Doctor of Philosophy

by

Linda Molla

June 2018

A HIGH-THROUGHPUT APPROACH TO UNCOVER NOVEL ROLES OF APOBEC2, A FUNCTIONAL ORPHAN OF THE AID/APOBEC FAMILY

Linda Molla, Ph.D.

The Rockefeller University 2018

APOBEC2 is a member of the AID/APOBEC cytidine deaminase family of proteins. Unlike most of AID/APOBEC, however, APOBEC2's function remains elusive. Previous research has implicated APOBEC2 in diverse organisms and cellular processes such as muscle biology (in *Mus musculus*), regeneration (in *Danio rerio*), and development (in *Xenopus laevis*). APOBEC2 has also been implicated in cancer. However the enzymatic activity, substrate or physiological target(s) of APOBEC2 are unknown. For this thesis, I have combined Next Generation Sequencing (NGS) techniques with state-of-the-art molecular biology to determine the physiological targets of APOBEC2. Using a cell culture muscle differentiation system, and RNA sequencing (RNA-Seq) by polyA capture, I demonstrated that unlike the AID/APOBEC family member APOBEC1, APOBEC2 is not an RNA editor. Using the same system combined with enhanced Reduced Representation Bisulfite Sequencing (eRRBS) analyses I showed that, unlike the AID/APOBEC family member AID, APOBEC2 does not act as a 5-methyl-C deaminase. Finally, using a combination of biochemical, Chromatin Immunoprecipitation Sequencing (ChiP-Seq) and polyA RNA-Seq analyses I show that APOBEC2 is a (negative) regulator of gene expression (at least in muscle cells) and binds chromatin

directly to inhibit transcription of genes involved in muscle cell differentiation. While the precise mechanism behind this activity is still a matter of investigation, this role of APOBEC2 in inhibiting genes involved in cell cycle exit, might have implications for its role in in cancer.

ACKNOWLEDGMENTS

"It takes a village to make a scientist"

I would like to thank my advisor Nina Papavasiliou for giving me the opportunity to be part of the lab. Thank you for your enthusiasm for science, encouragement, and support. Thank you, for generously spending a lot of time 'on call', especially during the last months of graduate school. You have taught me a lot about being persistent along the setbacks and triumphs of the scientific process. The mixture of your optimism and my overly skepticism was a good match to move this project forward.

I thank the members of my thesis committee: Mary Goll, Sanford Simon and Agata Smogorzewska, for their guidance and generously dedicating some of their precious time providing feedback, thoughtful comments, questions, and advice through the years. Special thanks, to Mary Baylies, for accepting to serve as the external examiner in my thesis defense, and for her thoughtful comments. I would like to thank George Cross for hosting me in his lab space during this last year at Rockefeller University.

I am indebted to my undergraduate advisor Diana Bratu at Hunter College for her support through the years and inspiring me with her strength, tenacity and good-hearted nature. I would also like to thank Shirley Raps, at Hunter College who was the first one to see scientific potential in me and encouraged me to keep going on this path. I am also thankful to Gerardo Morfini and Scott Brady for their support early on in this scientific journey and for allowing me to have a fantastic first science experience in their labs at the Marine Biological Laboratories (MBL). Many thanks to Juleen Zierath and Anna Krook for the opportunity of a short research exchange at the Karolinska Institutet (KI) and to so many people at KI who allowed me to 'pester' them with questions and helped me learn more about the model system of my research.

I am grateful for many people at the Rockefeller University and the German Cancer Research Center (DKFZ) who have helped me along the way: Diego Mourao-Sa, for the many scientific discussions, advice and thoughtful questions that helped me tremendously, and also for reviewing this thesis; Dewi Harjanto for being patient and thorough in answering many questions on bioinformatics to someone with zero coding background. I have learned a lot from you; Eric Fritz who gave me the lowdown on the lab, inspired me to be more strategic and structured in my experiments; Thomas Carroll, for the expertise and instrumental input in bioinformatics analyses; Pete Stavropoulos, for the feedback with the biochemistry related experiments and together with Nicholas Economos for purifying antibodies important for my work; Danae Schulz for her guidance in the ChIP-seq experiment and along with Monica Mugnier for their thoughtful questions and comments in lab meetings and practice talks; Erik Debler and Philipp Schmiege for their help in purifying protein important in future work; Sandra Ruf and Jose (Paulo) Lorenzo for being a pleasure to work with as my virtual labmates at DKFZ and for carrying the 'APOBEC2 torch' forward.

I am especially grateful to Violeta Rayon, Maryam Zaringhalam and Jason Pinger who luckily joined the lab alongside me, for their scientific discussion, friendship and much needed moral support during the many challenges. I would also like to thank Dimitrios Garyfallos for his friendship and company in the lab especially when I was working late at nights. Special thanks, to Deanna Belsky for her friendship and inspiring me with her positivity. Thanks to many other friends in New York, who I have met through the years for their support and for making the graduate school years fun and unforgettable.

Importantly I would like to thank all the people and groups who have also taught me important lessons beyond bench science: The colleagues and friends of INet NYC, for being such great team players, for their support and friendship; The Science Alliance Leadership Training (SALT) of the New York Academy of Sciences; the Fundamentals of Bioscience Industry Program of the Center of Biotechnology at Stony Brook University.

I am thankful for the financial and administrative support I have received through the David Rockefeller Graduate Program and the Nicholson Exchange fellowship. I am grateful to the staff in the Dean's Office, and especially Andrea Morris for ongoing career guidance and support.

I would have never completed this journey without the love, support, and encouragement from my family. I thank everyone in the Molla, Babali, and Blanco-Melo family: I thank my parents for their love, their trust in me, and for instilling in me the importance of hard work and integrity; I thank my sisters Olta and Jona for always being there for me, especially for being my support system in the challenging first years when we first moved to NY. *Faleminderit mami, babi, Olta, Jona! ;* My aunt Jeta and uncle Sulo for hosting my sisters and I, helping us get started in NY; Vinela and Ledion for their unconditional support and advice through the years; Elka for her warmth and affability. *Faleminderit nga zemra!*; Hector and Aurora, for embracing me with your support and showing me the beautiful Mexico. *Gracias por su apoyo!*

Last but not least I would like to thank my husband Daniel Blanco-Melo. I am so happy this graduate school journey brought us together, thank you for your love, support, patience (especially when 10 more minutes of work turned into 2 hours), listening at my long rants during the tougher times and for always finding a way to make me laugh.
Te quiero mucho Daniel!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. Introduction

1.1 Beyond the central dogma of Molecular Biology

DNA was determined to be the chemical basis of heredity and life processes through the work by Avery et al. in 1944 (Avery et al., 1944) and subsequently confirmed in 1952 by Hershey and Chase (Hershey and Chase, 1952). The collective efforts of many scientists such as Miescher (Miescher, 1871), Avery et al. (Avery et al., 1944), Chargaff et al. (Chargaff et al., 1951), Wilkins et al. (Wilkins et al., 1953), Franklin and Gosling (Franklin and Gosling, 1953) and many others behind the scenes (Judson, 1996; Maddox, 2003) produced many pieces of the puzzle that lead Watson and Crick to successfully put them together into the discovery of the double helical structure of the DNA in 1953 (Watson and Crick, 1953). These important scientific advances combined with knowledge about mRNAs and protein synthesis lead Crick to propose the theory of the "Central Dogma of Molecular Biology" in 1958 (Crick, 1958) and later in 1970 (Crick, 1970). The general concept was that information flows from DNA to RNA to protein, which determines the cellular and organismal phenotype.

The "Central Dogma of Molecular Biology" is a significant biological idea that led biologists to an informatics perspective on living organisms. Life itself is coded (DNA>RNA>Protein) and the code is heritable. To this day the central dogma informs our understanding of the role of heredity in deciphering biological functions, disease, and the process of evolutionary change. Nonetheless as it is the case for most breakthroughs as they open the door to many discoveries, they also provide many unanswered questions. For example some of these questions are: if genomic information is the basis of

biological function how can we explain the pronounced diversity among genomically identical cells in a multicellular organism? Given that identical twins have the same genetic make up how can we explain that they do not behave as "clones" of each other? Given the limited amount of genomic DNA, how can we explain our ability to make such a diverse number of antibodies responding to a vast number of foreign pathogens? Today, about 60 years after the conception of the "Central Dogma of Molecular Biology", the picture has gotten more complicated, and we have discovered many mechanisms that lead to biological diversity such as epigenetics, regulatory noncoding RNAs, alternative splicing, RNA modifications (epitranscriptomics), protein post-translational modifications and DNA/ RNA editing (Figure1.1). Therefore the central dogma of molecular biology has 'evolved'. Below I will give an overview of polynucleotide cytidine deamination as an example of how biological diversity that is not encoded in the genome can be generated.

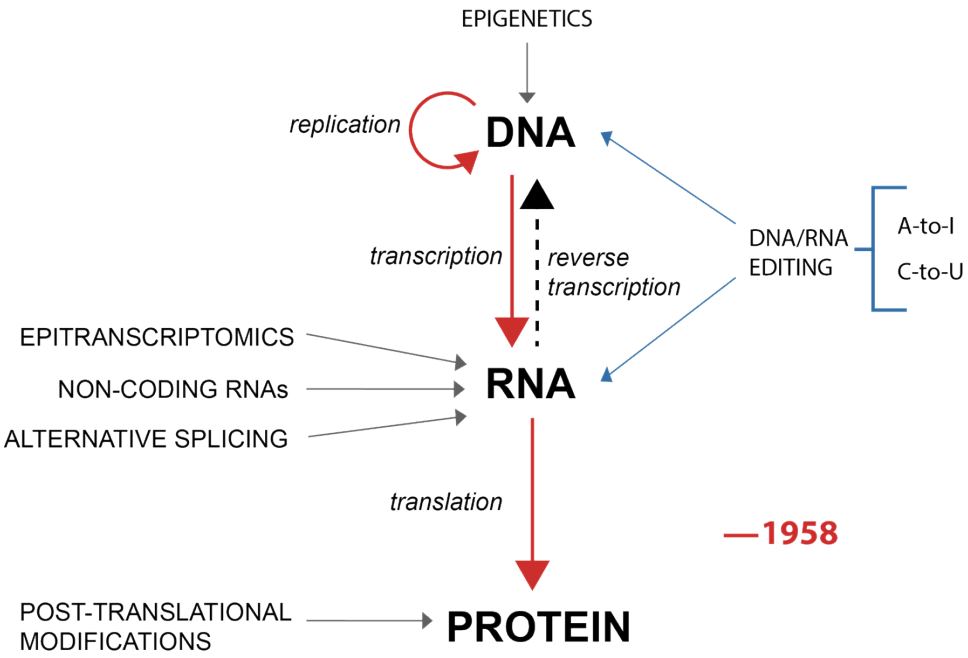


Figure 1.1 The expansion of the central dogma of molecular biology

In red are shown processes of genetic information flow that were known as of 1958 based on the "Central Dogma of Molecular biology". Current known types of processing in DNA, RNA in higher eukaryotes that cannot be deduced from what is coded in the genome (in gray), including DNA/RNA editing (in blue)

1.2 Polynucleotide Cytidine Deaminases: The AID/APOBEC family of enzymes

Biological information is encoded in DNA and RNA. Normally organisms tend to avoid mutations during DNA replication and transcription through DNA repair mechanisms. This allows the information encoded in DNA to be faithfully inherited to the next generation of cells and organisms. Problems in controlling mutations could lead to severe consequences as DNA and RNA are decoded. The AID (activation induced deaminase)/APOBEC (apolipoprotein B mRNA editing catalytic polypeptide-like) family of proteins was a surprising discovery that uncovered an unexpected level of gene expression control. They induce mutations on single stranded (ss) DNA or RNA by

catalyzing the removal of the amino group from a cytidine base in the context of polynucleotides, resulting in specific gene products being expressed or restricted (through mutations). Their enzymatic activity results in a cytidine (C) to uridine (U) transition (Figure 1.2). Functional studies have demonstrated how they are cleverly used by the cell to induce nucleic acid changes in polynucleotides in a targeted way and with a specific biological goal, but they can also be considered a threat to the stability of the genome or transcriptome when not properly regulated. The members of the family are very closely related to one another based on homology and conservation of the enzymatic domains, but they have different tissue-specific expression, substrates, and biological functions.

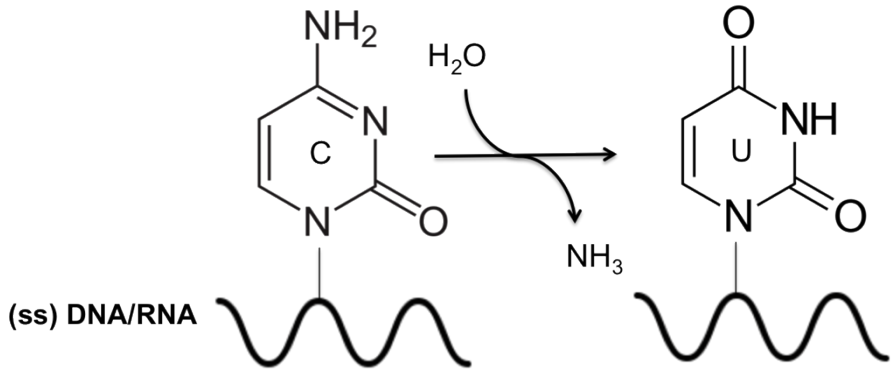


Figure 1.2 Deamination of (deoxy)cytidine by AID/APOBEC

1.2.1 The startling discovery of the AID/APOBEC family

The story of the AID/APOBEC family of enzymes began in 1987 with the discovery that the transcript of apolipoprotein B (*apoB*) mRNA contained a C to U base modification that was not coded in the corresponding genomic DNA (Chen et al., 1987; Powell et al., 1987). This posttranscriptional modification of the *apoB* mRNA at nucleotide position 6666 changes the genomically encoded glutamine codon (CAA) into a translational stop codon (UAA) that leads to an early translation termination. The editing of *apoB* mRNA

explained the formation of the truncated isoform. Hence two proteins apoB-100 (the full length form) and apoB-48 (the short form) can be generated from the same primary transcript.

ApoB is a structural component of lipoproteins and is important for carrying lipid and cholesterol in circulation from the digestive tract and liver to the rest of the tissues. At the time ApoB was known to be a protein related to the risk for atherosclerosis (Breslow, 1988). The potential that the editing of *apoB* mRNA could be related to atherogenic disease lead to a lot of interest. The cDNA encoding a protein that is essential for the deamination of C6666 was discovered in 1993 (Teng et al., 1993) and was named in 1995 as apolipoprotein B editing catalytic subunit 1 (APOBEC1), with number 1 indicating the possibility that other related genes may exist (Davidson et al., 1995). Indeed that was the case, many other members of the APOBEC family of enzymes have been discovered since then.

1.2.2 The common evolutionary origin of the AID/APOBEC family

In humans the AID/APOBEC family of enzymes includes 11 members of primary and alternatively spliced variants: Activation-induced deaminase (AID), APOBEC2, APOBEC3 (A-D, F–H), and APOBEC4 proteins. AID/APOBEC are apart of large superfamily of zinc-dependent deaminases that act on free cytidine, cytosine or deoxycytidine triphosphate (dCTP) and are involved in the metabolism of purines and pyrimidines, as well as deaminases that act on adenosine in the context of RNAs. Based on phylogenetic analysis, gene organization and catalytic domain the AID/APOBEC family probably originated from the Tad (tRNA adenosine deaminase)/ADAT2

(adenosine deaminase, tRNA- Specific 2), which edit adenosine to inosine at the anticodon of various tRNAs (Conticello, 2008, 2012). The AID/APOBEC family originated alongside the appearance of the vertebrate lineage, where deaminases are suggested to have split into two primary branches in the common ancestor of jawed and jawless vertebrates: APOBEC4-like and the AID-like branch (Figure 1.3). The APOBEC4-like branch thus arose independently of AID and is thought to be a result of a retrotransposition event due to the absence of introns in the deaminase-like region of APOBEC4 (Iyer et al., 2011), (Conticello, 2008). The other APOBECs are suggested to have originated from ancestral AID-like genes in jawless fish (PmCDA1 and PmCDA2, *P.marinus* cytosine deaminase). The AID-like branch further diversified giving rise to AID itself and APOBEC2, the most ancient members of the family (in bony and cartilaginous fish) about 500 million years ago. APOBEC2 was suggested to have arisen as a result of a retrotranspositional event given its different gene structure from that of other members, where APOBEC2 has an N-terminal exon with no similarity to any other known protein and the absence of introns in the deaminase-like region (Conticello et al., 2007a). APOBEC2 has been under purifying selection through evolution suggesting functional importance. The duplication of the AID locus about 300-400 years ago lead to the evolution of APOBEC1 in mammals. APOBEC3s arose in placental from duplication of AID after the divergence of the placentals from the marsupials about 170 million years ago, being under strong positive selection which lead to its rapid expansion in primates which have seven APOBEC3s (Conticello et al., 2007a).

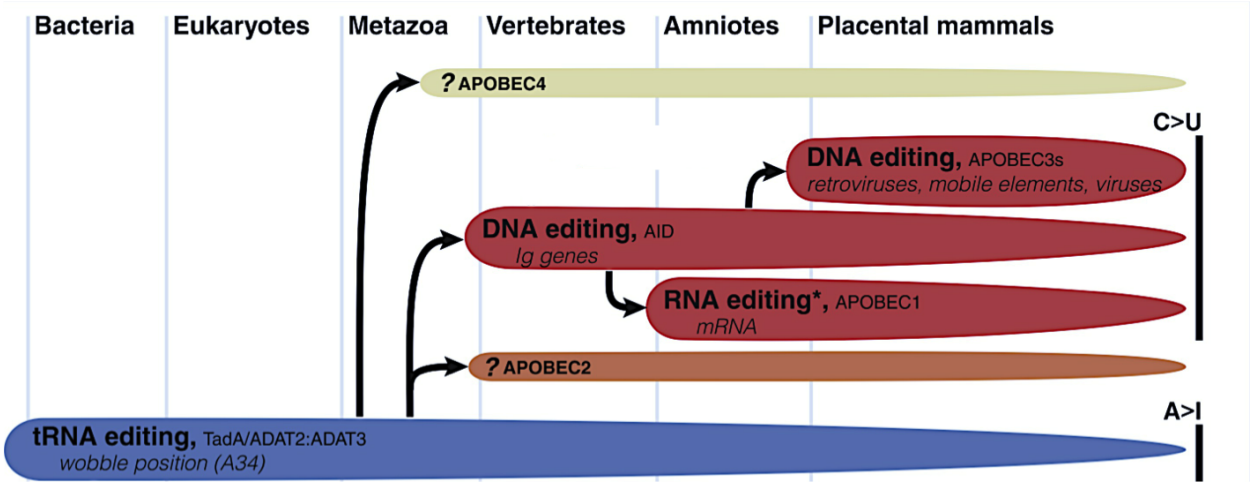


Figure 1.3 The evolutionary path of the AID/APOBEC

The plot indicates the phylogenetic relations among the various AID/APOBECs (and their common ancestral tRNA editor enzymes) plotted against the time of their emergence. The enzymes are shown beside the known function with the physiologic targets listed beneath. The branches of the zinc-dependent deaminases are colored based on type of deamination: A>I deamination (blue); C>U deamination (red). Branches of unknown catalytic activity are indicated by the name of the protein group prefixed by a question mark. APOBEC1 RNA editing is marked with an asterisk to indicate that APOBEC1 acts on DNA as well, but this activity has not yet been linked to a physiological role. *Figure adapted from Conticello, 2012*

1.2.3 AID/APOBECs are similar by structure homology

Insights about the structure of many of the members of AID/APOBEC have been gained using NMR spectroscopy and/or X-ray crystallography of full length, truncated or mutated versions of the proteins. All AID/APOBECs share the same sequential arrangement of secondary structural features (α-helices and β-sheets) forming a cytidine deaminase CDA domain (Figure 1.4a). It is thought that the differences in length, composition, location of these secondary structural features and the loops connecting them may determine the variability in the biological functions, substrate selection, localization, oligomerization and regulation of deamination activity. Moreover the catalytic function and substrate binding of the AID/APOBECs could be regulated through the binding of other protein co-factors, other RNAs, *cis/trans* regulatory DNA elements or oligomerization (Salter et al., 2016).

All APOBECs have at least one CDA domain, which contains a highly conserved zinc-dependent deaminase (ZDD) sequence motif (H-X-E-$X_{[24-36]}$--P-C-$X_{[2-4]}$-C, where X= any amino acid) responsible for catalyzing the deamination reaction. This motif forms the catalytic pocket where a Zn^{2+} metal ion is coordinated by 3 amino acids (two cysteines (C) and one histidine (H)) and is bound to a water molecule that gets activated resulting in a zinc hydroxide group ($ZnOH^-$). The mechanism of deamination is thought to be the same as the one for bacterial cytidine deaminase (Betts et al., 1994). Cytidine/deoxycytidine binds within the catalytic pocket and gets deaminated through a nucleophilic attack on carbon 4 of cytidine (C-4) by $ZnOH^-$. The conserved glutamate (E) acts a proton (H+) donor to the leaving amino group (NH_2) of the

cytidine/deoxycytidine. Overall the loss of ammonia (NH_3) leads to a net conversion of cytosine to uracil (Figure 1.4b).

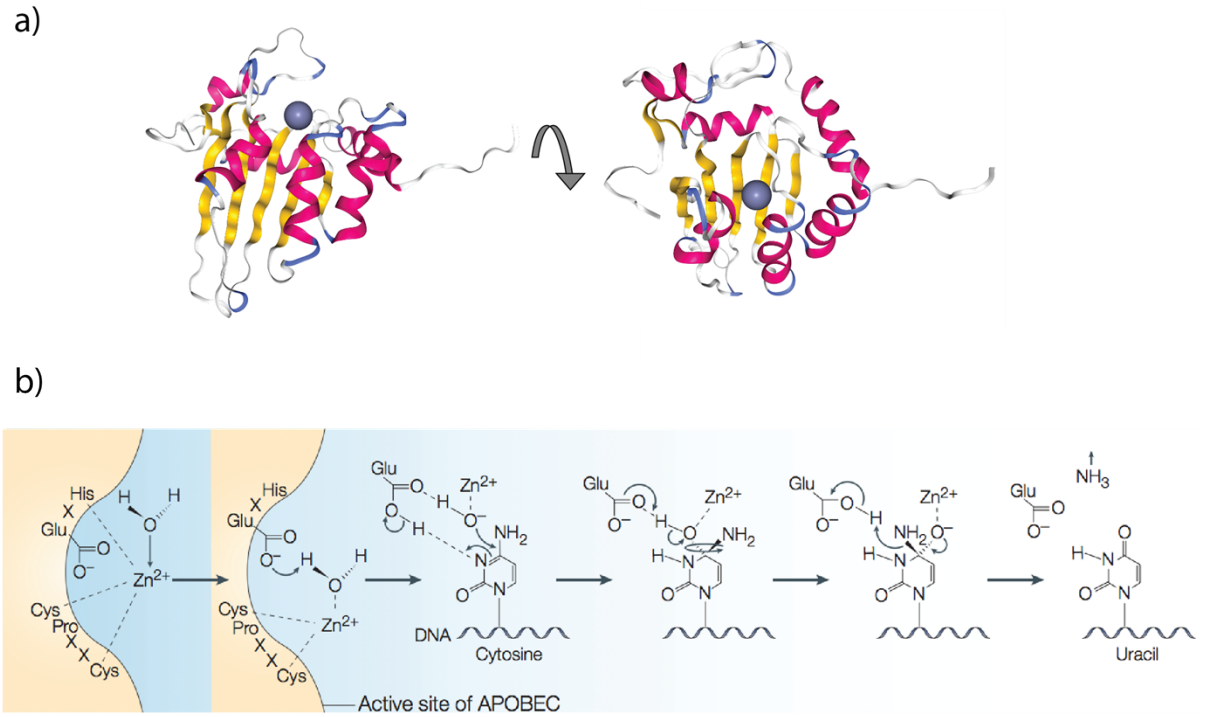


Figure 1.4 Diagram of the AID/APOBEC cytidine deaminase (CDA) fold

a) An example of the conserved CDA fold. Cartoon representation of the solution NMR structure of APOBEC2 (Δ40) (PDB 2RPZ). The structures are colored based on the secondary structure. The Zinc ion is shown as a sphere in purple

b) The proposed mechanism of cytidine deamination. *Image from Harris and Liddament, 2004*

1.2.4 AID/APOBECs are very diverse in their biological function and physiological substrates

Even though all of the AID/APOBECs have at least one conserved ZDD motif and common CDA fold, they are very diverse in terms of the tissue specific expression and biological functions that they perform. Here I summarize some of the physiological

substrates and biological roles of the AID/APOBEC members: AID, APOBEC1 and APOBEC3 family.

Role in RNA editing

RNA editing is defined as alteration of an RNA sequence through nucleotide insertion, deletion or modification mechanisms that leads to variations in sequence compared to what is encoded in the genome (Wedekind et al., 2003).

APOBEC1, as already mentioned above is an RNA editing enzyme that edits mRNAs through base modification. It was the first of the AID/APOBEC proteins shown to have a biological role, enzymatic activity, and physiological substrate. APOBEC1 was shown to catalyze a C to U change in the *apoB* mRNA, creating a premature stop codon and leading to the generation of a shorter transcript in the small intestine, a novel mRNA important for lipid metabolism (Teng et al., 1993). Thus here, APOBEC1 operates by creating mRNAs with differential function from a single locus. Transcriptome wide analyses in mice have also identified several APOBEC1-dependent editing sites mainly in the 3' untranslated regions (3'UTR) of various genes in small intestine and/or liver (Blanc et al., 2014; Rosenberg et al., 2011). A portion of these edited sites alter mRNA levels and protein translation efficiency (Blanc et al., 2014) in the specific tissues but the function of most of the sites is not known. Additional studies using transcriptome analysis in mice have shown that there are several transcripts that are edited by APOBEC1 (mainly 3'UTRs) beyond the liver or small intestine, in macrophages of the immune system (Harjanto et al., 2016; Rayon-Estrada et al., 2017) and also in microglia of the central nervous system (CNS) (Cole et al., 2017). Some of these edited sites correlate with gene expression changes through the modulation of translation (but not

mRNA stability). Lack of APOBEC1 mediated editing in macrophages has functional consequences in phagocytosis and migration (Rayon-Estrada et al., 2017); while lack of editing in microglia of the CNS leads to overall microglia functional dysregulation, which lead to behavioral and motor deficiencies (Cole et al., 2017). Two co-factors have been shown to be important in APOBEC1 directed editing: A1 complementation factor (ACF or A1CF) was shown to direct APOBEC1 to an 11nt mooring sequence (Mehta et al., 2000) and RNA binding motif protein 47(RBM47) (Fossat et al., 2014). The ancestral function of APOBEC1 involves editing of the genomic DNA, since when expressed in *E. coli* it results in mutations in genomic DNA and in *ex vivo* assays it restricts retroelements (Harris et al., 2002; Ikeda et al., 2011; Severi et al., 2011). It is not known whether under physiological conditions, APOBEC1 can also act on DNA in mammals.

Role in DNA editing

AID/APOBEC can also alter a DNA sequence by enzymatically deaminating deoxytcytidine, converting it to deoxyuridine (here termed as DNA editing). There are two members that can mediate DNA editing physiologically: AID and APOBEC3 proteins.

AID is highly expressed in activated B cells and is essential in adaptive immunity by mediating antibody affinity maturation and diversification (Muramatsu et al., 2000). The initial hypothesis was that AID acts on RNA, given its sequence similarity with APOBEC1, which was the only AID/APOBEC member characterized at the time (Muramatsu et al., 1999). Further research (*reviewed in* (Delker et al., 2009; Di Noia and Neuberger, 2007)), showed that AID edits the genomic immunoglobulin (Ig) loci and triggers class-switch recombination (CSR), somatic hypermutation (SHM) and

sometimes gene conversion (GCV). These processes that are important for Ig gene diversification, where during CSR, AID leads to double-strand breaks and genomic recombination, while during SHM, AID leads to point mutations. AID can target genes beyond the Ig loci in B cells when overexpressed (Robbiani et al., 2008), but it is unknown whether there is any other physiological targets in B cells. Additionally AID has been shown to inhibit retrotransposition through DNA deamination-independent mechanisms (MacDuff et al., 2009) and also targets viruses (Liang et al., 2013), but it is unknown whether these are physiological roles.

APOBEC3 proteins are known for their role in innate immunity through the hypermutation and restriction of a wide range of viruses and retroelements *reviewed in* (Harris and Liddament, 2004). The arms race between APOBEC3 and the viruses that it targets has led to the expansion of APOBEC3 from single copy in mouse to seven paralogs in primates (A-D, F-H) (Conticello, 2008). The importance of APOBEC3 in viral defense was realized with the identification of APOBEC3G as the protein involved in the restriction of HIV that is antagonized by the HIV protein Vif (Sheehy et al., 2002). This is done through hypermutation (through APOBEC3G DNA editing) of the nascent viral DNA genome during reverse transcription (Harris et al., 2003; Lecossier et al., 2003; Mangeat et al., 2003; Zhang et al., 2003). Since this discovery APOBEC3G, all the paralogs have been shown to be important in innate defense mechanism against diverse viruses, endogenous transposable elements and foreign DNA *reviewed in* (Harris and Dudley, 2015; Knisbacher et al., 2016).

Role in epigenetics

Besides the very well characterized roles in RNA editing and DNA editing AID/APOBEC family of enzymes have also been linked to epigenetic changes through active DNA demethylation (5mC removal). 5mC modification of gene elements is widely shown to correlate with gene expression changes and is crucial in mammalian development (Smith and Meissner, 2013). This modification is maintained during cell division through DNA methyl transferase (Dnmt) enzymes. There are two mechanisms proposed for DNA demethylation: passive and active. Passive demethylation refers to the dilution of DNA methylation during cell division in the absence of maintenance methylation, while active methylation involves removal of methylation marks through an active process in the absence of cell division. There are multiple studies supportive of a potential role of AID in active DNA demethylation, yet the genetic evidence for a such a claim has been challenged (*reviewed in* (Bochtler et al., 2017; Ramiro and Barreto, 2015)). Nonetheless there is strong evidence that AID might be important in affecting methylation status of target genes in B cells during activation in the mouse (Dominguez and Shaknovich, 2014). It is unlikely that AID has any genome wide effects in active DNA methylation during development as it has been previously proposed (Bhutani et al., 2010). However it is remains possible that AID is important in gene specific methylation, but the mechanisms of how this happens are unknown.

The physiological functions of APOBEC4, which is expressed in testes (Rogozin et al., 2005) is still unknown and APOBEC4 does not deaminate DNA in mutation assays in *E. coli* and yeast (Lada et al., 2011). There is some recent evidence that it might act to enhance transcription of host promoters and endogenous LTR promoters (Marino et al.,

2016). Whether this is a physiological relevant role and dependent on ZDD motif remain to be elucidated.

Role in cancer

The ability of AID/APOBEC family of enzymes to induce RNA/DNA editing is a double-edged sword. While these enzymes are important in many biological functions as mentioned above, if not properly expressed, they can be mutagenic and lead to cancer progression. Mouse models where AID/APOBEC(s) are overexpressed exhibit cancer development in various tissues (Swanton et al., 2015). An important example is how the increase of the levels of AID correlates with c-Myc oncogenic translocation in B cell tumors (Robbiani et al., 2008; Takizawa et al., 2008). Interestingly, the severity of lymphomas recently has been linked to the epigenetic role of AID in B cells (Teater et al., 2018). Analysis of human cancer genome data has revealed that APOBEC signatures are evident in many tumor genomes (Roberts et al., 2013), in the form of strand-coordinated C-to-T hypermutation and there are direct links of the misregulation of APOBEC3A and APOBEC3B(Henderson and Fenton, 2015). APOBEC1 RNA editing activity has been linked to oncogenesis due to the hyperediting of a novel APOBEC1 target mRNA (Nat1), encoding a translational repressor in mouse models (Hersberger et al., 2003; Yamanaka et al., 1997). It also has been linked to peripheral nerve sheath tumors in humans due to abnormal editing of the tumor suppressor neurofibromatosis type 1 RNA (NF1) (Mukhopadhyay et al., 2002). Moreover lack of APOBEC1 has been shown to decrease the progression of intestinal tumors, suggesting a potential role of APOBEC1 RNA editing in cancer progression (Blanc et al., 2007). It is possible that the misregulated expression of APOBEC1 can also affect the DNA editing of the genome given

APOBEC1's known role in mutagenesis when overexpressed in *E.coli* (Harris et al., 2002).

In the last 30 years since the discovery of AID/APOBEC(s), *cytidine deamination* has emerged to be a prevailing driver of biological diversity in mammals. These proteins lead to changes not encoded in the genome through a range of processes: acting as RNA editors to create novel mRNAs and sequence changes in 3'UTRs of mRNA that affect translation stability; acting as DNA editors to create novel genes, and to restrict viruses and retrotransposons; and changing the DNA 5mC modification levels which leads to transcript abundance differences. Their ability to edit transposable elements and genomic DNA can lead to increase in beneficial/protective mutations, on one hand, and cancer on the other.

1.3 APOBEC2, the 'orphan' deaminase

APOBEC2 was discovered to be a member of the AID/APOBEC family based on sequence similarity (Anant et al., 2001; Liao et al., 1999). It is evolutionary well-conserved through the vertebrate lineages (similar to AID) and even in bony fish (where two copies of APOBEC2 are present, thus alleviating the purifying selection), both paralogues show little sequence divergence. Through evolution, the amino acid sequence of APOBEC2 is constrained potentially by the need to preserve its function (as opposed to APOBEC3 who has been under positive selection and whose rate of sequence evolution has diverged due to the arms race with the target sequence) (Conticello et al., 2005, 2007b). The structure of APOBEC2 has been elucidated and even though there are

disparities on whether APOBEC2 is a rod-shaped tetramer (Prochnow et al., 2007) or a monomer in solution (Krzysiak et al., 2012), the presence of a CDA fold characteristic of the cytidine deaminases is well supported (Figure 1.4a). A number of phenotypes have been reported in the absence of APOBEC2 suggesting that APOBEC2 has a biological function, however there has been no demonstration of its predicted cytidine deaminase enzymatic activity or potential physiological substrates. APOBEC2 is deemed an orphan deaminase because it shows no catalytic activity and binding toward single stranded DNA or RNA. This gap in knowledge of the specific substrate or enzymatic activity could explain why APOBEC2 is one of the least studied AID/APOBEC(s) (Figure 1.5). In the sections below I will review the current knowledge in the field about APOBEC2's biological role and activity on polynuclotides at the time when I started this project.

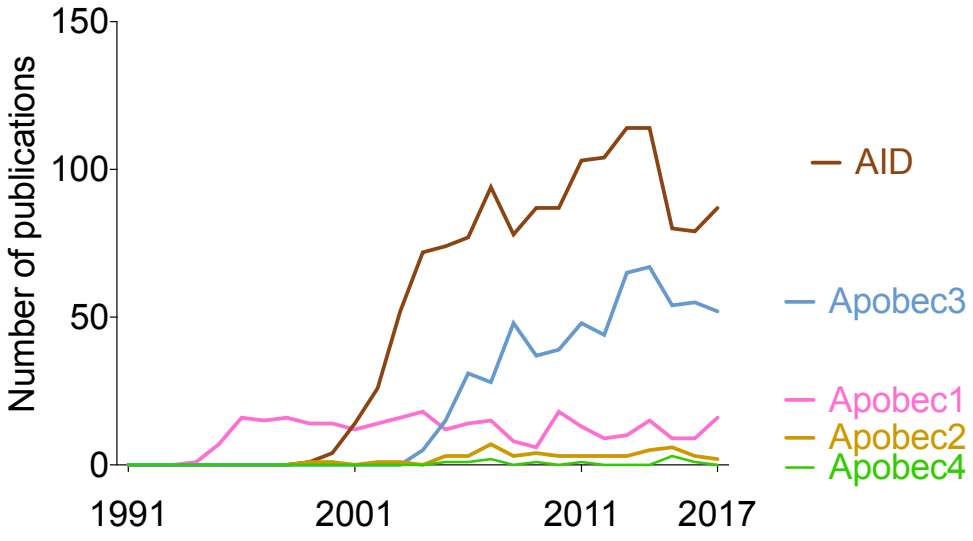


Figure 1.5 The number of pubmed publications on AID/APOBEC, 1991-2017
APOBEC2 and APOBEC4 are the least studied of the AID/APOBECs, mainly due to the lack of specific substrate or enzymatic activity linked to either protein.

1.3.1 No study has demonstrated enzymatic activity for APOBEC2

Traditionally two types of assays have been used to demonstrate that AID/APOBEC proteins are enzymatically active on DNA(Coker et al., 2006):

(A) The mutation assay tests if cytidine deaminases can act as DNA mutators when expressed in unnatural hosts. It involves overexpressing the protein of interest in unicellular organisms such as *Escherichia coli (E. coli)* plated in media containing a lethal drug that targets an essential protein. If the AID/APOBEC protein induces mutations that restore the function of the host's essential protein by changing the binding affinity of the drug, the cells survive. The number of drug resistant mutant cell colonies serves as a measure of the mutator activity of the enzyme. Cytosine deamination produces uracil, which is mostly removed via the uracil DNA glycosylase (UDG/*ung*)-dependent base excision repair pathway. Therefore assaying DNA deaminase activity in *E.coli* in the absence of *ung* increases the number of C-to-T mutations observed in the assay supporting the deaminase action on cytidine in DNA. Studies using this method has shown that APOBEC2 does not exhibit a mutator effect when expressed in *E. coli* (Harris et al., 2002) or in yeast (Lada et al., 2011). Absence of the mutation activity might mean that either (1) the cytidine deaminase does not act on the DNA used as a substrate, (2) the host environment is not ideal for appropriate folding of the cytidine deaminase, (3) cofactors or other proteins that might be necessary for the activity of the enzyme are lacking, (4) the proper substrate sequence specificity is not found in the reporter system, (5) the protein is not enzymatically active, and (6) a combination of some or all of the above.

(B) The biochemical DNA deamination assay tests if cytidine deaminases could induce

mutations in a synthetic single stranded DNA oligonucleotide that contains a deoxycytidine (or modified deoxycytidine). It involves incubating the artificial oligonucleotide substrate with the cytidine deaminase purified from *E. coli* (or other source) and then allowing the template to bind to its complimentary strand. The product is then treated with DNA glycosylases that will cleave the glycosidic bond if they find mismatches, yielding an abasic site. Cleavage of the abasic sites under alkaline conditions allows for specific detection of product after separation on a denaturing gel. The extent of deamination is a measure of the amount of the cleaved product. Previous studies showed no detectable cytidine deaminase activity for APOBEC2 (purified from *E.coli*) on cytidine, 5-methylcytidine (5mC) or 5-hydroxymethylcytidine (5hmC) using DNA oligonucleotides (Mikl et al., 2005; Nabel et al., 2012). Absence of the deamination activity could mean that: (1) APOBEC2 has not cytidine deaminase activity (2) APOBEC2 does not target DNA, (3) other cofactors are lacking, (4) sequence context used to test deamination is not appropriate, and (4) a combination of the above.

Other investigations into the biochemical properties of the protein (Sato et al., 2010) do not provide support for the idea that APOBEC2 acts on RNA. They suggest that mouse APOBEC2 present in muscle is unlikely to be associated with RNA because when treated with RNAse A its molecular weight in a Western blot analysis does not shift. Given that they do not use a crosslinking method, the results from this experiment are questionable since potential RNAs bound to APOBEC2 might have been degraded or lost during the handing of the extracts. They also show that APOBEC2 is not able to bind AU-rich or *apoB-100* transcript after UV-crosslinking, unlike APOBEC1. The latter could be due to inappropriate target tested.

All the experiments published looking for APOBEC2 activity use protein purified from *E. coli* or test its activity in a host cell that does not normally express APOBEC2. These unnatural environments might be lacking important required cofactors or physiological substrates. Additionally, such assays are non-stoichiometric and may not be representative of what happens physiologically even when they show activity. For example, the DNA mutator assay shows that APOBEC1 is capable of mutating DNA (Lada et al., 2011), yet no study so far has been able to show that this happens naturally in tissues where APOBEC1 is expressed at physiological levels.

In conclusion, published work on APOBEC2 indicates that the lack of a substrate for APOBEC2 makes it difficult to prove its enzymatic activity and elucidate its potential contributions in biology. Since APOBEC2 is a functional protein in many organisms, the identification of its specific physiological function is most likely to be achieved from the identification of its substrate. Thus, identifying the potential nucleic acid substrate(s) in cells or tissues that normally express APOBEC2 before recapitulating the activity of APOBEC2 biochemically might be instrumental in determining its enzymatic activity.

1.3.2 Many studies demonstrate that APOBEC2 has a biological function

APOBEC2 is widely expressed in cardiac and skeletal muscle in humans and mice (Liao et al., 1999) and is conserved in all jawed vertebrates including bony fish and mammals (Conticello, 2008). In mice APOBEC2 is more highly expressed in slow-twitch fibers. In APOBEC2 null mice, there is a shift from fast to slow fibers, a development of mild myopathy with age, a reduction in the total body mass correlating with a decrease in muscle mass (Sato et al., 2010). Furthermore APOBEC2 is predicted to be a novel

regulator of skeletal muscle growth during postnatal muscle development (Yang et al., 2015), and it is upregulated in aging skeletal muscle providing evidence for a role in sarcopenia (degenerative loss of skeletal muscle mass during aging) (Piec et al., 2005). Moreover APOBEC2 expression increases with denervation of mouse soleus and extensor digitorum longus muscle (Sato et al., 2010), suggesting that the protein might be important for postnatal adaptation following injury or during muscle homeostasis. Induction of neonatal mouse myoblast and C2C12 myoblast cell line to differentiate in culture results in an increase in the expression of APOBEC2, suggesting a role in muscle differentiation (Sato et al., 2010; Vonica et al., 2011).

APOBEC2 has also been implicated in DNA demethylation. The best evidence for a possible function of APOBEC2 in DNA demethylation comes from zebrafish. The zebrafish *Apobec2* genes (*Apobec2a* and *Apobec2b*) together with AID were shown to be required for the demethylation of a DNA fragment injected into the single cell embryo of the zebrafish, which is in turn accompanied by hypomethylation of the zebrafish genome (Rai et al., 2008). The *Apobec2a/2b* genes are also crucial for optic nerve regeneration in zebrafish, which is achieved through the return of the neurons and glia to a stem cell state (Powell et al., 2012). Given that DNA demethylation is important during reprogramming and that regeneration in zebrafish is dependent on *Apobec2a/2b*, these data suggest that APOBEC2 might be important in DNA demethylation in zebrafish. Subsequent studies from the same group showed that APOBEC2 had little impact on site-specific DNA demethylation during the process of retina regeneration suggesting that its role still remains to be determined (Powell et al., 2013).

APOBEC2 has also been shown to be important during development. In zebrafish

APOBEC2 has been implicated with muscle development potentially through a mechanism that involves binding to Unc45b, a chaperone protein. Knockdown of APOBEC2a/b in zebrafish leads to a dystrophic phenotype in skeletal musculature and impairs heart function (Etard et al., 2010). A study in *Xenopus* demonstrated that APOBEC2 is regulated by transforming growth factor β (TGF-β) signaling and that its activity is crucial for the determination of left-right axis specification during early embryogenesis (Vonica et al., 2011). Additionally APOBEC2 was shown to inhibit the TGF-β pathway, upstream of Smad2 transcriptional response. The inhibitory role or left-right phenotype was not observed when mutations were introduced in the conserved amino acids of the enzymatic domain required for the activity of other well-characterized cytidine deaminases, strongly suggesting that in *Xenopus* APOBEC2 has cytidine deaminase activity. Further evidence shows that the inhibitory role in the TGF-β pathway is conserved in mouse APOBEC2. Vonica et al. shows that APOBEC2 mediates inhibition of TGF-β signaling and that this might be necessary for proper differentiation of the myoblast cell line C2C12.

Changes in expression levels of APOBEC2 have also been linked to cancer development. Transgenic mice constitutively expressing APOBEC2 in all tissues develop liver and lung cancers by 1.5 years of age. These tumors correlate with mutations in the mRNAs (but not DNA) of two tumor suppressors genes, Pten and Tp53 (Okuyama et al., 2012). In this study only specific tumor related genes were sequenced, thus there is a possibility that APOBEC2 could target more genes in the transcriptome/genome. Moreover *Apobec2* loss of function deletions are discovered as candidate cervical cancer driver events (Bierkens et al., 2013).

Overall the above-mentioned studies provide evidence that APOBEC2 has biological role(s) in diverse species. However there is no agreement on how APOBEC2 achieves these effects. It has been hypothesized that APOBEC2 may be involved in RNA editing (Liao et al., 1999; Okuyama et al., 2012), DNA demethylation (Guo et al., 2011; Powell et al., 2012; Rai et al., 2008), or that it has lost its deaminase activity altogether and may act by a different mechanism (Etard et al., 2010; Sato et al., 2010; Vonica et al., 2011). Determining what the direct physiological targets are, whether APOBEC2's biological function is mediated through deamination, whether it deaminates cytidines and whether this activity is conserved remains to be determined.

Studies spanning different model organisms confirm that APOBEC2 is important in skeletal muscle biology and development. In the following sections of this chapter I will be briefly describing myogenesis, the process of creating muscle, which is the model I will use for my experimental work in the chapters that follow.

1.4 Myogenesis, the process of cell differentiation to generate muscle

The process of cell differentiation entails the transition of a stem cell (a cell capable of giving rise to many cell types, thus high in differentiation potential/"stemness") to a more specialized cell through the activation and maintenance of specific gene expression program(s) characteristic for that specific cell type. During development cell differentiation is important for cell fate (or lineage) specification allowing the pluripotent embryo to give rise to the diverse somatic cell types of a multicellular complex organism. In the adult, cell differentiation of multipotent adult stem cell (or progenitors) occurs

when new specialized cells need to be replenished during homeostasis maintenance, growth, or tissue regeneration. In response to specific environmental or developmental cues the cellular identity is directed through transcription factor(s) (TF) binding to DNA that establish specific gene expression patterns and is reinforced through mitotically inheritable chromatin states, which determine TF accessibility to DNA (Atlasi and Stunnenberg, 2017).

Skeletal muscles representing the largest tissue mass important for movement and metabolism (Braun and Gautel, 2011). The adult skeletal muscle is mainly composed of terminally differentiated, contractile (due to the sarcomere), multinucleated and postmitotic myofibers, and a small pool of quiescent muscle stem cells, known as satellite cells that can be 'activated' to form myoblasts (undifferentiated committed cells that can specialize to form muscle cell). This process of generating muscle during differentiation-myogenesis- occurs early during development (embryonic and - stages), and postnatally to allow muscle growth, tissue homeostasis and muscle regeneration upon injury (Bentzinger et al., 2012).

During developmental myogenesis, mesoderm-derived structures generate the first muscle fibers (primary fibers). In subsequent waves, additional fibers (secondary fibers) are generated along the initial fibers, which serve as scaffolds. Initially myogenic progenitors increase in numbers through proliferation, but later on decrease as the number of myonuclei (nuclei of myofibers) reaches a steady state. Once the muscle has matured, they enter quiescence and reside as satellite cells. Fully formed skeletal muscle in adults relies on satellite cells, which can self-renew and replenish muscle with new

terminally differentiated myofibers during cell turnover or regeneration (Dumont et al., 2015).

1.5 Master regulators controlling differentiation of muscle cells

Muscle cell lineage progression is associated with hierarchical and coordinated expression of various TF(s) that lead to cell specific gene expression (Figure 1.6) (Bentzinger et al., 2012). Higher up in the hierarchy are sine oculis homeobox transcription factors 1/4 (Six1/4) and paired-homeobox transcription factors 3/7 (Pax3/7), which are master regulators of early lineage specification in the embryo. In the postnatal organism, Pax3 and Pax7 mark the presence of satellite cells located underneath the basal lamina of adult myofibers. Pax3 is mainly involved in the regulation of embryonic functions and maintenance of an undifferentiated phenotype, while Pax7 appears to regulate proliferation and inhibition of differentiation, through binding to genes involved in the maintenance of adult satellite cell phenotype (Soleimani et al., 2012).

Myogenic formation in different scenarios (e.g. embryonically or postnatally in the adult) is defined by the expression of the myogenic regulatory factors (MRFs), a group of basic helix–loop–helix (bHLH) transcription factors that include myogenic factor 5(Myf5), myoblast determination protein (MyoD), muscle-specific regulatory factor 4 (Mrf4 also known as Myf6) and Myogenin (Myog). MRFs bind to E-protein family of bHLH proteins forming heterodimers and bind to E-box consensus sequence (CANNTG) present in the regulatory regions of muscle-specific genes (Berkes and Tapscott, 2005). MRFs act at multiple points in the muscle lineage to coordinately establish the skeletal muscle phenotype through regulation of proliferation, irreversible cell cycle withdrawal of myoblasts, and fusion of myoblasts into myotubes. Moreover

regulation of other muscle specific genes facilitates the differentiation and assembly of the sarcomere, the contractile units of the muscle mostly made up of actin and myosin filaments and arranged in highly ordered arrangement (Hernández-Hernández et al., 2017). Mouse genetic models demonstrate that MyoD and Myf5 function in determination of myogenic cells, Myog and MyoD function on terminal differentiation and Mrf4 has a role in both determination and differentiation. The Myf5$^{-/-}$MyoD$^{-/-}$ mice show postnatal lethality, due to a complete absence of skeletal myoblasts or myofibers (Rudnicki et al., 1993), suggesting they are important in specifying the myogenic lineage in embryos. Myog $^{-/-}$ mice continue to specify the muscle lineage through the formation of myoblasts but show perinatal lethality due to defects in muscle fiber formation (Hasty et al., 1993; Nabeshima et al., 1993) suggesting importance during differentiation. Mrf4 is considered both a determination and a differentiation factor. MyoD$^{-/-}$Mrf4$^{-/-}$ show postnatal lethality and a phenotype similar to that of the Myog $^{-/-}$ mice (Rawls et al., 1998), suggesting that Mrf4 and MyoD play a redundant role in mediating skeletal muscle differentiation during development.

Temporally, Myf5 and MyoD expression is induced in proliferating myoblasts (or activated satellite cells), which is followed by the simultaneous downregulation in expression of Myf5 and induction of Myog during cell cycle exit, leading to a 'point of no return' where cells are committed to go down the path of differentiation. Downstream activity of MyoD and Myog leads to the expression of the Mrf4 gene and other late muscle differentiation genes to allow the formation of multinucleated fibers. Lastly in mature muscle fibers, MyoD and Myog are downregulated, whereas Mrf4 continues to be expressed at high levels acting as the main MRF in mature differentiated muscle. Studies

looking at genome wide binding of MyoD have shown that it stably binds in genes that are known to be downregulated or upregulated during differentiation at both the myoblast and the differentiated phase and additionally the majority of the binding sites contain E-box sites (a DNA response element that acts as a protein-binding site and has been found to regulate gene expression) (Cao et al., 2010; Mousavi et al., 2013). More recently it was shown that the genome-wide binding profiles of Myf5 and MyoD were identical, and the studies support a model where Myf5 functions to facilitate chromatin remodelling in muscle progenitors by inducing histone H4 acetylation while MyoD recruits RNA Pol II and activates transcription of the same target genes (Conerly et al., 2016). Another study examining the role of Myog in differentiation identified genes involved in cell cycle progression as key transcriptional targets that are downregulated by Myog (Liu et al., 2012), suggesting that Myog reduces the expression of genes that mediate cell cycle progression.

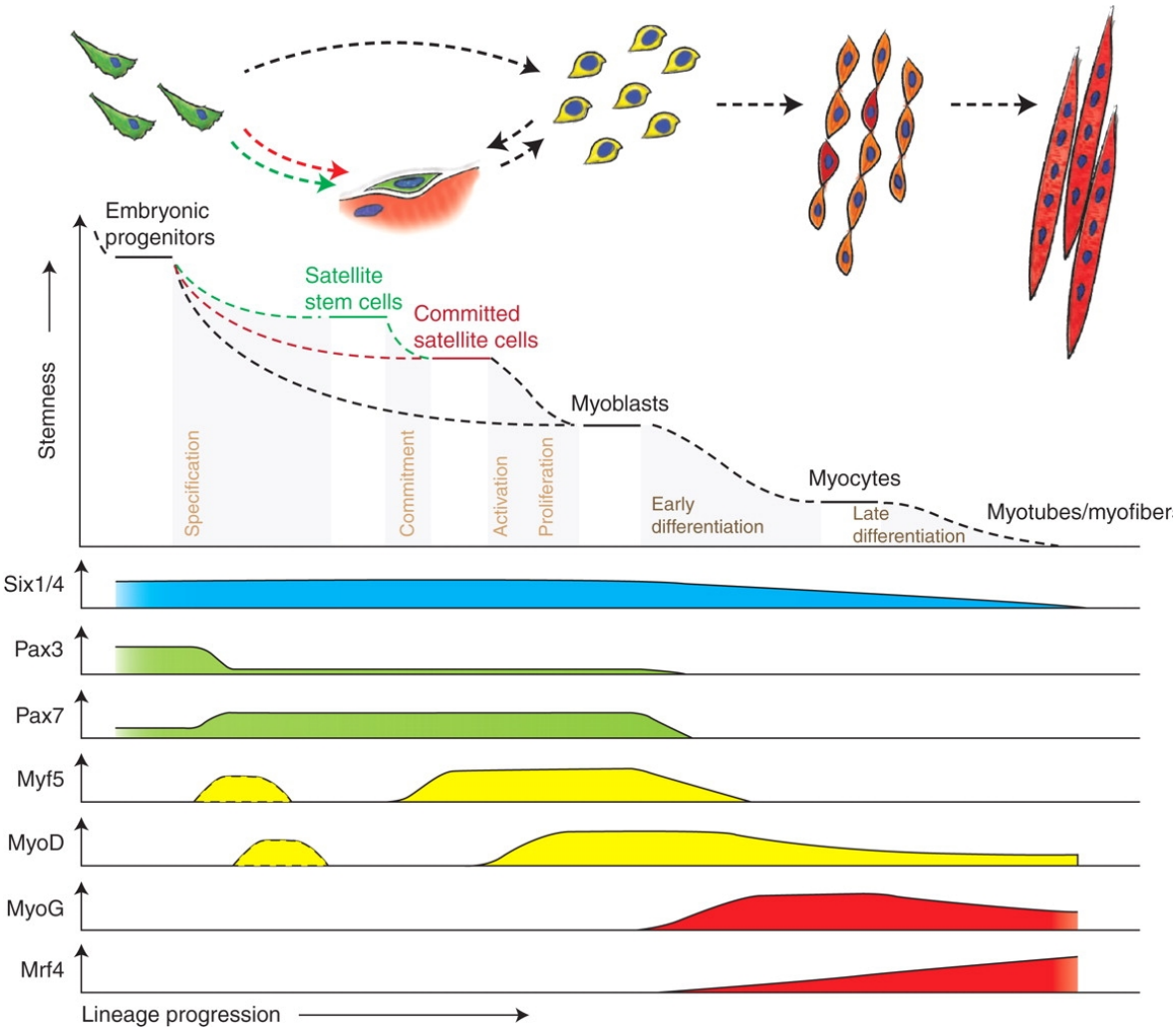


Figure 1.6 Master regulators controlling myogenesis

Scheme of the expression pattern of the main myogenic transcription factor during postnatal myogenesis. *Image from (Bentzinger et al., 2012)*

1.6 The role of myogenic regulatory factors in cell cycle progression

To successfully differentiate, myoblasts must withdraw from cell cycle. Besides their role in differentiation, MRFs have been shown to modulate cell cycle progression (Singh Kulwant and Dilworth F. Jeffrey, 2013) (Figure 1.7). Both MyoD and Myf5 can promote expansion of the muscle progenitor population (Megeney et al., 1996; Ustanina et al.,

2007; Zhang et al., 2010) and show opposing expression patterns during the different phases of the cell cycle (Kitzmann et al., 1998). Myf5 protein level peak in G0, decrease at G1, and go up again at the end of G1, remaining stable through mitosis. MyoD protein level peaks in mid-G1, decreases to its minimum level in the G1/S transition, and go up again from S to M. The levels of both proteins during cell cycle are controlled by phosphorylation-dependent degradation via the 26S proteasome. The importance and consequences of changing levels of MyoD and Myf5 during cell cycle progression is not completely understood. What is knows so far is that MyoD binds to the transcriptional regulatory region of genes with roles in the cell cycle in proliferating myoblasts (Blais et al., 2005; Cao et al., 2006, 2010), and activates expression of Cdc6 and Mcm2, involved in preparing chromatin for DNA replication and progression through S-phase of the cell cycle (Zhang et al., 2010). In response to differentiation signals, MyoD induces expression of Myog and initiates a gene expression program that commits myoblast to exit the cell cycle through the activation (and potentially through direct binding) of p21/*cdkn1a*, p57/*cdkn1c* and retinoblastoma protein (pRB) (Falco et al., 2006; Figliola and Maione, 2004; Zhang et al., 1999).

MyoD and Myog synergize to lead to cell cycle exit through 3 main mechanisms (1) upregulation of p21/*cdkn1a* and p57/*cdkn1c* which suppress the activity of Cdk(s) and cyclins (Zhang et al., 1998), (2) upregulation of pRB which establishes repressive histone methylation at cell cycle genes (Blais et al., 2007), (3) the suppression of E2F family members, which are major regulators of the expression of Cdks and cyclins. Inhibition of E2F family members has been shown to be achieved through multiple mechanisms: Myog upregulates the expression of miR-20a targeting transcription factors E2F1, E2F2,

E2F3 (Sylvestre et al., 2007); Myog reguates Lats2 (Liu et al., 2012), a protein kinase implicated in targeting of the transcriptional repressor complex DREAM and pRB to E2F target genes to block cell cycle progression (Tschöp et al., 2011).

In summary, current research suggests that the MRFs that play a role in determination (MyoD and Myf5) facilitate cell cycle progression, whereas the MRFs that mediate differentiation (Myog) induce cell cycle exit. Current research shows that Myog mediates cell cycle exit indirectly, through regulating the expression of target genes involved in blocking cell cycle progression.

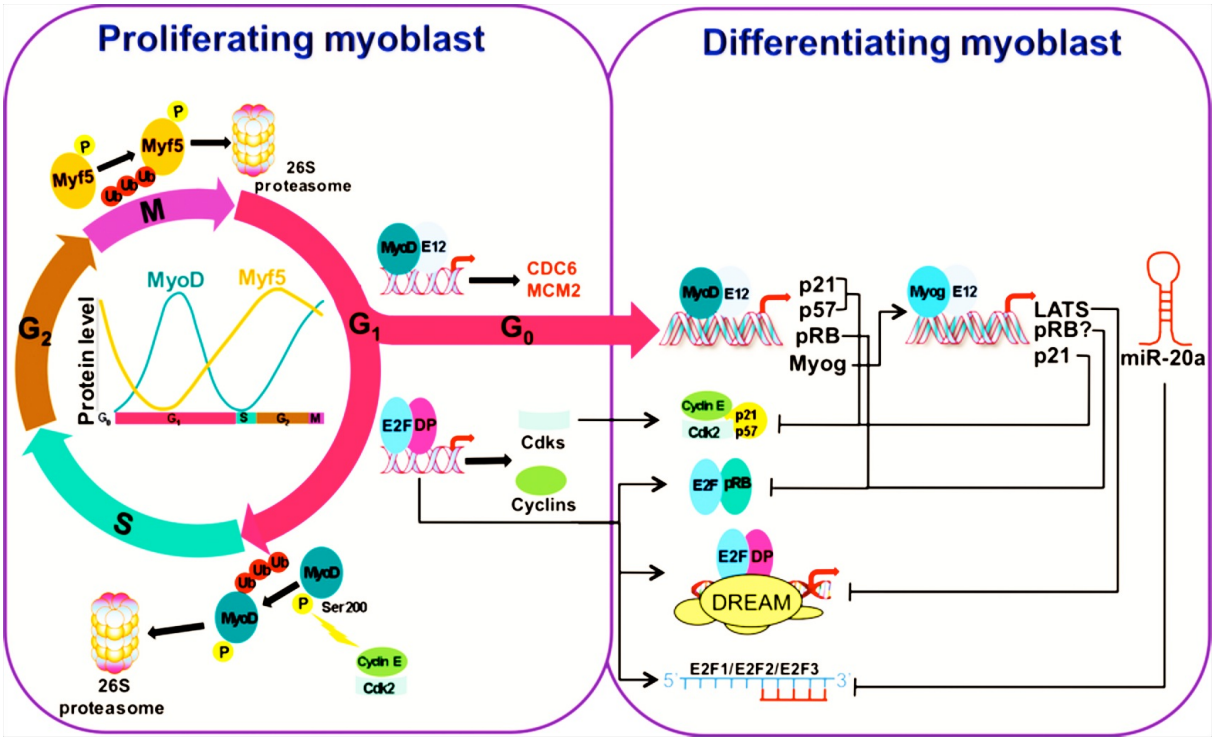


Figure 1.7 The role of myogenic regulatory factors in cell cycle progression
Scheme showing how MRFs affect cell cycle arrest necessary for proper myogenesis.
Image from Singh Kulwant and Dilworth F. Jeffrey, 2013

1.7 Statement of the problem

APOBEC2 is one of the most ancestral members of the AID/APOBEC family and it has been under purifying selection suggesting that it has an important function (Conticello et al., 2005, 2007b). Previous research has demonstrated that APOBEC2 has a biological role in various organisms. So far it is unknown how APOBEC2 is mediating its biological function(s), what its physiological targets are, whether it is acting as a deaminase in polynucleotides, or whether it is deaminating cytidine.

The main goals of the research presented here is to utilize high throughput next generation sequencing methods to elucidate how APOBEC2 mediates its physiological role(s) and whether it is important for (1) gene expression regulation, (2) DNA methylation, (3) RNA editing and (4) whether it can bind DNA. For these studies I have chosen to focus on muscle, since it is the tissue where APOBEC2 is expressed physiologically, highly, and in an inducible manner. The ultimate goal of my thesis work is to identify physiological binding targets that can be utilized to shed light into the existence (or not) of an enzymatic activity for APOBEC2, and thus decipher its broader role beyond muscle, with potential implications in gene expression regulatory mechanisms and/or in driving genome biological diversity.

CHAPTER 2. Characterizing the role of APOBEC2 during muscle differentiation

As described in the introduction, the role of APOBEC2 in muscle biology has been widely described. Mouse APOBEC2 is highly expressed in skeletal muscle. Its deficiency results in a mild phenotype in adult skeletal muscle, where APOBEC2$^{-/-}$ mice are smaller in size, have smaller muscles, show higher levels of slow-twitch fibers, and with age, show signs of myopathy through the presence of abnormally nucleated myofibers in 6 month old mice (Sato et al., 2010). Additionally APOBEC2 expression levels go up in the aging skeletal muscle in rats, linking APOBEC2 to sarcopenia (muscle wasting during aging) (Piec et al., 2005). Despite many studies supporting APOBEC2's biological function in muscle, there is still a lack in knowledge about its physiological target(s) or presence of enzymatic activity (Harris et al., 2002; Lada et al., 2011). As mentioned the primary goal of my thesis work is to discover the direct substrates of APOBEC2 with the hope that this would give us insights and bring us closer to the identification of its potential enzymatic activity. To do this I chose a simple system of muscle differentiation in culture as a model for investigating substrate(s) of APOBEC2. In the first part of this chapter, I describe phenotypic data related to the cell culture differentiation model and in the second part of the chapter, I describe further studies using APOBEC2 null mice.

2.1 C2C12 cell culture system as a model to study APOBEC2 function

C2C12 is a widely used model to study skeletal myogenesis. C2C12s were originally derived from mouse satellite cells activated to proliferate after muscle injury in adult mice (Blau et al., 1985; Yaffe and Saxel, 1977). The advantages of using this system are

that these cells proliferate as myoblasts in high-serum conditions and can be induced to differentiate into myotubes in cell culture in low-serum conditions. Unlike primary myoblasts, which have a limited proliferative capacity, C2C12 is an immortalized cell line that provides an alternative and widely used model to study myogenesis. Because C2C12s proliferate indefinitely while retaining the ability to differentiate, they provide a convenient and faster system to use. The limitations of the system are (1) the cells are intrinsically different from primary myoblasts metabolically and in their ability to contract forming myofibrils and (2) you lack the cell niche (potential other external factors) compared to using a mouse model. The system is ideal for our studies on the identification of APOBCEC2 substrates because (1) it had previously been demonstrated that *Apobec2* mRNA expression is induced upon differentiation in C2C12s (Vonica et al., 2011) and (2) levels of the *Apobec2* mRNA increase during differentiation suggesting APOBEC2 is exerting its physiological effects and (3) these cells are a commonly used system for recapitulating the first steps of muscle differentiation in culture.

I first confirmed the progressive increase in the levels of APOBEC2 protein during "myogenesis" in the C2C12 model system (Figure 2.1a), which would suggest a potential biological and/or enzymatic activity of the protein during C2C12 differentiation. I also confirmed that under the conditions I used, I could reliably induce terminal differentiation of myoblasts into myotubes, by calculating the fusion index (percentage MyHC-positive myotubes with > 2 nuclei). As previously demonstrated, I showed that the number of MyHC multinucleated fibers increase after induction of differentiation as indicated by an increase in the fusion index (Figure 2.1b). I could thus recapitulate previous findings in the literature.

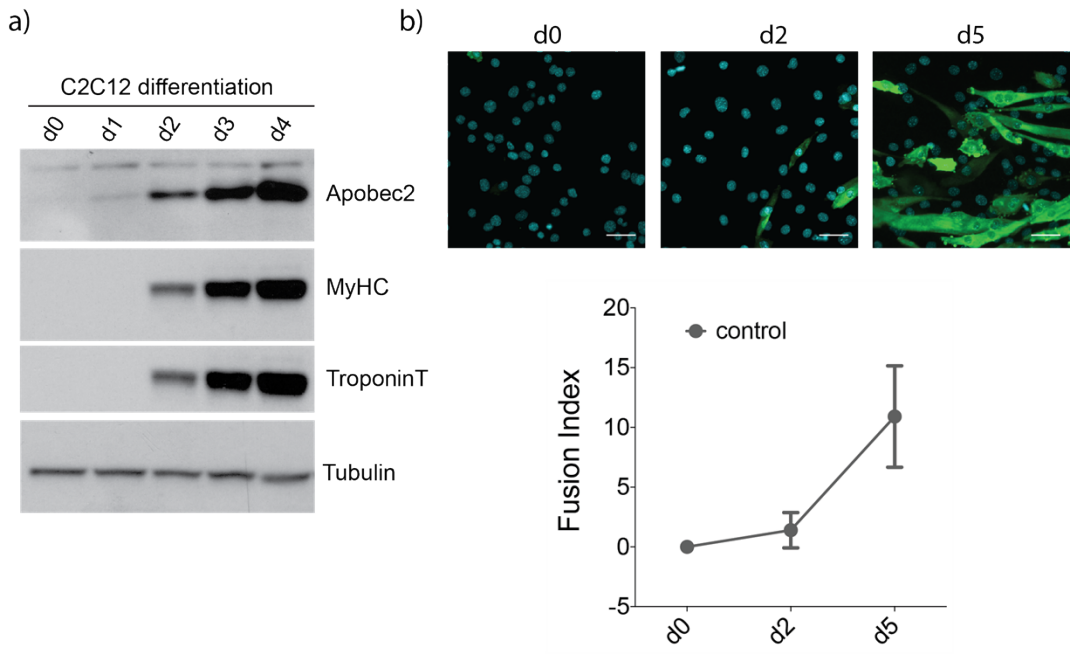


Figure 2.1 APOBEC2 expression in C2C12 increases during differentiation

Myogenic differentiation was induced by switching the cells to the differentiation medium(DM). Days indicate the time the cells were in DM. a) Whole cell extracts of mouse C2C12 myoblasts and myotubes were analyzed by Western blotting using anti-Apobec2 antibodies. MyHC and TroponinT were used as markers of late differentiation, alpha-tubulin was used as loading control b) C2C12 cells were cultured in DM for 0,2 and 5 days, fixed, and stained with antibody to MyHC (green). Nuclei were visualized by DAPI staining (blue). Below the quantification of differentiation expressed as fusion index, which is the percentage MyHC-positive myotubes with >2 nuclei. Results are presented as means from quantification of at least 6 images/sample. Error bars indicate SD. Image scale 50um.

2.2 Knockdown of APOBEC2 during C2C12 myogenesis interferes with differentiation

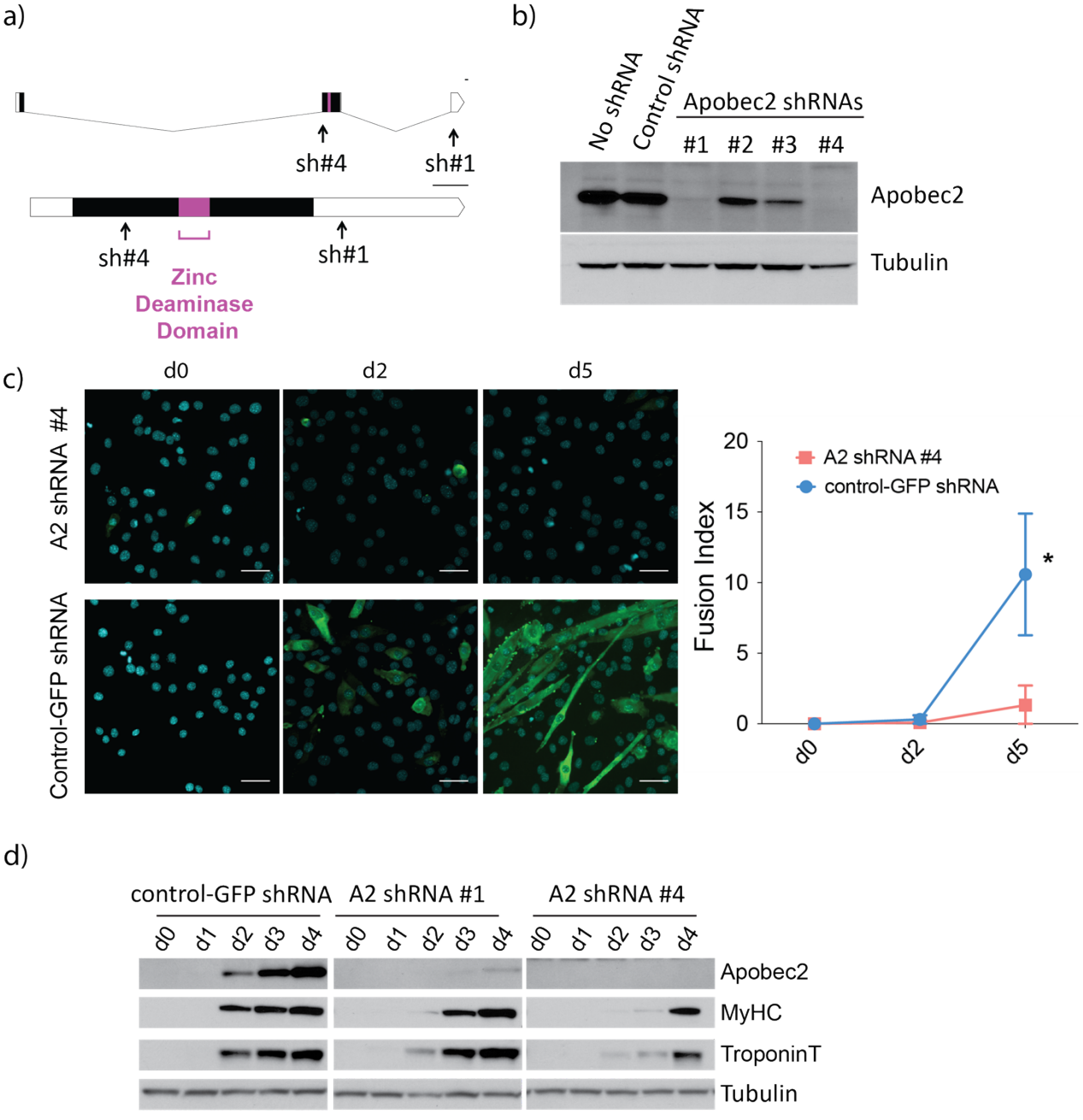
To explore APOBEC2's biological role and enzymatic activity during myogenesis I depleted APOBEC2's protein levels, using short hairpin RNAs (shRNA) that are constitutively expressed once successfully integrated in the genome. When successfully bound to their target mRNA (the mRNA for *Apobec2*), these shRNAs will lead to its degradation.

I tested four hairpins, two of which (#1 and #4) showed abundant knockdown at the protein level when compared to the knockdown controls (Figure 2.2a,b). The target of shRNA#1 on *Apobec2* mRNA is at the 3' untranslated region (3'UTR) while the target of shRNA#4 on *Apobec2* mRNA is within the second exon. Cells constitutively expressing the hairpins were selected, seeded in equal numbers and 12 hours after seeding were induced to differentiate in cell culture. My results demonstrate that knock-down of APOBEC2 leads to a reduction in the fusion of myotubes as indicated by decreased fusion index (Figure 2.2c) and a delay in the expression of differentiation markers, TroponinT and MyHC (Figure2.2d). Thus overall proper formation of myotubes is impaired. These results are corroborated by a recent publication where it was shown that expression of APOBEC2 increases during differentiation of embryonic stem cell-derived myogenic precursors into myotubes in cell culture. Similarly to my data, knockdown of APOBEC2 reduced differentiation and the expression of differentiation markers such as Myog (Carrió et al., 2016).

Figure 2.2 Knockdown of APOBEC2 in C2C12 cells leads to problems in differentiation

a) Schematic representation of *Apobec*2 gene. Lines as introns and boxes as exons, black boxes as coding regions. Image scale=100bases. Location of the zinc deaminase domain and the site of the binding of A2sh#1 and A2sh#4 shRNAs are shown. Produced using http://wormweb.org/ b) Lysates of C2C12 from cells with no hairpin (no shRNA), control shRNA (GFPsh) and APOBEC2 shRNA (#1-#4) cultured in DM for 4 days were analyzed by Western blotting (WB) using anti-Apobec2 antibodies. Alpha-tubulin was used as loading control c) C2C12s were fixed and immunostained using antibodies specific MyHC (green), DAPI (blue) was used to stain for DNA. Quantification of differentiation expressed as fusion index is shown. Error bars indicate SD d) C2C12 lysates in DM at day 0-4 were analyzed by WB. MyHC and TroponinT were used as markers of late differentiation; alpha-tubulin, as loading control.

Statistics: t test. At least 6 fields of view were measured and data is shown as means. Error bars indicate SD (n=3). *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$


a)
sh#4
sh#1
sh#4
sh#1
Zinc
Deaminase
Domain
b)
No shRNA
Control shRNA
Apobec2 shRNAs
#1 #2 #3 #4
Apobec2
Tubulin
c)
d0
d2
d5
A2 shRNA #4
Control-GFP shRNA
A2 shRNA #4
control-GFP shRNA
Fusion Index
20
15
10
5
0
*
d0
d2
d5
d)
control-GFP shRNA
A2 shRNA #1
A2 shRNA #4
d0 d1 d2 d3 d4
Apobec2
MyHC
TroponinT
Tubulin

Muscle differentiation is closely coordinated with cell cycle progression. During the differentiation of myogenic precursors into mature myotubes, cells withdraw irreversibly from the cell cycle. Cell cycle exit occurs early during the differentiation program and is required for fusion and proper myotube formation (Walsh and Perlman, 1997). Cells commit to differentiation as indicated by Myog induction and they commit to irreversible cell cycle arrest as indicated by induction of p21/*cdkn1a*, the cyclin-dependent kinase inhibitor (CDI) whose expression levels correlate with blockage of DNA synthesis and cell cycle exit (Guo et al., 1995)(Andrés and Walsh, 1996). This is followed by expression of myosin heavy chain (MHC) and fusion. Given the importance of cell cycle withdrawal in proper differentiation, I next investigated whether APOBEC2 affects cell cycle withdrawal through differentiation.

Therefore, I quantified the proportion of cells in G0/G1, entering S, and in G2/M phases of the cell cycle as they are induced to differentiate. I used a modified thymidine analogue (EdU or 5-ethynyl-2'-deoxyuridine) in combination with PI (propidium iodide) or PI(alone). Both molecules incorporate into the DNA, EdU is incorporated into newly synthesized DNA and can be fluorescently labeled, and PI binds to DNA by intercalating between the bases. The combination of the two allows for the FACS-based, clear differentiation between all the different cell cycle phases proportional to the amount of DNA stained. As expected from the literature, during differentiation of C2C12 the proportion of the cells in G0/G1 increases while fewer cells are entering S phase and are in G2/M phase. Interestingly knocking down APOBEC2 in C2C12s leads to a decrease in the proportion of cells in G0/G1 phase and an increase in the proportion of cells entering S phase during differentiation (Figure 2.3a,b). The data shown here represent a mixture of

cells at different stages of the cell cycle. In the future, we would like to repeat the cell cycle experiment following C2C12 cell cycle synchronization, so that the cells at different stages of the cell cycle in a culture are brought to the same phase. Overall the data suggests that APOBEC2 is necessary for timely cell cycle withdrawal and consequently for terminal differentiation.

In the future, we like to confirm that the phenotype observed due to knockdown of APOBEC2 is protein specific and reproduce the cell cycle phenotype using the A2sh#1 hairpins and rescue the phenotype observed to rule out off target affects. My initial attempts to rescue the phenotype of delayed differentiation (using synonymous mutations in the construct in a constitutive expression and an inducible expression system) showed conflicting results, even though western blot analysis confirmed that this construct is resistant to shRNA-mediated depletion. Moreover it is not uncommon for RNAi rescue experiments to fail for multiple reasons (known and unknown) unrelated to the off-target affects (Datler and Grimm, 2013). Some of the challenges in the C2C12 cell culture system are being able to achieve expression of the rescue construct at the levels and dynamics of the endogenous protein. Thus since during differentiation a lot of events need to be coordinated in a timely manner, the amount of construct expressed might be too much or too little and not at the optimum time points. In the future these systems will be optimized to test the rescue of the cell cycle phenotype observed.

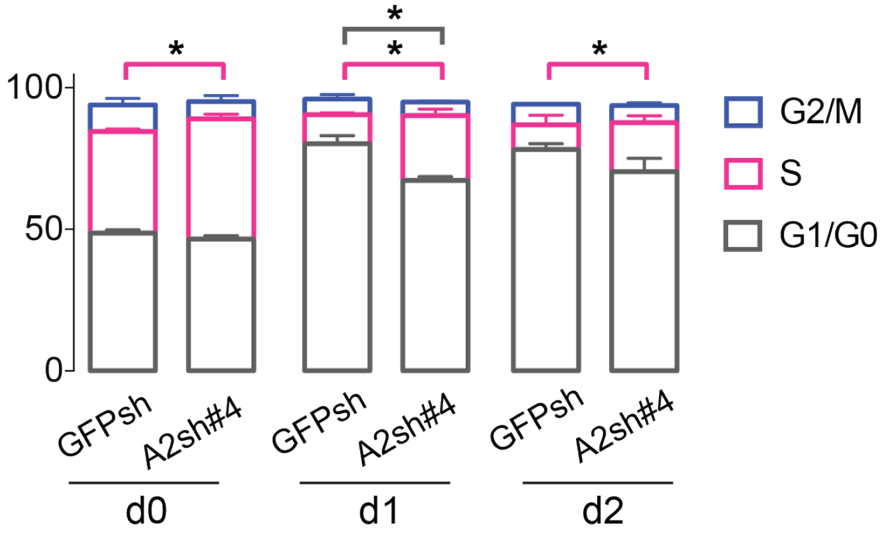
Overall, I demonstrated that APOBEC2 in C2C12s is important for timely progression to cell cycle exit and terminal differentiation of myoblasts into myotubes. This provides a good model system to further investigate the potential substrates and enzymatic activity of APOBEC2.

Figure 2.3 Knockdown of APOBEC2 in C2C12 cells leads to defects in cell cycle withdrawal during differentiation

a) C2C12 stably expressing A2sh#4 or GFPsh (control) cultured in DM for different days (d0,1,2) were stained with PI and EdU. FACS plots shows the time-course effect of differentiation on cell cycle profile in C2C12 cells. Results are presented as means. Error bars indicate SD (n = 3, corresponding to three independent experiments). Data shown as means ± SD. *p < 0.05; **p < 0.01; ***p < 0.001. Color of the line corresponds to the cell cycle phase shown.

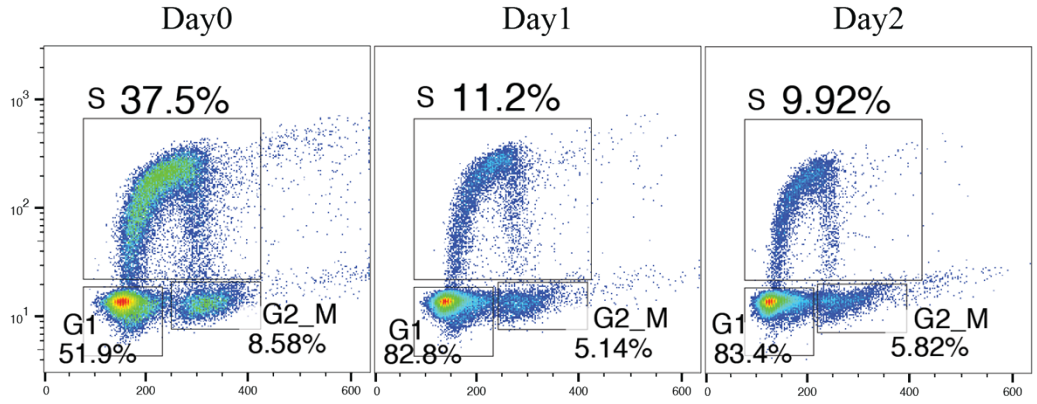
b) C2C12 cultured in DM for different days (d0, 1, 2) were stained with EdU and PI to mark cells in the different stages of the cell cycle (G0/G1, entering S and G2/M). The FACS plots show the time-course effect of differentiation on cell cycle profile in C2C12 cells.

a)

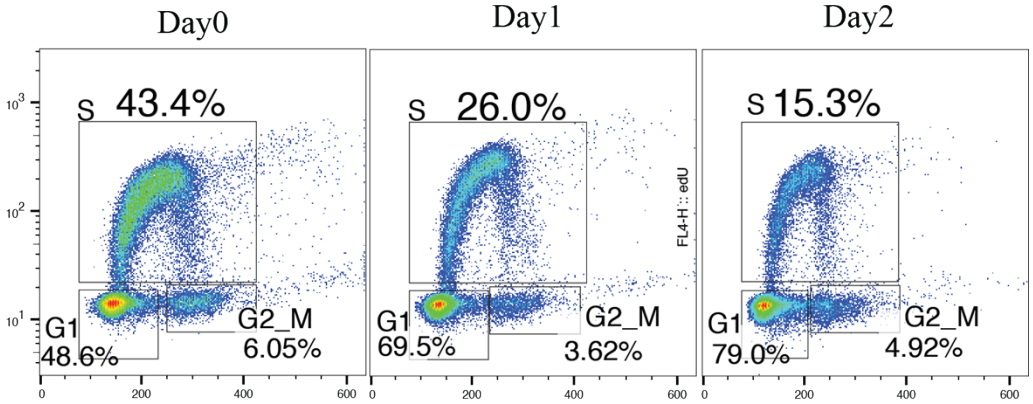


b)

Control-shGFP



A2-shRNA 4



2.3 APOBEC2 deficiency leads to defects in fiber maturation and nuclear positioning

Previous research on APOBEC2 in muscle has shown minor phenotypic affects due to APOBEC2 deficiency but no major overall defects in muscle development, health, fertility, or survival (Mikl et al., 2005; Sato et al., 2010). This could be due to genetic redundancy where another (or more) gene(s) compensates (partly or fully) for its absence by performing a similar function and as a result APOBEC2's deficiency has no drastic effect on the biological phenotype. Indeed there are many examples of this occurring through development including in muscle (Haldar et al., 2008; Kafri et al., 2009). An example is Myf-5 and MyoD that functionally substitute for one another (at least partly) during myogenesis and only when both of them are deleted does one observe a block in muscle development (Rudnicki et al., 1993). Another possibility is that APOBEC2 could have a potential role mainly during postnatal muscle growth, maintenance of tissue homeostasis or muscle regeneration after injury. Given that previous work on the APOBEC2 null mice does not show major defects in embryonic muscle development, I next investigated whether there are any defects during myotube formation during *de novo* myogenesis in adult mice.

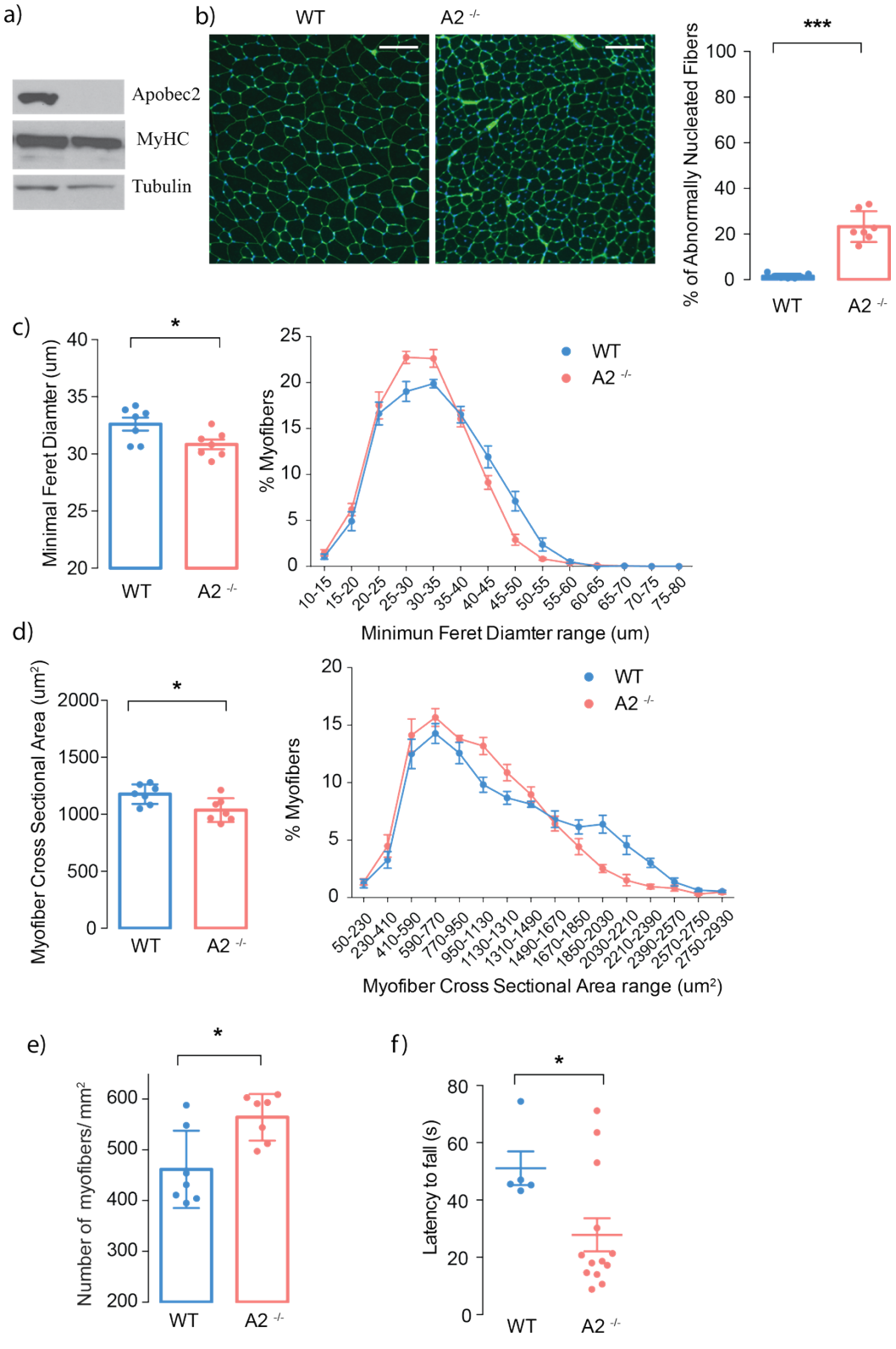
Nuclear migration is an important event throughout development and it is characteristic of muscle fibers. During muscle formation nuclei undergo mechanistically and temporally distinct movements which leads to nuclei being evenly spaced at the periphery of the myofiber to maximize the distance between them (Bone and Starr, 2016). For our studies we use a different APOBEC2 null mouse not previously published where no APOBEC2 protein is detected (Figure 2.4a). First I characterized muscles from

the APOBEC2$^{-/-}$ in the centrally nucleated phenotype. Previous work has linked APOBEC2 with muscle wasting and an increase in the proportion of slow to fast fiber type ratio, suggesting that the fast fiber types might be affected due to APOBEC2 deficiency. Therefore I carried experiments using the *tibialis anterior* (TA) muscle, which is a fast-twitch muscle. My experiments show that in the TA muscle there is a higher proportion of abnormally nucleated fibers in 10 week old APOBEC2$^{-/-}$ mice (Figure 2.4b), suggesting that abnormalities in muscle from APOBEC2$^{-/-}$ mice start much earlier than what was previously shown (Sato et al., 2010). Moreover mice deficient in APOBEC2 have smaller myofibers as shown by measuring fiber cross section area (CSA) and minimum feret diameter accompanied but increase in the number of myotubes (Figure 2.4c-e), in accordance with a recent publication (Ohtsubo et al., 2017a). Furthermore, muscle strength was tested by grip test where time to release from an inverted grid was recorded. My data show that in the four-limb grip test, APOBEC2 deficient mice show a reduction in the total hang time in the grid indicative of reduction in muscle strength (Figure 2.4f). Overall, steady state muscle is qualitatively and quantitatively different in APOBEC2$^{-/-}$ animals.

Figure 2.4 APOBEC2$^{-/-}$ muscle fibers are small and abnormally nucleated in mice

a) Expression of APOBEC2 and MyHC, alpha-tubulin, as loading control is shown through Western Blot analysis in skeletal muscle b) Representative image of the cross-sections of the *Tibialis Anterior* (TA) in 10-12 week old WT and APOBEC2 $^{-/-}$ (A2$^{-/-}$) mice. WGA in green (delineates the myofiber boundaries) and DAPI in blue stains the nuclei. The images are used to automate the calculation of the number of fibers with nuclei not present around the fiber (abnormally nucleated) using ImageJ macros. Plotted on the right is the percent of abnormally nucleated fibers. Data are represented as means ± SD c) TA fiber size quantification using minimal feret diameter (in μm). Mean values on the left, distribution of diameter on the right d) Fiber size quantification using cross sectional area (in μm^2). Mean values on the left, distribution of area on the right f) Number of myofibers in TA for equal areas analyzed e) An inverted grip-hanging test was performed on in 3-5 month old WT and APOBEC2$^{-/-}$ mice. Average values of 5 trials are shown

Statistics: Unpaired t test. N > 5. At least 250 individual myofibers are analyzed for each cross section of TA muscle and average values for each mouse are plotted here. Unless noted otherwise data are represented as means ± SEM. *p < 0.05; **p < 0.01; ***p < 0.001; no star, statistically non significant. Scale bar = 100um. Unless noted otherwise statistical test: Unpaired t test with Welch's correction.


a)
Apobec2
MyHC
Tubulin
b)
WT
A2 -/-
***
% of Abnormally Nucleated Fibers
100
80
60
40
20
0
WT
A2 -/-
c)
*
Minimal Feret Diamter (um)
40
35
30
25
20
WT
A2 -/-
% Myofibers
25
20
15
10
5
0
WT
A2 -/-
10-15
15-20
20-25
25-30
30-35
35-40
40-45
45-50
50-55
55-60
60-65
65-70
70-75
75-80
Minimun Feret Diamter range (um)
d)
*
Myofiber Cross Sectional Area (um²)
2000
1500
1000
500
0
WT
A2 -/-
% Myofibers
20
15
10
5
0
WT
A2 -/-
50-230
230-410
410-590
590-770
770-950
950-1130
1130-1310
1310-1490
1490-1670
1670-1850
1850-2030
2030-2210
2210-2390
2390-2570
2570-2750
2750-2930
Myofiber Cross Sectional Area range (um²)
e)
*
Number of myofibers/ mm²
600
500
400
300
200
WT
A2 -/-
f)
*
Latency to fall (s)
80
60
40
20
0
WT
A2 -/-

The regenerative capacity of muscle is due to satellite cells, the myogenic stem cell population residing in adult skeletal muscles (C2C12 cells derive from satellite cells). In healthy adult muscle the satellite cells are quiescent (G0 state, non-dividing). Following injury, the cells become activated and re-renter cell cycle (G0 to G1/S/G2/M) increasing the number of myoblasts, and subsequently exit the cell cycle to self-renew (G1 to G0) or to differentiate (Relaix and Zammit, 2012). I used a model of muscle regeneration following injury to determine whether deficiency of ABOBEC2 leads to any defects in differentiation during myogenesis in the adult mouse.

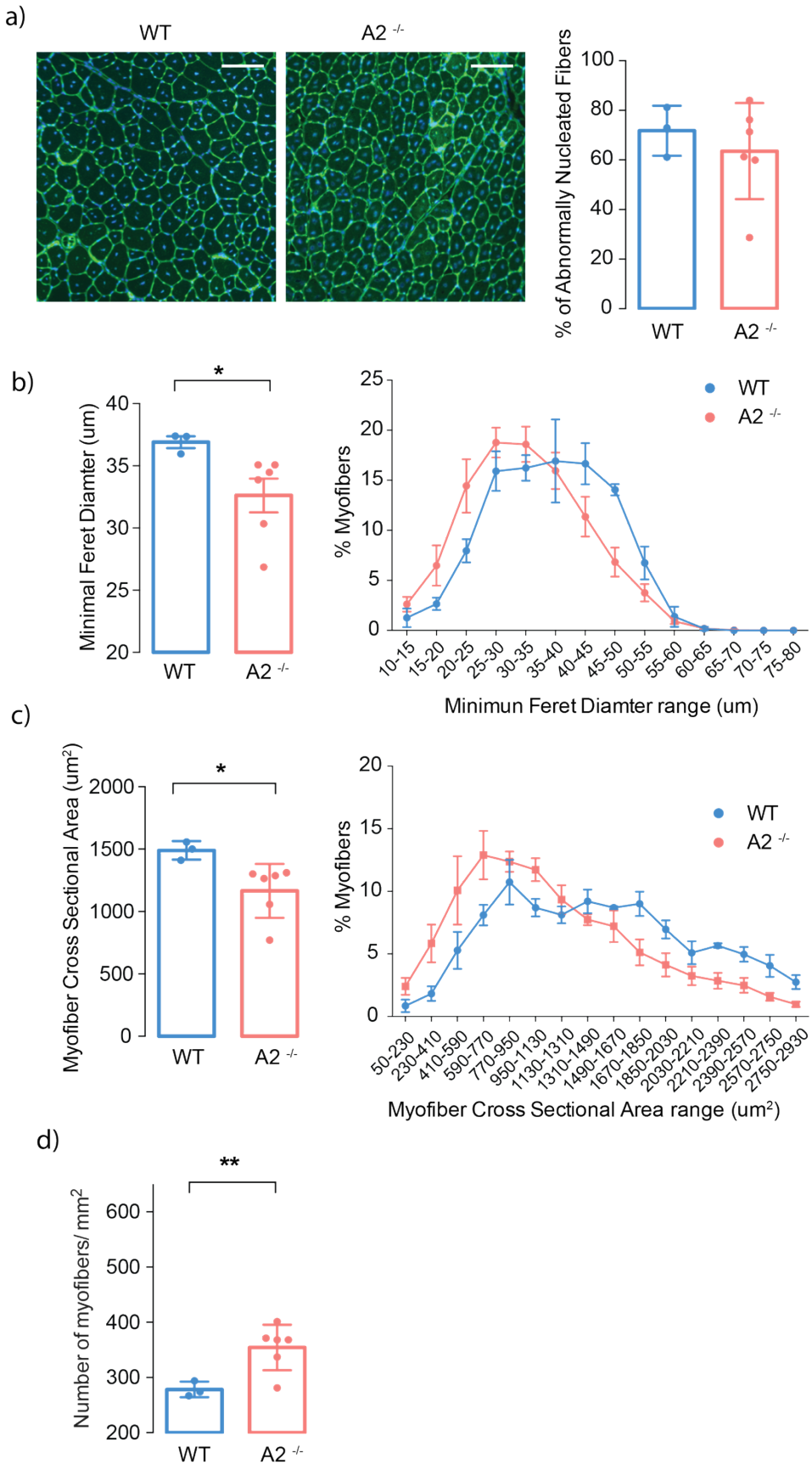
I induced injury in the TA muscle by injecting $BaCl_2$ in both wild type and APOBEC2$^{-/-}$ mice, using one of the two TA muscles as control (inject saline). Muscle cross sections were analyzed 14 days after inducing injury, during which I expected myotube formation to be complete in wild type mice (Hardy et al., 2016). My results show that mice deficient in APOBEC2 form myotubes at 14 days after injury but they are smaller in size when measuring fiber cross section area (CSA) and minimum feret diameter and have more fibers per area analyzed (Figure 2.5). Therefore APOBEC2$^{-/-}$ muscle fibers start off different in size and number, and re-set to that baseline, even after injury. These data suggest that lack of APOBEC2 does not delay myofiber formation after injury, but might be relevant for proper muscle maturation and growth.

Collectively, the data so far suggests a role of APOBEC2 in cell cycle regulation. In C2C12s failure to properly exit cell cycle is accompanied with defects in proper differentiation. I decided to use the C2C12 system to further investigate potential APOBEC2 substrates.

Figure 2.5 APOBEC2$^{-/-}$ muscle fibers are smaller in mice 14 days post-injury

a) Cross-section of the *Tibialis Anterior* (TA) muscle in WT and APOBEC2$^{-/-}$ (A2$^{-/-}$) mice. WGA in green (delineates the myofiber boundaries) and nuclei were visualized by DAPI staining. Scale bar is 100um b) Fiber size quantification using minimal feret diameter (in μm). Mean values on the left, distribution of diameter on the right c) Fiber size quantification using cross sectional area (in μm^2). Mean values on the left, distribution of area on the right. N = 7

At least 250 individual myofibers are analyzed for each cross section and average values for each mouse are plotted here. Data are represented as means ± SEM. *p < 0.05; **p < 0.01; ***p < 0.001; no star, statistically non significant. Statistical test: Unpaired t test with Welch's correction. Scale bar in images is 100um


a)
WT
A2 -/-
% of Abnormally Nucleated Fibers
100
80
60
40
20
0
WT
A2 -/-
b)
*
Minimal Feret Diamter (um)
40
35
30
25
20
WT
A2 -/-
% Myofibers
25
20
15
10
5
0
WT
A2 -/-
10-15
15-20
20-25
25-30
30-35
35-40
40-45
45-50
50-55
55-60
60-65
65-70
70-75
75-80
Minimun Feret Diamter range (um)
c)
*
Myofiber Cross Sectional Area (um²)
2000
1500
1000
500
0
WT
A2 -/-
% Myofibers
20
15
10
5
0
WT
A2 -/-
50-230
230-410
410-590
590-770
770-950
950-1130
1130-1310
1310-1490
1490-1670
1670-1850
1850-2030
2030-2210
2210-2390
2390-2570
2570-2750
2750-2930
Myofiber Cross Sectional Area range (um²)
d)
**
Number of myofibers/ mm²
600
500
400
300
200
WT
A2 -/-

CHAPTER 3. Evaluating the role of APOBEC2 in gene expression and RNA editing.

Thus far, through cell biological observations I have demonstrated that APOBEC2 has a role in differentiation to muscle cells because when deficient there is a delay in differentiation concurrent with an inability to properly exit cell cycle. To begin to assess how loss of APOBEC2 resulted in this phenotype, and keeping in mind the postulated role of this protein in CpG demethylation, I decided to look at transcriptional changes that would correlate with the cell biological findings.

Toward this, I performed poly (A)+ RNA high throughput sequencing (or mRNA-Seq). This method can be used to quantify the changing expression levels of each transcript (perhaps allowing me to establish transcriptional outcomes of the putative demethylase). It additionally provides information about the transcript's sequence at base pair resolution. Thus mRNA-Seq would allow me to investigate another possible role of this orphan deaminase in RNA editing. RNAseq analysis of APOBEC2 dependent changes would provide initial evidence toward which of the two potential molecular mechanisms (DNA demethylation versus RNA editing) established for other members of the AID/APOBEC family might be descriptive of APOBEC2 activity.

3.1 RNAseq analysis pipeline and validations

To generate libraries toward mRNA-seq, I first generated single cell clones (3 replicates) of C2C12 cells transfected with a control knockdown vector (shGFP). I also generated additional single cell clones (3 replicates) expressing a hairpin, which could efficiently knock down APOBEC2 (A2sh4).

I used the RNA to prepare stranded libraries using the TruSeq RNA-Seq library preparation protocol, which allows us to differentiate between sense and antisense transcription. Multiplexed samples underwent paired-end sequencing, which allows for sequencing in both ends of a fragment, thus improving the coverage of the transcripts and the likelihood of detecting splicing events or genomic rearrangements. In brief, as shown in the diagram in (Figure 3.1) the subsequent analysis pipeline involves filtering for reads that do not pass quality control measures, adaptor trimming and alignments of the reads to a reference genome and transcriptome. Only reads that uniquely map to exactly one location in the genome are considered for differential gene expression analysis. Reads are assigned to a specific genomic feature such as genes and counted for downstream gene expression analysis (Liao et al., 2013). The total number of counts for each gene is normalized by total library size for each sample before being used to determine gene expression fold change differences for each of the genes (Anders et al., 2012; Love et al., 2014). To estimate the relative abundance of a gene in a sample the number of read counts is normalized by the effective transcript length, which accounts for the length of the transcripts and the differences in read distribution along the transcript to compute transcripts per million (TPM). The latter represents the number of copies of a transcript that would be expected in a collection of one million transcripts and serves as a relative measure of the amount of transcript in the sample (Patro et al., 2017).

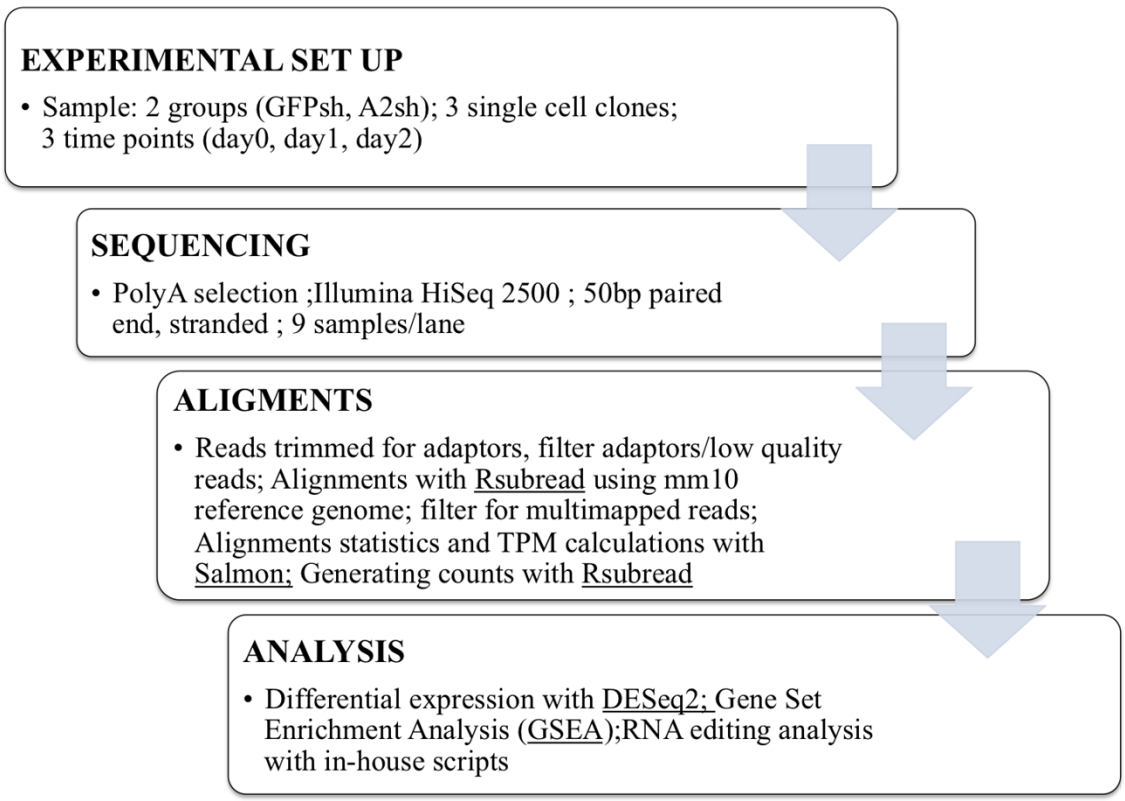


Figure 3.1 Schematic representation of the RNAseq analysis pipeline

The efficiency of ABOBEC2 knockdown for each of the single cell clones was evaluated using a western blot on the C2C12 lysates at different time points in differentiation medium. APOBEC2 itself was clearly expressed at the expected level for each clone analyzed. The results indicate cells differentiating as expected and a successful knockdown (Figure 3.2a). These single cell clones were induced to differentiate, and I collected RNA from all 6 replicates at day 0, 1, and 2 post differentiation. The days selected represent time points where the expression levels of APOBEC2 are highly changing from very low at day 0 to high at day 2. As a first quality control I assayed the levels of the *Apobec2* mRNA transcript (using the RNA-Seq analysis pipeline) and confirmed the expected knockdown of APOBEC2 (Figure 3.2b).

Moreover I compared the gene expression changes between the GFPsh control and C2C12s that are not expressing any shRNAs and no differences were detected (data not shown).

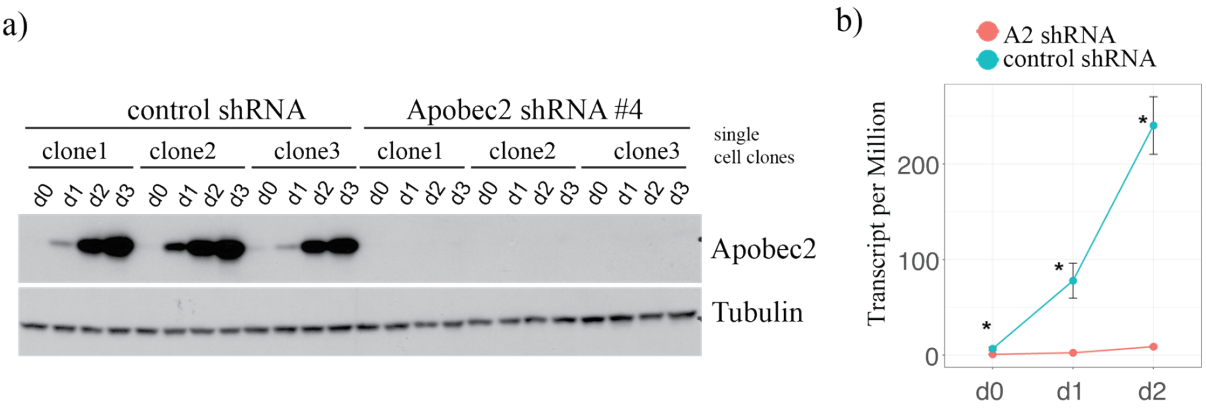


Figure 3.2 RNAseq analysis pipeline and validations

a) C2C12 lysates from single cell clones that were used in the RNAseq experiment in DM at day 0,1,2,3 were analyzed by WB using anti-APOBEC2 antibodies. alpha-tubulin was used as loading control. 3 clones were used for each group b) Graph representing the relative levels of APOBEC2 during d 0, 1, 2 time points, using the RNAseq data. Transcripts per million (TPM) is a measurement of the relative abundance of genes within the sample. Data represented as mean and SEM of the 3 biological replicates. * p-adjusted (as determined by DESeq2) < 0.1

3.2 Analysis of gene expression changes and pathway enrichment analysis during myogenesis

As a second quality control I identified genes and gene sets that change through differentiation in the control context. Because I already knew that APOBEC2 protein levels increase with differentiation from very low (day 0) to highly expressed (day 2), I chose day0, day1 and day2 as the relevant time points for the RNA seq analysis. Looking at the RNA-Seq data set from the control samples (GFPsh), I could identify many gene expression changes through differentiation. Specifically from day 0 to day 1 after

inducing differentiation there are a total of 3919 genes that are significantly differentially expressed (significance: adjusted p value <0.1 as defined from DESeq2) from which 1955 are up-regulated and 1964 are down-regulated. From day 1 to day 2 of the differentiation there are a total of 2421 significantly differentially expressed from which 1376 are upregulated and 1045 are downregulated (Figure 3.3).

To extract biological meaning from these RNA-Seq datasets, I used Gene Set Enrichment Analysis (GSEA), an approach that identifies groups of genes that share common biological function or regulation, or pathways that are coordinately changing due to APOBEC2 knockdown (Mootha et al., 2003; Subramanian et al., 2005). The advantages of the method are that (1) produces results that can more easily be interpreted in regard to the relevant biological processes (2) it is designed to detect subtle but coordinated changes in the expression of a group of functionally related genes that would be missed otherwise. The strength of this method also lies in the availability the Molecular Signatures Database (MSigDB), which contains a large number of repositories of gene sets (extracted from research publications or other specialized resources), acquired through manual and computational means (Liberzon et al., 2011). MSigDB includes the "Hallmark" gene sets, that forms a refined collection of gene sets that is computationally defined and manually curated and annotated, thus reducing potential variation and redundancy by summarizing common information across multiple gene sets (Liberzon et al., 2015).

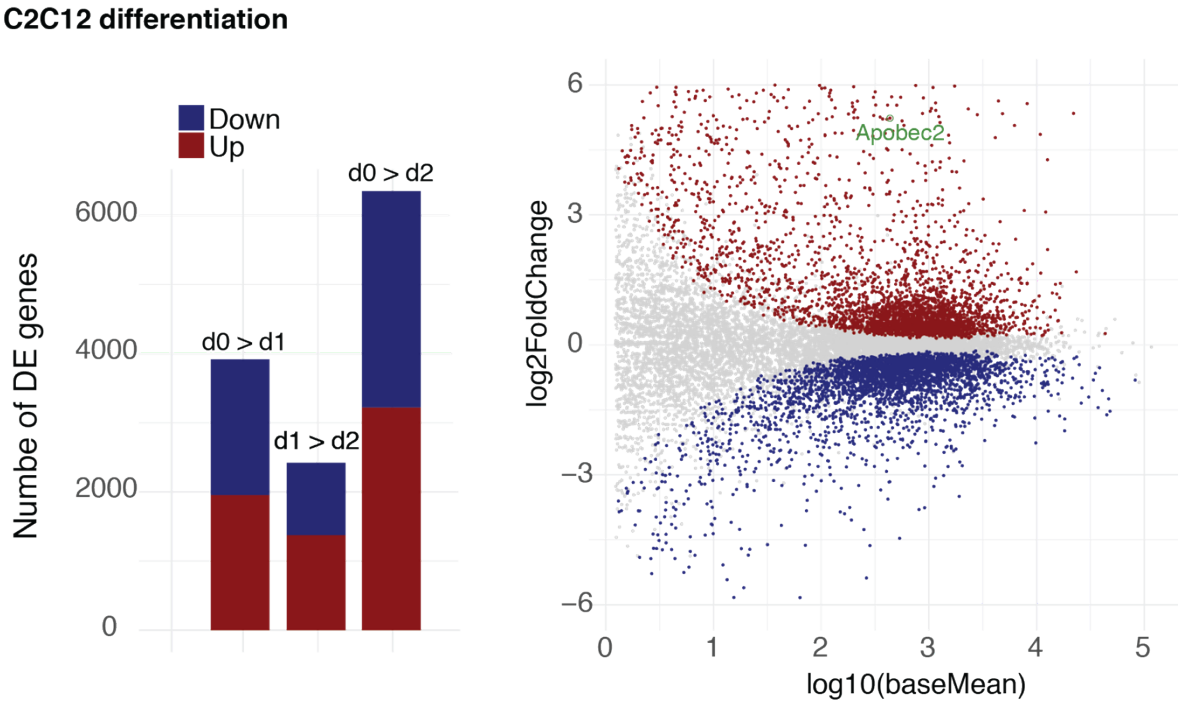


Figure 3.3 Gene expression changes of myogenesis

Barplot on the left shows the number genes significantly up/downregulated through differentiation. Scatter plot on the right shows the log2 fold changes between da0 and day2 time point, over the log10 of the mean of normalized counts for all the sample. Significantly differentially expressed genes with p adjusted value <0.1 are shown in red (upregulated) or blue (downregulated) and not significantly differentially expressed genes in gray. APOBEC2 is shown in green

To determine what biological pathways are important during the differentiation process (at the chosen time points), I examined what MSigDB *Hallmark* gene sets are enriched in the list of genes that are differentially expressed using GSEA. Some of the top significant gene sets are listed in Table 3.1. They include groups of genes important in myogenesis (HALLMARK_MYOGENESIS gene set), which are enriched in the list of genes that are upregulated during day0 to day2 transition confirming the importance of this gene set network during C2C12s differentiation in cell culture (Figure 3.4a). Interestingly genes encoding targets of E2F transcription factors (HALLMARK_E2F_TARGETS), important for the regulation of cell cycle are enriched in the list of genes whose expression is down regulated through differentiation, during time points day 0 to day2 (Figure 3.4b). The E2F transcription factors have been widely shown to be required for the regulation of DNA replication and cell cycle regulation (Bracken et al., 2004), which are in turn important requirements for the progression of proper differentiation for muscle fibers. More statistically significant gene sets in the list of genes that are upregulated during differentiation include: MYC targets and metabolic gene sets, while gene sets that are enriched in genes downregulated during differentiation include pathways in cell cycle checkpoints and interferon alpha and gamma response (Table 3.1).

One of the advantages of the MSigDB is that it contains many gene sets that represent many biological processes, but this also brings challenges. For example, the growth of the number of the gene sets has also lead to the problem of (1) redundancy (e.g. gene sets with many overlapping genes can dominate the top of the significant sets hiding other relevant hits) and (2) heterogeneity (a gene set not always behaving

consistently based on biological context or due to poor biological resolution). So in addition to utilizing the '*Hallmark'* gene sets (Liberzon et al., 2015), and to better establish the most biologically relevant results while reducing gene set heterogeneity I decided to create a list of customized gene sets. Toward that, from the MSigDB gene sets I extracted the subset of genes that are significantly unregulated, downregulated or non-significant (ns) during the differentiation process of control C2C12 cells generating C2C12-specific pathways. I used these, to determine the C2C12 differentiation-related gene sets significantly enriched due to APOBEC2 knockdown.

Table 3.1 List of Gene Sets significantly changing during myogenesis

Gene lists, ranked by expression changes during differentiation from d0 to d2, were fed to the GSEA. Top significant gene sets from MSigDB Hallmark database are shown. Gene listed are only those with adjusted p value <0.1. The false discovery rate (FDR) is the estimated probability that a gene set with a given NES represents a false positive finding.

NAME	GENES	NES	FDR
HALLMARK_E2F_TARGETS	Tfrc, Cdkn1a, Ipo7, Cbx5, Rad51c, Mthfd2, Prdx4, Msh2, Rpa1, Tbrg4, Tubb5, Rfc3, Tk1, Pa2g4, Lig1, Slbp, E2f8, Dut, Eif2s1, Tcf19, Prim2, Rfc2, Cit, Rad50, H2afx, Shmt1, Atad2, Cks1b, Melk, Cdkn2c, Mcm6, Aurkb, Mcm4, Hn1, Spc25, Donson, Chek2, Smc4, Cdkn1b, Zw10, Wee1, Nup153, Cdk1, Birc5, Hmgb2, Ube2t, Tmpo, Wdr90, Cse1l, Mxd3, Nop56, Tipin, Plk1, Rfc1, Dek, Ctps, Pold1, Kif2c, Stmn1, Timeless, Gins1, Aurka, Ccp110, Trip13, Rrm2, Dck, Gins3, Brca2, Tra2b, Ccne1, Pold3, Lbr, Brca1, Cdca3, Tacc3, Bub1b, Plk4, Cdc20, Orc2, Mybl2, Mki67, Pcna, Spag5, Mcm3, Hmmr, Lmnb1, Mcm5, Ctcf, Racgap1, Diaph3, Ranbp1, Xpo1, Dnmt1, Cdca8, Cenpe, Gspt1, Mms22l, Rpa3, Chek1, Spc24, Hnrnpd, Top2a, Depdc1a, Cdk4, Kif22, Cks2, Hus1, Pold2, Bard1, Mcm7, Phf5a, Ube2s, Pnn, Pds5b, Cdc25a, Hmga1, Tubg1, Paics	-6.57	<0.001
HALLMARK_G2M_CHECKPOINT	Slc12a2, E2f2, Hoxc10, Top1, Meis2, Slc7a5, Wrn, E2f4, Ccnd1, Slc7a1, Kif5b, Dmd, Efna5, Prmt5, Gins2, Notch2, Odc1, Ccnt1, Pml, Prim2, Abl1, G3bp1, Mapk14, Exo1, H2afx, Smarcc1, Pbk, Hif1a, Dkc1, Cul5, Bcl3, Cks1b, Cdkn2c, Mcm6, Aurkb, Cul4a, E2f1, Hn1, Smc4, Cdkn1b, Pafah1b1, Cdk1, Birc5, Tmpo, Tnpo2, Plk1, Stil, Odf2, Cdc45, Ndc80, Ncl, Kif2c, Stmn1, Rbl1, Mtf2, Aurka, Dbf4, Rbm14, Cul1, Cdc27, Brca2, Smc2, Tra2b, Ccnf, Kif15, Knl1, Lbr, Chaf1a, Tacc3, Plk4, Cdc20, Mybl2, Kif11, Mki67, Prpf4b, Prc1, Mcm3, Mnat1, Hmmr, Lmnb1, Mcm5, Ythdc1, Ctcf, Racgap1, Rps6ka5, Xpo1, Upf1, Kif20b, Cenpe, Gspt1, Chek1, Egf, Fbxo5, Sap30, Hnrnpd, Top2a, Cdk4, Kif22, Cks2, Rad54l, Hus1, Cdc7, Bard1, Incenp, Ube2c, Ube2s, Ttk, Pds5b, Cdc25a, Hmga1, Bub1, Arid4a, Nek2	-4.78	<0.001
HALLMARK_INTERFERON_ALPHA_RESPONSE	Ripk2, Helz2, Cd47, Ifitm2, Ifitm3, Mov10, Ube2l6, Trim25, Mvb12a, Parp12, Csf1, Wars, Ncoa7, Psme1, Casp8, Elf1, Tap1, Il7, B2m, Trim26, Procr, Uba7, Eif2ak2, Psme2, Ifi35, Trim21, Rtp4, Parp14, Irf1, Lgals3bp, Isg20, Parp9, Ddx60, Fam46a, Adar, Gmpr, Cd74, Lap3, Ifit2, Irf2, Ifi30, Irf9, Ifih1	-4.20	<0.001
HALLMARK_INTERFERON_GAMMA_RESPONSE	Ripk2, Cdkn1a, Pnp, Samhd1, Helz2, Vamp5, Sri, Nfkbia, Btg1, Irf5, Ifitm2, Ifitm3, Mthfd2, Itgb7, Ube2l6, Isoc1, Trim25, Nampt, Ripk1, Mvp, Parp12, Wars, Cmklr1, Sod2, Pml, Psme1, Jak2, Casp8, Ddx58, Tap1, Il7, Socs1, Eif4e3, B2m, Il6, Trim26, Hif1a, Fas, Nod1, Gch1, Vamp8, Vcam1, Myd88, Eif2ak2, Psme2, H2-Q7, H2-D1, Ifi35, Csf2rb, Pde4b, C1ra, Trim21, Rtp4, Stat3, Parp14, Irf1, Ncoa3, Upp1, Znfx1, Lgals3bp, Pla2g4a, Isg20, Stat1, Zbp1, Tapbp, Ddx60, C1rb, Bpgm, Adar, H2-M3, St3gal5, Rapgef6, Ptpn2, Cd74, Lap3, Ifit2, Il2rb, Irf2, Ifi30, H2-K1, Socs3, Nfkb1, Irf9, Ifih1, Auts2, Arl4a	-3.94	<0.001

NAME	GENES	NES	FDR q-val
HALLMARK_MYOGENESIS	Itga7, Pygm, Tnnc2, Myl6b, Atp2a1, Pgam2, Ckm, Acta1, Erbb3, Casq2, Tnnt3, Actn2, Myom2, Tnni2, Chrng, Myh2, Myh1, Mylpf, Tnnt2, Tead4, Ache, Myh3, Tnnc1, Ryr1, Myom1, Dtna, Cox6a2, Mybph, Mef2a, Ncam1, Ckb, Actc1, Myh8, Sgcg, Fst, Mef2c, Tnni1, Pfkm, Myog, Mef2d, Akt2, Cav3, Nqo1, Myl1, Hspb2, Csrp3, Klf5, Cdkn1a, Myh7, Mb, Myh4, Myl4, Tcap, Gadd45b, Eno3, Bin1, Ankrd2, Chrnb1, Cryab, Pkia, Tnnt1, Pde4dip, Fabp3, Nav2, Wwtr1, Agl, Svil, Camk2b, Cd36, Eif4a2, Myl2, Chrna1, Igfbp7, Ldb3, Rb1, Myoz1, Gsn, Kifc3, Dmpk, Bhlhe40, Casq1, Mapk12, Acsl1, Gja5, Reep1, Dmd, Notch1, Adam12, Schip1, Prnp, Hbegf, Ptp4a3, Bag1, Des, Cacng1, Hspb8, Stc2, Sirt2, Large1, Dapk2, Myh9, Ak1, Tpm3, Myo1c, Apod, Ckmt2, Aebp1, Ablim1, Pcx, Flii, Cdh13, Cacna1h, Sh2b1, Efs, Gnao1, Ephb3, Syngr2, Sgca, Myf6, Kcnh1, Fgf2, Pvalb, Igfbp3, Adcy9, Fdps, Pdlim7, Sptan1, Mras, Crat, Fxyd1, Foxo4, Clu, Pick1, Tsc2, Col4a2	7.79	<0.001
HALLMARK_OXIDATIVE_PHOSPHORYLATION	Ndufs4, Timm10, Pdhx, Rhot1, Echs1, Nqo2, Atp6v1d, Slc25a11, Cox15, Atp5d, Tomm70a, Isca1, Dld, Afg3l2, Lrpprc, Acadm, Idh1, Cyb5a, Cyc1, Timm9, Ndufs1, Opa1, Aco2, Cox6a1, Sucla2, Pdk4, Hspa9, Oat, Atp1b1, Ndufs2, Atp5g1, Atp6v1e1, Ndufs3, Ndufs8, Pdp1, Ndufv1, Cox17, Atp5o, Bax, Mrpl15, Uqcrfs1, Acaa1a, Ldha, Glud1, Alas1, Cpt1a, Mtx2, Iscu, Uqcrq, Pdhb, Prdx3, Ndufab1, Ndufa6, Idh3g, Idh3a, Pdha1, Mrps22, Atp6v0b, Htra2, Suclg1, Uqcr11, Timm17a, Mtrr, Ndufa3, Ndufa7, Supv3l1, Aifm1, Tomm22, Pmpca, Retsat, Uqcrc2, Phyh, Hccs, Uqcr10, Atp5g3, Slc25a5, Sdhd, Ndufs7, Hadha, Cycs, Abcb7, Cox6b1, Fdx1, Bdh2, Phb2, Atp5j, Hsd17b10, Atp6v1g1, Polr2f, Atp5a1, Mdh1, Cox5a	4.55	<0.001
HALLMARK_MYC_TARGETS_V2	Rrp12, Nip7, Gnl3, Rrp9, Sord, Mybbp1a, Grwd1, Wdr43, Ppan, Pes1, Aimp2, Bysl, Nop16, Wdr74, Slc29a2, Pus1, Nop2, Noc4l, Mphosph10, Phb, Pprc1, Tbrg4, Tfb2m, Pa2g4, Utp20, Ipo4, Mrto4, Ndufaf4, Mcm4, Ddx18, Farsa, Las1l, Nop56, Plk1, Supv3l1, Imp4, Cbx3, Plk4, Srm, Hk2, Mcm5, Map3k6, Cdk4	3.86	<0.001
HALLMARK_MYC_TARGETS_V1	Gnl3, Rrp9, Nhp2, Rsl1d1, Ddx21, Aimp2, Tomm70a, Iars, Cdk2, Nop16, Prdx4, Eif2s2, C1qbp, Eprs, Ywhaq, Cyc1, Abce1, Xpot, Phb, Eif3d, Eif3j1, Cct5, Etf1, Odc1, Pa2g4, Pabpc4, Cops5, Dut, Eif2s1, Tyms, G3bp1, Ldha, Ppia, Smarcc1, Eif3b, Ncbp2, Mcm6, Eef1b2, Mcm4, Ap3s1, Prdx3, Ddx18, Mrps18b, Cnbp, Ndufab1, Stard7, Cct3, Rplp0, Eif4g2, Pwp1, Psmb3, Nop56, Psmd7, Dek, Cdc45, Psmc6, Ctps, Psmc4, Sf3b3, Apex1, Cul1, Uba2, Prps2, Cbx3, Pgk1, Tra2b, Ncbp1, Psmd8, Cdc20, Cct4, Orc2, Srm, Ssbp1, Eif1a, Pcna, Psmd1, Mcm5, Rpl14, Ranbp1, Xpo1, Pabpc1, Acp1, Gspt1, Rps2, Hnrnpd, Psmd3, Cdk4, Rfc4, Pold2, Phb2, Mcm7, Mrpl9, Cox5a	3.11	<0.001

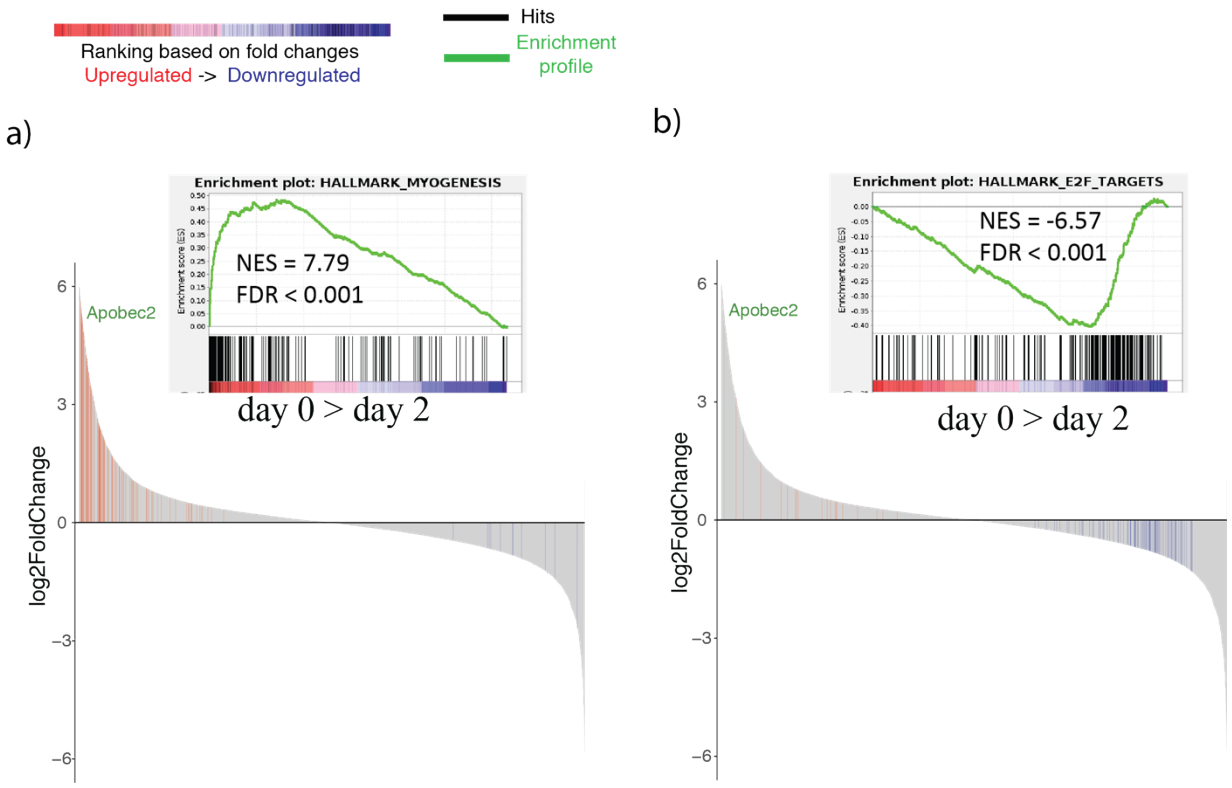


Figure 3.4 Gene set enrichment analysis for gene expression changes during myogenesis

GSEA score curves of the RNA-seq output. Two genesets that were significantly enriched through C2C12 differentiation are shown. False Discovery Rate (FDR) is calculated by comparing the actual data with 1000 Monte-Carlo simulations. Black bars represent the position of members of the category in the ranked list together with the running enrichment score (plotted in green). Bar plots below shows log2 fold changes between day0 and day2 time point for all expressed genes (grey). Highlighted are genes that are part of the geneset and upregulated (red) or downregulated (blue) through differentiation. In green is APOBEC2 a) Myogenesis signature genes enriched in the list of genes that are significantly upregulated during differentiation b) E2F targets enriched in the list of genes that are significantly downregulated during differentiation

3.3 Identification of APOBEC2 dependent gene expression changes

Having derived a list of gene expression changes that correlate with the differentiation of C2C12 cells into myotubes, I then defined the cohort of gene expression changes that were likely APOBEC2 dependent.

As expected, mRNA levels for APOBEC2 itself were clearly downregulated in the APOBEC2-knockdown cells. In addition there were many other genes that were also significantly differentially expressed with adjusted p-value < 0.1. That list includes: for day0, 878 genes (330 unregulated and 548 downregulated); for day1, 2505 genes (1106 unregulated and 1399 downregulated); for day2, 2525 genes (1118 unregulated and 1407 downregulated) (Figure 3.5). To determine the biological pathways affected by APOBEC2 knockdown, I ran GSEA on the list of genes that are differentially expressed at any of the time points and used the customized *Hallmark* MSigDB gene sets (the C2C12 differentiation specific ones which were generated as described above). Some of the top gene sets that are affected due to APOBEC2 knockdown are listed in Table 3.2).

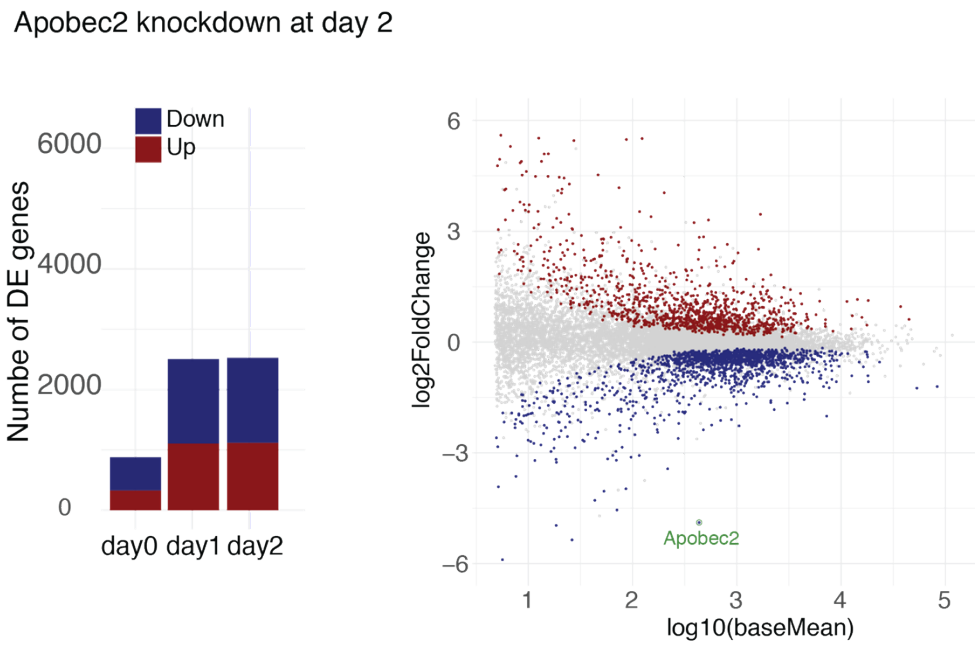


Figure 3.5 Lack of APOBEC2 in C2C12 cells leads to gene expression changes

Barplot on the left shows the number genes significantly up/downregulated when we knockdown APOBEC2. Scatter plot on the right shows the log2 fold changes due to APOBEC2 knockdown at day 2 after differentiation, over the log10 of the mean of normalized counts for all the samples. timepoint. Significantly differentially expressed genes with p adjusted value <0.1 are shown in red (upregulated) or blue (downregulated) and not significantly differentially expressed genes in gray. APOBEC2 is shown in green

Table 3.2 List of Gene Sets significantly changing due to APOBEC2 knockdown

Gene lists, which were ranked based on expression changes due to APOBEC2 knockdown at day2, were fed to the GSEA. Top significant gene sets from MSigDB Hallmark database are shown. Gene listed are only those with adjusted p value <0.1. The false discovery rate (FDR) is the estimated probability that a gene set with a given NES represents a false positive finding.

NAME	GENES	NES	FDR
HALLMARK_MYOGENESIS_D2D0DEG_UP	Itga7, Stc2, Pfkm, Pde4dip, Dtna, Ncam1, Gadd45b, Sgcg, Bhlhe40, Fst, Prnp, Hbegf, Cdkn1a, Myoz1, Myh4, Eno3, Atp2a1, Ckb, Mef2d, Wwtr1, Myh2, Tead4, Syngr2, Gja5, Pgam2, Sgca, Actn2, Myl6b, Des, Tsc2, Bin1, Mef2a, Schip1, Ptp4a3, Tnni2, Acta1, Adam12, Cdh13, Akt2, Ak1, Myh7, Tnnt3, Tnnc2, Mb, Ryr1, Myom1, Myh1, Pick1	-6.55	<0.001
HALLMARK_OXIDATIVE_PHOSPHORYLATION_D2D0DEG_UP	Rhot1, Ndufs4, Atp5d, Atp5o, Atp6v1d, Pmpca, Tomm70a, Cox17, Ndufs3, Aco2, Tomm22, Atp5j, Cox5a, Dld, Afg3l2, Isca1, Slc25a5, Atp5g1, Aifm1, Mtrr, Uqcr11, Atp5a1, Sdhd, Uqcr10, Cyc1, Iscu, Mdh1, Timm10, Timm9	-5.40	<0.001
HALLMARK_MYC_TARGETS_V1_D2D0DEG_UP	Tomm70a, Eif2s1, Rpl14, Smarcc1, Eprs, Cox5a, Pa2g4, Eif3d, Eif2s2, Rsl1d1, Cct5, Eif3b, Abce1, Eif3j1, Rrp9, Eif1a, Xpot, Eif4g2, Pwp1, Psmd3, Cyc1, Gspt1, Aimp2, Ncbp2, Psmb3, Nop16, Phb, Ctps, Nhp2	-4.87	<0.001
HALLMARK_TNFA_SIGNALING_VIA_NFKB_D2D0DEG_UP	Junb, Egr3, Egr2, Mcl1, Pmepa1, Lif, Nr4a1, Ier2, Egr1, Klf10, Slc2a6, Trib1, Tnc, Gadd45b, Bhlhe40, Hbegf, Btg1, Cdkn1a, Bcl6, Rhob, Spsb1, Fos, Fosb, B4galt5, Rcan1, Panx1, Hes1	-4.72	<0.001
HALLMARK_MTORC1_SIGNALING_D2D0DEG_UP	Tfrc, Egln3, Atp2a2, Atp6v1d, Acsl3, Bhlhe40, Eprs, Wars, Cdkn1a, Hspa4, Eif2s2, Mllt11, Psmc2, Abcf2, Rrp9, Pno1, Atp5g1, Uso1, Shmt2, Cct6a, Psmd13, Psmd12, Ufm1	-4.56	<0.001
HALLMARK_MYC_TARGETS_V1_D2D0DEG_NS	Hspd1, Psma1, Ptges3, Hdac2, Psma6, Lsm7, Ube2l3, Ifrd1, Psma4, Kpnb1, Syncrip, Cct2, Snrpd3	-3.79	<0.001
NAME	**GENES**	**NES**	**FDR**
HALLMARK_INTERFERON_GAMMA_RESPONSE_D2D0DEG_DOWN	Helz2, Parp12, Ube2l6, Pml, Ifih1, Samhd1, Trim21, Psme1, C1ra, Eif2ak2, Tap1, Pnp, C1rb, Trim26, Nfkbia, Lgals3bp, Parp14, Tapbp, Ripk2, Stat1, Ddx58, H2-M3, Casp8, Sri, Vcam1, Psme2, Adar, Nfkb1, Irf9, B2m, Isg20, Socs1, Upp1, H2-Q7, Ncoa3, Cd74	5.36	<0.001
HALLMARK_INTERFERON_ALPHA_RESPONSE_D2D0DEG_DOWN	Helz2, Parp12, Ube2l6, Ifih1, Trim21, Psme1, Eif2ak2, Tap1, Uba7, Trim26, Lgals3bp, Parp14, Mov10, Ripk2, Elf1, Casp8, Psme2, Parp9, Adar, Irf9, B2m, Isg20, Csf1, Cd47, Cd74	5.17	<0.001
HALLMARK_XENOBIOTIC_METABOLISM_D2D0DEG_DOWN	Ptgr1, Aldh9a1, Aldh3a1, Aldh2, Atoh8, Ugdh, Car2, Nqo1, Pdk4, Idh1, Ets2, Casp6, Il1r1, Bphl, Xdh, Crot, Ddah2, Dcxr, Epha2, Maoa, Upp1, Dhrs1, Retsat	4.61	<0.001
HALLMARK_P53_PATHWAY_D2D0DEG_DOWN	Vwa5a, Tap1, Tspyl2, Eps8l2, Apaf1, Tm7sf3, S100a10, Ddb2, Ptpn14, Cebpa, Socs1, Hdac3, Dcxr, Epha2, Upp1, Ada, Retsat	4.32	<0.001
HALLMARK_G2M_CHECKPOINT_D2D0DEG_DOWN	Ccnd1, Pml, Efna5, Tmpo, Wrn, Tra2b, Cdkn1b, Exo1	4.27	<0.001
HALLMARK_MITOTIC_SPINDLE_D2D0DEG_DOWN	Cep250, Fscn1, Hdac6, Cd2ap, Net1, Arhgef11, Llgl1, Synpo, Tubgcp2, Alms1, Farp1, Arl8a	4.16	<0.001
HALLMARK_E2F_TARGETS_D2D0DEG_DOWN	Tmpo, Chek2, Tra2b, Cdkn1b	4.01	<0.001

The genes that are involved in development of skeletal muscle during myogenesis and that are significantly upregulated during C2C12 differentiation were enriched in the list of genes that are downregulated at day2 due to APOBEC2 knockdown. This is in accordance with the phenotype observed of a delay in differentiation. Furthermore, the subset of gene targets of E2F transcription that are significantly downregulated during differentiation are enriched in the list of genes that are upregulated due to APOBEC2 knockdown at day2. A similar inverse pattern is observed with genes important in the G2/M checkpoint (Figure 3.6a). Therefore APOBEC2 expression correlates with cell cycle regulation and myogenesis.

Given the strong correlation between the gene sets that we've identified whose expression was being modulate in the absence of APOBEC2, and the cellular phenotypes we observed (e.g. inability to properly exit cell cycle in the absence of APOBEC2), we decided to take a closer look at the genes of the E2F target gene set and the Myogenesis geneset. Interestingly *Myog* and p21/*cdkn1a*, crucial during the induction of myogenesis and cell cycle withdrawal are both downregulated due to APOBEC2 deficiency and could account at least partially for the phenotypes observed (Figure 3.6b).

Overall, deficiency in APOBEC2 leads to substantial gene expression changes upon initiation of differentiation (and even prior) affecting gene networks important in myogenesis and cell cycle regulation. This result is unique to other cytidine deaminases knockouts, none of which lead to alterations of gene expression levels (Salter et al., 2016). The rest of the thesis will be focused on our attempt to determine how these gene expression changes come about.

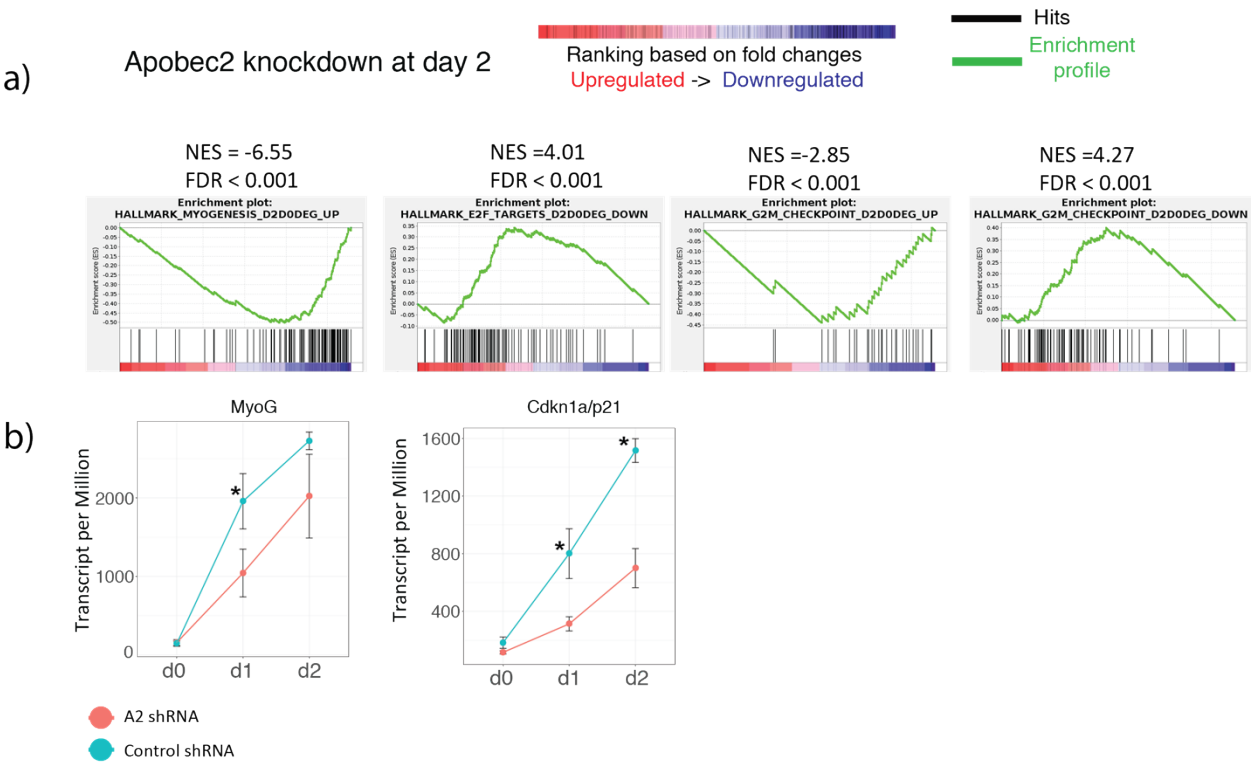


Figure 3.6 Lack of APOBEC2 in C2C12 cells leads to changes in gene networks

a) Gene set enrichment analysis (GSEA) score curves of the RNAseq output. Groups of genes that normally are upregulated as a set (Myogenesis Gene set) are enriched in the list of genes downregulated when APOBEC2 is deficient. Groups of genes that normally are downregulated as a set (E2F target Gene set) are enriched in the list of genes upregulated when APOBEC2 is deficient. While for the G2M gene set, the subset of genes that go up is enriched in the list of genes that are downregulated while the subset of genes that go down with differentiation are enriched in the list of genes that are upregulated due to APOBEC2 knockdown b) Graph representing the amount of Myog and *Cdkn1a*/p21 using the RNAseq data. Transcripts per million (TPM) is a measurement of the relative proportion of transcripts within the sample. Data represented as mean and SEM of the 3 biological replicates.

3.4 Identification of APOBEC2 dependent RNA editing

An RNA editing event is one where the RNA's sequence is changed, excluding changes due to processes such as RNA splicing and alternative polyadenylation. The sequence is altered through deamination, converting an A to an I or a C to a U. A-to-I editing (catalyzed by the ADAR family of deaminases) has been correlated to loss of RNA stability leading to apparent gene expression differences (Wang et al., 2013). In contrast, C-to-U editing (catalyzed by the AID/APOBEC family member APOBEC1) does not lead to gene expression changes at the level of mRNA but seems important in regulating genes by modulating translation (Chen et al., 1987; Powell et al., 1987; Rayon-Estrada et al., 2017; Rosenberg et al., 2011). Since certain proteins which contain cytidine deaminase motifs mediate A to I rather than C to U RNA editing (Rubio et al., 2007), I decided to assess if APOBEC2 could function as an RNA editor (on A or C)

The RNA-Seq dataset described in the previous sections also contains information about the mRNA transcripts at single nucleotide resolution, thus making it optimal for answering this question in a rigorous manner. I used the same analysis pipeline that has been previously described (Harjanto et al., 2016) to examine whether there are any APOBEC2 dependent RNA editing events (either C-to-U or A-to-I) in the C2C12 differentiation model. In brief, we identify putative RNA editing events by searching for single nucleotide mismatches to the reference genome. The mismatch datasets are then filtered against several parameters (based on sequencing depth and editing rates). The analysis is run using different filter parameters ranging from no filters at all to very stringent filters. The ideal potential RNA editing hits are sites that are present in control mRNA sequences but absent from deaminase-deficient mRNA sequences. Importantly,

the inverse comparison is also performed; the number of editing events that were present in the deaminase-deficient (APOBEC2 knockdown) sample but not in the control sample. This inverse comparison provides a measurement of the background noise of the technique (as a result of mismapping or reference genome sequence differences). This also allows for computation of an inferred false positive rate (IFPR), as number of events in the deaminase-deficient sample / number of events in the control sample. The number of the putative RNA editing hits being similar or lower than the number of the inverse hits suggests the presence of noise and lack of editing.

The RNA editing analysis showed that the number of putative editing sites detected for A-to-I or C-to-U in any of the time points (control vs. APOBEC2 knockdown comparison), is not significantly different from the sites detected in the inverse analysis (APOBEC2 knockdown vs. controls comparison). Figure 3.7 shows the numbers of hits for the minimally filtered analysis. By comparison APOBEC1$^{+/+}$ versus APOBEC1$^{-/-}$ analysis shows generally higher number of hits compared to the inverse analysis (APOBEC1$^{-/-}$ versus APOBEC1$^{+/+}$) even when stringent filters are applied. If A-to-I or C-to-U editing had a dependence on APOBEC2, we would expect fewer hits consistently in the inverse analysis. The latter is not what we observe.

Depending on the time point analyzed and the stringency of the filters used for APOBEC2- dependent C-to-U editing the IFPR rate was between 98-107%, while for A-to-I editing the IFRP rate was between 50% and 200%. On the other side, using the same analysis pipeline on the APOBEC1 dependent C-to-U editing, the IFRP rate was between 2-4%. This suggests that the few potential editing sites that were picked up in our analysis are most likely to be background noise.

Based on these data, we can conclude that in the C2C12 differentiation model there are no detectable C-to-U or A-to-I RNA editing events dependent on APOBEC2, in confirmation of a recent publication which demonstrates that muscle is one of the few tissues with very low levels of A-to-I RNA editing (Tan et al., 2017). Given our analysis, it is highly unlikely that APOBEC2 is acting as an A-to-I and C-to-U RNA editing enzyme during C2C12 differentiation.

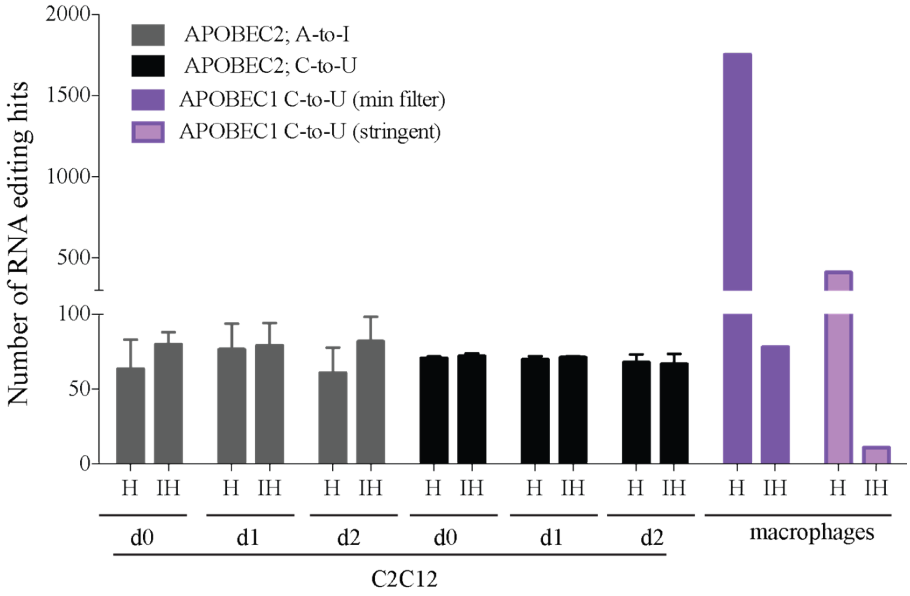


Figure 3.7 Evaluation APOBEC2's potential role in RNA editing

Candidate RNA editing sites called from APOBEC2 knockdown samples, control (GFPsh) at day 0, 1, and 2 in DM in C2C12s and wild-type and APOBEC1$^{-/-}$ macrophages. Putative hits (A-to-I and C-to-U) undergo various filtering parameters. For C2C12s, the minimum filtered sites are shown; for macrophages both minimum and stringently filtered sites are shown. Hits (H) represent candidate editing sites present in control (GFPsh) but not in APOBEC2 knockdown dataset, or hits present in the wild-type but not in the APOBEC1-/- dataset. Inverse hits (IH) represent hits yielded when the inverse comparison is made, thus edit sites present in the APOBEC2 knockdown dataset but not in the control (GFPsh) and edit sites that are present in APOBEC1$^{-/-}$ dataset but not in wild- type. Data are represented as means ± SD using outputs of 3 RNA-Seq datasets.

CHAPTER 4. Assessing the involvement of APOBEC2 in methylome changes

As demonstrated in the previous chapter, the transcriptome analyses showed that there are many APOBEC2 dependent gene expression changes during myogenesis in C2C12s. Furthermore, I showed that these changes are unlikely to result from RNA editing activity. I wanted to assess whether loss of APOBEC2 correlated with changes in DNA methylation.

The DNA mark 5mC has been widely described as an important regulator of transcription and is very important in mammalian development (Smith and Meissner, 2013). More recently, 5hmC has been discovered as significant in epigenetic regulation (Branco et al., 2012). Interestingly AID, a member of the AID/APOBEC family of enzymes has been implicated in multiple studies in gene-specific DNA demethylation (Ramiro and Barreto, 2015). Previous data in zebrafish (Rai et al., 2008) and in an embryonic stem cell (ES) -derived myogenic progenitors model (Carrió et al., 2016) had also connected APOBEC2 with active 5mC demethylation, while overexpression experiments demonstrated some ability to mediate 5hmC demethylation when expressed in HEK293 cells (Guo et al., 2011). In contrast, other studies do not support the idea that APOBEC2 mediates its biological role through active 5mC demethylation during regeneration in zebrafish (Powell et al., 2013, 2014) or in biochemical studies (Nabel et al., 2012). However, whether APOBEC2 is capable of directly mediating active DNA demethylation through cytidine deamination under natural conditions, remained an open question.

To determine whether the function of APOBEC2 during differentiation could be mediated through the demethylation of 5mC we wanted to identify changes in the levels

of modified cytosines in the DNA within C2C12 cells as they develop from myoblasts to myotubes. This is an ideal system to study a possible role of APOBEC2 in active DNA demethylation because differentiation requires cell cycle arrest. Thus any potential APOBEC2-dependent changes in amounts of modified cytosines could be linked to *active* DNA demethylation (as opposed to passive changes propagated through cell division). I therefore proceeded to compare the genome wide cytosine modifications between control C2C12s to those where APOBEC2 is knocked down at different time points during the differentiation.

4.1 RRBS analysis pipeline and validations

Methods of methylome analysis (*reviewed in* (Plongthongkum et al., 2014)) involve genomic DNA undergoing bisulfite conversion, where cytosines (C) are converted to thymines (T), but modified cytosines (such as 5hmC and 5mC) are left unchanged. The bisulfite converted DNA is then sequenced and mapped to a reference genome. The percent of non-converted cytosines (modified Cs) at a given site is calculated as the ratio of reads containing Cs over the total number of reads covering the specific site. To identify genome scale methylation changes we choose to use enhanced Reduced Representation Bisulfite Sequencing (eRRBS). This method is a modified version of the original RRBS approach (Gu et al., 2011; Meissner et al., 2005) and leads to an increase in the detection and coverage of CpG sites (Akalin et al., 2012a; Garrett-Bakelman et al., 2015). RRBS includes an initial step that enriches for CG-rich genomic fragments by cutting genomic DNA with the MspI (C^CGG) restriction enzyme, a methylation-insensitive enzyme (thus making it ideal to unbiasedly cut fragments containing modified

or unmodified Cs) resulting in short fragments that contain CpG dinucleotides at the ends. This is followed by size selection of the fragments, bisulfite conversion, PCR amplification and sequencing. eRRBS on the other hand has improved the original RRBS method by increasing the number of CpGs detected and improving coverage of genomic regions.

The advantages of RRBS and eRRBS compared to whole genome bisulfite sequencing (WGBS) is the increase in coverage of the CpG dense regions (such as promoters and CpG islands), the depth of sequencing required is less, thus making this choice more cost effective while also providing single-nucleotide resolution of the well covered areas. The limitation of the method is that the coverage in CpG poor regions is still low, limiting the analysis and interpretation only to the well-covered regions. Another limitation of the method, which in our experiment serves to our advantage, is that RRBS does not discriminate between 5mC and other DNA modifications such as 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC). This would allow us to simultaneously examine cytosine modification changes but without distinguishing among them.

The RRBS analysis pipeline (Figure 4.1) outlines the steps taken for this analysis. We used 5 different single cell clones for each group that is being compared (C2C12s expressing GFPsh as a control and A2sh#4 for knocking down *Apobec2*) at 3 different time points after differentiation. eRRBS library preparation and alignments of the reads were done by the Weill Cornell Medical College Epigenomics Core as previously described (Garrett-Bakelman et al., 2015). In brief, multiplexed samples underwent single-end sequencing. Sequencing reads then passed through quality control filters and

adaptor trimming. These reads were then aligned to a bisulfite converted reference mouse genome with specialized mapping tools that allow for mapping of bisulfite converted reads (Krueger and Andrews, 2011). The methylation context for each cytosine was determined with scripts from the core facility. Then we analyze modification differences of well-covered CpG sites using published and validated bioconductor packages (Akalin et al., 2012b; Li et al., 2013). For more details about the methylation analysis pipeline refer to the Materials and Methods chapter.

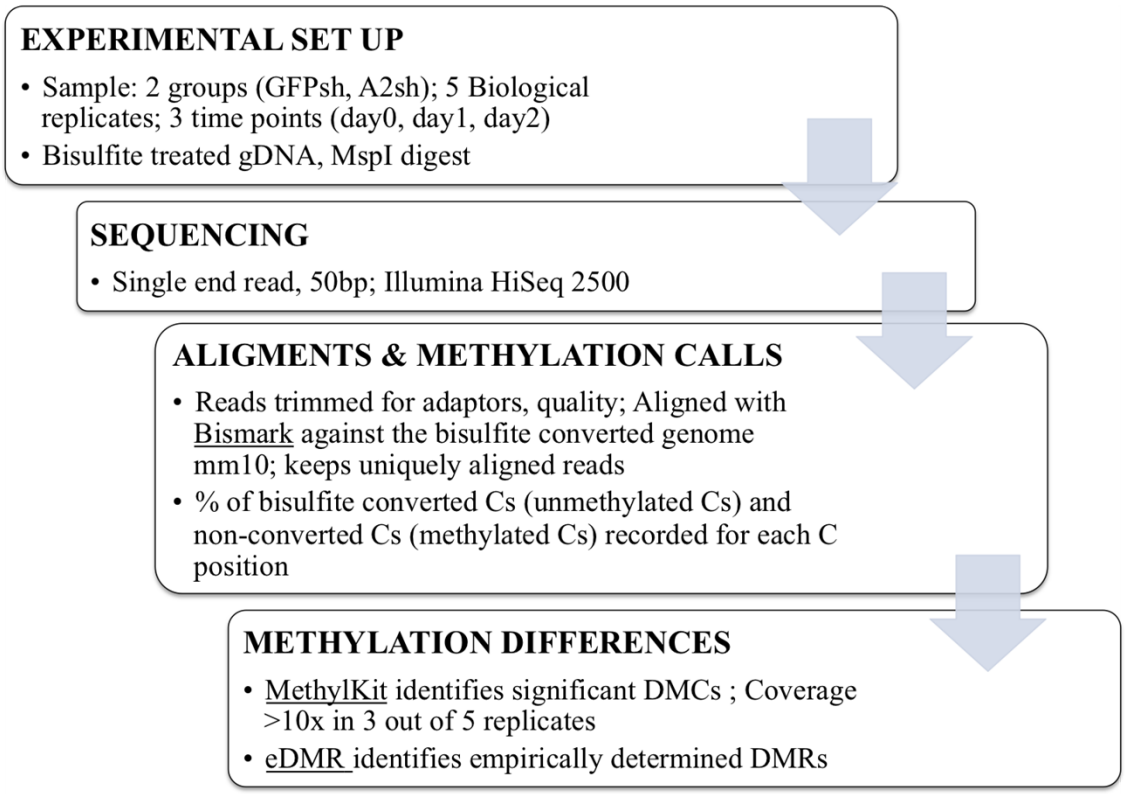


Figure 4.1 Schematic representation of the ERRBS analysis pipeline

As starting material for this analysis, I used C2C12s expanded from individual single cell clones (which I also validated for knockdown efficiency using Western blotting (Figure 4.2a). eRRBS resulted in 67-73% mapping efficiency for trimmed reads and bisulfite conversion rates were ~99% for both forward and reverse strands, suggesting that this potential source of error is not a factor in the interpretation of the dataset. The enhanced RRBS (eRRBS) method yielded the expected high coverage for CpG islands and promoters (Akalin et al., 2012a). On average about ~89% of CpG islands, ~55% of CpG island shores (+/- 2kb from the CpG island) and ~62% of the annotated promoters (+/- 2kb from the TSS) have at least one well-covered (>10x) CpG site (Figure 4.2b). To be included in subsequent analyses it was required that a CpG be covered at least 10x in at least 3 out of 5 biological replicates which resulted in ~1 M represented CpGs for the dataset. These values agree with the DNA methylation standards recommended from ENCODE (Encyclopedia of DNA elements) (ENCODE and modENCODE, 2011).

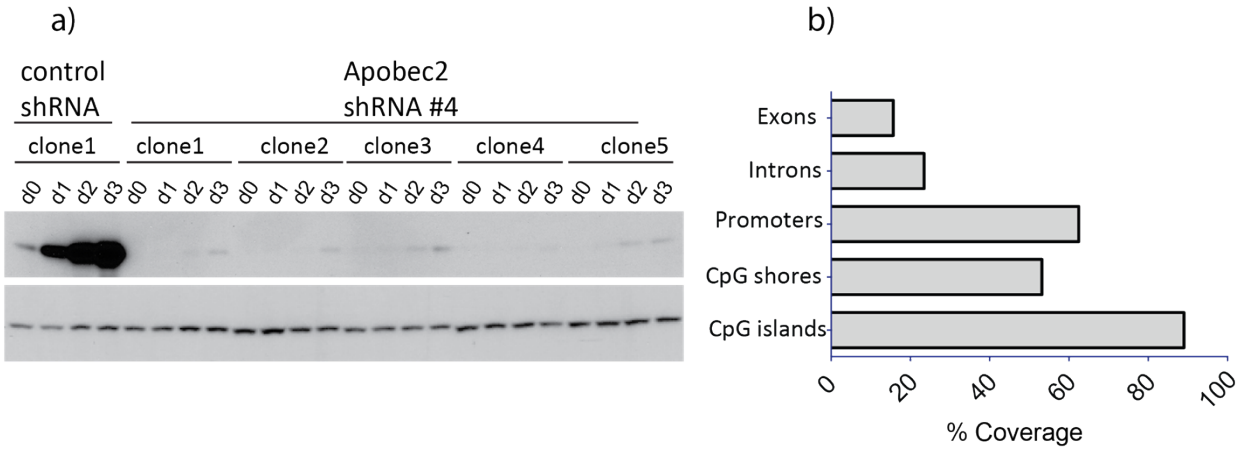


Figure 4.2. RRBS analysis pipeline, validations and coverage

a) C2C12 lysates from single cell clones that were used in the ERBBS sequencing experiment in DM at day0,1,2,3 were analyzed by Western blotting using anti-Apobec2 antibodies. alpha-tubulin was used as loading control. Only one representative clone is shown for the control GFPsh samples b) Average percent coverage of different genomic regions by ERRBS (n = 3, used CpGs covered >10x in at least 3 out of 5 biological replicates during comparisons between the groups for 3 different timepoints (d0, d1, d2)

4.2 Analysis of APOBEC2 dependent differences in DNA methylation

To determine whether APOBEC2 has genome wide or gene-specific effects on the methylome, the distribution and frequency of methylation for each well-covered CpG site was compared between control C2C12s (expressing GFPsh) and APOBEC2 knockdown C2C12s (expressing sh#4), at various time points after induction of differentiation. Additionally, I used the Methylkit (Akalin et al., 2012b) package to determine significant differentially methylated cytosines (DMCs). Overall for all CpGs that were well covered, the expected bimodal distribution of methylation was observed and there was no apparent difference in DNA methylation distributions associated with APOBEC2 expression. The mean methylation frequency for CpGs was highly similar for each genotype (Figure 4.3a). Here out of the ~1 million well-covered CpG sites only about 1% (~11000) of CpGs were differentially methylated at any of the time points and such changes mostly represent hypomethylated CpGs (Figure 4.3b). Annotation of the DMCs based on gene feature (promoter, 5'UTR, 3'UTR, exon, intron, downstream of the gene, intergenic) shows that most of these methylation changes happen in intergenic, introns and promoters. Also most of them do not fall in CpG islands or CpG shores (Figure 4.3c).

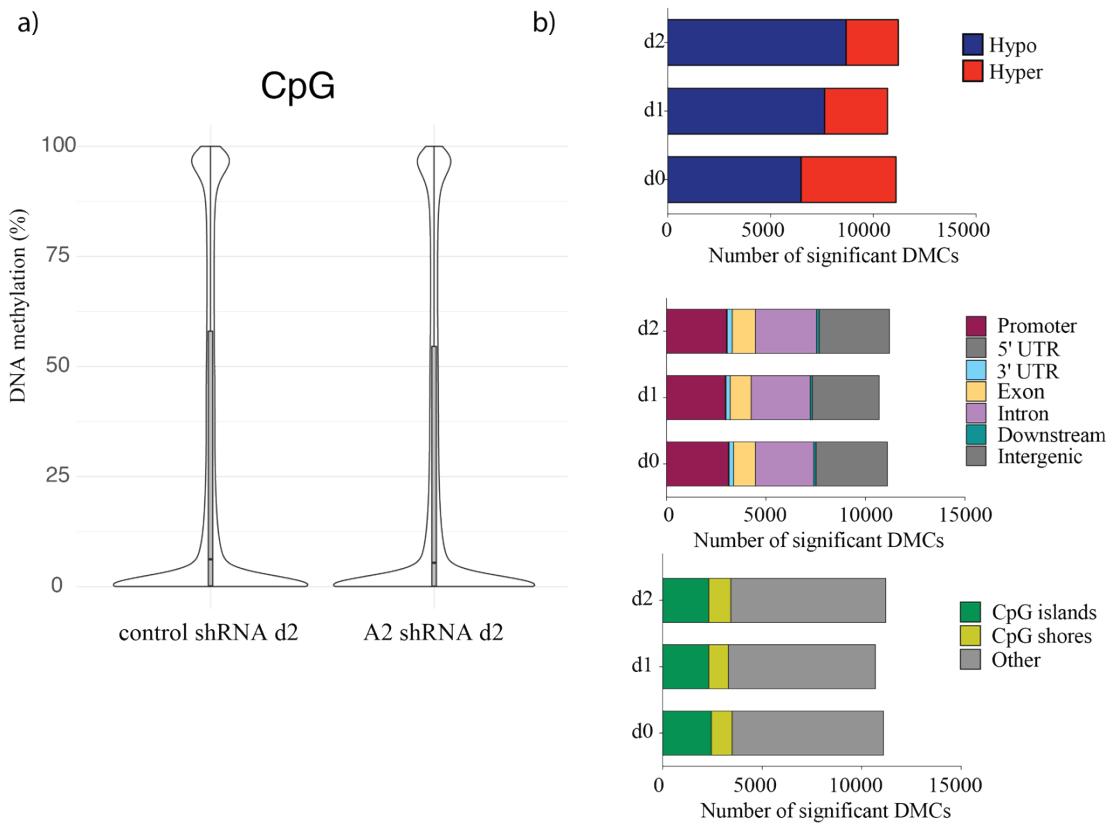


Figure 4.3 APOBEC2 dependent differences in DNA methylation in C2C12s across CpGs

a) Distribution of DNA methylation frequencies in C2C12s as determined by eRRBS for individual CpGs. Violin plots represent the distribution of DNA methylation frequencies for each feature. Median and first and third quartiles are shown with the box plots b) Number of DMC(s). Significant Hypomethylated DMCs (dark blue), significant hypermethylated DMCs (red), nonsignificant DMCs (gray) c) Graphs representing annotations of the DMCs with genomic features or with CpGisland/shores

To further analyze the biological meaning of the DMCs and detect any gene-specific methylation changes I chose two strategies. First, I summarized the methylation values over a specific region such as a CpG island or a promoter (containing a minimum of 3 of well-covered CpGs) and determined if there are any region specific significant methylation changes. The mean methylation frequency for each feature was highly similar for each genotype (Figure 4.4a). Here, out of ~14 thousand CpG islands included in the analysis about 3-9, depending on the time point analyzed were differentially methylated. While out of the ~50 thousand promoters analyzed, about ~5-8 were differentially methylated. Very few of these promoters show significant gene expression differences with hypomethylation correlating with a reduction in gene expression, suggesting that the gene expression changes observed do not correlate with methylation changes in the corresponding genes (Figure 4.4b).

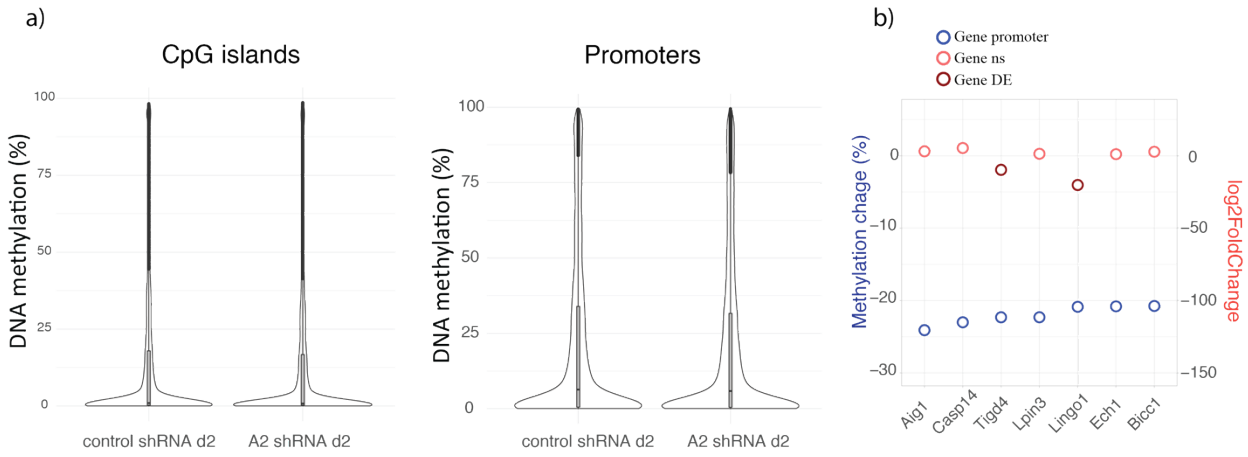


Figure 4.4 APOBEC2 dependent differences in DNA methylation across promoters and CpG islands

a) Distribution of DNA methylation frequencies (weighted means of all replicates) in C2C12s as determined by eRRBS for individual, CpG islands and promoters. Violin plots represent the distribution of DNA methylation frequencies for each feature. Median and first and third quartiles are shown with the box plots. Promoters are defined at -/+ 2Kb around the TSS in Ensemble annotations. CpG islands were taken from the cpgIslandExt track of the UCSC table browser b) Graph representing genes with significant methylation changes in the promoter genes and the corresponding log2 fold change at day2 post differentiation. Dark red is significantly differentially expressed (p adjusted < 0.1)

Secondly, I wanted to determine whether these DMCs fall in close proximity to one another, which would make them potential interesting gene regulatory regions to follow up on. For that I used eDMR (Li et al., 2013) to empirically determine differentially methylated regions (DMRs). DMRs are traditionally defined as genomic regions that have different methylation statuses among multiple samples, contain clusters of CpGs and are potential functional regions involved in gene transcriptional regulation. The eDMR method takes into consideration the distribution of the distances between well-covered CpGs in the dataset across the genome to optimize the definition of an empirical determined region size. The latter determines the minimum distance adjacent CpGs can be separated by, in order to define them as being part of the same DMR.

My analysis detected very few significant DMRs, which are mainly hypomethylated due to APOBEC2 knockdown. When the significant DMRs are annotated based on gene feature (promoter, 5'UTR, 3'UTR, exon, intron, downstream of the gene, intergenic), they mainly fall in promoters or intergenic regions depending on the timepoint analyzed (Figure 4.5a). For the statistically significant hypomethylated DMRs the average percent hypomethylation is between 20%-23% and the regions contain on average 4-5 DMCs depending on the time point analyzed. Only 2 of the DMRs reside in genes that show significant gene expression changes. (Figure 4.5b) Interestingly DNMT3, a *de novo* methyltransferase, is downregulated due to APOBEC2 knockdown throughout differentiation suggesting that it could account for the few hypomethylation events detected (Figure 4.5c).

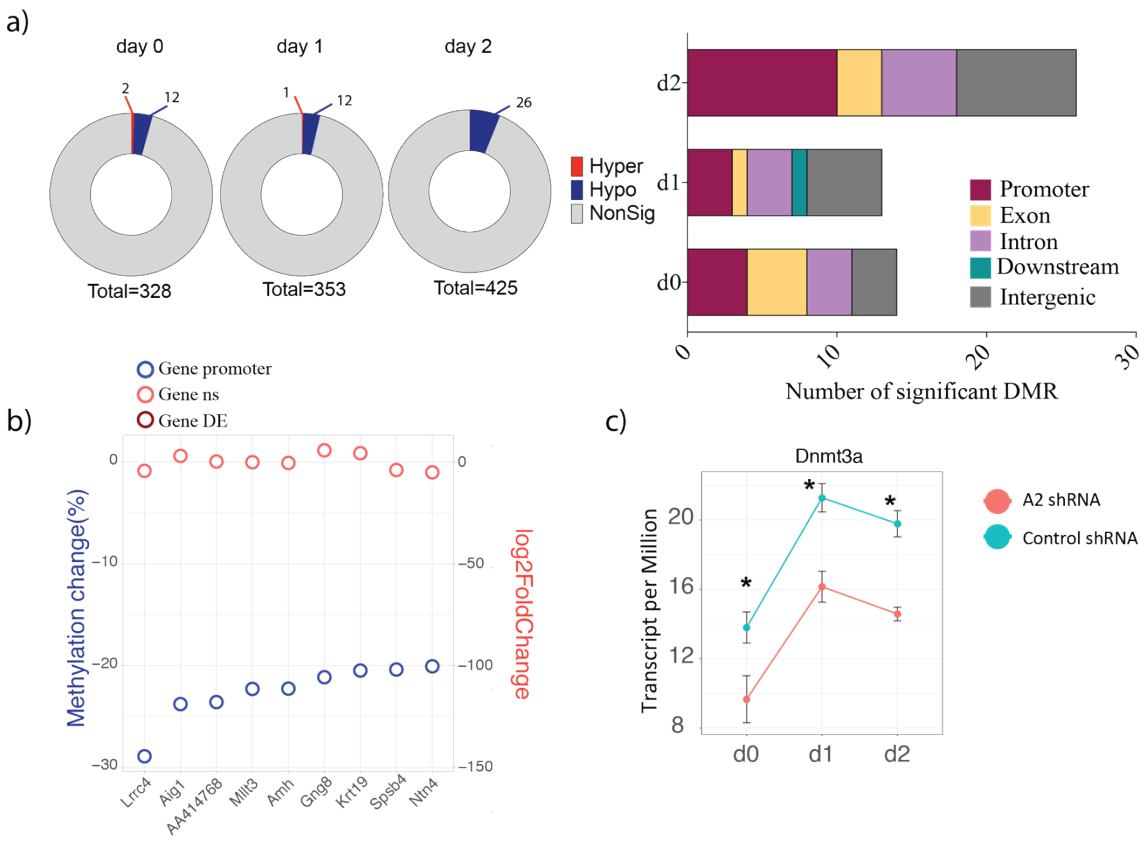


Figure 4.5 APOBEC2 dependent differences in DNA methylation across DMRs

a) Number of Differentially Methylated Cytosines was calculated using the eDMR package. Significant Hypomethylated DMR (dark blue), significant hypermethylated DMR(red), nonsignificant DMR (gray). A DMR is empirically determined and it requires at least 1 DMC in the region, as determined using methylKit, at least 3 CpGs included in the region, and at least an absolute mean methylation difference greater than 20%.)The significant DMR should have at least 5 CpGs where at least 3 of them are DMCs as determined by methylKit and there is a minimum 20% methylation change for the region and q.value < 0.001. Bargraph on the right representing annotations of the DMCs with genomic features or with CpG island/shores b) Graph representing genes with significant DMRs in promoter regions that corresponding log2 fold change at day2 post differentiation. Dark red is significantly differentially expressed (p adjusted < 0.1) c) Graph representing the amount of Dnmt3a using the RNAseq data. Transcripts per million (TPM) is a measurement of the proportion of transcripts within the sample. Data represented as mean and SEM of the 3 biological replicates.

Given the very few changes I detected, I asked if they were simply due to inherent variability between biological replicates. Therefore, I estimated the level of signal to noise ratio as a measure of false positive rate. To do this I utilized the methylation dataset and made two types of comparison using the same methylation analysis pipeline mentioned above. (1) I randomly picked biological replicates from only the control sample and analyzed methylation changes including only 2 biological replicates for each group being compared (within group comparison) to determine the amount of noise in the system. (2) I repeated the same analysis using randomly picked samples from both control and *knockdown* samples (between groups comparison) including 2 biological replicates for each group, to determine the amount of signal in the system. The results of this analysis indicated that the percent of significant DMCs is very similar in both types of comparisons (Table 4.1) and that the signal to noise ratio is close to 0.98.

Table 4.1 Estimating the false positive rate

Proportion of significant DMCs for multiple group comparisons (between controls with knockdown) and within group comparisons (between biological replicates within the same group)

Number of CpGs differentially Methylated (Thousand)	Number of CpGs covered (Thousand)	% sig DMCs	
33	1057	3.2%	**Between groups**
38	1082	3.5%	
55	1061	5.1%	
44	1083	4.1%	**Within groups**
43	1068	4.0%	
41	1063	3.8%	

4.3 Correlation of APOBEC2-dependent changes in DNA methylation and mRNA expression

As final analysis, I compared the methylation changes across all the represented promoters in the eRRBS dataset with the expression changes of the same genes in RNA-Seq dataset for each of the time points. Each pairwise comparison between the control and knockdown datasets showed that there is no linear relationship between the variables ($|r| < 0.06$ in all cases) (Figure 4.6). This suggested that the observed modest differences in DNA methylation of the promoters were not directly associated with changes in the expression of the same genes. It is unlikely that these methylation changes are biologically relevant.

Overall my analyses suggested that the level of "signal" in my dataset does not surpass the level of "noise". This conclusion is strengthened by the notion that even changes that seem to occur on relevant promoters, do not lead to consistent gene expression differences. Thus if any APOBEC2 dependent methylation changes do exist, they are likely buried under the inherent noise in the dataset. Alternatively, APOBEC2 dependent methylation changes could exist in regions of the genome not captured by eRRBS. Nevertheless, my data thus far, together with the fact that any changes I see tend toward hypomethylation (which is inconsistent with the notion of APOBEC2 functioning as an active DNA demethylase) lead me to negate the hypothesis that APOBEC2 functions as a CpG demethylase, at least in C2C12 cells.

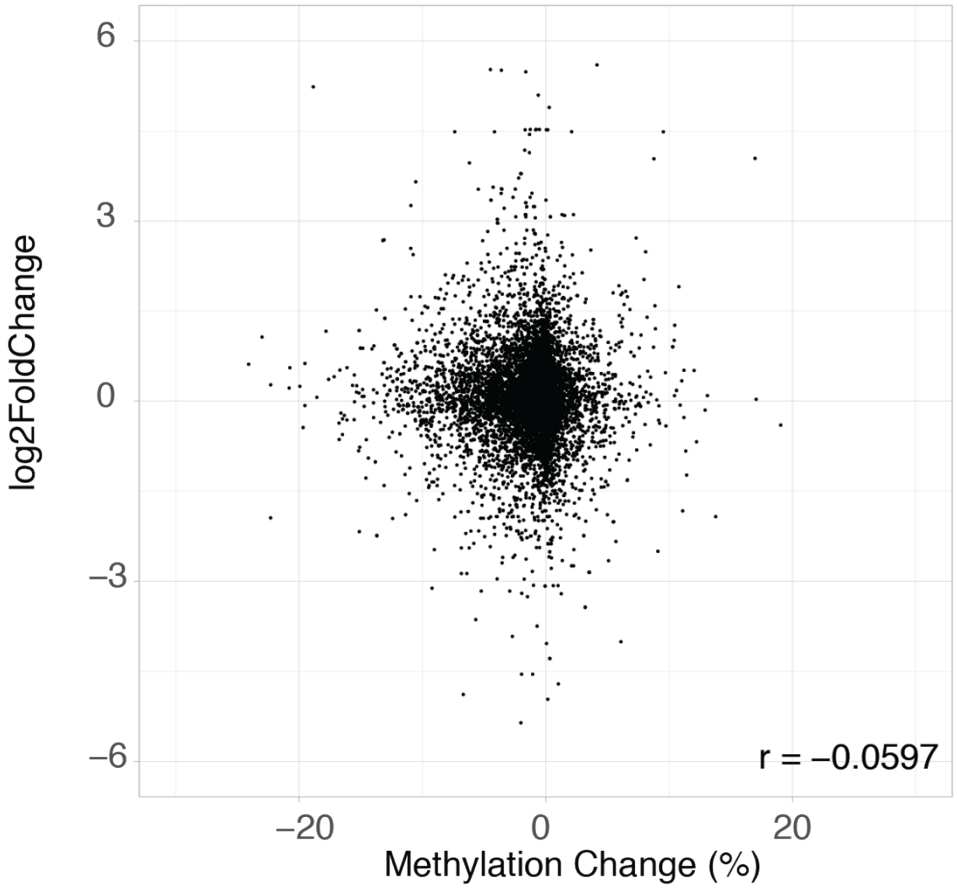


Figure 4.6 Comparison of differences in DNA methylation with gene expression

Methylation changes across all the represented promoters in the eRRBS dataset compared with the expression changes of the same genes in RNA-Seq dataset. Shown here are datasets from the day 2 timepoint. r = Pearson's correlation coefficient

CHAPTER 5. Assessment of the ability of APOBEC2 to associate with and affect chromatin

APOBEC2 is deemed an orphan deaminase because it showed no catalytic activity (or even binding) toward single stranded DNA (methylated or not), or RNA. However, functional data indicated that it is involved in muscle biology. Previous studies hypothesized it might be involved in DNA demethylation or RNA editing. My data clearly discount both of these possibilities. However, my data also show that deficiency of APOBEC2 during myogenesis, substantially affects gene expression, suggesting a direct effect on gene regulation. Thus, I decided to evaluate the hypothesis that APOBEC2 might be regulating transcriptional activity of target genes, either directly by binding to DNA genetic elements or indirectly through interaction with other direct chromatin regulators. To do this, I attempted to firstly, determine the subcellular localization of APOBEC2 in differentiating cells and its binding affinity to chromatin. Secondly, I utilized chromatin immunoprecipitation, followed by high throughput sequencing (ChIP-Seq) to identify genomic regions that might be preferentially bound by APOBEC2. In this chapter I will present data from these studies, which bring us closer to the identification of the mechanism for its transcriptional effects.

5.1 Supporting evidence for APOBEC2's association with the chromatin

To determine the subcellular localization of APOBEC2 I first searched for the presence of a putative nuclear localization signal (NLS) on the APOBEC2 protein sequence. I used cNLS Mapper, a prediction system for nuclear localization signal (Kosugi et al., 2009). The algorithm predicted the sequence: DPEKLKELIDLPPFEIVTGVRLPVNFFKFQFR

with a score of 3.7 which indicates a low activity NLS and thus predicts that APOBEC2 will be localized at both the nucleus and the cytoplasm. I followed this up with immunostaining of APOBEC2, which supported the notion that it is localized at both nuclei and cytoplasm (Figure 5.1a).

Having shown that at least a fraction of APOBEC2 is nuclear, I wanted to determine whether APOBEC2 bound chromatin. To do this I utilized sequential salt extraction of chromatin proteins (Figure 5.1b). Nuclear proteins bind to chromatin with different affinities and thus require different salt concentrations to be dissociated. At a low salt concentration, loosely bound proteins will be easily dissociated from chromatin, whereas tightly bound proteins will not (Porter et al., 2017). The salt-extraction profile of APOBEC2 from nuclei of differentiated C2C12s indicates its presence in eluates where high concentration of NaCl (0.75 and 1M) is used to wash the chromatin. As a comparison H4 histone, which is normally strongly bound to DNA, is shown in the same western blot to dissociate completely with 0.75M NaCl. (Figure5.1b). This data suggests a strong association of a portion of nuclear APOBEC2 with chromatin in differentiated C2C12s.

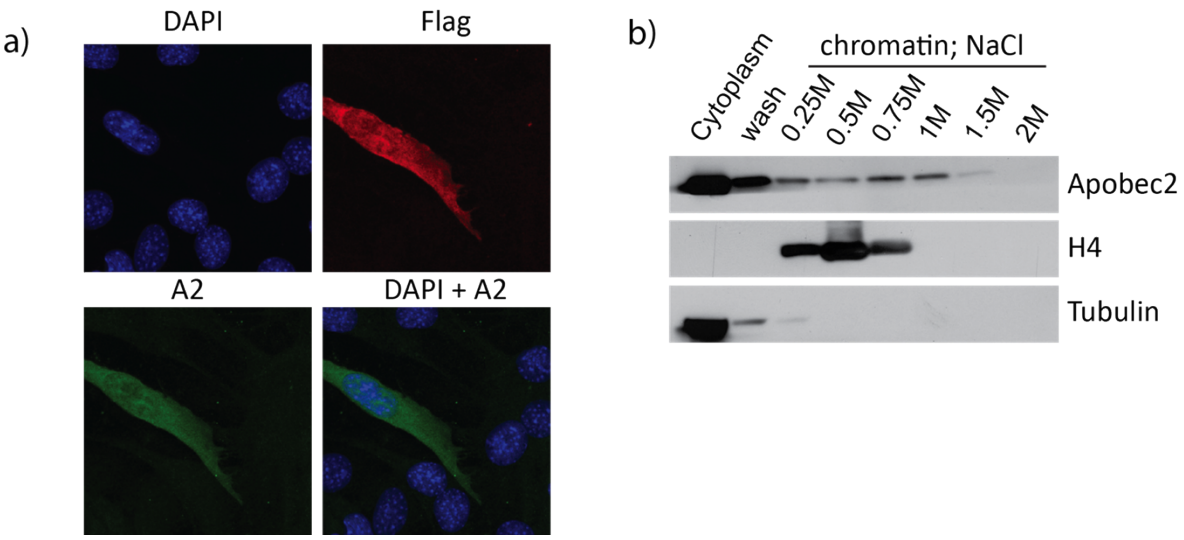


Figure 5.1 Cellular localization and chromatin association of APOBEC2

a) Immunostaining of overexpressed Flag-Apobec2, Flag (red), Apobec2 (green) and DAPI-positive (blue) nuclei in C2C12s cultured for 5 days in DM b) The NaCl-elution profiles of endogenous Apobec2 and histone H4 are shown. alpha-tubulin is a cytoplasmic marker. The amount of indicated proteins in eluates was measured by Western blotting.

5.2 Validation and ChIP-Seq analysis pipeline

To assess whether chromatin-bound APOBEC2, occupies specific regions of the C2C12 genome, I performed a ChIP-Seq (Chromatin Immunoprecipitation Sequencing) experiment. First, I validated the efficiency of the APOBEC2 antibody by doing an immunoprecipitation, which shows that the antibodies are effective in depleting the input (Figure5.2a). As the first step in ChIP-Seq, I stabilized the interactions by cross-linking DNA and bound proteins with formaldehyde. I sheared the cross-linked chromatin into small fragments by sonication and then immunoprecipitated the DNA-protein complexes using APOBEC2 specific antibodies. The goal was to get an enrichment of DNA fragments that are bound by APOBEC2. Given that shearing of the genomic DNA is not uniform and downstream sequencing does not produce even coverage of the genomic

DNA, a portion of the fragmented DNA, not immunoprecipitated by APOBEC2, is kept to be used for the assessment of the background signal (the input control). I then reversed the cross-linking to release the DNA fragments from the binding complex. For ChIP-Seq, I used C1C12s at 2 different time points after inducing differentiation (14 hours and 34 hours post differentiation). I picked those time points because they (1) slightly precede the RNA-Seq time points, where we observed many changes in gene expression and (2) represent time points of low and high APOBEC2 protein abundance. The goal was to maximize the chances of finding APOBEC2 occupancy (for example in case of transient binding) and assess if the expression levels of APOBEC2 determine where the protein binds. I used the input control fragments of DNA and APOBEC2-enriched DNA fragments, to prepare libraries using the library preparation protocol from NEBNext DNA library prep kit for high throughput sequencing. The latter includes PCR amplification to increase the amount of starting DNA and multiplexing of the samples to be able to combine multiple samples for sequencing.

Multiplexed samples underwent single end sequencing. In brief, as shown in the diagram (Figure 5.2b) the subsequent analysis pipeline involves filtering for reads that do not pass quality control measures, adaptor trimming and alignments of the reads to a reference genome. Quality metrics for the ChIP-Seq data were assessed using ChIPQC bioconductor package (Carroll et al., 2014), according to Encyclopedia of DNA Elements (ENCODE) working standards and guidelines for ChIP experiments (Landt et al., 2012).

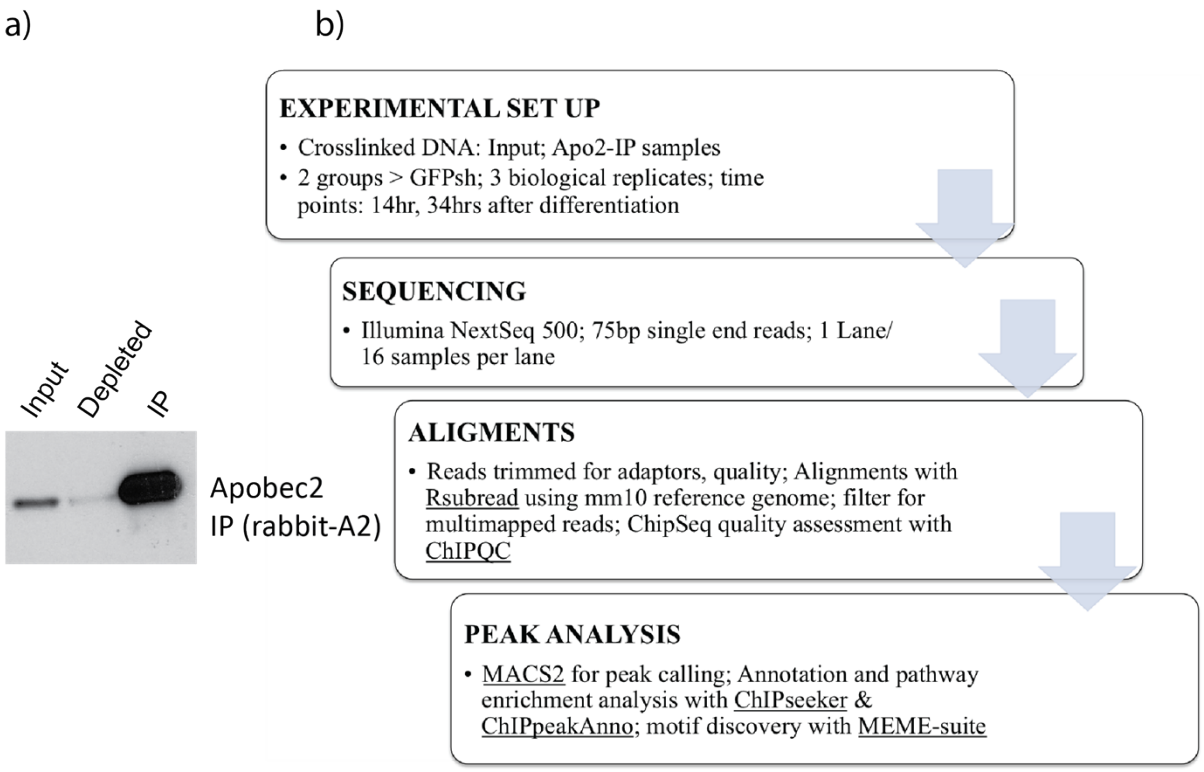


Figure 5.2 Validation and Chip-Seq analysis pipeline

a) Validation of Apobec2 antibodies using C2C12s whole cell lysates (at DM day5); Lysates are immunoprecipitated (IP) with anti-APOBEC2 antibodies. Input (lysate pre-IP), depleted (lysate post-IP) and immunoprecipitated fractions are shown b) Schematic of ChipSeq analysis pipeline

The mapping efficiency range was 88-90%, resulting in 18M-52M mapped reads depending on the sample. The proportion of multimapped reads (reads that can be assigned to more than one location) was between 14.9%-24.9% of all the mapped reads. These numbers show high multimapping rates and low reads in peaks (RiP) values (not shown), which are generally though to reflect low signal to noise ratios (and thus high background). This suggests one of two things: either that binding observed is at the background level (i.e. random) or that the binding complex/protein has very few true

binding sites at the time points analyzed (and thus is highly specific). Moreover, high multimapping reads could also be representative of true binding events in repeat regions.

Next I looked at the complexity of the libraries by checking duplication rates. Library complexity is the fraction of DNA fragments that are non-redundant (example due to PCR amplification). The proportion of duplicated reads (reads that map to the exact position in the genome as at least one other read) ranges from 2.99% to 4.58% of all the mapped reads. ChIP-Seq datasets for narrowly binding transcriptions factors are expected to have regions of very high enrichment, which will include true fragments originating at the same location. Sequencing of such "duplicated" fragments is expected and biologically relevant in the regions, where the protein complex binds with high affinity. On the other hand, when the depth of the sequencing library is high one can reach a point where the complexity is exhausted and duplication events represent PCR-amplified DNA fragments that are sequenced repeatedly rather than true identical fragments that are biologically relevant. Thus, very high duplication rates sometimes are indicative of low complexity libraries. The percent of non-redundant reads among unique mapped reads (not multimapped) in our experiment ranges from 94% to 96%. These rates are very high (relative to the recommended ENCODE standard of non-redundant proportion of greater than 80%) and this could suggest high library complexity but also potential issues with enrichment of potential APOBEC2 bound DNA.

Overall, given the high library complexity (low duplication rates), together with the low signal to noise (low proportion of RiP observe), future ChIP-Seq experiments with APOBEC2 could benefit from increasing sequencing depth or optimizing the immunoprecipitation to increase enrichment of APOBEC2 bound fragments.

Peaks of protein localization were called using Model-based Analysis of ChIP-Seq (MACS2) (Feng et al., 2011; Zhang et al., 2008) where the input DNA sample was used as a control for the background noise. Each experiment was performed in three biological replicates. According to a previous study when multiple replicates are generated using ChIP-Seq, selecting the peaks that are the same in the majority of the biological replicates yields more reliable peaks, than the absolute match with fewer replicates (Yang et al., 2014). Therefore I defined most reliable and reproducible consensus peaks of localization, those peaks that were called by the MACS2's algorithm in at least 2 out of 3 replicates for each time point analyzed. For the 14-hour time point, there are 1317, 2445 and 2370 localization peaks in each of the biological replicates with 859 consensus peaks (peaks in common in at least 2 out of 3 replicates). For the 34-hour time point, there were 2208, 5024 and 1067 localization peaks in each of the biological replicates with 2016 consensus peaks.

Previous publications have shown the existence of signal artifact blacklist regions: these are genomic regions that contain high artificial read coverage in ChIP-seq experiments that are conserved across different cell lines and conditions (The ENCODE Project Consortium, 2012). These have been shown to influence the quality assessment metrics (Carroll et al., 2014). We decided to filter blacklisted regions from our ChIP-Seq localization peaks for downstream analysis of APOBEC2 binding regions. After filtering, we had 818 consensus peaks from the 14-hour time point and 1980 for the 34-hour time point. Therefore only a very small proportion of the peaks (2-5%), represent blacklisted repeat regions.

For downstream analysis, only those consensus peaks in 2 out of 3 replicates cleaned for blacklisted regions were used. I used the ChIPSeeker (Yu et al., 2015) and ChIPpeakAnno (Zhu, 2013; Zhu et al., 2010) bioconductor packages for downstream annotation and clusterProfiler (Yu et al., 2012) for pathway enrichment analysis. For more details about the ChIP-Seq analysis pipeline and methods refer to the Materials and Methods chapter.

5.3 APOBEC2 occupies mostly promoter regions in the genome

First I wanted to determine the overall signal around the binding locations of APOBEC2, which is called the peak summit (i.e. the location with the highest read coverage in the peak(Zhang et al., 2008). Here I used the mean location of peak summits across the different biological replicates. I then generated the mean signal expressed in number of read counts per million mapped reads (normalized per library size) around all of the binding sites in any of the time points (union of the peaks at the two different time points). Overall the signal around the binding sites is higher in the 34-hour time point. This could be due to an increase in APOBEC2 protein abundance, resulting in increase in overall binding sites and/or increase in the amount of protein bound leading to an increase in the proportion of RiP. The input, used as a control, shows no enrichment over the peak summits (Figure 5.3a,b).

I annotated the peaks of localization (peaks that are present in any of the time points) based on the nearest genes and the genomic feature they overlap with (such as promoter, 5’ UTR, 3’ UTR, exon, intron, downstream of the gene and intergenic). For both time points most of these binding regions, fall within promoter regions, defined as regions 4kb (-2kb to +2kb) around the TSS. The proportion of APOBEC2 binding

regions that are in promoters is 93% and 75%, for the 14-hour and 34-hour time point, respectively. As expected given the enrichment in signal overall there are more binding regions in the 34-hour time point (Figure 5.3c).

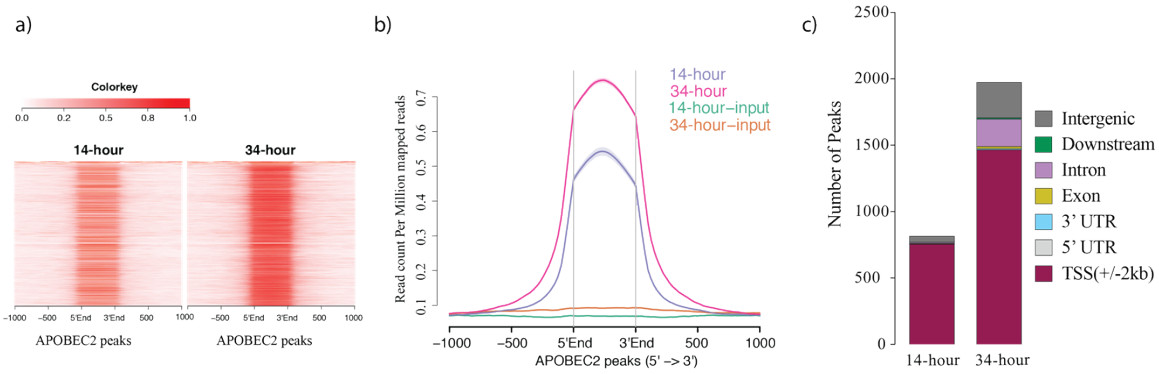


Figure 5.3 Annotation of APOBEC2 occupied regions

a) On the left heatmap of normalized signal 1100bp around peak summits. APOBEC2 binding regions were established by grouping consensus peaks (peaks in the majority of replicates), filtered for blacklisted regions and called in any of the time points. Average binding site location was determined from mean location of peak summits across biological replicates and intensity of signal (shown in the Colorkey) is the mean of the RPM normalized read counts. Both time points are in biological triplicates b) The mean normalized APOBEC2 signal (plotted as read counts per million mapped reads) across all the binding sites shown in a) This plot shows the global differences between the two time points. Both time points are in biological triplicates c) Genomic annotations of the APOBEC2 of the consensus binding regions in each of the time points (14hr and 34 hr). Binding regions are annotated with based on genomic feature. The priority of assignment to a genomic feature when there is annotation overlap is: Promoter (2kb around the TSS), 5' UTR, 3' UTR, Exon, Intron, Downstream (within 3kb downstream of the end of the gene).

Given that the majority of the binding sites fall near the TSS of genes I decided to focus on this subset. My analysis showed that there are about ~1500 genes that are bound by APOBEC2 near their transcription start sites in any of the time points. Of those, 580 genes are in common between both time points. Finally, APOBEC2 shows increased signal over the TSS in the 34-hour timepoint (Figure 5.4a,b).

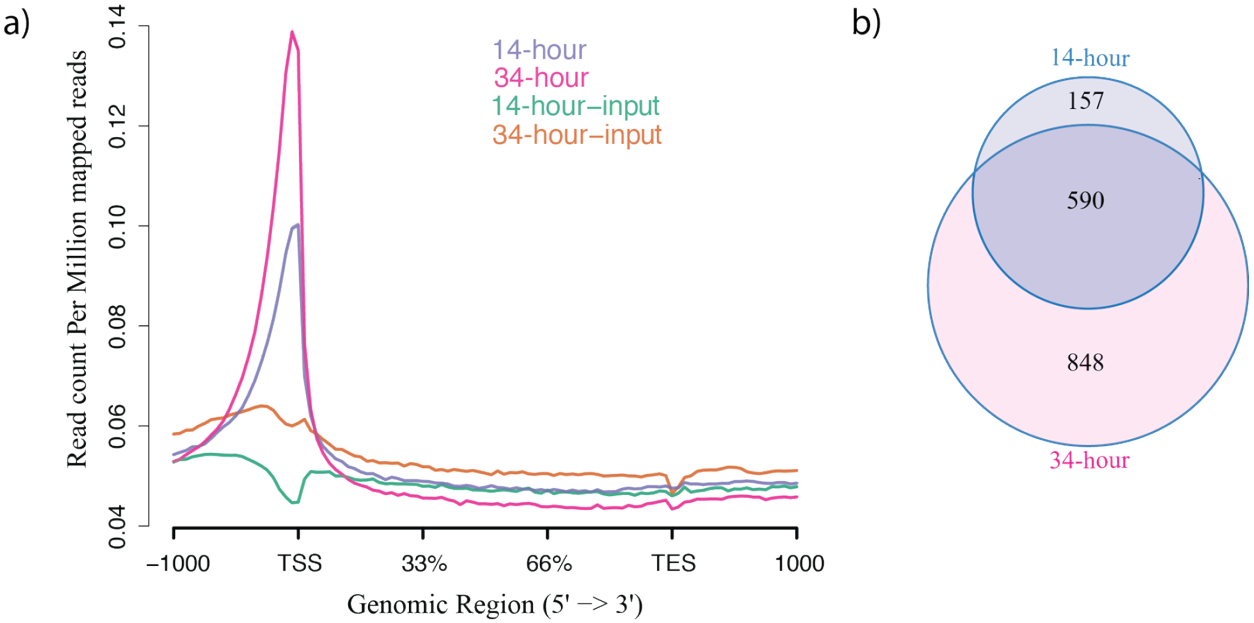


Figure 5.4 Comparison of APOBEC2 occupied genes

a) The mean normalized APOBEC2 signal (plotted as read counts per million mapped reads) across all annotated genes. This plot shows the global differences in APOBEC2 binding between the two time points in TSS. Both time points are in biological triplicates b) The venn diagram represents the number of genes that show APOBEC2 occupancy in their promoters at 14 and 34 hour time points and their overlap.

I then utilized the MSigDB datasets (the *Hallmark* and *Reactome pathways*) to determine the top enriched gene sets among the genes that are bound in any of the time points. This analysis indicated that APOBEC2 bound genes show a high and significant enrichment of cell cycle related pathways (E2F targets and Cell cycle) (Figure 5.5), Table

5.1-5.2. Overall this analysis is suggestive of a direct role of APOBEC2 in affecting gene expression changes through binding in promoter regions of the genome, specifically in cell cycle related genes and E2F target genes. Given the observed phenotype of the delay in cell cycle exit during differentiation due to APOBEC2 deficiency, I found this intriguing and followed up on these findings.

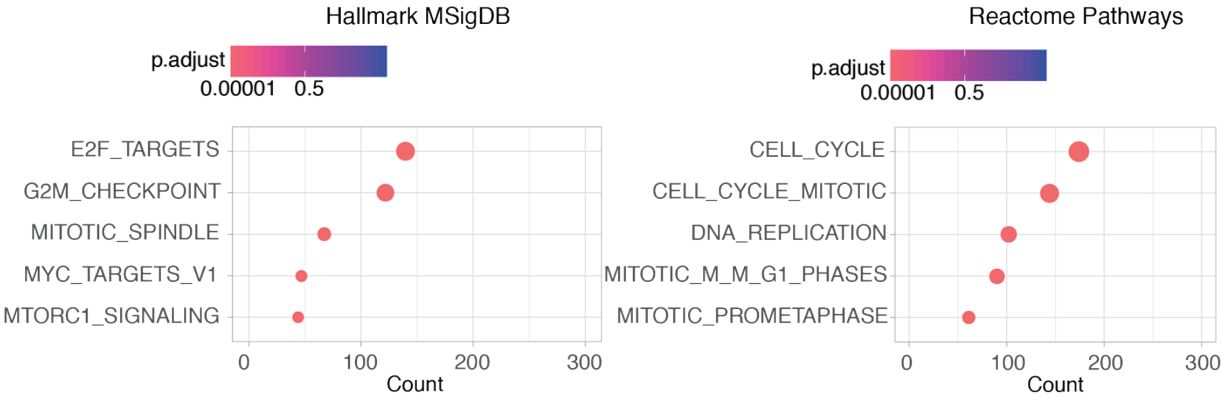


Figure 5.5 Pathway enrichment analysis of APOBEC2 occupied genes

Genes that were bound by APOBEC2 in their promoters were established by grouping binding regions called in any of the time points. Pathway enrichment analysis of these genes using overrepresentation test on both MSigDB *Hallmark* and MSigDB *Reactome Pathways* gene set collections. The enriched genesets are ranked by smallest adjusted p value and largest overlap size and the top 5 gene sets are shown. The dots in the plot have sizes proportional to the number of overlapping genes and colored according to the adjusted p value.

Table 5.1 Pathway enrichment analysis using APOBEC2 bound genes with MSigDB *Hallmark* gene set

Pathway enrichment analysis of genes that are bound by APOBEC2 in their promoters in any of the time points. Results show overrepresentation test on MSigDB *Hallmark* gene set collections with the top 5 gene sets shown. Table of gene sets as shown in Figure 5.5.

Geneset	pvalue	p.adjust	Count	Genes
HALLMARK_E2F_TARGETS	3.98E-78	1.99E-76	140	Mcm3/Prim2/Bard1/Cnot9/Mcm6/Ube2t/Lbr/Spc25/Bub1b/Pcna/Mybl2/Cse1l/Plk4/Exosc8/Smc4/Cks1b/Dclre1b/Cenpe/H2afz/Depdc1a/Nbn/Mms22l/Melk/Usp1/Cdkn2c/Nasp/Kif2c/Cdc20/Cdca8/Ak2/Rpa2/Stmn1/Slbp/Tacc3/Dck/Pole/Chek2/Ung/Cit/Ran/Rfc2/Mcm7/Pms2/Brca2/Rfc3/Lsm8/Ing3/Nup205/Ezh2/Mad2l1/Pole4/Mthfd2/Cdca3/Ncapd2/Rad51ap1/Ube2s/Lig1/Ccne1/E2f8/Pold3/Wee1/Plk1/Kif22/Dctpp1/Mki67/Gins4/Hmgb2/Mcm5/Asf1b/Rnaseh2a/Orc6/Nudt21/Gins3/Ctcf/Mre11a/Spc24/Chek1/H2afx/Zw10/Tipin/Ccnb2/Stag1/Cdc25a/Mlh1/Asf1a/Cdk1/Tmpo/Nup107/Timeless/Pan2/Pa2g4/Hus1/Xpo1/Hmmr/Rad50/Shmt1/Aurkb/Spag5/Srsf1/Nme1/Top2a/Psmc3ip/Brca1/Kif18b/Kpna2/Jpt1/Birc5/Rrm2/Brms1l/Ubr7/Dek/Cks2/Mxd3/Cdkn3/Dlgap5/Diaph3/Rad1/Rad21/Dscc1/Atad2/Myc/Xrcc6/Cenpm/Racgap1/Espl1/Cbx5/Mcm4/Ranbp1/Tra2b/Wdr90/Tcf19/Tubb5/Msh2/Lmnb1/Pola2/Hells/Smc3/Suv39h1/Kif4/Smc1a
HALLMARK_G2M_CHECKPOINT	6.14E-59	1.53E-57	122	Mcm3/Prim2/Bard1/Dtymk/Mcm6/Rasal2/Exo1/Lbr/Cenpf/Nek2/Map3k20/Meis2/Knl1/Bub1/Tpx2/E2f1/Rbl1/Top1/Mybl2/Ube2c/Ccna2/Plk4/Smc4/Cks1b/Notch2/Cenpe/H2afz/Casp8ap2/Smc2/Cdkn2c/Stil/Rad54l/Nasp/Kif2c/Cdc20/Sfpq/Rpa2/Stmn1/Srsf10/Dbf4/Cenpa/Tacc3/Nsd2/Ythdc1/Cdc7/Mtf2/Pole/Kmt5a/Brca2/Ezh2/Mad2l1/Ube2s/Prc1/Numa1/Plk1/Kif22/Mki67/Tfdp1/Mcm5/Orc6/Ctcf/Gins2/Chek1/Hspa8/H2afx/Kif23/Smad3/Ccnb2/Ttk/Stag1/Traip/Cdc25a/Kif15/Fbxo5/Cdk1/Tmpo/H2afv/Hus1/Xpo1/Hmmr/G3bp1/Aurkb/Pafah1b1/Srsf1/Kpnb1/Cdc6/Top2a/Kpna2/Jpt1/Birc5/Mnat1/Hif1a/E2f3/Prpf4b/Cks2/Fancc/Papd7/Cdkn3/Prmt5/Pbk/Rad21/Myc/Nup50/Ccnt1/Troap/Racgap1/Espl1/Tra2b/Polq/Ccnf/Chaf1a/Ndc80/Snrpd1/Ss18/Lmnb1/Pola2/Incenp/Kif20b/Kif11/Suv39h1/Kif4/Smc1a
HALLMARK_MITOTIC_SPINDLE	1.55E-13	2.58E-12	67	Clasp1/Rasal2/Kifap3/Cenpf/Nek2/Cntrl/Ckap5/Bub1/Bcl2l11/Tpx2/Ect2/Smc4/Notch2/Sass6/Cenpe/Cdk5rap2/Kif2c/Kntc1/Brca2/Alms1/Plekhg2/Prc1/Numa1/Plk1/Kif22/Katnb1/Anln/Kif23/Pif1/Myo1e/Ccnb2/Ttk/Kif15/Fbxo5/Wasf1/Cdk1/Fgd6/Ccdc88a/Rapgef6/Cntrob/Pafah1b1/Top2a/Sept9/Birc5/Dock4/Arf6/Net1/Nedd9/Cep72/Dlgap5/Cenpj/Myh9/Racgap1/Espl1/Dlg1/Tiam1/Itsn1/Ezr/Ndc80/Lmnb1/Sac3d1/Incenp/Kif20b/Kif11/Smc3/Kif4/Smc1a
HALLMARK_MYC_TARGETS_V1	0.00020918	0.0026148	47	Hspe1/Ddx18/Mcm6/Pcna/Ccna2/H2afz/Pole3/Usp1/Cdc20/Ran/Mcm7/Ssbp1/Hnrnpa2b1/Mad2l1/Serbp1/Psmc4/Rrm1/Psma1/Tfdp1/Mcm5/Xpot/Ptges3/Pa2g4/Ppia/Xpo1/Canx/G3bp1/Eif4a1/Srsf1/Nme1/Kpnb1/Psmb3/Psmd3/Kpna2/Dek/Pabpc1/Myc/Xrcc6/Mcm4/Ranbp1/Tra2b/Rfc4/Tcp1/Glo1/Lsm2/Snrpd1/Prdx3
HALLMARK_MTORC1_SIGNALING	0.00199915	0.01999154	44	Tcea1/Hspe1/Actr3/Cxcr4/Uchl5/Arpc5l/Bub1/Ube2d3/Pgm2/Pik3r3/Nfyc/Abcf2/Insig1/Pitpnb/Ung/Nup205/Mthfd2/Slc6a6/Edem1/Tomm40/Psmc4/Sytl2/Plk1/Calr/Hmbs/Cdc25a/Txnrd1/Hsp90b1/Ppia/Ccng1/Canx/Tmem97/Acaca/Atp5g1/Nmt1/Fdxr/Rrm2/Egln3/Mcm4/Ccnf/Hspa9/Fads2/Fads1/Psat1

Table 5.2 Pathway enrichment analysis using APOBEC2 bound genes with MSigDB *Reactome Pathways*

Pathway enrichment analysis of genes that are bound by APOBEC2 in their promoters in any of the time points. Results show overrepresentation test on MSigDB *Reactome Pathways* gene set collections with the top 5 gene sets shown. Table of gene sets as shown in Figure 5.5.

Geneset	pvalue	p.adjust	Count	Genes
REACTOME_CELL_CYCLE	3.83E-63	2.09E-60	174	Mcm3/Prim2/Sgo2a/Hjurp/Clasp1/Mcm6/Cenpl/Nuf2/Ahctf1/Lin9/Nsl1/Nek2/Mcm10/Tubb4b/Cntrl/Spc25/Ckap5/Kif18a/Bub1b/Knl1/Oip5/Bub1/Pcna/Mcm8/E2f1/Dsn1/Rbl1/Mybl2/Ube2c/Dido1/E2f5/Ccna2/Plk4/Pmf1/Cks1b/H2afz/Cdk5rap2/Itgb3bp/Orc1/Cdkn2c/Kif2c/Cdc20/Cdca8/Rpa2/Nudc/Rcc2/Cenps/Dbf4/Cenpa/Cep135/Lin54/Cdc7/Pole/Chek2/Rfc5/Anapc5/Kntc1/Rfc2/Mcm7/Mad1l1/Rfc3/Pot1a/Mad2l1/Alms1/Lig1/Psmc4/Ccne1/Pold3/Numa1/Wee1/Psma1/Plk1/Tfdp1/Gins4/Cenpu/Mcm5/Orc6/Rbl2/Cenpt/Cenpn/Gins2/Cdt1/Nup133/Cdkn2d/Spc24/Chek1/H2afx/Zw10/Ppp2r1b/Kif23/Zwilch/Ccnb2/Atr/Stag1/Cdc25a/Syne1/Fbxo5/Nup43/Cdk1/Zwint/Nup37/Nedd1/Mdm2/Nup107/Hus1/Xpo1/Spdl1/Aurkb/Wrap53/Mis12/Pafah1b1/Ska2/Psmb3/Psmd3/Cdc6/Brca1/Nup85/Birc5/Anapc11/Rrm2/Mis18bp1/Pole2/Mnat1/Syne2/Hsp90aa1/Hist1h2ai/Hist1h2bm/Hist1h4k/Hist1h2bn/Hist1h4i/Hist1h2ag/Hist1h4f/Hist1h2be/Hist1h2bc/Hist1h4a/Gmnn/E2f3/Cenpp/Cep72/Cenph/Ccnb1/Tinf2/Cenpj/Rb1/Skp2/Rad1/Rad21/Myc/Rangap1/Cenpm/Tuba1a/Mcm4/Rfc4/Mis18a/Dyrk1a/Pkmyt1/Tubb5/Cenpq/Sgo1/Ndc80/Kif20a/Cdc23/Cdc25c/Lmnb1/Cep76/Ska1/Pola2/Incenp/Fen1/Smc3/Pola1/Ercc6l/Cenpi/Smc1a
REACTOME_CELL_CYCLE_MITOTIC	2.88E-56	7.87E-54	144	Mcm3/Prim2/Sgo2a/Clasp1/Mcm6/Cenpl/Nuf2/Ahctf1/Lin9/Nsl1/Nek2/Mcm10/Tubb4b/Cntrl/Spc25/Ckap5/Kif18a/Bub1b/Knl1/Bub1/Pcna/Mcm8/E2f1/Dsn1/Rbl1/Mybl2/Ube2c/E2f5/Ccna2/Plk4/Pmf1/Cks1b/Cdk5rap2/Itgb3bp/Orc1/Cdkn2c/Kif2c/Cdc20/Cdca8/Rpa2/Nudc/Rcc2/Cenps/Dbf4/Cenpa/Cep135/Lin54/Cdc7/Pole/Rfc5/Anapc5/Kntc1/Rfc2/Mcm7/Mad1l1/Rfc3/Mad2l1/Alms1/Lig1/Psmc4/Ccne1/Pold3/Numa1/Wee1/Psma1/Plk1/Tfdp1/Gins4/Cenpu/Mcm5/Orc6/Rbl2/Cenpt/Cenpn/Gins2/Cdt1/Nup133/Cdkn2d/Spc24/Zw10/Ppp2r1b/Kif23/Zwilch/Ccnb2/Stag1/Cdc25a/Fbxo5/Nup43/Cdk1/Zwint/Nup37/Nedd1/Nup107/Xpo1/Spdl1/Aurkb/Mis12/Pafah1b1/Ska2/Psmb3/Psmd3/Cdc6/Nup85/Birc5/Anapc11/Rrm2/Pole2/Mnat1/Hsp90aa1/Gmnn/E2f3/Cenpp/Cep72/Cenph/Ccnb1/Cenpj/Rb1/Skp2/Rad21/Myc/Rangap1/Cenpm/Tuba1a/Mcm4/Rfc4/Dyrk1a/Pkmyt1/Tubb5/Cenpq/Sgo1/Ndc80/Kif20a/Cdc23/Cdc25c/Cep76/Ska1/Pola2/Incenp/Fen1/Smc3/Pola1/Ercc6l/Cenpi/Smc1a
REACTOME_DNA_REPLICATION	6.79E-47	1.24E-44	102	Mcm3/Prim2/Sgo2a/Clasp1/Mcm6/Cenpl/Nuf2/Ahctf1/Nsl1/Mcm10/Spc25/Ckap5/Kif18a/Bub1b/Knl1/Bub1/Pcna/Mcm8/E2f1/Dsn1/Ccna2/Pmf1/Itgb3bp/Orc1/Kif2c/Cdc20/Cdca8/Rpa2/Nudc/Rcc2/Cenps/Dbf4/Cenpa/Cdc7/Pole/Rfc5/Kntc1/Rfc2/Mcm7/Mad1l1/Rfc3/Mad2l1/Lig1/Psmc4/Pold3/Psma1/Plk1/Gins4/Cenpu/Mcm5/Orc6/Cenpt/Cenpn/Gins2/Cdt1/Nup133/Spc24/Zw10/Ppp2r1b/Kif23/Zwilch/Stag1/Fbxo5/Nup43/Zwint/Nup37/Nup107/Xpo1/Spdl1/Aurkb/Mis12/Pafah1b1/Ska2/Psmb3/Psmd3/Cdc6/Nup85/Birc5/Pole2/Gmnn/E2f3/Cenpp/Cenph/Rb1/Rad21/Rangap1/Cenpm/Mcm4/Rfc4/Cenpq/Sgo1/Ndc80/Kif20a/Ska1/Pola2/Incenp/Fen1/Smc3/Pola1/Ercc6l/Cenpi/Smc1a
REACTOME_MITOTIC_M_M_G1_PHASES	8.85E-41	1.21E-38	90	Mcm3/Prim2/Sgo2a/Clasp1/Mcm6/Cenpl/Nuf2/Ahctf1/Nsl1/Mcm10/Spc25/Ckap5/Kif18a/Bub1b/Knl1/Bub1/Mcm8/E2f1/Dsn1/Pmf1/Itgb3bp/Orc1/Kif2c/Cdc20/Cdca8/Rpa2/Nudc/Rcc2/Cenps/Dbf4/Cenpa/Cdc7/Pole/Kntc1/Mcm7/Mad1l1/Mad2l1/Psmc4/Psma1/Plk1/Cenpu/Mcm5/Orc6/Cenpt/Cenpn/Cdt1/Nup133/Spc24/Zw10/Ppp2r1b/Kif23/Zwilch/Stag1/Fbxo5/Nup43/Zwint/Nup37/Nup107/Xpo1/Spdl1/Aurkb/Mis12/Pafah1b1/Ska2/Psmb3/Psmd3/Cdc6/Nup85/Birc5/Pole2/Gmnn/E2f3/Cenpp/Cenph/Rad21/Rangap1/Cenpm/Mcm4/Cenpq/Sgo1/Ndc80/Kif20a/Ska1/Pola2/Incenp/Smc3/Pola1/Ercc6l/Cenpi/Smc1a
REACTOME_MITOTIC_PROMETAPHASE	3.06E-38	3.34E-36	61	Sgo2a/Clasp1/Cenpl/Nuf2/Ahctf1/Nsl1/Spc25/Ckap5/Kif18a/Bub1b/Knl1/Bub1/Dsn1/Pmf1/Itgb3bp/Kif2c/Cdc20/Cdca8/Nudc/Rcc2/Cenps/Cenpa/Kntc1/Mad1l1/Mad2l1/Plk1/Cenpu/Cenpt/Cenpn/Nup133/Spc24/Zw10/Ppp2r1b/Zwilch/Stag1/Nup43/Zwint/Nup37/Nup107/Xpo1/Spdl1/Aurkb/Mis12/Pafah1b1/Ska2/Nup85/Birc5/Cenpp/Cenph/Rad21/Rangap1/Cenpm/Cenpq/Sgo1/Ndc80/Ska1/Incenp/Smc3/Ercc6l/Cenpi/Smc1a

5.4 Comparing the APOBEC2 occupied regions in DNA with expression changes

As I demonstrated above, APOBEC2 mostly binds to promoters of genes. Next, I wanted to know whether APOBEC2 occupancy correlates with gene expression changes. I did this by comparing APOBEC2 occupancy at 14-hour (or 34-hour) with gene expression changes at 24-hour (or 48-hour), on the premise that occupancy will not result in immediate gene expression changes (e.g. at 14-hour or 34-hour), but rather changes later in time (though the precise time point of maximal gene expression changes that would correlate with occupancy remains undetermined).

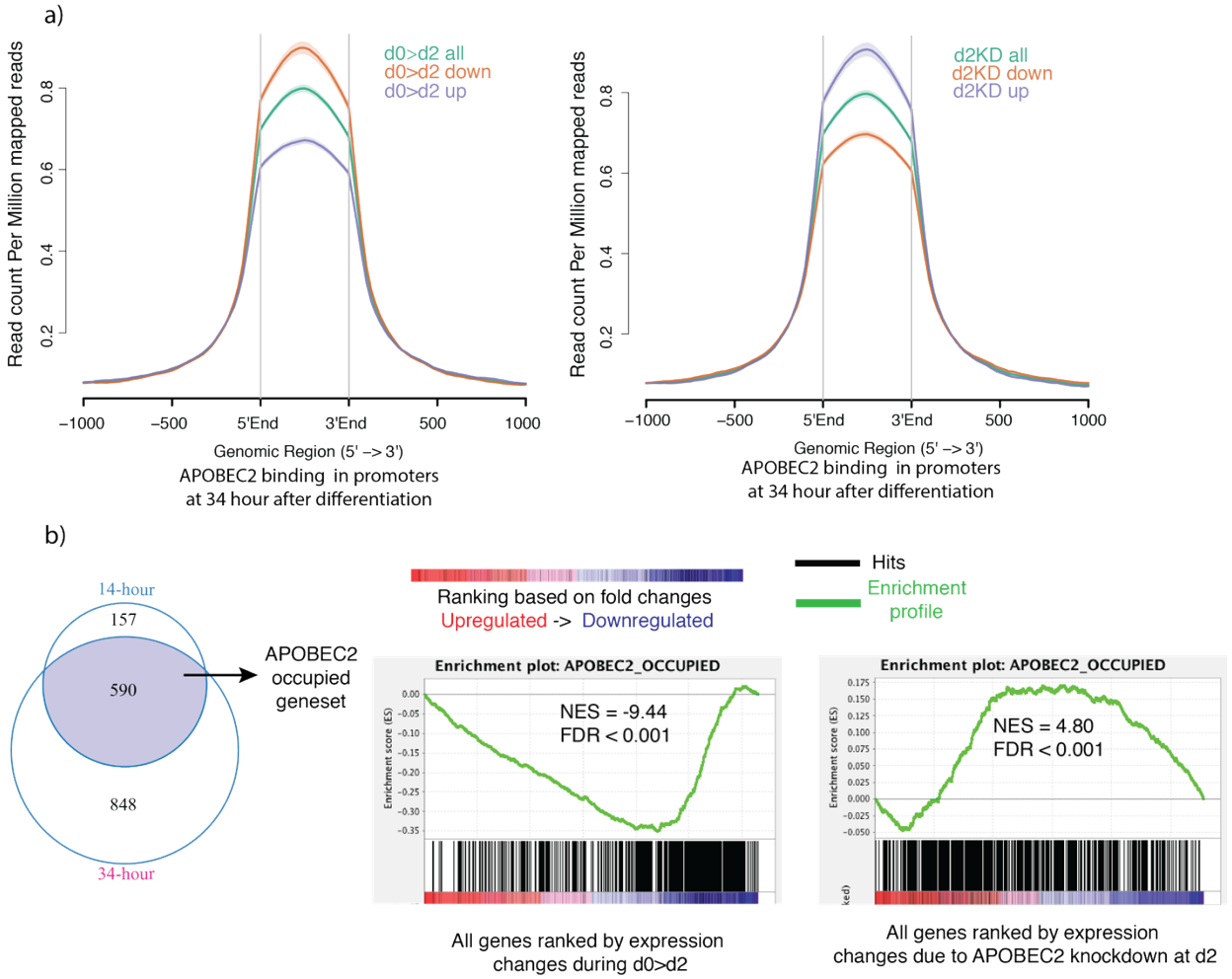
Within these constraints, I assessed the average APOBEC2 binding in genes that show increase or decrease in expression through differentiation. I determined the average profile of APOBEC2 signal (expressed as read counts per million mapped reads) on binding regions that map in all promoters (up to 4kb around TSS), the ones that map in promoters of genes that are upregulated through differentiation and the ones that map in promoters of genes that are downregulated through differentiation. The results show that APOBEC2 is bound to both genes that are upregulated and downregulated through differentiation. Interestingly genes that are downregulated through differentiation show higher APOBEC2 binding signal (Figure 5.6a). This result implies that APOBEC2 might be primarily involved in inhibiting transcription and consequently increasing concentrations of APOBEC2 would be needed to optimally inhibit transcription (this could explain why there is lower binding in genes that still continue to show increase in expression through differentiation). Alternatively, it is possible that APOBEC2 could have different effects on gene transcription depending on other protein binding factors/ regulatory elements, with specific interactions leading to higher binding signal and gene

inhibition and others with low abundance leading to gene activation. Also the APOBEC2 binding is higher in promoters of genes that are upregulated when the protein is knocked down, suggesting that normally during differentiation the binding of APOBEC2 is important for their inhibition (Figure 5.6a). Overall this analysis is suggestive of a role of higher APOBEC2 binding signal in promoters leading to inhibition of gene activation.

The analysis above showed the average APOBEC2 binding profile in promoters of occupied genes and how higher binding correlates with gene expression inhibition through differentiation and activation when the protein is absent. I was interested in further analyzing the APOBEC2 bound genes in relationship to the whole list of genes that are expressed in C2C12s. I decided to determine if the APOBEC2 occupied genes are randomly distributed in a list of genes that are ranked by expression changes (through differentiation or through APOBEC2 deficiency) or if they are preferentially enriched at the top or at the bottom of such ranked list. For this I performed a GSEA analysis. I created an *APOBEC2 occupied gene set*, using genes that show consistent APOBEC2 occupancy at both 14-hour and 34-hour time points. My analysis showed that the *APOBEC2 occupied gene set* is significantly enriched in the list of genes that are inhibited through differentiation. Moreover, this trend is reversed when I looked at the enrichment of the *APOBEC2 occupied gene set* in the ranked list of genes that are differentially expressed due to APOBEC2 deficiency, with the gene set showing enrichment in the list of genes that are upregulated (Figure 5.6b). This result reinforces the hypothesis that APOBEC2 binding in promoters correlates with gene expression inhibition. Overall our ChIP-Seq analysis indicates that APOBEC2 has a direct role in gene inhibition.

Figure 5.6 Correlation of APOBEC2 occupancy in promoters with subtle but coordinated gene expression changes

a) The mean normalized APOBEC2 signal (plotted as read counts per million mapped reads) across all binding regions at 34-hour, that fall in gene promoters. All of the peaks (green) were split between the ones that fall in genes that are upregulated (log2 fold change >0, purple), downregulated (log2 fold change<0, brown) through C2C12 differentiation (left) or due to APOBEC2 knockdown at day2 (right). Plot shows the global differences in APOBEC2 binding in promoters of genes grouped by gene expression differences. All conditions are in biological triplicates b) The genes that show consistent APOBEC2 occupancy in their promoters at both time points were used to create an APO2_OCCUPIED gene set as shown in the venn diagram. We used GSEA to test the enrichment of the APO2_OCCUPIED gene set in the list of genes that are differentially expressed through differentiation (left) or the ones that are differentially expressed due to APOBEC2 knockdown at day2 (right). GSEA score curves of the gene expression changes through differentiation from day0 to day2 (left) and due to APOBEC2 knockdown at day2 (right). NES and FDR values are shown. The enrichment profile over the whole ranked gene set is shown in green. Overlaps (hits) are shown in black lines. A positive ES indicates gene set enrichment at the top of the ranked list; a negative ES indicates gene set enrichment at the bottom of the ranked list.


a)
Read count Per Million mapped reads
0.2
0.4
0.6
0.8
d0>d2 all
d0>d2 down
d0>d2 up
−1000
−500
5'End
3'End
500
1000
Genomic Region (5' –> 3')
APOBEC2 binding in promoters
at 34 hour after differentiation
Read count Per Million mapped reads
d2KD all
d2KD down
d2KD up
Genomic Region (5' –> 3')
APOBEC2 binding in promoters
at 34 hour after differentiation
b)
14-hour
157
590
848
34-hour
APOBEC2
occupied
geneset
Ranking based on fold changes
Upregulated -> Downregulated
Hits
Enrichment
profile
Enrichment plot: APOBEC2_OCCUPIED
NES = -9.44
FDR < 0.001
Enrichment score (ES)
All genes ranked by expression
changes during d0>d2
Enrichment plot: APOBEC2_OCCUPIED
NES = 4.80
FDR < 0.001
All genes ranked by expression
changes due to APOBEC2 knockdown at d2

Previously I showed that the cell cycle related pathways are significantly enriched in the list of APOBEC2 bound genes. Similarly to the analysis above, I decided to determine the trends of gene expression changes and APOBEC2 overall occupancy in these specific group of genes. I extracted the genes from *the MSigDB Reactome cell cycle* (Croft et al., 2014; Fabregat et al., 2018; Liberzon et al., 2011) gene set and mapped the average APOBEC2 binding on the promoters of these genes. As predicted, overall the cell cycle genes that are downregulated through differentiation show higher APOBEC2 binding signal. Additionally, the average APOBEC2 binding is higher in cell cycle genes that show increase in expression due to APOBEC2 deficiency (Figure 5.7a).

The GSEA analysis shows significant enrichment of the *APOBEC2 occupied geneset* in the list of cell cycle genes that are downregulated during differentiation and in the list of cell cycle genes that are upregulated due to APOBEC2 knockdown. Overall this analysis indicates that APOBEC2 has an active role in the inhibition of cell cycle genes during differentiation and it suggests that in its absence, these genes are not downregulated properly, which could directly lead to the phenotype observed. (Figure 5.7b). Taking a closer look at the cell cycle genes that are upregulated in the absence of APOBEC2 we detect (among many others) Rbl1, a pocket protein whose inactivation facilitates G1/S transition; Cdc6 and Cdc7, which are important in G1/S transition and initiation of DNA replication; Mcm3/4/6/5, important in DNA replication; Ckd1, Cdc25c, important for the onset of mitosis; Ccnb1/Ccnb2 important in G2/M transition; Bub1b and Mad2l1, important in mitotic checkpoint that ensure proper chromosome segregation (Figure 5.7c). This suggests that normally APOBEC2 binds to and leads to inhibition of a

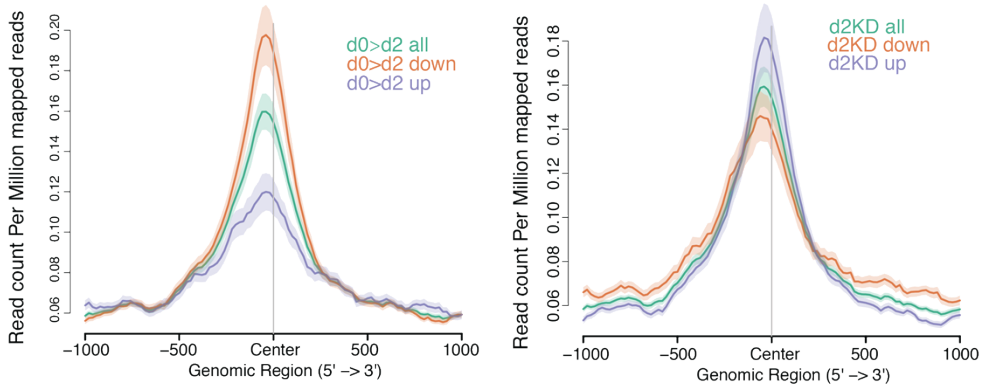
diverse number of genes in multiple steps of the cell cycle. My data support the idea that APOBEC2 is important in the inhibition of genes across all stages of cell cycle.

Collectively my experiments and analysis so far suggest a role of APOBEC2 in gene expression inhibition during myoblast (C2C12) differentiation.

Figure 5.7 Correlation of APOBEC2 occupancy in promoters with subtle but coordinated gene expression changes in cell cycle pathway

a) The mean normalized APOBEC2 signal (as read counts per million mapped reads) across all binding regions at 34 hours, located in promoters of the cell cycle related genes as defined from the MSigDB Reactome Pathways gene set collection. All of the peaks (green) were split between the ones that fall in genes that are upregulated (log2 fold change >0, purple), downregulated (log2 fold change<0, brown) through C2C12 differentiation (upper graph) or due to due to APOBEC2 knockdown (lower graph). Plot shows the global differences in APOBEC2 binding between the different gene lists. All conditions are in biological triplicates b) The genes with consistent APOBEC2 occupancy in their promoters at both time points were used to create an *APOBEC2 occupied geneset*. We used GSEA to test the enrichment of *APOBEC2 occupied geneset* in the list of the expressed genes of the MSigDB Reactome Cell Cycle gene set that are differentially expressed through differentiation (left) or the ones that are differentially expressed due to APOBEC2 knockdown at day2 (right). GSEA score curves of the gene expression changes through differentiation from day0 to day2 (left) and due to APOBEC2 knockdown at day2 (right). NES and FDR are shown. The enrichment profile over the whole ranked gene set is shown in green. Overlaps (hits) are shown in black lines. A positive ES indicates gene set enrichment at the top of the ranked list; a negative ES indicates gene set enrichment at the bottom of the ranked list c) The bar plot shows gene expression changes (by log2 fold changes) due to APOBEC2 knockdown at day 2 of differentiation for all of the expressed genes of the MSigDB Reactome Cell Cycle gene set (in gray). Genes that are bound in their promoters at 34hr after inducing differentiation and that are upregulated (red) or downregulated (blue) in the absence of APOBEC2.

a)



APOBEC2 binding at 34 hour after differentiation In promoters of cell cycle genes

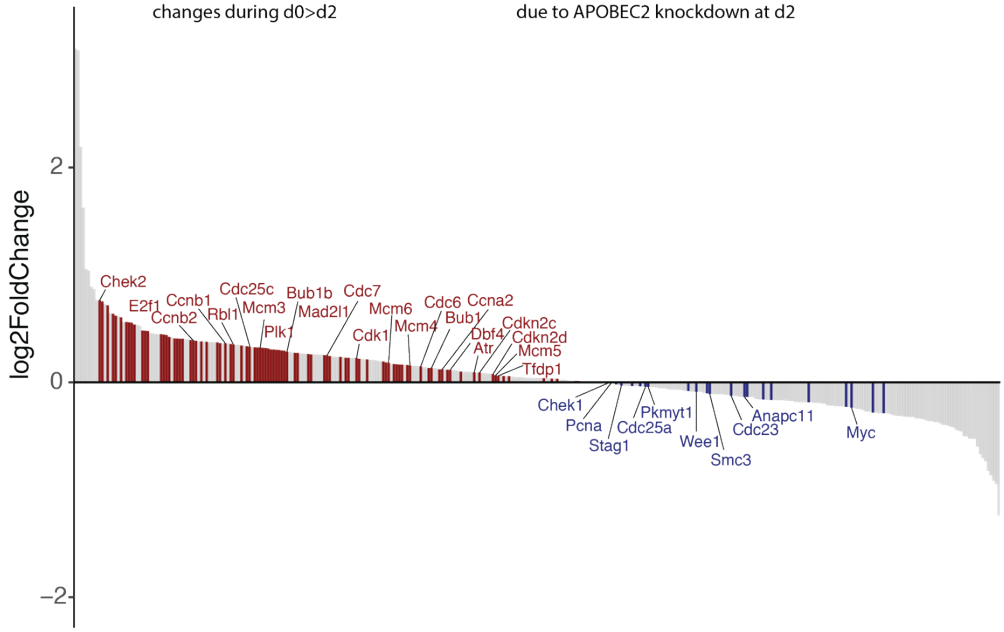
APOBEC2 binding at 34 hour after differentiation In promoters of cell cycle genes

b)

All cell cycle genes ranked by expression changes during d0>d2

All cell cycle genes ranked by expression changes due to APOBEC2 knockdown at d2

c)



5.5 Bioinformatics analysis of APOBEC2's consensus DNA binding motif

When I started my PhD very little was known about the enzymatic activity of APOBEC2 and its targets. APOBEC2's binding activity, deaminase substrate, sequence specificity and the physiological genetic element(s) it binds to were not known (and remain so in the literature, to date). Over the course of my project, I have identified that APOBEC2 binds to DNA and also established a list of genes that are highly likely to be substrates of APOBEC2 (either direct or indirect through binding with other chromatin factors). This opens up a new avenue for us to explore any potential APOBEC2 sequence specificity. I next determined the presence of any sequence patterns (or motifs) in the APOBEC2 binding regions. The goal of doing this is to find the most enriched motifs that could be best representing the potential enzymatic substrate of APOBEC2. This is important for future work in the field in determining the mechanism of how APOBEC2's binding to these genomic regions could lead to gene expression changes.

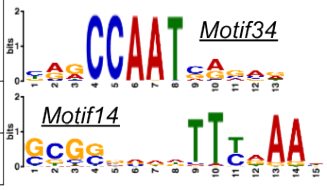
I used MEME-ChIP (Bailey et al., 2009; Machanick and Bailey, 2011) to predict *de novo* DNA binding motifs on the binding sites of the 14-hour and 34-hour time point. The two of the most significant binding motifs are *Motif 14*: GCGSSVRDTTYRAAH (identified by MEME scanning the binding regions at the 14-hour time point and *Motif 34*: YRGCCAATSRGMR (identified by MEME scanning the binding regions at the 34-hour time point). Furthermore I used FIMO (Grant et al., 2011) to scan the APOBEC2 binding regions for the presence of the identified motifs. I determined that *Motif 14* is present in ~38 and 13% of the binding sites, for the 14-hour and 34-hour time point respectively. While *Motif 34* is present in 25% and 24% of the binding sites for the 14-hour and 34-hour time point respectively (Figure 5.8a).

Both of the motifs are distributed within 1kb of the TSS for both time points (Figure 5.8b). Overall, about 79% of the genes that are bound by APOBEC2 in the 14-hour time point continue to remain bound at the 34-hour time point. This is also evident in the overlap in the occurrence of the motifs discovered from binding locations. *Motif* 14 that is enriched in the 14-time point binding regions is also present in the 34-hour time point but at a lower overall percentage. *Motif 34* that is enriched in the 34-hour time point is found in both the 14-hour and 34-hour time point.

The data overall is suggestive of persistent APOBEC2 binding in at least a portion of the occupied genes. This is an interesting result suggesting that APOBEC2 could be a static inhibitor of transcription through differentiation.

a)

MOTIF	% of occupied genes with at least one motif		E Value
	14hr	34hr	
Motif34 YRGCCAATSRGMR	25%	24%	6.4e-336
Motif14 GCGSSVRDTTYRAAH	38%	13%	4.10E-303



b)

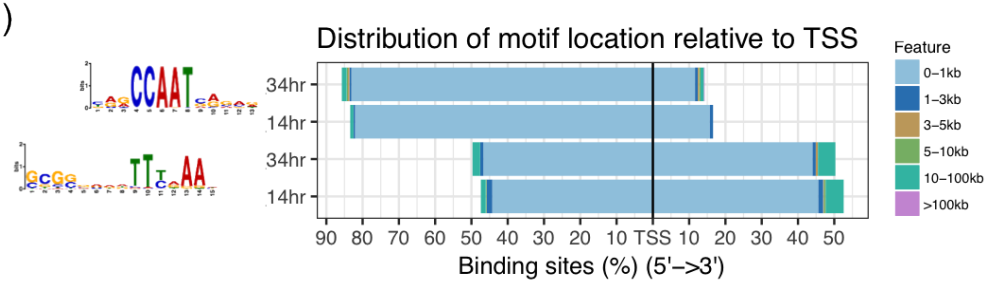


Figure 5.8 Predictions and validations of APOBEC2's consensus DNA binding motif

Top motifs –sequence patterns, that are the most enriched in APOBEC2 consensus peaks (peaks from the majority of biological repeats) were identified using MEME-chip. E-value represents the statistical significance of the de novo motif identification. Table shows percent of motifs found in genes with peaks in their promoters using FIMO.

The bargraph shows the distribution of the motif location relative to the TSS for both timepoints for each motif

CHAPTER 6. Discussion

When this thesis project was starting many questions existed about APOBEC2. First, there was no evidence as to whether or not APOBEC2 at physiological levels can act as an RNA editor or as a DNA demethylator as has been previously hypothesized. Secondly, neither the substrate of APOBEC2 (presumably nucleic acid) nor the physiological target (in RNA or DNA) was known. Consequently, neither APOBEC2's potential enzymatic activity nor its mechanism of action was known. Through my thesis work, I was able to tackle both of these questions. This work was mainly possible because of the advent of high throughput sequencing technologies, impacting the process of searching for gene targets in an exhaustive manner through transcriptome and genome wide analysis. The results, which were presented during the previous chapters, will be discussed in more depth here.

6.1 APOBEC2's importance in muscle biology

6.1.1 Results using immortalized satellite cells

My experiments provide some genetic evidence that APOBEC2 is important in muscle biology in a cell culture model and also in an animal model. Working with myoblasts (C2C12s), I demonstrated that APOBEC2 is an important factor for timely progression to cell cycle exit and proper muscle differentiation. Working with the APOBEC2$^{-/-}$ mouse model I showed that even though there are no apparent changes in the rate of myogenesis during *de novo* muscle formation, steady state muscle is qualitatively and quantitatively different. APOBEC2$^{-/-}$ animals have more muscle fibers that are smaller in size, show a

reduction in muscle strength and have muscle with mispositioned nuclei much earlier in life than previously described (Sato et al., 2010). These results collectively link APOBEC2 to postnatal muscle maturation and maintenance.

During the course of this work other research groups have provided additional evidence that corroborate some of my findings. *Carrió et al* utilized embryonic stem cells that can be transformed to take a muscle-specific cell fate through inducible overexpression of Pax7 (iPax7 ES). In these cells, upon transformation with Pax7 and induction of differentiation, the levels of APOBEC2 expression increase - similarly to my findings using the C2C12 model. Conversely when APOBEC2 is knocked down in the iPax7 ES cells, this dramatically impairs the expression of differentiation markers such as Myog and, ultimately, muscle differentiation (Carrió et al., 2016). A similar result was published by *Vonica et al* induced APOBEC2 knockdown (in C2C12s) using siRNAs and also observed repression of Myog expression (Vonica et al., 2011). In contrast to these studies, *Ohtsubo et al.* used knockdown experiments in primary myoblast cells but did not observe the same delay in differentiation phenotype (Ohtsubo et al., 2017a). A significant factor that could account for such differences in the cell culture systems is the timing of differentiation induction through shift in low serum. In my experiments, as well as those by *Carrió et al.* (where the same siRNA sequences were used for APOBEC2 knockdown as in my studies) and *Vonica et al*., APOBEC2 is knocked down at the myoblast stage prior to the induction of differentiation. Therefore there is little APOBEC2 either before or after the differentiation stimulus. This clearly results in an inhibition in expression of Myog ((Carrió et al., 2016; Vonica et al., 2011) and this thesis) supporting the notion that APOBEC2 is important for proper expression of Myog

during myoblast differentiation. Conversely, in *Ohtsuboa et al.* initiation of the APOBEC2 knock down is concurrent with the induction of differentiation (reducing protein abundance probably within the first day of induction, a time at which cells are exiting cell cycle) (Ohtsubo et al., 2017a). Surprisingly this resulted in the opposite phenotype: upregulation of Myog and other differentiation markers and no delay in differentiation. At face value, this discrepancy might simply suggest that APOBEC2 is required earlier during myogenesis for proper cell cycle exit (something that will be discussed more extensively below).

6.1.2 Results using the APOBEC2$^{-/-}$ mouse model

In contrast to these discrepancies in cell line derived data, my results of a role of APOBEC2 in fiber size but absence of a delay in *de novo* myogenesis phenotype in APOBEC2$^{-/-}$ are completely congruent with *Ohtsubo et al*. The latter study additionally shows an increase in the number of smaller primary myotubes formed in APOBEC2$^{-/-}$ mice during the early stages of regeneration and also during the initial days of differentiation of APOBEC2$^{-/-}$ myoblasts in cell culture (Ohtsubo et al., 2017a).

During *de novo* myogenesis activated satellite cells (or proliferating myoblasts) have a choice: either exit cell cycle and differentiate or exit cell cycle and return to quiescence. Given this dual potential of the cells here and my results linking APOBEC2 to cell cycle regulation during C2C12 differentiation it would be interesting to look for any APOBEC2 dependent changes in cell cycle regulation in the earlier stages of regeneration in the mouse model. Indeed a recent report used single-myofiber cultures from the APOBEC2$^{-/-}$ mice and demonstrated a reduction in the pool of quiescent satellite

cells, suggesting that APOBEC2 could be regulating the withdrawal of activated cells from cell cycle and their return to quiescence (G0) (Ohtsubo et al., 2017b). This also agrees with the phenotype of increased fiber number that both *Ohtsubo et al*. and I observe, potentially due to less cells returning to quiescence thus tipping the balance towards more proliferating myoblasts and more fibers formed. It would also be interesting to determine if the defects in cell cycle potentially resulting in hyper proliferation could lead to satellite cell exhaustion in aging mice or following repeated cycles of muscle degeneration- regeneration. This could explain the development of myopathy and atrophy in the mice lacking APOBEC2 (Sato et al., 2017).

6.1.3 Overall role of APOBEC2 in muscle development and maintenance

At the organismal level, it is clear that APOBEC2 is important in muscle health. APOBEC2$^{-/-}$ lose muscle strength as shown through the inverted grip test (Figure 2.4f) and APOBEC2$^{-/-}$ mice display impaired exercise capacity and mitochondrial defects that increase muscle mitophagy, ultimately leading to myopathy and muscle atrophy (Sato et al., 2017).

At the cellular level, APOBEC2$^{-/-}$ mice have mispositioned nuclei in muscle cells, which is considered a pathological marker of muscle disorders and their presence has been associated with ongoing muscle repair. Nevertheless it remains to be determined if this is the consequence or the cause of myopathy (Folker and Baylies, 2013). Recent research has shown that muscle contraction is necessary and sufficient for peripheral nuclear movement (Roman et al., 2017; Rosen and Baylies, 2017). It would be informative to determine if APOBEC2 depletion is associated with defects in muscle

contraction. Regardless, a likely hypothesis to explain mispositioned nuclei, based on my data, would be satellite cell dysfunction: specifically, C2C12 (satellite) cells have a defect in cell cycle exit, and if I extrapolate from that to the mouse, I can imagine a scenario where lack of cell cycle exit correlates with a reduction in self renewal (a reduction of cells that return to quiescence). This would then result in more proliferating cells that are constantly fusing with the myotubes - resulting in the phenotype of haphazardly nucleated fibers, observed in the mice.

Also at the cellular level, APOBEC2$^{-/-}$ animals are thought to have a higher proportion of slow muscle fibers (as opposed to fast) suggesting that normally the protein is important in muscle remodeling - allowing the muscle to adapt to specific conditions. It might be interesting to check whether during *de novo* myogenesis there are any defects in the proportion of fiber types (slow vs. fast), which would suggest that APOBEC2 might be influencing cell fate (i.e. the differentiation towards a specific muscle fiber type).

The cellular and organismal data generated thus far speak to a role for APOBEC2 in muscle maintenance. It is possible that APOBEC2 is really not involved in the development of muscle tissue. Conversely, it is also possible that the lack of a phenotype in APOBEC2$^{-/-}$ during development is due to redundancy with another factor. For example Myf5 and MyoD can functionally substitute for one another during myogenesis and inhibition in muscle development is observed only when both genes are null (Rudnicki et al., 1993). It would be interesting to look for upregulation of any of the MRFs or AID/APOBECs during development in the APOBEC2$^{-/-}$ mice and also follow this up with mouse genetics studies where multiple MRFs (or other AID/APOBEC) and APOBEC2 are deleted. Interestingly in the C2C12 differentiation model, when

APOBEC2 is knocked down, APOBEC3 (which is normally downregulated through differentiation) is significantly upregulated (log2 fold change 1.03, adjusted p value 6.7e-05). It is unclear whether APOBEC3 is compensating for the lack of APOBEC2 or if its expression levels are influenced by the upregulation of the interferon response due to APOBEC2 knockdown (Table 3.2). If inhibition of both APOBEC2 and APOBEC3 could lead to a more drastic phenotype then it could be supportive evidence for this possibility.

In conclusion I would like to emphasize that the collective work presented here and previous published work suggest that APOBEC2 is likely fulfilling multiple roles in muscle at different stages of muscle development, for example in cell cycle regulation during myogenesis (or possibly satellite cell renewal) or in muscle remodeling or growth in fully formed muscle.

6.2 APOBEC2 does not fit in the traditional mold of AID/APOBEC family

AID/APOBEC family members (as discussed in the introduction chapter) have been shown to be perform diverse biological roles through various mechanisms: RNA editing, DNA editing and modifying DNA 5mC levels. My analysis using next generation sequencing methods addressed some of these potential roles for APOBEC2.

A direct conclusion from my data is that APOBEC2 is not acting as an RNA editor (in either C-to-U or A-to-I). In both proliferating myoblasts and differentiating myotubes no such activity could be observed. Moreover previous work in the lab looking at potential RNA editing in mature muscle tissue using APOBEC2$^{-/-}$ (Fritz, 2014)

supports the same conclusion: no changes in RNA editing are observed, while using pipelines fitted to detect RNA editing.

Furthermore, the methylation studies presented here rule out the possibility that APOBEC2 acts as a DNA demethylase in myoblasts (C2C12s) since only sporadic hypermethylated events (which do not correlate with expression changes in the corresponding genes) were detected at any of the timepoints after differentiation when APOBEC2 is absent. However, the caveat remains that the C2C12 model might not be representative of what happens *in vivo* in mouse. I cannot completely rule out the possibility that APOBEC2 could mediate some methylation changes during muscle development. For example previous work on AID shows no affects in DNA methylation in an *ex vivo* stimulated B cell model (Fritz et al., 2013) but another study showed that AID is essential for demethylation and diversification of methylome in germinal center B cells (GCBs) *in vivo* (Dominguez et al., 2015). As well, for my work I had to pick timepoints - which I did based on expression levels of APOBEC2 (I hypothesized that high expression levels would most optimally correlate with function). However I may have missed timepoints where methylation change would transiently occur. I believe that this is unlikely since recent reports looking at genome scale DNA methylation changes during muscle- lineage determination and terminal differentiation (fully formed muscle) demonstrate the presence of significant DMRs (majority hypermethylated) but no methylation changes between primary myoblasts and myotubes when differentiated in culture (Carrió et al., 2015; Tsumagari et al., 2013). This is in accordance with what I observe in my cell culture system. It is also possible that eRRBS might not be capturing regions in which methylation changes dependent on APOBEC2 do occur (RRBS tends to

cover most deeply promoter and coding DNA sequence-CDS regions with very little coverage of intergenic sites and potential enhancer sites). As well it is also possible that very few APOBEC2 dependent changes occur but that those are crucial: for instance, work with the iPax7 ES cell model (described above) suggested that APOBEC2 reduces Moyg-associated DNA demethylation through differentiation (Carrió et al., 2016) which would explain Myog's repression. This site is not well covered in my eRRBS data and there maybe others that behave similarly. Thus, there might be few key DNA demethylation events that could be mediated by APOBEC2 and that can be undetectable in our RRBS analysis because of low signal to noise ratio or lack of coverage in the affected regions (low sequencing read representation). Moreover, in the future it would be important to check whether there are any significant DNA methylation changes, specifically in the promoters of genes that were shown through ChIP-Seq to be directly bound by APOBEC2. In all, my work does not support the hypothesis of a role of APOBEC2 in DNA deamination.

Lastly as discussed more extensively below, APOBEC2 seems to be a chromatin-binding factor. It is unclear whether APOBEC2 can bind to DNA directly or whether it does so through binding to a cofactor/complex. Currently we are working on determining whether there is physical direct binding and detectable DNA editing activity in regions of the DNA where APOBEC2 is enriched. This would be an interesting avenue to follow up on since deamination at TSS can induce DNA breaks (which in the satellite cell system have been linked to differentiation -specifically, muscle differentiation is presumed to be dependent on targeted DNA strand breaks through the action of nuclease caspase-activated DNase (CAD) that lead to gene expression changes (Larsen et al., 2010). It is

tempting to speculate that APOBEC2 could be implicated in this process. Beyond these early findings, recent research has shown that transient DNA strand breakage exerts an effect on cell fate, acting to limit stem cell self-renewal and stimulate differentiation, which has been shown to be important in case of many tissues, including muscle (Al-Khalaf et al., 2016; Behrens et al., 2014; Larsen and Megeney, 2010; Larsen et al., 2010; Narciso et al., 2007). Overall, determining if APOBEC2 has a DNA editing role would be an important next step and could bring us closer to understanding better APOBEC2's activity.

6.3 APOBEC2 adds a novel role to the AID/APOBEC family of enzymes

My data provide molecular evidence of the importance of APOBEC2 at the transcriptional level. I have demonstrated that APOBEC2 is a regulator of gene expression, which is a novel role, previously undescribed for any of the AID/APOBEC family members. APOBEC2 (directly or indirectly) affects gene networks important for muscle differentiation and cell cycle regulation as demonstrated by the deregulation of the myogenesis pathway genes and the E2F target genes. I showed that two central genes important in myogenesis and cell cycle exit, Myog and *Cdkn1a*/p21 are downregulated due to APOBEC2 knockdown, suggesting that normally APOBEC2 might be acting upstream of these factors and affecting their proper induction during cell differentiation. It is still unclear how APOBEC2 might be mediating this process. In the future it would also be interesting to determine if expressing Myog or *Cdkn1a*/p21 would be sufficient to rescue the phenotype observed.

But how does APOBEC2 mediate transcriptional regulation? Since it is unlikely to be a demethylase or an RNA editor, this outcome can only be due to its direct or

indirect binding to DNA, and some catalytic activity that remains unresolved but which can mediate the transcriptional changes I observe. In that context, I have shown APOBEC2 acting as direct chromatin binder. In differentiated myotubes (Figure 5.1) I showed that even though the majority of the protein resides in the cytoplasm I could detect some nuclear APOBEC2 tightly bound to the chromatin. The discovery that APOBEC2 localizes in the nucleus and binds to chromatin is supported by recent studies in zebrafish where it was shown that sumoylation regulates APOBEC2 nuclear exclusion but once in the nucleus APOBEC2 can regulate the binding of POU6F2 (a homeodomain transcription factor associated with retinal differentiation) to DNA (Powell et al., 2014). For future work, it would be interesting to determine the dynamics and timing of APOBEC2 nuclear localization during the differentiation process as this could provide additional insights about the regulation of APOBEC2's activity and functional consequences of inducing its shuttling to the nucleus.

I have shown that APOBEC2 binds to genes (close to their TSS) consistently at both time points assessed (14-hour and 34-hour post differentiation) suggesting substantial duration of occupancy at the relevant promoters. It would be interesting to see if this result would hold true if we assess APOBEC2 binding during the myoblast proliferation stage or at later time points in the terminal differentiation or in the mature muscle tissue. The number of true APOBEC2 binding sites might be even greater than what we have reported here since our analysis showed that further optimization of the ChIP-Seq or increase in sequencing depth could increase the signal detected. Moreover the analysis shown here is used to detect narrow binding regions which are characteristic of transcription factor binding (narrow peaks), it would be interesting to look for other

types of binding such as broader binding regions characteristic of bigger complexes or histones (broad peaks).

I have also shown that the top enriched pathways among the APOBEC2 bound genes are cell cycle related pathways. This is in accordance with our phenotype of defects in cell cycle exit during differentiation when APOBEC2 is depleted. I also show that APOBEC2 occupied genes (including the cell cycle related genes) are overrepresented in the list of genes that are normally inhibited through differentiation (and upregulated when APOBEC2 is depleted) suggesting that APOBEC2 might have a role in transcription inhibition (Figure 5.6; 5.7) The latter is also corroborated with the conclusion that higher APOBEC2 binding correlates with gene repression through differentiation. Tying together the transcriptome data to the APOBEC2 genome wide binding data turned out to be fruitful here in generating hypothesis as to how APOBEC2 is operating. However, the constrain in such comparison is that expression changes and occupancy might not be exactly congruent. It is not known how much correlation we should expect between APOBEC2 binding to a gene estimated by ChIP-Seq peaks and transcript abundance for the gene estimated by RNA-Seq (Cheng et al., 2012). It would be interesting to see the consequences of APOBEC2 binding at later time points through differentiation, which could provide additional evidence for APOBEC2's role in gene repression. Finally, I have yet to perform validation experiments on the ChiP-Seq data, mostly because I am waiting to generate a CRISPR-knockout of APOBEC2 in C2C12 cells. The reason I want to compare the control to a knockout and not a knockdown is that the antibody I used is efficient enough to pick up even small amounts of protein that might still be present in the knockdown samples, resulting in DNA fragments that after amplification rounds are

still highly represented (and very small differences between wildtype and knockdown samples via ChIP-Seq data). Thus validation experiments are still ongoing in the lab.

Overall, more work is required to determine exactly how APOBEC2 inhibits transcription. One possibility is that APOBEC2 is part of a repressive epigenetic complex; unpublished Bio-ID data from a colleague (Javier Marcelo Di Noia, Poorani Ganesh Subramani, McGill University) who used APOBEC2 as a control for a study in AID in protein:protein interaction experiments in 293T and B cells demonstrates that APOBEC2's neighbors include exactly such a complex, comprising HDAC1, PRMT7, SMARCA5, MBD3, RIF1 and intriguingly, H2AF(data not shown). It would be worth to do similar experiments in C2C12. Another possibility is that APOBEC2 is an inactive deaminase, which recognizes and binds a specific nucleotide sequence or perhaps a specific nucleotide modification, thus recruiting such a complex to specific promoters. Whether the specific promoters are recognized by APOBEC2 and also by the repressor complex, or independently by APOBEC2 remains to be determined.

6.4 Getting closer to understanding how APOBEC2 is mediating its biological role

The discovery of potential promising APOBEC2 binding targets presents an opportunity to gain insights into APOBEC2's activity. One way I attempted to do this is by bioinformatically searching for overrepresented sequence patterns (motifs) that could best represent the enzymatic substrates. Of all of the predicted motifs the top enriched ones were shown (Figure 5.8) containing CCAAT and TTYRAA. Interestingly most of these motifs are within 1kb of the transcription binding sites and some are very close to the start (within 50bp). These motifs could be crucial tools in *in vitro* (biochemical) analyses

of function. Another interesting question regarding APOBEC2 is how does it determine which promoter regions to bind to. Its binding to promoters could be determined by (1) a specific DNA motif (2) other interacting proteins (3) a combination of the above. It would be interesting to determine if the motifs that we have identified are significantly enriched in the APOBEC2 bound genes as compared to all of the promoters. This could give us more insights into APOBEC2's mechanism of mediating gene repression.

There are some important questions that still remain to be answered and help elucidate how APOBEC2 is mediating its function. *Is the ZDD domain necessary for APOBEC2's function during myoblast differentiation?* One way to answer this question is to try to recapitulate the activity of APOBEC2 using a catalytic mutant version of the protein. *Does APOBEC2 directly bind to the DNA?* An electrophoretic mobility shift assay (EMSA) using the substrates or the motif sequence could provide some answers. Additionally, utilizing the substrates in a mutation assay could provide evidence for mutator activity. As mentioned, our sequencing datasets also provide an avenue, where we can scan for possible mutations. *Does APOBEC2 bind to other cofactors or epigenetic remodelers?* Immunoprecipitation of such complexes combined with mass spectrometry could corroborate the BioID data. *Is the function of APOBEC2 in cell cycle applicable to other systems?* There is increasing evidence that APOBEC2 is expressed in systems beyond the cardiac and skeletal muscle and that in some of these it mediates functional effects (Powell et al., 2014). APOBEC2 is also expressed in mouse neurons (Sharma et al., 2016), in placenta (not published, personal communication), immune cells (Immgen database). Moreover, changes in APOBEC2 expression (either upregulated or downregulated) in humans also correlate with breast cancer, sarcoma, colorectal

carcinoma, non-small cell lung carcinoma (Expression Atlas, EMBL-EBI) etc. Therefore elucidating whether or not APOBEC2 has a mutator or epigenetic role would be important for elucidating important gene regulatory mechanisms that additionally might have a role in cancer.

6.5 Concluding remarks

The excitement around AID/APOBEC exists (at least partly) because of how they mediate alterations (resulting in mutations or editing events) at the level of polynucleotides to fulfill a biological role. Their discovery presented a novel, surprising (and fascinating!) mechanism of expanding genome diversity. Further research through the years has shown that AID/APOBEC(s) are important in diverse biological systems with implications in health and disease. At the time I started working on this thesis project, APOBEC2 was considered the most mysterious and obscure protein of the APOBEC/AID family. Few things were already known about APOBEC2's biological role but nothing was known as to how it mediates its role and as to whether the mechanism is similar to the other AID/APOBEC enzymes. The collective work in my thesis disproves the hypothesis that APOBEC2 is mediating DNA demethylation or RNA editing (on C or in A) similar to other members of AID/APOBEC. Instead, my work demonstrates that APOBEC2 has a clear role (direct or indirect) as a transcriptional repressor through binding (directly or possibly as part of a complex) to promoters of genes relevant to muscle differentiation (and specifically to cell cycle exit). I am hopeful that the specific binding targets I discovered will be utilized to elucidate the precise mechanism of how APOBEC2 affects gene expression and thus to decipher its broader

role (potentially beyond muscle) in gene expression regulatory mechanisms and/or in driving genome biological diversity.

CHAPTER 7. Materials and Methods

7.1 C2C12 cell culture

C2C12 cells (CRL-1772, ATCC) were maintained in DMEM (30-2002, ATCC) with 10% fetal bovine serum (FBS) and feed every two days. To differentiate equal number of cells (0.25 x 10^6) were seeded in 6-well plates followed by media change to DMEM with 2% horse serum after 12 hours. For generating single cell clones for RNA-seq and RRBS experiments C2C12s were sorted using fluorescence-activated cell sorting (FACS) and seeded into a 96 well plate. Each clone was expanded and tested for successful knockdown through immunoblotting.

7.2 Protein knockdown using lentivirus infection

C2C12s were infected with lentiviruses carrying shRNAs, targeting APOBEC2 (or GFP as a control), which allows for infection of both proliferating and non-proliferating cells and provides a stable repressive effect. All APOBEC2 shRNAs were obtained from The Broad Institute's Mission TRC-1 mouse library and present in pLKO.1-puro construct, which carries the puromycin-resistance gene and drives shRNA transcription from a human U6 promoter. Plasmids used: pLKO.1 - TRC cloning vector (Addgene, # 10878) (Moffat et al., 2006); pLKO.1 puro GFP siRNA (Addgene, # 12273) (Orimo et al., 2005). The design of shRNAs and cloning in pLKO.1-TRC, were done according to the Addgene protocol (Protocol Version 1.0. December 2006). The following siRNAs sequences were used for APOBEC2 knockdown:

shRNA#1	CCTGGCTTCCTGATTCTACTT
shRNA#2	GCCTCTCAGAATGGAGATGAT
shRNA#3	CTTCCTATACTATGAGGAGAA
shRNA#4	GCTACCAGTCAACTTCTTCAA
GFPsiRNA,	GCAAGCTGACCCTGAAGTTCAT

Lentiviruses were produced by co-transfection of pLKO.1-puro short hairpins containing construct, packaging plasmid psPAX2 (Addgene, #12260) and envelope plasmid pMD2.g (Addgene, #12259) in HEK 293T cells. Transfections were done using Lipofectamine 2000 Reagent (Lifetechnologies) as per manufacturer instructions. Supernatants with lentiviral particles were collected at 24 and 48 hours after transfection, passed through a 0.45 mm filter and applied to C2C12s. For APOBEC2 constitutive knockdown 30% confluent cells were infected with pLKO.1 containing lentiviruses in growth media containing 8ug/mL polybrene for 12 hours. Two days after lentiviral infection cells were cultured with 4ug/ml virus-free puromycin containing media for two more days to select cells stably expressing the shRNAs.

7.3 Cell cycle and cell proliferation analysis

Equal number of cells (0.125 x 10^6) were seeded in 6-well plates followed by media change to DMEM with 2% horse serum after 12 hours. C2C12s were harvested by trypsinization and stained with EdU (Thermofisher, Clik-iT EdU, C10337) as per manufacturer instructions. Following washes of the Click-it solution cells were resuspended with 0.05 mg/ml Propidium Iodide (PI) and 0.1 mg/ml RNAseA mix, transfer into Eppendorf tubes and stained for 30 min at room temperature in the dark and analyzed by flow cytometry. Data were gated to exclude apoptotic cells.

7.4 Immunoblotting

For immunoblotting experiments, the C2C12s were first washed with cold 1xPBS and lysed in 100ul RIPA lysis buffer (santacruz, sc-24948) in 6-well plates. They were incubated at 4°C for 15 min and then extracts were scrapped into an eppendorf tube. Lysates were snap frozen in liquid nitrogen to improve efficiency of lysis. After thawing

the lysates on ice and clearing out cell debris by centrifugation, same amount of total protein (ranges between 10-30μg) was loaded onto each lane of a Polyacrylamide gel (Criterion XT Bis-Tris Gel 12%, Bio-Rad). Following electrophoresis, the resolved protein was transferred to a polyvinylidene difluoride (PVDF) (Bio-Rad) membrane and subjected to western blot (WB) analysis. The source and dilution for each antibody used were: polyclonal rabbit-APOBEC2 (gift from Alin Vonica MD, PhD) (1:1000), monoclonal mouse-APOBEC2 (clone 15E11, homemade) 1:5000, TroponinT, clone JLT-12 (T6277, Sigma-Aldrich) 1:500, alpha tubulin DM1A (Abcam, ab7291) 1:5000, MyHC, MF-20 (DSHB) 1:20.

7.5 Immunofluorescence staining and fusion index of C2C12s

C2C12 cells ($0.05^{\wedge}10^{6}$) were seeded in collagen coated coverslips (BD Biosciences, 356450) in 12- well plate the day before inducting of differentiation with 2% horse serum. They were washed with cold PBS and fixed with paraformaldehyde (4%) in PBS for 10 min at 4°C. This was followed by 2 washes, 5 min each at room temperature and blocking solution (0.5% BSA, 1% gelatin, 5% normal goat serum, 0.1% Triton) in PBS for 1 hour at room temperature. This was followed by overnight stain with antibodies in a humidified chamber at 4°C, three washes with cold PBS 5 min each at room temperature. Coverslips were then incubated with secondary antibodies for 1 hour at room temperature, followed by three washes with PBS 5min in at room temperature. Immunofluorescence staining of C2C12 cells was carried out with antibodies specific for MyHC (MyHC MF20, DSHB), DAPI (Vectoshield Antifade Mounting Medium with DAPI, H-1200), Flag (Sigma-Aldrich, F1804). Images were taken using confocal Leica

TCS SP5 II or widefield Zeiss Cell Observer and image analysis was done with Fiji (ImageJ).

7.6 APOBEC2$^{-/-}$ mice

C57BL/6 wild-type and APOBEC2$^{-/-}$ (provided by Lawrence Chan and Benny Hung-Junn Chan, Baylor College of Medicine) were used for the regeneration experiments. All mice were bred and maintained under specific pathogen-free conditions at the Rockefeller University Animal Care Facility and The Rockefeller University Institutional Animal Care and Use Committee (IACUC) approved all procedures involving mice.

7.7 $BaCl_2$ induced muscle injury and regeneration

To induce muscle injury we performed intramuscular injection of $BaCl_2$ in the *Tibialis Anterior* (TA) muscle as previously described. $BaCl_2$ induces muscle fiber necrosis while completely preserving the basement membrane (Caldwell et al., 1990) followed by regeneration. We used 8-10 week old female mice in all of the experiments. According to approved IACUC protocols mice were anaesthetized by isoflurane inhalation. $BaCl_2$ (50 uL of 1.5% w/v in saline) or 0.9% sodium chloride solution (as a control) was injected into the TA muscle of hind limb. For pain relief mice were given acetaminophen in the drinking solution (4.48mg.ml, acetaminophen) as described (Mickley et al., 2006). Injected animals were caged singly throughout the experiments. On day 14 after injury, the mice were euthanized and the TA muscles were collected, followed by cryosection and staining.

7.8 Muscle tissue cryosection, staining and image analysis

TA muscles were frozen in liquid nitrogen cooled in 2-methylabutane and embedded in optimal cutting temperature compound (OCT). 10 um-thick cross sections were made using Microm HM505E cryostat at -20C, air dried and stained with Wheat Germ Agglutinin, Alexa Fluor 488 Conjugate (1ug/mL) and DAPI (1ug/ml). Images were acquired using an inverted fluorescent microscope (Eclipse Ti, Nikon) coupled to a Neo sCMOS camera (Andor Technology), and multiple images of whole TA muscle cross section were stitched with the software supplied by the manufacturer (NIS-Elements AR, Nikon). Images were acquired using 10x objective with a scale of 0.65um/pixel. For each TA muscle cross-section 3 non-overlapping sections (close to the muscle center) were chosen. For quantification of the number of fibers with mispositioned nuclei ImageJ macros were used.

7.9 Inverted grip-hanging test

Experiment was performed by Luendreo Barboza. Animals were places in grip apparatus lifted 30cm from the ground. The time of hanging was measured for 5 successive trials. Experiment was done with one mouse cohort.

7.10 Salt-extraction profiling

C2C12 cells were seeded in equal numbers (2x10^6) and induced to differentiate after 12 hours. 5 days after differentiation they were lysed in the plate with 100ul sucrose lysis buffer (320mM sucrose, 0.5% NP-40, 10mM Tris pH8, 3mM $CaCl_2$, 2mM Mg acetate, 0.1mM EDTA). Extracts were incubated for 5 minutes on ice and spun at 500g for 5 minutes to collect the nuclear pellet and supernatant as the cytosol fraction. Nuclear pellets were washed with no-salt Nuclei Buffer (50mM Tris pH8, 1%NP-40, 10%

glycerol). Following the washes the nuclear proteins were extracted at increasing concentrations of NaCl from 250 mM up to 2 M in Nuclei Buffer during which they are homogenized using dounce tissue grinders (Fisher, K8853000000), incubated on ice for 10min and spun at 4°C for 10 additional minutes. Eluted materials was collected, resolved on polyacrylamide gel electrophoresis (Criterion XT Bis-Tris Gel 12%, Bio-Rad) and immunoblotted with specific antibodies: H4 antibody (cat, company) 1:5000, monoclonal mouse-APOBEC2 (clone 15E11, homemade) 1:5000, alpha tubulin DM1A (Abcam, ab7291) (1:5000).

7.11 RNA analysis

7.11.1 RNA expression analysis

Library preparation and sequencing were done by Rockefeller University Genomics Resource Center [https://www.rockefeller.edu/genomics/] using TruSeq Stranded mRNA Sample Prep kit as per manufacturer's instruction. The procedure includes purification of poly-adenylated RNAs. Libraries were sequenced with 50bp paired-read sequencing on the HiSeq2500 (Illumina).

Paired end read alignments and gene expression analysis were performed with the Bioinformatics Resource Center at Rockefeller University. Paired-end reads were aligned to mm10 genome using the subjunc function in the Bioconductor Rsubread (Liao et al., 2013) package and bigWig files for visualization were generated from aligned reads using the Bioconductor rtracklayer (Lawrence et al., 2009) and GenomicAlignments packages (Lawrence et al., 2013). For analysis of differential expression, transcript quantifications were performed using Salmon (Patro et al., 2017) in quasi-mapping mode. Gene expression values were calculated from transcript quantifications using tximport

(Soneson et al., 2016). Gene expression changes were identified at a cut off of 5% FDR (benjamini-hockberg correction) using the Wald test implemented in DESeq2 (Love et al., 2014). Annotation files used:

BSgenome.Mmusculus.UCSC.mm10(v1.4.0);org.Mm.db(v3.5.0);

TxDb.Mmusculus.UCSC.mm10.knownGene.gtf.gz(v3.4.0)

7.11.2 Gene Set Enrichment Analysis

GSEA (v3.0) (Subramanian et al., 2005) was ran on a list of genes that are pre-ranked based on Wald statistic, which is the log2 fold change value divided by the standard error according to GSEA recommendations. Molecular Signature Database (MSigDB) Gene Sets were used (Liberzon et al., 2015), and converted from human to mouse using the entrezID, using Mouse Genome Informatic (MGI) Vertebrate homology (HOM_MouseHumanSequence) (Blake et al., 2017) and HGNC Comparison of Orthology Predictions (HCOP) orthology (Eyre et al., 2007; Gray et al., 2015; Wright et al., 2005). The following MSigDB gene sets were used: c2.cp.reactome.v5.0.entrez.gmt, Reactome gene sets, Entrez IDs (Croft et al., 2014; Fabregat et al., 2018) and H.v5.0.entrez.gmt, Hallmark gene sets Entrez IDs, (Liberzon et al., 2015).

GSEA's parameters:

java -cp <path>/gsea-3.0.jar -Xmx512m xtools.gsea.GseaPreranked -gmx <GenesetFile.gmt> -norm meandiv -nperm 1000 -rnk <Preranked_gene.rnk> -scoring_scheme classic -rpt_label <Label> -create_svgs false -make_sets true -plot_top_x 20 -rnd_seed timestamp -set_max 500 -set_min 15 -zip_report false -out <outputFolder> -gui false

For each of the MSigDB gene sets listed above we generated our own collection by overlapping the MSigDB gene sets with the genes that are coordinately differentially expressed during the C2C12 differentiation at the specific time points chosen. We ended up with the following collection of gene sets of genes that are significantly unregulated, downregulated or not significantly changed during differentiation from day 0 to day 2: OriginalGeneSet_d2d0DEG_UP;OriginalGeneSet_d2d0DEG_DOWN; OriginalGeneSet_d2d0DEG_NS.

The False Discovery Rate (FDR) is calculated by comparing the actual data with 1000 Monte-Carlo simulations. The NES (Normalized Enrichment Score) computes the density of modified genes in the dataset with the random expectancies, normalized by the number of genes found in a given gene cluster, to take into account the size of the cluster.

7.11.3 RNA editing analysis

RNA editing analysis was performed by Dewi Harjanto, PhD as previously reported elsewhere (Harjanto et al., 2016). Editing detection was performed by comparing C2C12 control samples (GFPsh) to APOBEC2 knockdown samples using RNA-seq datasets in triplicates for each sample. Minimum filters include quality control thresholds (minimum of five reads covering the putative site with at least two reads supporting the editing event; filtering of reads that contain indels or support an edit in the first or last two base pairs of a read). Stringent filters applied to the APOBEC1 dependent C-to-U edited sites include all of the above and additionally the magnitude of the control vector was at least 15 and the angle between the wild-type and knockout vectors was at least 0.11 radians, as described in the paper referenced in this section.

7.12 DNA methylation analysis

7.12.1 Enhanced Reduced representation bisulfite sequencing (eRRBS)

eRRBS library preparation, sequencing and read alignment was performed by the Epigenomics Core Facility of Weill Cornell Medicine [epicore.med.cornell.edu/] as previously described (Akalin et al., 2012a; Garrett-Bakelman et al., 2015). The procedure includes bisulfite conversion of the DNA. Libraries were sequenced with 50bp single reads (SR) in HiSeq2500 (Illumina). Reads were aligned to a bisulfite converted reference mouse genome with Bismark (Krueger and Andrews, 2011). The methylation context for each cytosine was determined with scripts from the core facility.

Here coverage of specific genomic regions by eRRBS dataset, refers to the percent of features (eg percent of promoters, CpG islands) that contain at least one CpG that is well covered (> 10x). For gene specific annotations the mm10 UCSC knownGene annotations from the UCSC table browser were used and for CpG islands the mm10 cpgIslandExt track of the UCSC table browser. Genomic features were defined as: CpG islands, CpG island shores were defined as 2kb upstream and downstream of a CpG island; Gene promoters (region 2kb upstream and 2kb downstream of the TSS), exons, introns and intergenic regions.

7.12.2 Differential methylation analysis

MethylKit (v1.3.8) (Akalin et al., 2012b) was used to identify differentially methylated cytosines (DMCs) with q-value less than 0.01 and methylation percentage difference of at least 25% after filtering eRRBS dataset by coverage, normalizing by median and including CpG sites that are covered >10x, in 3 out of 5 biological replicates (lo.count = 10, lo.perc = NULL, hi.count = 1000, hi.perc = 99.9),

(destrand=TRUE,min.per.group=3L). eDMRs (v0.6.4.1) (Li et al., 2013) was used to empirically determine differentially methylated regions, using the DMCs identified with methylKit. In order for a region to be defined as a DMR, default parameters (num.DMCs=1, num.CpGs=3, DMR.methdiff=20) of eDMR were used, so that each region has: (1) at least 1 DMC in the region, as determined using methylKit, (2) at least 3 CpGs included in the region and (3) absolute mean methylation difference greater than 20%. For a region to be defined as a significant DMR, default parameters were used (DMR.qvalue = 0.001, mean.meth.diff = 20, num.CpGs = 5, num.DMCs = 3) so that each significant DMRs has (1) 5 CpGs where at least 3 of them are significant DMCs as determined by methylKit (2) have a minimum 20% methylation change for the region.

7.13 Chromatin binding analysis

7.13.1 Chromatin immunoprecipitation method

C2C12s were plated at ~70% confluence 12 hours prior to inducing differentiation (seed ~2x10^6 cells) maintained in DMEM (ATCC, 30-2002) with 10%FBS. This was followed by media change to DMEM with 2% horse serum (Life Biotechnologies, 26050-088) to induce differentiation. The cells (~5x10^6 /10cm plate) were harvested at 24-hour or 34-hour after inducing differentiation. They were fixed on plate with 1% PFA in PBS for 10 minutes at RT. Glycine was added to a final concentration of 125mM. Cells were washed 2 times with 1x PBS with protease inhibitor cocktail (PIC, Roche, 11836170001). They were lysed on the plate with cold Farnham lysis buffer to ~10x10^6 cells /mL (5mM PIPES pH 8.0, 0.5% NP-40, 85mM KCl,1mM EDTA, PIC) and incubated rotating for 15min at 4°C . Lysates were scraped off the plates, pelleted and resuspended in LB2 (10 mM Tris pH 8.0, 200 mM NaCl, 1 mM

EDTA, 0.5 mM EGTA, PIC) and incubated rotating for 15 minutes at 4°C and then centrifuged. Pellets were resuspended to 5x10^7 cells/mL in LB3 (10 mM Tris pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% sodium-deoxycholate, 0.5% sodium lauroyl sarcosinate, PIC) until suspension was homogenized. Samples were then sonicated using Covaris ultrasonicator model S220 for 15 minutes with the following settings: 140W peak power, 5% duty, 200 cycles per burst. TritonX-100 to a final concentration of 1% was added to the samples. Samples were clarified by centrifugation at 20,000 g for 10 minutes at 4°C. The supernatant is the soluble chromatin extract. The soluble fragmented chromatin from ~2.5x10^7 was used for each IP. For each IP 100ul Dynabeads (Thermofisher anti-rabbit M280, 11203D) were mixed with 10ul polyclonal rabbit-APOBEC2 antibodies (gift from Alin Vonica MD, PhD) incubating overnight (~16 hours). A magnetic stand was used to separate beads from the lysate and beads were washed one time each with for 5min in: low salt wash (0.1%SDS, 1%Triton X-100, 2mM EDTA, 20mM Tris pH8, 150mM NaCl, PIC), high salt wash (0.1%SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris pH8, 500mM NaCl, PIC), lithium chloride wash (150mM LiCl, 1% NP-40, 1% NaDOC, 1mM EDTA, 10mM TrispH8, PIC), TE wash (10mM Tris-HCl pH8, 1mM EDTA, 50mM NaCl, PIC). Beads were resuspended in 52 ul of elution buffer (50mM Tris-HCl pH8, 10mM EDTA, 1%SDS) and incubated at 30min at 65°C while shaking to prevent beads from settling. The eluate was transferred to a new tube, inputs of the same volume were incubated for 8 hours at 65°C to reverse the crosslink. The samples were treated with RNAse (Roche, Cat. No. 11 119 915 001) for 1 hour at 37°C, and with Proteinase K for 2 hours at 55°C. Fragmented DNA was purified

with Ampure beads (Agencourt AMPure XP beads A63881) as per the manufacter's instructions.

7.13.2 Chromatin immunoprecipitation sequencing and analysis

The ChiP-Seq included biological triplicates for each group. ChIP-Seq libraries were prepared using NEBNext Ultra DNA Library Prep Kit as per manufacturer's instructions. Libraries were sequenced with 75 base pair single read sequencing on the NextSeq 500 (Illumina). Read alignments and initial analysis were performed with the Bioinformatics Resource Center at Rockefeller University. Single-end reads were aligned to mm10 genome using the subread function in the Bioconductor Rsubread (Liao et al., 2013) package and bigWig files for visualization were generated from aligned reads using the Bioconductor rtracklayer (Lawrence et al., 2009) and GenomicAlignments packages (Lawrence et al., 2013). Quality metrics for the ChIP-Seq data were assessed using ChIPQC bioconductor package (Carroll et al., 2014), according to Encyclopedia of DNA Elements (ENCODE) working standards and guidelines for ChIP experiments (Landt et al., 2012). Reads mapping to more than one genomic location were filtered prior to peak calling using Model-based Analysis of ChIP-Seq (MACS2) (Feng et al., 2011; Zhang et al., 2008) with duplicate filtering applied and input DNA sample as a control. Peaks that are reproducible in the majority of the replicates (present in 2 out of 3) were filtered for known artifact or blacklisted regions (The ENCODE Project Consortium, 2012). For each of the peaks a weighted mean location of peak summits across biological replicates is calculated (Yang et al., 2014). The list of binding regions 100 base pairs around the mean peak summits was used for downstream analysis. Ngs.plot (v2.61) was used with specific parameters (-G mm10 -D refseq -C -L 1000 -FL

150 -P 4 -SC 0,1 -GO none -RB 0.05) to generate average profiles of ChIP-Seq reads (Shen et al., 2014). ChIPSeeker (v1.14.2) (Yu et al., 2015) and ChIPpeakAnno (3.12.7) (Zhu, 2013; Zhu et al., 2010) were used for downstream analysis after peak calling for annotation of the binding regions to the nearest gene. clusterProfiler (3.6.0) (Yu et al., 2012) was used for the pathway enrichment analysis of the genes with peaks using MSigDB database. The MSigDB gene sets are first converted from human to mouse entrezID, using Mouse Genome Informatics (MGI) Vertebrate homology (HOM_MouseHumanSequence) (Blake et al., 2017) and HGNC Comparison of Orthology Predictions (HCOP) orthology (Eyre et al., 2007; Gray et al., 2015; Wright et al., 2005). The following MSigDB gene sets were used: c2.cp.reactome.v5.0.entrez.gmt, Reactome gene sets Entrez IDs (Croft et al., 2014; Fabregat et al., 2018) and H.v5.0.entrez.gmt, Hallmark gene sets Entrez IDs (Liberzon et al., 2015). GSEA (Subramanian et al., 2005) (v3.0) was used for testing the enrichment of the *APOBEC2 occupied geneset* in the list of genes that are differentially expressed. Annotation files used: BSgenome.Mmusculus.UCSC.mm10 (v1.4.0) org.Mm.db (v3.5.0), TxDb.Mmusculus.UCSC.mm10.knownGene.gtf.gz(v3.4.0);

7.13.3 Prediction of binding motifs

The MEME-ChIP pipeline within the MEME suite of motif tools (v4.12.0) was used for motif based analysis (Ma et al., 2014; Machanick and Bailey, 2011). Peaks that are reproducible in the majority of the replicates (present in 2 out of 3) (Yang et al., 2014) were used for motif discovery after being filtered for blacklisted regions (The ENCODE Project Consortium, 2012). For each of the peaks a weighted mean location of peak summits across biological replicates is calculated (Yang et al., 2014). The list of

binding regions 100 base pairs around the mean peak summits was used for motif analysis. The MEME-ChIP pipeline was run with parameter *--parse-genomic-coord* to allow for the retrieval of discovered motif genomic locations. Within the MEME-ChIP pipeline, a background file is initially created by shuffling input Fasta sequences. MEME and DREME were used to predict *de novo* DNA binding motifs. De novo motif assignment to known motifs within the Jaspar 2014 database was performed using TomTom (Gupta et al., 2007). FIMO (Grant et al., 2011) was used for scanning for the occurrences of a given motif. The distribution of the motif location relative to the nearest TSS was done using ChipSeeker (v1.14.2)(Yu et al., 2015).

REFERENCES

Akalin, A., Garrett-Bakelman, F.E., Kormaksson, M., Busuttil, J., Zhang, L., Khrebtukova, I., Milne, T.A., Huang, Y., Biswas, D., Hess, J.L., et al. (2012a). Base-Pair Resolution DNA Methylation Sequencing Reveals Profoundly Divergent Epigenetic Landscapes in Acute Myeloid Leukemia. PLOS Genetics *8*, e1002781.

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A., and Mason, C.E. (2012b). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biology *13*, R87.

Al-Khalaf, M.H., Blake, L.E., Larsen, B.D., Bell, R.A., Brunette, S., Parks, R.J., Rudnicki, M.A., McKinnon, P.J., Dilworth, F.J., and Megeney, L.A. (2016). Temporal activation of XRCC1-mediated DNA repair is essential for muscle differentiation. Cell Discovery *2*, 15041.

Anant, S., Mukhopadhyay, D., Sankaranand, V., Kennedy, S., Henderson, J.O., and Davidson, N.O. (2001). ARCD-1, an apobec-1-related cytidine deaminase, exerts a dominant negative effect on C to U RNA editing. Am J Physiol Cell Physiol *281*, C1904–C1916.

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. Genome Res. *22*, 2008–2017.

Andrés, V., and Walsh, K. (1996). Myogenin expression, cell cycle withdrawal, and phenotypic differentiation are temporally separable events that precede cell fusion upon myogenesis. J Cell Biol *132*, 657–666.

Atlasi, Y., and Stunnenberg, H.G. (2017). The interplay of epigenetic marks during stem cell differentiation and development. Nature Reviews Genetics *18*, 643–658.

Avery, O.T., MacLeod, C.M., and McCarty, M. (1944). Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types. Journal of Experimental Medicine *79*, 137–158.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME Suite: tools for motif discovery and searching. Nucleic Acids Res *37*, W202–W208.

Behrens, A., Deursen, J.M. van, Rudolph, K.L., and Schumacher, B. (2014). Impact of genomic damage and ageing on stem cell function. Nature Cell Biology *16*, 201–207.

Bentzinger, C.F., Wang, Y.X., and Rudnicki, M.A. (2012). Building Muscle: Molecular Regulation of Myogenesis. Cold Spring Harb Perspect Biol *4*, a008342.

Berkes, C.A., and Tapscott, S.J. (2005). MyoD and the transcriptional control of myogenesis. Seminars in Cell & Developmental Biology *16*, 585–595.

Betts, L., Xiang, S., Short, S.A., Wolfenden, R., and Carter, C.W. (1994). Cytidine Deaminase. The 2·3 Å Crystal Structure of an Enzyme: Transition-state Analog Complex. Journal of Molecular Biology *235*, 635–656.

Bhutani, N., Brady, J.J., Damian, M., Sacco, A., Corbel, S.Y., and Blau, H.M. (2010). Reprogramming towards pluripotency requires AID-dependent DNA demethylation. Nature *463*, 1042–1047.

Bierkens, M., Krijgsman, O., Wilting, S.M., Bosch, L., Jaspers, A., Meijer, G.A., Meijer, C.J.L.M., Snijders, P.J.F., Ylstra, B., and Steenbergen, R.D.M. (2013). Focal aberrations indicate EYA2 and hsa-miR-375 as oncogene and tumor suppressor in cervical carcinogenesis. Genes Chromosomes Cancer *52*, 56–68.

Blais, A., Tsikitis, M., Acosta-Alvear, D., Sharan, R., Kluger, Y., and Dynlacht, B.D. (2005). An initial blueprint for myogenic differentiation. Genes Dev. *19*, 553–569.

Blais, A., van Oevelen, C.J.C., Margueron, R., Acosta-Alvear, D., and Dynlacht, B.D. (2007). Retinoblastoma tumor suppressor protein-dependent methylation of histone H3 lysine 27 is associated with irreversible cell cycle exit. J. Cell Biol. *179*, 1399–1412.

Blake, J.A., Eppig, J.T., Kadin, J.A., Richardson, J.E., Smith, C.L., and Bult, C.J. (2017). Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. Nucleic Acids Res *45*, D723–D729.

Blanc, V., Henderson, J.O., Newberry, R.D., Xie, Y., Cho, S.-J., Newberry, E.P., Kennedy, S., Rubin, D.C., Wang, H.L., Luo, J., et al. (2007). Deletion of the AU-Rich RNA Binding Protein Apobec-1 Reduces Intestinal Tumor Burden in Apcmin Mice. Cancer Res *67*, 8565–8573.

Blanc, V., Park, E., Schaefer, S., Miller, M., Lin, Y., Kennedy, S., Billing, A.M., Hamidane, H.B., Graumann, J., Mortazavi, A., et al. (2014). Genome-wide identification and functional analysis of Apobec-1-mediated C-to-U RNA editing in mouse small intestine and liver. Genome Biology *15*, R79.

Blau, H.M., Pavlath, G.K., Hardeman, E.C., Chiu, C.P., Silberstein, L., Webster, S.G., Miller, S.C., and Webster, C. (1985). Plasticity of the differentiated state. Science *230*, 758–766.

Bochtler, M., Kolano, A., and Xu, G.-L. (2017). DNA demethylation pathways: Additional players and regulators. BioEssays *39*, 1–13.

Bone, C.R., and Starr, D.A. (2016). Nuclear migration events throughout development. J Cell Sci *129*, 1951–1961.

Bracken, A.P., Ciro, M., Cocito, A., and Helin, K. (2004). E2F target genes: unraveling the biology. Trends in Biochemical Sciences *29*, 409–417.

Branco, M.R., Ficz, G., and Reik, W. (2012). Uncovering the role of 5-hydroxymethylcytosine in the epigenome. Nature Reviews Genetics *13*, 7–13.

Braun, T., and Gautel, M. (2011). Transcriptional mechanisms regulating skeletal muscle differentiation, growth and homeostasis. Nat Rev Mol Cell Biol *12*, 349–361.

Breslow, J.L. (1988). Apolipoprotein genetic variation and human disease. Physiol. Rev. *68*, 85–132.

Caldwell, C. j., Mattey, D.L., and Weller, R.O. (1990). Role of the basement membrane in the regeneration of skeletal muscle. Neuropathology and Applied Neurobiology *16*, 225–238.

Cao, Y., Kumar, R.M., Penn, B.H., Berkes, C.A., Kooperberg, C., Boyer, L.A., Young, R.A., and Tapscott, S.J. (2006). Global and gene-specific analyses show distinct roles for Myod and Myog at a common set of promoters. EMBO J. *25*, 502–511.

Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G.J., Parker, M.H., MacQuarrie, K.L., Davison, J., Morgan, M.T., Ruzzo, W.L., et al. (2010). Genome-wide MyoD Binding in Skeletal Muscle Cells: A Potential for Broad Cellular Reprogramming. Developmental Cell *18*, 662–674.

Carrió, Díez-Villanueva Anna, Lois Sergi, Mallona Izaskun, Cases Ildefonso, Forn Marta, Peinado Miguel A., and Suelves Mònica (2015). Deconstruction of DNA Methylation Patterns During Myogenesis Reveals Specific Epigenetic Events in the Establishment of the Skeletal Muscle Lineage. STEM CELLS *33*, 2025–2036.

Carrió, E., Magli, A., Muñoz, M., Peinado, M.A., Perlingeiro, R., and Suelves, M. (2016). Muscle cell identity requires Pax7-mediated lineage-specific DNA demethylation. BMC Biology *14*, 30.

Carroll, T.S., Liang, Z., Salama, R., Stark, R., and de Santiago, I. (2014). Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. Front. Genet. *5*.

Chargaff, E., Lipshitz, R., Green, C., and Hodes, M.E. (1951). The Composition of the Desoxyribonucleic Acid of Salmon Sperm. J. Biol. Chem. *192*, 223–230.

Chen, S.H., Habib, G., Yang, C.Y., Gu, Z.W., Lee, B.R., Weng, S.A., Silberman, S.R., Cai, S.J., Deslypere, J.P., and Rosseneu, M. (1987). Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. Science *238*, 363–366.

Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K.Y., Rozowsky, J., Yan, K.-K., Dong, X., Djebali, S., Ruan, Y., et al. (2012). Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Res. *22*, 1658–1667.

Coker, H.A., Morgan, H.D., and Petersen-Mahrt, S.K. (2006). Genetic and In Vitro Assays of DNA Deamination. In Methods in Enzymology, (Academic Press), pp. 156–170.

Cole, D.C., Chung, Y., Gagnidze, K., Hajdarovic, K.H., Rayon-Estrada, V., Harjanto, D., Bigio, B., Gal-Toth, J., Milner, T.A., McEwen, B.S., et al. (2017). Loss of APOBEC1 RNA-editing function in microglia exacerbates age-related CNS pathophysiology. PNAS *114*, 13272–13277.

Conerly, M.L., Yao, Z., Zhong, J.W., Groudine, M., and Tapscott, S.J. (2016). Distinct Activities of Myf5 and MyoD Indicate Separate Roles in Skeletal Muscle Lineage Specification and Differentiation. Developmental Cell *36*, 375–385.

Conticello, S.G. (2008). The AID/APOBEC family of nucleic acid mutators. Genome Biol *9*, 229.

Conticello, S.G. (2012). Creative deaminases, self-inflicted damage, and genome evolution. Annals of the New York Academy of Sciences *1267*, 79–85.

Conticello, S.G., Thomas, C.J.F., Petersen-Mahrt, S.K., and Neuberger, M.S. (2005). Evolution of the AID/APOBEC Family of Polynucleotide (Deoxy)cytidine Deaminases. Mol Biol Evol *22*, 367–377.

Conticello, S.G., Langlois, M.-A., and Neuberger, M.S. (2007a). Insights into DNA deaminases. Nat Struct Mol Biol *14*, 7–9.

Conticello, S.G., Langlois, M., Yang, Z., and Neuberger, M.S. (2007b). DNA Deamination in Immunity: AID in the Context of Its APOBEC Relatives. In Advances in Immunology, Frederick W. Alt and Tasuku Honjo, ed. (Academic Press), pp. 37–73.

Crick, F. (1958). On Protein Synthesis. The Symposia of the Society for Experimental Biology 12 138–163.

Crick, F. (1970). Central Dogma of Molecular Biology. Nature *227*, 561–563.

Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., et al. (2014). The Reactome pathway knowledgebase. Nucleic Acids Res. *42*, D472-477.

Datler, C., and Grimm, S. (2013). Reconstitution of CKMT1 expression fails to rescue cells from mitochondrial membrane potential dissipation: Implications for controlling RNAi experiments. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research *1833*, 2844–2855.

Davidson, N.O., Innerarity, T.L., Scott, J., Smith, H., Driscoll, D.M., Teng, B., and Chan, L. (1995). Proposed nomenclature for the catalytic subunit of the mammalian apolipoprotein B mRNA editing enzyme: APOBEC-1. RNA *1*, 3–3.

Delker, R.K., Fugmann, S.D., and Papavasiliou, F.N. (2009). A coming-of-age story: activation-induced cytidine deaminase turns 10. Nat Immunol *10*, 1147–1153.

Di Noia, J.M., and Neuberger, M.S. (2007). Molecular Mechanisms of Antibody Somatic Hypermutation. Annual Review of Biochemistry *76*, 1–22.

Dominguez, P.M., and Shaknovich, R. (2014). Epigenetic function of activation-induced cytidine deaminase and its link to lymphomagenesis. Front. Immunol. *5*, 642.

Dominguez, P.M., Teater, M., Chambwe, N., Kormaksson, M., Redmond, D., Ishii, J., Vuong, B., Chaudhuri, J., Melnick, A., Vasanthakumar, A., et al. (2015). DNA Methylation Dynamics of Germinal Center B Cells Are Mediated by AID. Cell Reports *12*, 2086–2098.

Dumont, N.A., Wang, Y.X., and Rudnicki, M.A. (2015). Intrinsic and extrinsic mechanisms regulating satellite cell function. Development *142*, 1572–1581.

ENCODE, and modENCODE (2011). ENCODE and modENCODE Guidelines For Experiments Generating ChIP, DNase, FAIRE, and DNA Methylation Genome Wide Location Data.

Etard, C., Roostalu, U., and Strähle, U. (2010). Lack of Apobec2-related proteins causes a dystrophic muscle phenotype in zebrafish embryos. J Cell Biol *189*, 527–539.

Eyre, T.A., Wright, M.W., Lush, M.J., and Bruford, E.A. (2007). HCOP: a searchable database of human orthology predictions. Brief Bioinform *8*, 2–5.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018). The Reactome Pathway Knowledgebase. Nucleic Acids Res. *46*, D649–D655.

Falco, G.D., Comes, F., and Simone, C. (2006). pRb: master of differentiation. Coupling irreversible cell cycle withdrawal with induction of muscle-specific transcription. Oncogene *25*, 5244.

Feng, J., Liu, T., and Zhang, Y. (2011). Using MACS to Identify Peaks from ChIP-Seq Data. Curr Protoc Bioinformatics *CHAPTER*, Unit2.14.

Figliola, R., and Maione, R. (2004). MyoD induces the expression of p57Kip2 in cells lacking p21Cip1/Waf1: overlapping and distinct functions of the two cdk inhibitors. J. Cell. Physiol. *200*, 468–475.

Folker, E.S., and Baylies, M.K. (2013). Nuclear positioning in muscle development and disease. Front Physiol *4*.

Fossat, N., Tourle, K., Radziewic, T., Barratt, K., Liebhold, D., Studdert, J.B., Power, M., Jones, V., Loebel, D.A.F., and Tam, P.P.L. (2014). C to U RNA editing mediated by APOBEC1 requires RNA-binding protein RBM47. EMBO Reports *15*, 903–910.

Franklin, R.E., and Gosling, R.G. (1953). Molecular Configuration in Sodium Thymonucleate. Nature *171*, 740–741.

Fritz, E. (2014). Genome-Wide Characterization of the Effects of Nucleic Acid Modifying Enzymes: Cytidine Deaminases and DNA Methylation. Student Theses and Dissertations.

Fritz, E.L., Rosenberg, B.R., Lay, K., Mihailović, A., Tuschl, T., and Papavasiliou, F.N. (2013). A comprehensive analysis of the effects of the deaminase AID on the transcriptome and methylome of activated B cells. Nat. Immunol. *14*, 749–755.

Garrett-Bakelman, F.E., Sheridan, C.K., Kacmarczyk, T.J., Ishii, J., Betel, D., Alonso, A., Mason, C.E., Figueroa, M.E., and Melnick, A.M. (2015). Enhanced Reduced Representation Bisulfite Sequencing for Assessment of DNA Methylation at Base Pair Resolution. J Vis Exp.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018.

Gray, K.A., Yates, B., Seal, R.L., Wright, M.W., and Bruford, E.A. (2015). Genenames.org: the HGNC resources in 2015. Nucleic Acids Res *43*, D1079–D1085.

Gu, H., Smith, Z.D., Bock, C., Boyle, P., Gnirke, A., and Meissner, A. (2011). Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat. Protocols *6*, 468–481.

Guo, J.U., Su, Y., Zhong, C., Ming, G., and Song, H. (2011). Hydroxylation of 5-Methylcytosine by TET1 Promotes Active DNA Demethylation in the Adult Brain. Cell *145*, 423–434.

Guo, K., Wang, J., Andrés, V., Smith, R.C., and Walsh, K. (1995). MyoD-induced expression of p21 inhibits cyclin-dependent kinase activity upon myocyte terminal differentiation. Mol Cell Biol *15*, 3823–3829.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biology *8*, R24.

Haldar, M., Karan, G., Tvrdik, P., and Capecchi, M.R. (2008). Two Cell Lineages, myf5 and myf5-Independent, Participate in Mouse Skeletal Myogenesis. Developmental Cell *14*, 437–445.

Hardy, D., Besnard, A., Latil, M., Jouvion, G., Briand, D., Thépenier, C., Pascal, Q., Guguin, A., Gayraud-Morel, B., Cavaillon, J.-M., et al. (2016). Comparative Study of Injury Models for Studying Muscle Regeneration in Mice. PLOS ONE *11*, e0147198.

Harjanto, D., Papamarkou, T., Oates, C.J., Rayon-Estrada, V., Papavasiliou, F.N., and Papavasiliou, A. (2016). RNA editing generates cellular subsets with diverse sequence within populations. Nature Communications *7*, 12145.

Harris, R.S., and Dudley, J.P. (2015). APOBECs and virus restriction. Virology *479–480*, 131–145.

Harris, R.S., and Liddament, M.T. (2004). Retroviral restriction by APOBEC proteins. Nat Rev Immunol *4*, 868–877.

Harris, R.S., Petersen-Mahrt, S.K., and Neuberger, M.S. (2002). RNA Editing Enzyme APOBEC1 and Some of Its Homologs Can Act as DNA Mutators. Molecular Cell *10*, 1247–1253.

Harris, R.S., Bishop, K.N., Sheehy, A.M., Craig, H.M., Petersen-Mahrt, S.K., Watt, I.N., Neuberger, M.S., and Malim, M.H. (2003). DNA deamination mediates innate immunity to retroviral infection. Cell *113*, 803–809.

Hasty, P., Bradley, A., Morris, J.H., Edmondson, D.G., Venuti, J.M., Olson, E.N., and Klein, W.H. (1993). Muscle deficiency and neonatal death in mice with a targeted mutation in the myogenin gene. Nature *364*, 501–506.

Henderson, S., and Fenton, T. (2015). APOBEC3 genes: retroviral restriction factors to cancer drivers. Trends Mol Med *21*, 274–284.

Hernández-Hernández, J.M., García-González, E.G., Brun, C.E., and Rudnicki, M.A. (2017). The myogenic regulatory factors, determinants of muscle development, cell identity and regeneration. Seminars in Cell & Developmental Biology *72*, 10–18.

Hersberger, M., Patarroyo-White, S., Qian, X., Arnold, K.S., Rohrer, L., Balestra, M.E., and Innerarity, T.L. (2003). Regulatable liver expression of the rabbit apolipoprotein B mRNA-editing enzyme catalytic polypeptide 1 (APOBEC-1) in mice lacking endogenous APOBEC-1 leads to aberrant hyperediting. Biochemical Journal *369*, 255–262.

Hershey, A.D., and Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. J Gen Physiol *36*, 39–56.

Ikeda, T., Galil, K.H.A.E., Tokunaga, K., Maeda, K., Sata, T., Sakaguchi, N., Heidmann, T., and Koito, A. (2011). Intrinsic restriction activity by apolipoprotein B mRNA editing enzyme APOBEC1 against the mobility of autonomous retrotransposons. Nucleic Acids Research *39*, 5538.

Iyer, L.M., Zhang, D., Rogozin, I.B., and Aravind, L. (2011). Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. Nucleic Acids Res *39*, 9473–9497.

Judson, H.F. (1996). The Eighth Day of Creation: Makers of the Revolution in Biology, Commemorative Edition (Plainview, N.Y: Cold Spring Harbor Laboratory Press).

Kafri, R., Springer, M., and Pilpel, Y. (2009). Genetic Redundancy: New Tricks for Old Genes. Cell *136*, 389–392.

Kitzmann, M., Carnac, G., Vandromme, M., Primig, M., Lamb, N.J., and Fernandez, A. (1998). The muscle regulatory factors MyoD and myf-5 undergo distinct cell cycle-specific expression in muscle cells. J. Cell Biol. *142*, 1447–1459.

Knisbacher, B.A., Gerber, D., and Levanon, E.Y. (2016). DNA Editing by APOBECs: A Genomic Preserver and Transformer. Trends in Genetics *32*, 16–28.

Kosugi, S., Hasebe, M., Tomita, M., and Yanagawa, H. (2009). Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite motifs. PNAS *106*, 10171–10176.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572.

Krzysiak, T.C., Jung, J., Thompson, J., Baker, D., and Gronenborn, A.M. (2012). APOBEC2 Is a Monomer in Solution: Implications for APOBEC3G Models. Biochemistry *51*, 2008–2017.

Lada, A.G., Krick, C.F., Kozmin, S.G., Mayorov, V.I., Karpova, T.S., Rogozin, I.B., and Pavlov, Y.I. (2011). Mutator effects and mutation signatures of editing deaminases produced in bacteria and yeast. Biochemistry Moscow *76*, 131–146.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. *22*, 1813–1831.

Larsen, B.D., and Megeney, L.A. (2010). Parole terms for a killer. Cell Cycle *9*, 2940–2945.

Larsen, B.D., Rampalli, S., Burns, L.E., Brunette, S., Dilworth, F.J., and Megeney, L.A. (2010). Caspase 3/caspase-activated DNase promote cell differentiation by inducing DNA strand breaks. Proc. Natl. Acad. Sci. U.S.A. *107*, 4230–4235.

Lawrence, M., Gentleman, R., and Carey, V. (2009). rtracklayer: an R package for interfacing with genome browsers. Bioinformatics *25*, 1841–1842.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for Computing and Annotating Genomic Ranges. PLOS Computational Biology *9*, e1003118.

Lecossier, D., Bouchonnet, F., Clavel, F., and Hance, A.J. (2003). Hypermutation of HIV-1 DNA in the absence of the Vif protein. Science *300*, 1112.

Li, S., Garrett-Bakelman, F.E., Akalin, A., Zumbo, P., Levine, R., To, B.L., Lewis, I.D., Brown, A.L., D'Andrea, R.J., Melnick, A., et al. (2013). An optimized algorithm for detecting and annotating regional differential methylation. BMC Bioinformatics *14*, S10.

Liang, G., Kitamura, K., Wang, Z., Liu, G., Chowdhury, S., Fu, W., Koura, M., Wakae, K., Honjo, T., and Muramatsu, M. (2013). RNA editing of hepatitis B virus transcripts by activation-induced cytidine deaminase. PNAS *110*, 2246–2251.

Liao, W., Hong, S.-H., Chan, B.H.-J., Rudolph, F.B., Clark, S.C., and Chan, L. (1999). APOBEC-2, a Cardiac- and Skeletal Muscle-Specific Member of the Cytidine Deaminase Supergene Family. Biochemical and Biophysical Research Communications *260*, 398–404.

Liao, Y., Smyth, G.K., and Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res *41*, e108–e108.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. Cell Systems *1*, 417–425.

Liu, Q.-C., Zha, X.-H., Faralli, H., Yin, H., Louis-Jeune, C., Perdiguero, E., Pranckeviciene, E., Muñoz-Cànoves, P., Rudnicki, M.A., Brand, M., et al. (2012). Comparative expression profiling identifies differential roles for Myogenin and p38α MAPK signaling in myogenesis. J Mol Cell Biol *4*, 386–397.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology *15*, 550.

Ma, W., Noble, W.S., and Bailey, T.L. (2014). Motif-based analysis of large nucleotide data sets using MEME-ChIP. Nature Protocols *9*, 1428–1450.

MacDuff, D.A., Demorest, Z.L., and Harris, R.S. (2009). AID can restrict L1 retrotransposition suggesting a dual role in innate and adaptive immunity. Nucleic Acids Res. *37*, 1854–1867.

Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics *27*, 1696–1697.

Maddox, B. (2003). Rosalind Franklin: The Dark Lady of DNA (New York: Harper Perennial).

Mangeat, B., Turelli, P., Caron, G., Friedli, M., Perrin, L., and Trono, D. (2003). Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. Nature *424*, 99–103.

Marino, D., Perković, M., Hain, A., Jaguva Vasudevan, A.A., Hofmann, H., Hanschmann, K.-M., Mühlebach, M.D., Schumann, G.G., König, R., Cichutek, K., et al. (2016). APOBEC4 Enhances the Replication of HIV-1. PLoS ONE *11*, e0155422.

Megeney, L.A., Kablar, B., Garrett, K., Anderson, J.E., and Rudnicki, M.A. (1996). MyoD is required for myogenic stem cell function in adult skeletal muscle. Genes Dev. *10*, 1173–1183.

Mehta, A., Kinter, M.T., Sherman, N.E., and Driscoll, D.M. (2000). Molecular Cloning of Apobec-1 Complementation Factor, a Novel RNA-Binding Protein Involved in the Editing of Apolipoprotein B mRNA. Mol. Cell. Biol. *20*, 1846–1854.

Meissner, A., Gnirke, A., Bell, G.W., Ramsahoye, B., Lander, E.S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res *33*, 5868–5877.

Mickley, A.G., Hoxha, Z., Biada, J.M., Kenmuir, C.L., and Bacik, S.E. (2006). Acetaminophen Self-administered in the Drinking Water Increases the Pain Threshold of Rats (Rattus norvegicus). Journal of the American Association for Laboratory Animal Science *45*, 48–54.

Miescher, F. (1871). Medicinisch-chemische Untersuchungen: 4. Med-Chem Unters 4: 441-460.

Mikl, M.C., Watt, I.N., Lu, M., Reik, W., Davies, S.L., Neuberger, M.S., and Rada, C. (2005). Mice Deficient in APOBEC2 and APOBEC3. Mol Cell Biol *25*, 7270–7277.

Moffat, J., Grueneberg, D.A., Yang, X., Kim, S.Y., Kloepfer, A.M., Hinkle, G., Piqani, B., Eisenhaure, T.M., Luo, B., Grenier, J.K., et al. (2006). A Lentiviral RNAi Library for Human and Mouse Genes Applied to an Arrayed Viral High-Content Screen. Cell *124*, 1283–1298.

Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., et al. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet *34*, 267–273.

Mousavi, K., Zare, H., Dell'Orso, S., Grontved, L., Gutierrez-Cruz, G., Derfoul, A., Hager, G.L., and Sartorelli, V. (2013). eRNAs Promote Transcription by Establishing Chromatin Accessibility at Defined Genomic Loci. Molecular Cell *51*, 606–617.

Mukhopadhyay, D., Anant, S., Lee, R.M., Kennedy, S., Viskochil, D., and Davidson, N.O. (2002). C-->U editing of neurofibromatosis 1 mRNA occurs in tumors that express both the type II transcript and apobec-1, the catalytic subunit of the apolipoprotein B mRNA-editing enzyme. Am. J. Hum. Genet. *70*, 38–50.

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

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., and Honjo, T. (2000). Class Switch Recombination and Hypermutation Require Activation-Induced Cytidine Deaminase (AID), a Potential RNA Editing Enzyme. Cell *102*, 553–563.

Nabel, C.S., Jia, H., Ye, Y., Shen, L., Goldschmidt, H.L., Stivers, J.T., Zhang, Y., and Kohli, R.M. (2012). AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. Nat Chem Biol *8*, 751–758.

Nabeshima, Y., Hanaoka, K., Hayasaka, M., Esumi, E., Li, S., Nonaka, I., and Nabeshima, Y. (1993). Myogenin gene disruption results in perinatal lethality because of severe muscle defect. Nature *364*, 532–535.

Narciso, L., Fortini, P., Pajalunga, D., Franchitto, A., Liu, P., Degan, P., Frechet, M., Demple, B., Crescenzi, M., and Dogliotti, E. (2007). Terminally differentiated muscle cells are defective in base excision DNA repair and hypersensitive to oxygen injury. Proc. Natl. Acad. Sci. U.S.A. *104*, 17010–17015.

Ohtsubo, H., Sato, Y., Suzuki, T., Mizunoya, W., Nakamura, M., Tatsumi, R., and Ikeuchi, Y. (2017a). APOBEC2 negatively regulates myoblast differentiation in muscle regeneration. The International Journal of Biochemistry & Cell Biology *85*, 91–101.

Ohtsubo, H., Sato, Y., Suzuki, T., Mizunoya, W., Nakamura, M., Tatsumi, R., and Ikeuchi, Y. (2017b). Data supporting possible implication of APOBEC2 in self-renewal functions of myogenic stem satellite cells: Toward understanding the negative regulation of myoblast differentiation. Data in Brief *12*, 269–273.

Okuyama, S., Marusawa, H., Matsumoto, T., Ueda, Y., Matsumoto, Y., Endo, Y., Takai, A., and Chiba, T. (2012). Excessive activity of apolipoprotein B mRNA editing enzyme catalytic polypeptide 2 (APOBEC2) contributes to liver and lung tumorigenesis. Int. J. Cancer *130*, 1294–1301.

Orimo, A., Gupta, P.B., Sgroi, D.C., Arenzana-Seisdedos, F., Delaunay, T., Naeem, R., Carey, V.J., Richardson, A.L., and Weinberg, R.A. (2005). Stromal Fibroblasts Present in Invasive Human Breast Carcinomas Promote Tumor Growth and Angiogenesis through Elevated SDF-1/CXCL12 Secretion. Cell *121*, 335–348.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods *14*, 417–419.

Piec, I., Listrat, A., Alliot, J., Chambon, C., Taylor, R.G., and Bechet, D. (2005). Differential proteome analysis of aging in rat skeletal muscle. FASEB J.

Plongthongkum, N., Diep, D.H., and Zhang, K. (2014). Advances in the profiling of DNA modifications: cytosine methylation and beyond. Nat Rev Genet *15*, 647–661.

Porter, E.G., Connelly, K.E., and Dykhuizen, E.C. (2017). Sequential Salt Extractions for the Analysis of Bulk Chromatin Binding Properties of Chromatin Modifying Complexes. JoVE (Journal of Visualized Experiments) e55369–e55369.

Powell, C., Elsaeidi, F., and Goldman, D. (2012). Injury-Dependent Müller Glia and Ganglion Cell Reprogramming during Tissue Regeneration Requires Apobec2a and Apobec2b. J. Neurosci. *32*, 1096–1109.

Powell, C., Grant, A.R., Cornblath, E., and Goldman, D. (2013). Analysis of DNA methylation reveals a partial reprogramming of the Müller glia genome during retina regeneration. PNAS *110*, 19814–19819.

Powell, C., Cornblath, E., and Goldman, D. (2014). Zinc-binding Domain-dependent, Deaminase-independent Actions of Apolipoprotein B mRNA-editing Enzyme, Catalytic Polypeptide 2 (Apobec2), Mediate Its Effect on Zebrafish Retina Regeneration. J. Biol. Chem. *289*, 28924–28941.

Powell, L.M., Wallis, S.C., Pease, R.J., Edwards, Y.H., Knott, T.J., and Scott, J. (1987). A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell *50*, 831–840.

Prochnow, C., Bransteitter, R., Klein, M.G., Goodman, M.F., and Chen, X.S. (2007). The APOBEC-2 crystal structure and functional implications for the deaminase AID. Nature *445*, 447–451.

Rai, K., Huggins, I.J., James, S.R., Karpf, A.R., Jones, D.A., and Cairns, B.R. (2008). DNA Demethylation in Zebrafish Involves the Coupling of a Deaminase, a Glycosylase, and Gadd45. Cell *135*, 1201–1212.

Ramiro, A.R., and Barreto, V.M. (2015). Activation-induced cytidine deaminase and active cytidine demethylation. Trends Biochem. Sci. *40*, 172–181.

Rawls, A., Valdez, M.R., Zhang, W., Richardson, J., Klein, W.H., and Olson, E.N. (1998). Overlapping functions of the myogenic bHLH genes MRF4 and MyoD revealed in double mutant mice. Development *125*, 2349–2358.

Rayon-Estrada, V., Harjanto, D., Hamilton, C.E., Berchiche, Y.A., Gantman, E.C., Sakmar, T.P., Bulloch, K., Gagnidze, K., Harroch, S., McEwen, B.S., et al. (2017). Epitranscriptomic profiling across cell types reveals associations between APOBEC1-mediated RNA editing, gene expression outcomes, and cellular function. PNAS *114*, 13296–13301.

Relaix, F., and Zammit, P.S. (2012). Satellite cells are essential for skeletal muscle regeneration: the cell on the edge returns centre stage. Development *139*, 2845–2856.

Robbiani, D.F., Bothmer, A., Callen, E., Reina-San-Martin, B., Dorsett, Y., Difilippantonio, S., Bolland, D.J., Chen, H.T., Corcoran, A.E., Nussenzweig, A., et al. (2008). AID is required for the chromosomal breaks in c-myc that lead to c-myc/IgH translocations. Cell *135*, 1028–1038.

Roberts, S.A., Lawrence, M.S., Klimczak, L.J., Grimm, S.A., Fargo, D., Stojanov, P., Kiezun, A., Kryukov, G.V., Carter, S.L., Saksena, G., et al. (2013). An APOBEC

cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet *45*, 970–976.

Rogozin, I.B., Basu, M.K., Jordan, I.K., Pavlov, Y.I., and Koonin, E.V. (2005). APOBEC4, a New Member of the AID/APOBEC Family of Polynucleotide (Deoxy)Cytidine Deaminases Predicted by Computational Analysis. Cell Cycle *4*, 1281–1285.

Roman, W., Martins, J.P., Carvalho, F.A., Voituriez, R., Abella, J.V.G., Santos, N.C., Cadot, B., Way, M., and Gomes, E.R. (2017). Myofibril contraction and crosslinking drive nuclear movement to the periphery of skeletal muscle. Nature Cell Biology *19*, 1189–1201.

Rosen, J.N., and Baylies, M.K. (2017). Myofibrils put the squeeze on nuclei. Nature Cell Biology *19*, 1148–1150.

Rosenberg, B.R., Hamilton, C.E., Mwangi, M.M., Dewell, S., and Papavasiliou, F.N. (2011). Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3′ UTRs. Nat Struct Mol Biol *18*, 230–236.

Rubio, M.A.T., Pastar, I., Gaston, K.W., Ragone, F.L., Janzen, C.J., Cross, G.A.M., Papavasiliou, F.N., and Alfonzo, J.D. (2007). An adenosine-to-inosine tRNA-editing enzyme that can perform C-to-U deamination of DNA. PNAS *104*, 7821–7826.

Rudnicki, M.A., Schnegelsberg, P.N.J., Stead, R.H., Braun, T., Arnold, H.-H., and Jaenisch, R. (1993). MyoD or Myf-5 is required for the formation of skeletal muscle. Cell *75*, 1351–1359.

Salter, J.D., Bennett, R.P., and Smith, H.C. (2016). The APOBEC Protein Family: United by Structure, Divergent in Function. Trends in Biochemical Sciences *41*, 578–594.

Sato, Y., Probst, H.C., Tatsumi, R., Ikeuchi, Y., Neuberger, M.S., and Rada, C. (2010). Deficiency in APOBEC2 Leads to a Shift in Muscle Fiber Type, Diminished Body Mass, and Myopathy. J. Biol. Chem. *285*, 7111–7118.

Sato, Y., Ohtsubo, H., Nihei, N., Kaneko, T., Sato, Y., Adachi, S., Kondo, S., Nakamura, M., Mizunoya, W., Iida, H., et al. (2017). Apobec2 deficiency causes mitochondrial defects and mitophagy in skeletal muscle. FASEB J fj.201700493R.

Severi, F., Chicca, A., and Conticello, S.G. (2011). Analysis of Reptilian APOBEC1 Suggests that RNA Editing May Not Be Its Ancestral Function. Mol Biol Evol *28*, 1125–1129.

Sharma, A., Klein, S.L., Barboza, L., Lodhi, N., and Toth, M. (2016). Principles Governing DNA Methylation during Neuronal Lineage and Subtype Specification. J. Neurosci. *36*, 1711–1722.

Sheehy, A.M., Gaddis, N.C., Choi, J.D., and Malim, M.H. (2002). Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature *418*, 646–650.

Shen, L., Shao, N., Liu, X., and Nestler, E. (2014). ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genomics *15*, 284.

Singh Kulwant, and Dilworth F. Jeffrey (2013). Differential modulation of cell cycle progression distinguishes members of the myogenic regulatory factor family of transcription factors. The FEBS Journal *280*, 3991–4003.

Smith, Z.D., and Meissner, A. (2013). DNA methylation: roles in mammalian development. Nature Reviews Genetics *14*, 204.

Soleimani, V.D., Punch, V.G., Kawabe, Y., Jones, A.E., Palidwor, G.A., Porter, C.J., Cross, J.W., Carvajal, J.J., Kockx, C.E.M., van IJcken, W.F.J., et al. (2012). Transcriptional dominance of Pax7 in adult myogenesis is due to high-affinity recognition of homeodomain motifs. Dev. Cell *22*, 1208–1220.

Soneson, C., Love, M.I., and Robinson, M.D. (2016). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research *4*, 1521.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS *102*, 15545–15550.

Swanton, C., McGranahan, N., Starrett, G.J., and Harris, R.S. (2015). APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity. Cancer Discov *5*, 704–712.

Sylvestre, Y., De Guire, V., Querido, E., Mukhopadhyay, U.K., Bourdeau, V., Major, F., Ferbeyre, G., and Chartrand, P. (2007). An E2F/miR-20a autoregulatory feedback loop. J. Biol. Chem. *282*, 2135–2143.

Takizawa, M., Tolarová, H., Li, Z., Dubois, W., Lim, S., Callen, E., Franco, S., Mosaico, M., Feigenbaum, L., Alt, F.W., et al. (2008). AID expression levels determine the extent of cMyc oncogenic translocations and the incidence of B cell tumor development. J. Exp. Med. *205*, 1949–1957.

Tan, M.H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A.N., Liu, K.I., Zhang, R., Ramaswami, G., Ariyoshi, K., et al. (2017). Dynamic landscape and regulation of RNA editing in mammals. Nature *550*, 249.

Teater, M., Dominguez, P.M., Redmond, D., Chen, Z., Ennishi, D., Scott, D.W., Cimmino, L., Ghione, P., Chaudhuri, J., Gascoyne, R.D., et al. (2018). AICDA drives

epigenetic heterogeneity and accelerates germinal center-derived lymphomagenesis. Nat Commun *9*, 222.

Teng, B., Burant, C.F., and Davidson, N.O. (1993). Molecular cloning of an apolipoprotein B messenger RNA editing protein. Science *260*, 1816–1819.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Tschöp, K., Conery, A.R., Litovchick, L., Decaprio, J.A., Settleman, J., Harlow, E., and Dyson, N. (2011). A kinase shRNA screen links LATS2 and the pRB tumor suppressor. Genes Dev. *25*, 814–830.

Tsumagari, K., Baribault, C., Terragni, J., Varley, K.E., Gertz, J., Pradhan, S., Badoo, M., Crain, C.M., Song, L., Crawford, G.E., et al. (2013). Early de novo DNA methylation and prolonged demethylation in the muscle lineage. Epigenetics *8*, 317–332.

Ustanina, S., Carvajal, J., Rigby, P., and Braun, T. (2007). The myogenic factor Myf5 supports efficient skeletal muscle regeneration by enabling transient myoblast amplification. Stem Cells *25*, 2006–2016.

Vonica, A., Rosa, A., Arduini, B.L., and Brivanlou, A.H. (2011). APOBEC2, a selective inhibitor of TGFβ signaling, regulates left–right axis specification during early embryogenesis. Developmental Biology *350*, 13–23.

Walsh, K., and Perlman, H. (1997). Cell cycle exit upon myogenic differentiation. Current Opinion in Genetics & Development *7*, 597–602.

Wang, I.X., So, E., Devlin, J.L., Zhao, Y., Wu, M., and Cheung, V.G. (2013). ADAR Regulates RNA Editing, Transcript Stability, and Gene Expression. Cell Reports *5*, 849–860.

Watson, J.D., and Crick, F.H.C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature *171*, 737–738.

Wedekind, J.E., Dance, G.S.C., Sowden, M.P., and Smith, H.C. (2003). Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. Trends in Genetics *19*, 207–216.

Wilkins, M.H.F., Stokes, A.R., and Wilson, H.R. (1953). Molecular structure of deoxypentose nucleic acids. Nature *171*, 738–740.

Wright, M.W., Eyre, T.A., Lush, M.J., Povey, S., and Bruford, E.A. (2005). HCOP: The HGNC comparison of orthology predictions search tool. Mamm Genome *16*, 827–828.

Yaffe, D., and Saxel, O. (1977). Serial passaging and differentiation of myogenic cells isolated from dystrophic mouse muscle. Nature *270*, 725–727.

Yamanaka, S., Poksay, K.S., Arnold, K.S., and Innerarity, T.L. (1997). A novel translational repressor mRNA is edited extensively in livers containing tumors caused by the transgene expression of the apoB mRNA-editing enzyme. Genes Dev. *11*, 321–333.

Yang, X., Koltes, J.E., Park, C.A., Chen, D., and Reecy, J.M. (2015). Gene Co-Expression Network Analysis Provides Novel Insights into Myostatin Regulation at Three Different Mouse Developmental Timepoints. PLOS ONE *10*, e0117607.

Yang, Y., Fear, J., Hu, J., Haecker, I., Zhou, L., Renne, R., Bloom, D., and McIntyre, L.M. (2014). Leveraging biological replicates to improve analysis in ChIP-seq experiments. Comput Struct Biotechnol J *9*.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS *16*, 284–287.

Yu, G., Wang, L.-G., and He, Q.-Y. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics *31*, 2382–2383.

Zhang, H., Yang, B., Pomerantz, R.J., Zhang, C., Arunachalam, S.C., and Gao, L. (2003). The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. Nature *424*, 94–98.

Zhang, K., Sha, J., and Harter, M.L. (2010). Activation of Cdc6 by MyoD is associated with the expansion of quiescent myogenic satellite cells. J. Cell Biol. *188*, 39–48.

Zhang, P., Wong, C., DePinho, R.A., Harper, J.W., and Elledge, S.J. (1998). Cooperation between the Cdk inhibitors p27KIP1 and p57KIP2 in the control of tissue growth and development. Genes & Development *12*, 3162–3167.

Zhang, P., Wong, C., Liu, D., Finegold, M., Harper, J.W., and Elledge, S.J. (1999). p21(CIP1) and p57(KIP2) control muscle differentiation at the myogenin step. Genes Dev. *13*, 213–224.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biology *9*, R137.

Zhu, L.J. (2013). Integrative Analysis of ChIP-Chip and ChIP-Seq Dataset. In Tiling Arrays, (Humana Press, Totowa, NJ), pp. 105–124.

Zhu, L.J., Gazin, C., Lawson, N.D., Pagès, H., Lin, S.M., Lapointe, D.S., and Green, M.R. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics *11*, 237.