2014

Genome-Wide Characterization of the Effects of Nucleic Acid Modifying Enzymes: Cytidine Deaminases and DNA Methylation

Eric Luke Fritz

GENOME-WIDE CHARACTERIZATION OF THE EFFECTS OF NUCLEIC ACID MODIFYING ENZYMES: CYTIDINE DEAMINASES AND DNA METHYLATION

A Thesis Presented to the Faculty of

The Rockefeller University

in Partial Fulfillment of the Requirements for

the degree of Doctor of Philosophy

by

Eric Luke Fritz

June 2014

GENOME-WIDE CHARACTERIZATION OF THE EFFECTS OF NUCLEIC ACID MODIFYING ENZYMES: CYTIDINE DEAMINASES AND DNA METHYLATION

Eric Luke Fritz, Ph.D.
The Rockefeller University 2014

Activation-induced cytidine deaminase (AID) is essential for two processes of immunoglobulin diversification in germinal center B cells: somatic hypermutation (SHM), in which mutations are introduced into immunoglobulin (Ig) genes, and class-switch recombination (CSR), in which genomic constant regions are recombined to encode antibodies of different isotypes. Both of these processes require AID-catalyzed C-to-U lesions at the Ig loci, which are resolved to generate point mutations or double-stranded DNA breaks in the cases of SHM and CSR, respectively. Despite over a decade of intense study, a number of open issues remain surrounding AID. The diversity of findings regarding AID's role in DNA demethylation raises the question of the scope of its involvement in this process. Additionally, while it is clear that AID-mediated damage occurs, the effects of this damage on the average B cell have not been characterized. Finally, the issue of whether AID is able to edit RNA *in vivo* has never been rigorously addressed in the literature.

In each of these cases, the advent of high-throughput sequencing provides methods for genome-wide characterization of AID's effects. This thesis presents the application of a number of genome-scale, sequencing-based methods to characterize the effects of AID deficiency and overexpression on the activated B cell: mRNA-Seq and miRNA-Seq allow for measurements of RNA expression and editing, while reduced-representation bisulfite sequencing (RRBS) assays DNA methylation. These analyses confirmed AID's known role in

immunoglobulin isotype switching, while also demonstrating that it has little other effect on gene expression. Additionally, no evidence of AID-dependent mRNA or miRNA editing could be detected. Finally, RRBS data failed to support a role for AID in the regulation of DNA methylation. Thus, despite evidence of its additional activities in other systems, antibody diversification appears to be AID's sole physiological function in activated B cells.

Following the conclusion of my studies of AID's effects in B cells, I applied similar genomics tools to two amenable topics in nucleic acid modifications. First, I used mRNA-Seq to attempt to determine the substrate of the orphan cytidine deaminase Apolipoprotein B mRNA-editing enzyme, catalytic polypeptide 2 (APOBEC2). Next, I used whole-genome bisulfite sequencing to explore the distribution of 5-methylcytosine in *Trypanosoma brucei*. In both of these cases, results were inconclusive but suggest future directions for investigation.

Acknowledgments

First I would like to thank my advisor, Nina Papavasiliou, for her enthusiasm, guidance, and support. I would like to thank the members of my faculty advisory committee for their time and thoughtful suggestions: Elaine Fuchs, Mary Goll, George Cross, and David Allis. I would also like to thank Reuben Harris for serving as my external examiner.

The Papavasiliou lab has been tremendously important to my development as a scientist and a person, and I would like to thank all of the lab members and rotation students that I have had the pleasure of working with over my time there: Peter Alff, Catharine Boothroyd, Jan Davidson-Moncada, Rebecca Delker, Nick Economos, Dimitris Garyfallos, Will Gibson, Paul Hakimpour, Claire Hamilton, Dewi Harjanto, Galadriel Hovel-Miner, Sasa Jereb, Roos Karssemeijer, Hee-sook Kim, Marianne Labriola, Kenneth Lay, Tanya Leonova, Linda Molla, Monica Mugnier, Jason Pinger, Anita Ramnarain, Violeta Rayon Estrada, Brad Rosenberg, Danae Schulz, Theodoros Sklaviadis, Pete Stavropoulos, Alex Strikoudis, Grace Teng, Gianna Triller, and Maryam Zaringhalam. In particular, I would like to thank Brad Rosenberg, without whose enthusiasm and advice this work would not have been possible, and Linda Molla, whose drive and dedication leave me no doubt that Apobec2 will be figured out very soon.

I would also like to thank the many people from outside of the laboratory that contributed to this work: Geulah Livshits for immunofluorescence reagents, Slobodan Beronja for shRNA constructs, Chris Mason for ERCC RNA spikes, Aleks Mihailovic for miRNA-Seq library preparation and sequencing, Klara Velinzon for cell sorting, Scott Dewell for sequencing advice, Yushan Li for performing Epityper assays, Alin Vonica for the APOBEC2 antibody, and the staff of the Genomics Resource Center for sequencing, QC, and qPCR assays. I would also like to thank the Cancer Research Institute, the National Institutes of Health, and the Starr Foundation for funding.

Finally and most importantly, I would like to thank Mom, Dad, Chrissy, Emily, and Claire for their love and encouragement.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction[*]

The different behavior of the cells of an organism despite their shared DNA sequence has been one of the great motivating questions of molecular biology. One of the key advances in understanding this process has been recognition of the influence of chemical modifications of nucleic acids themselves. Although it has long been known that such an absolutist view of the central dogma is a gross oversimplification (Crick, 1970), the characterization of nucleic acid modifications has made it clear that DNA and RNA are not passive one-dimensional encodings of protein sequences. Rather, their properties can be altered by a wide array of modifying enzymes. The resulting menagerie of modified bases have crucial consequences for nearly every aspect of cell function (Grosjean, 2009).

1.1 Nucleic acid modifications

Both DNA and RNA are biochemically modified in a staggering number of ways in the cell. RNA in particular displays great diversity in its modifications, with 144 identified to date (Limbach et al., 1994; Machnicka et al., 2012). It is likely that RNA's greater diversity is a consequence of both the larger number of roles that RNA plays in the cell and the importance of preventing deleterious changes in DNA.

These modifications are important for the function of nearly every type of RNA. Modifications of tRNA are both the most diverse and most frequent, with about 17% of bases modified in eukaryotic cells (Jackman and Alfonzo, 2012).

[*] Portions of this chapter were published in (Fritz and Papavasiliou, 2010) and (Fritz et al., 2013)

Perhaps most famously, the modified base inosine at position 34 of tRNAs is responsible for the "wobble" base pairing with the third position of the codon (Murphy and Ramakrishnan, 2004). In rRNA, modifications occur at a large number of functionally important regions and influence stability (Decatur and Fournier, 2002). Modifications, in particular pseudouridylation and 2'-O-methylation, play a similar stabilizing role in snRNAs (Karijolich and Yu, 2010). Modifications of mRNAs include N6-methylation of adenosine, which is dynamically regulated and important for modulating RNA stability (Wang et al., 2014). Also found within mRNA are the deamination modifications, which convert adenosine and cytosine to inosine and uracil respectively. This process, which also occurs in tRNA and miRNA, is termed RNA editing, and will be treated at greater length.

DNA modifications, while smaller in number, still impact cellular function in a variety of important ways. These modifications can be assigned to two classes: those that are introduced enzymatically, and those that are not. The second class, which includes bases such as 8-oxoguanine, occur as a result of oxidative damage and do not appear to be functional (Cooke et al., 2003).

With the exceptions of deamination of adenosine, cytosine, and its derivatives, all known enzymatic modifications of DNA bases do not affect base pairing. One well studied example is N6-methyladenosine, which among other functions is crucial in *E. coli* for discriminating between the template and newly synthesized strands for purposes of DNA repair (Barras and Marinus, 1989) and preventing re-replication of newly synthesized DNA (Russell and Zinder, 1987). Another notable DNA base is β-d-glucopyranosyloxymethyluracil, or base J,

which prevents transcriptional readthrough in *Leishmania* (van Luenen et al., 2012). However the most prominent class of modified DNA bases in vertebrates are 5-substituted cytosines. These bases, 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC), appear to be the only persistent modified bases with functional consequences found in mammalian DNA (Figure 1.1). Of these, 5mC was the first discovered (Hotchkiss, 1948) and is by far the most abundant, most versatile, and best studied.

1.1.1 DNA cytosine methylation

1.1.1.1 Functions of DNA methylation

5mC in DNA is common to all domains of life and serves a wide variety of functions in different organisms. In *E. coli*, it is important for the regulation of stationary phase transcription (Kahramanoglou et al., 2012). In the fungus *Neurospora crassa*, cytosine methylation is crucial for silencing repetitive elements (Selker et al., 2003). The cytosine methylation system is extremely complex in *Arabidopsis*, where it maintains a silenced state for transposons and regulates gene expression (Stroud et al., 2013). In the ciliated protist *Oxytricha trifallax*, 5mC marks DNA for degradation during a complicated set of genome rearrangements that accompany reproduction (Bracht et al., 2012). Methylation of cytosine also appears to be important for life cycle stage-specific gene expression in the parasitic nematode *Trichinella spiralis* (Gao et al., 2012) and caste-specific gene expression in the honeybee (Elango et al., 2009). In many other organisms, such as the kinetoplastid *T. brucei*, the existence of DNA methylation has been reported, but its function is unknown (Militello et al., 2008).

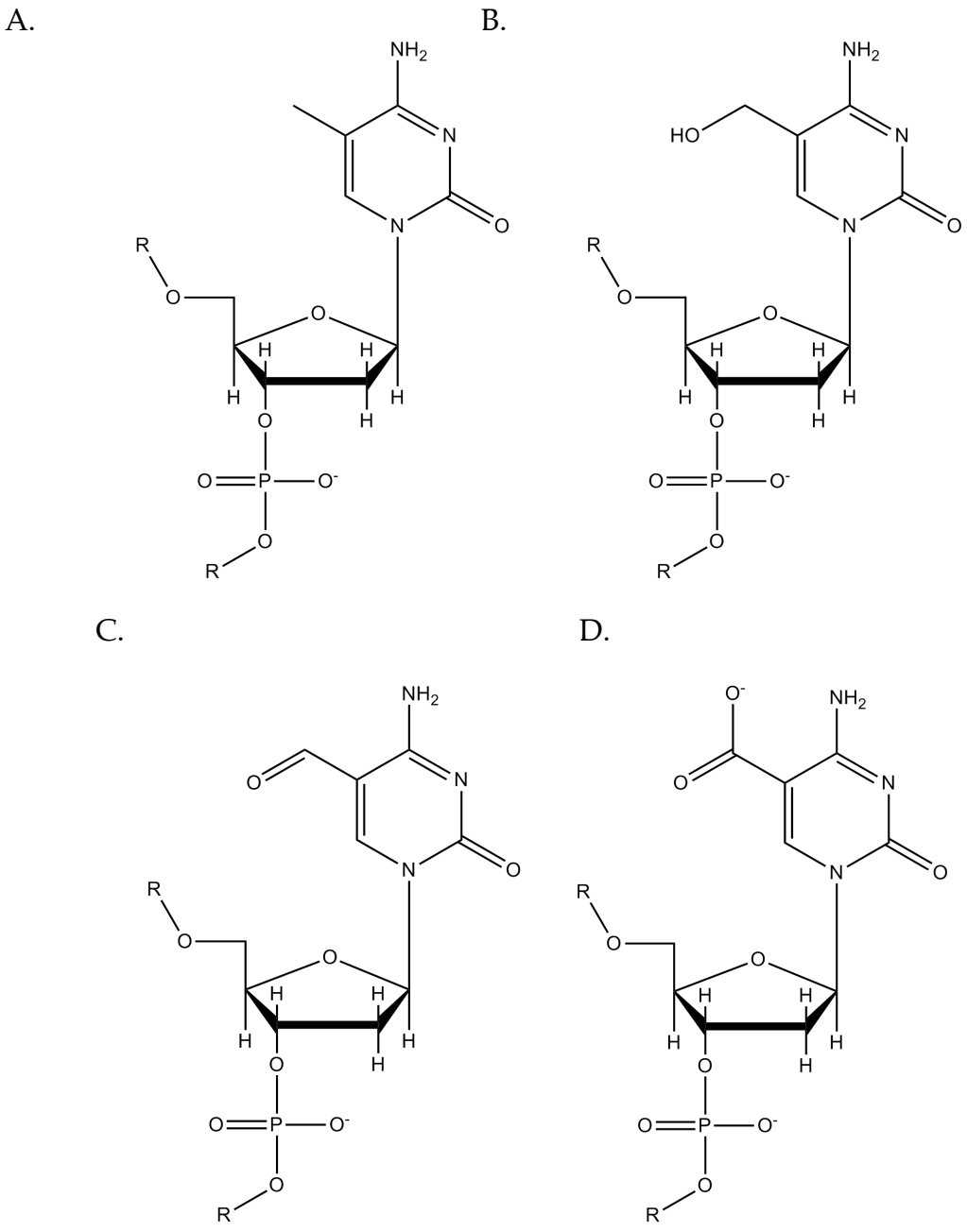


Figure 1.1. 5-substitued deoxycytidine derivatives. (A) 5-methylcytosine, (B) 5-hydroxymethylcytosine, (C) 5-formylcytosine, (D) 5-carboxycytosine.

In mammals, 5mC occurs predominantly in the sequence context CpG, which allows for symmetrical marking of both strands. Approximately 70-80% of cytosines in this context are methylated in mammalian DNA, which represents about 1% of the genome overall (Ehrlich et al., 1982). However there is also appreciable 5mC in non-CpG sequence contexts in certain mammalian cell types (Ramsahoye et al., 2000), and widespread methylation in other contexts in fungi (Rountree and Selker, 1997) and plants (Cokus et al., 2008). Notably, the dinucleotide CG is both globally depleted and unevenly distributed within mammalian genomes (Illingworth and Bird, 2009). This has been attributed to the higher inherent mutagenicity of 5mC compared to unmodified cytosine (Bird, 1980) as well as the functional importance of the CpG-rich regions, which are termed "CpG islands."

Originally, cytosine methylation in mammals was thought to function as a silencing mark that could explain stably inherited differences in gene expression between cellular lineages of an organism (Holliday and Pugh, 1975; Riggs, 1975). Many examples have since been found that clearly support the hypothesis that methylation of an element leads to its silencing. The inactive X chromosome is almost entirely methylated at CpG, and can be reactivated by loss of methylation (Mohandas et al., 1981). Retrotransposons are frequently densely methylated, and this methylation inhibits their mobilization (Yoder et al., 1997). Imprinted genes are also regulated by methylation, often with the silent allele modified (Li et al., 1993). Finally, many developmentally regulated genes are inhibited in a relatively straightforward manner by methylation, the prototypical case being *Oct4* during transition from the ES cell state (Hattori et al., 2004).

However, it has become clear that DNA methylation is not a simple and absolute bar to transcription. The model that has emerged holds that at some, but not all, CpG island-containing promoters, methylation leads to long-term silencing, while at CpG-poor promoters, methylation has little effect on transcription (Jones, 2012). Silencing of promoters by CpG island methylation also seems to function primarily in early development, and is a minor means of gene expression regulation in later differentiation, if it occurs at all (Bock et al., 2012). Additionally, it has been shown that DNA methylation is a part of a larger network of factors that are involved in silencing. It appears that silencing of a gene by other DNA-binding proteins generally precedes DNA methylation, which functions more as a heritable stabilizer of the silenced state (Cedar and Bergman, 2009).

DNA methylation also has several functions that do not fit neatly into the classical promoter-silencing model. It appears that most of the functional differences in DNA methylation between differentiated cells occur at enhancers and insulators rather than promoters (Ziller et al., 2013). Gene bodies also vary widely in their levels of methylation, but the significance of this is still unclear (Jones, 2012). It has also been proposed that the lower levels of methylation in introns compared to exons may serve a functional role in splicing (Laurent et al., 2010).

1.1.1.2 Establishment and maintenance of DNA methylation

Shortly after implantation, the embryo undergoes a period of rapid methylation that is catalyzed by the "*de novo*" DNA cytosine methyltransferases DNMT3A and DNMT3B (Okano et al., 1999). This leads to a pattern of near-

global methylation of CpG-poor sequences, with some CpG-rich regions protected by DNA binding protein such as SP1 (Macleod et al., 1994). After this initial pattern is established, it is maintained heritably by the action of the "maintenance" DNA cytosine methyltransferase DNMT1 (Li et al., 1992). This enzyme is able to faithfully reproduce DNA methylation patterns on newly synthesized DNA by means of its high specificity for the hemimethylated CpG duplex (Song et al., 2012).

Specific loci become methylated later in development after binding by one of a number of silencing factors, such as the H3 lysine methyltransferases G9A (Feldman et al., 2006) or EZH2 (Viré et al., 2006). These factors recruit DNMT3 enzymes, in addition to other silencing factors such as histone deacetylases. In turn, densely methylated sequences can recruit other silencing factors, such as MECP2, which can regenerate the repressive chromatin state following DNA replication (Nan et al., 1998). In this way, DNA methylation is better understood as a resilient scaffold upon which silencing occurs, rather than the effector of silencing itself.

1.1.1.3 DNA demethylation

While the means for the establishment and maintenance of DNA methylation in mammals is relatively well understood, the means by which the mark is removed from DNA is a far more controversial area. DNA demethylation occurs at several points during development. The most dramatic loss of methylation occurs immediately after zygote formation, when the paternally-derived genome undergoes almost complete demethylation in 6-8 hours (Mayer et al., 2000). The maternally-derived genome undergoes

demethylation of similar scale but with much slower kinetics, achieving full demethylation at the morula stage (Howlett and Reik, 1991). The other large-scale demethylation event that takes place is the "resetting" of the global somatic cell methylation in primordial germ cells (PGCs), the precursors to germ cells. Between E7.5 and E11.5, this population of cells lose most DNA methylation (Hajkova et al., 2002). Small scale, targeted demethylation events also take place during differentiation of somatic cells, which appear to be driven by proximal transcription factor binding (Stadler et al., 2011).

Two broad classes of mechanisms have been proposed for DNA demethylation: passive and active. Passive demethylation refers to the dilution of DNA methylation through replication in the absence of maintenance methylation activity. This appears to be the dominant mechanism for demethylation of maternal genome in the zygote: methylation gradually decreases at the same time that oocyte-derived DNMT1 is excluded from the nucleus (Howell et al., 2001). It may also account for some site-specific demethylation later in development, with various DNA-binding proteins occluding newly synthesized DNA and preventing access for DNMT1 (Hsieh, 2000).

While the literature dealing with passive DNA demethylation is relatively straightforward, the mechanisms of active demethylation in mammals have been considerably more controversial. While direct, one-step reversal of cytosine methylation is too energetically unfavorable a process to occur, the AlkB family of enzymes have been shown to demethylate 1-methyladenosine and 3-methylcytosine via oxidized intermediates, releasing the methyl carbon as formaldehyde and regenerating the original base (Falnes et al., 2002; Trewick et al., 2002). However, no comparable Alkb-family enzyme that accepts 5mC as a

substrate has been found in vertebrates. The methyl-CpG-binding protein MBD2 has been proposed as such a direct DNA demethylase (Bhattacharya et al., 1999), but these findings could not be replicated (Bird, 2002). A problem with the direct-demethylation candidates is that a number of studies indicate that DNA is broken and repaired in the course of demethylation (Barreto et al., 2007; Hajkova et al., 2010; Kress et al., 2006). A hypothesis not contradicted by these reports is that DNA demethylation in mammals is achieved by targeted removal of 5mC by a glycosylase without previous modification of the base (Jost, 1993). This is thought to be the dominant mechanism of active DNA demethylation in plants, with the DME/ROS1 family of glycosylases serving this function in *Arabidopsis* (Zhu, 2009). MBD4 has been proposed as this glycosylase in mammals (Kim et al., 2009b), although previous *in vitro* work found this enzyme to be far more active on thymidine than 5mC (Zhu et al., 2000). Other reports suggested that demethylation may proceed by a radical mechanism catalyzed by the elongator complex member ELP3 (Okada et al., 2010) or by conversion to thymidine by DNMT3-family enzymes under conditions of low S-adenosylmethionine (Kangaspeska et al., 2008; Métivier et al., 2008).

Among all of these proposed mechanisms, the evidence most strongly supports two: demethylation by oxidation, and demethylation by deamination. The characterization of the TET family of enzymes has given credence to the former. These Fe(II)- and α-ketoglutarate-dependent enzymes convert 5mC to 5hmC *in vitro*, and lead to global depletion of 5mC when overexpressed (Tahiliani et al., 2009). Subsequent work has shown that demethylation of the paternal zygotic genome (Gu et al., 2011b) and PGCs (Hackett et al., 2013) require

TET-family proteins, as do other examples of site-specific demethylation (Klug et al., 2013).

There have also been a number of proposals as to the post-oxidation steps in TET-mediated demethylation. It has been suggested that C is directly regenerated by decarboxylation of 5caC (Schiesser et al., 2012), but no decarboxylase with such an activity has been identified. It is also possible in some cases that a pseudo-passive mechanism may take place, with hemihydroxymethylated CpG failing to serve as a suitable substrate for DNMT1 (Inoue and Zhang, 2011), but the kinetics of some examples of active demethylation preclude this from serving as the only mechanism. A 5hmC-specific (or 5fC- or 5caC-specific) glycosylase could also lead to net demethylation by base excision repair, and it has been suggested that TDG could serve this purpose (Maiti and Drohat, 2011). Beyond the possibility that TET-mediated oxidation represents the mechanism of active DNA demethylation, the discovery of high levels of 5hmC in Purkinje neurons (Kriaucionis and Heintz, 2009) and ES cells (Pastor et al., 2011) as well as the existence of proteins that have very high affinity for 5hmC (Mellén et al., 2012) and 5fC (Iurlaro et al., 2013) suggests that oxidized cytosine derivatives may be functional marks in their own right and not simply intermediates.

The other well-supported mechanism for active demethylation in mammals involves deamination by cytidine deaminases. The observation that activation-induced cytidine deaminase (AID) and apolipoprotein B editing enzyme, catalytic polypeptide 1 (APOBEC1) can act on 5mC in DNA led to the hypothesis that the resulting T:G mismatch could be repaired in an error-free manor to yield net demethylation (Morgan et al., 2004). It has also been

hypothesized that deaminase-mediated damage could lead to processive-like demethylation by long-patch base excision repair, and could do so by acting on unmodified cytosine (Fritz and Papavasiliou, 2010). The evidence for and against deamination-mediated mechanisms of demethylation will be discussed at length in a later section.

1.1.2 RNA editing

The term RNA editing is used to describe two separate varieties of RNA modification: base-insertion/deletion editing, and base-modification editing. Insertion/deletion editing was the first type discovered, and involves the post-transcriptional addition or deletion of internal uridines to yield RNA sequences that differ from their source DNA (Benne et al., 1986). This process is crucial for proper expression of certain mitochondrial genes in kinetoplastids, and is crucial for their survival (Schnaufer et al., 2001).

While insertion/deletion editing is unique to kinetoplastids, base-modification editing is much more widespread. The two types of base-modification editing that occur in vertebrates are termed A-to-I editing, which is catalyzed by the ADAR family of proteins, and C-to-U editing, which is catalyzed by the AID/APOBEC family. A-to-I editing occurs extremely widely in mammalian transcriptomes, with more than 10^8 sites identified as being edited by ADAR at some level (Bazak et al., 2014). The majority of these events occur in Alu repeats in transcript 3'UTRs (Ramaswami et al., 2012).

Adar-catalyzed editing has been shown to have a number of functional roles. Because ADAR enzymes show a preference for double-stranded RNA (Nishikura et al., 1991) and A-to-I editing disrupts complementarity, it has been

suggested that the general function of this type of editing is modulating RNA secondary structure. In support of this hypothesis, ADAR and the RNAi pathway appear to compete for substrates (Wu et al., 2011), and lack of ADAR1 can lead to an inflammatory response consistent with overproduction of endogenous dsRNA (Hartner et al., 2009; Rice et al., 2012).

Because inosine has similar base-pairing properties as guanosine, A-to-I editing can have direct effects on coding sequences as well. This mode of regulation diversifies the coding sequences of a number of ion channels (Burns et al., 1997; Higuchi et al., 1993; Hoopengardner et al., 2003), and has been shown to be a mechanism of temperature adaptation in octopi (Garrett and Rosenthal, 2012). Such coding changes may also contribute to the anti-viral activity of ADAR (Hamilton et al., 2010).

1.2 The AID/APOBEC family of cytosine deaminases

The AID/APOBEC family of polynucleotide cytidine deaminases catalyze deamination of cytosine to yield uracil in single-stranded DNA, with some exceptions. They all possess at least one characteristic catalytic domain in which the histidine and cysteines of the motif H[AV]E-$X_{[24-36]}$-P$CX_{[2-4]}$C coordinate a zinc ion, which in turn activates a water molecule to add at the 4 position of cytosine (Conticello, 2008). Subsequent loss of ammonia yields net conversion of cytosine to uracil (Figure 1.2). While these proteins share a common mechanism, there is great diversity in the specific substrates and functional importance of the AID/APOBECs, which will be discussed in depth for each member of the family.

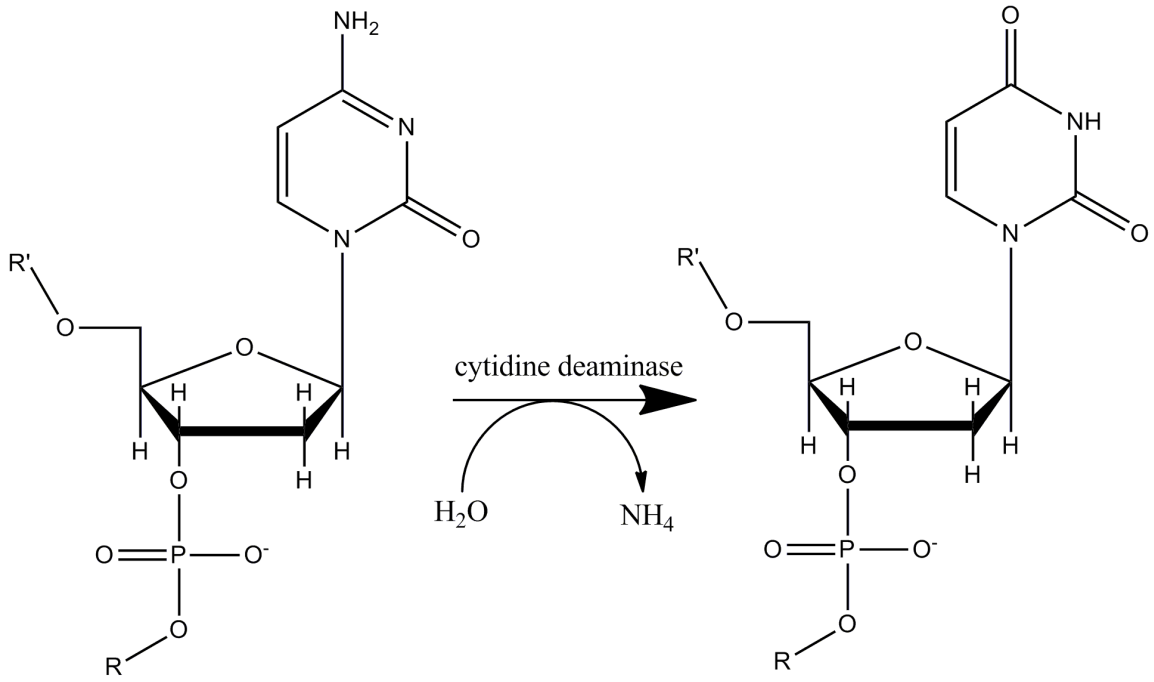


Figure 1.2. Mechanism of polynucleotide cytosine deamination.

1.2.1 APOBEC1

APOBEC1 was the first member of the AID/APOBEC family to be discovered (Teng et al., 1993). Its well-characterized activity is the site-specific editing of the *ApoB* transcript in small intestinal enterocytes. This editing results in a premature stop codon and translation of a truncated form of the protein with different properties in lipid metabolism (Chen et al., 1987; Powell et al., 1987).

APOBEC1 is unique among the AID/APOBEC family as the only member for which RNA editing has been demonstrated *in vivo*. It has also been shown that APOBEC1 can edit a large number of transcript 3'UTRs *in vivo* (Rosenberg et al., 2011), and that is overexpression in the liver causes tumor formation (Yamanaka et al., 1997). In DNA, APOBEC1 can act on cytosine in *E. coli* (Harris et al., 2002) and cytosine and 5-methylcytosine in DNA *in vitro* (Morgan et al., 2004). The latter observation led to the hypothesis that it may be involved in DNA demethylation, and there is some evidence that this may take place in neurons (Guo et al., 2011).

1.2.2 APOBEC2

APOBEC2 is clearly a member AID/APOBEC family by sequence similarity, and is evolutionarily well-conserved as far back as bony fish (Conticello, 2004). It has also been successfully crystalized, and displays folds characteristic of free nucleotide cytidine deaminases (Prochnow et al., 2006). Despite these features, no catalytic activity for this protein has been identified (Anant et al., 2001; Harris et al., 2002; Mikl et al., 2005; Nabel et al., 2012). However, a number of phenotypes have been reported in the absence of APOBEC2. It is expressed at high levels in skeletal and cardiac muscle (Liao et

al., 1999), with higher levels in slow-twitch fibers than in fast-twitch (Mikl et al., 2005). Mice deficient in APOBEC2 display a higher proportion of slow-twitch muscle fibers, as well as decreased body mass and myopathy in later life (Sato et al., 2010). Mice that constitutively overexpress APOBEC2 may also develop lung and liver tumors at an elevated rate (Okuyama et al., 2012). In other species, more dramatic phenotypes have been observed. Knockdown of *apobec2* in early development results in impaired left/right axis definition in *Xenopus laevis*, which was attributed to inhibition of TGF-β signaling (Vonica et al., 2011). In zebrafish, morpholino knockdown of *apobec2* at the single-cell stage led to severe muscle dystrophy (Etard et al., 2010) or neuron loss, which was attributed to hypermethylation of the *neurod2* promoter (Rai et al., 2008). Knockdown in adult zebrafish was also found to impair nerve regeneration following injury (Powell et al., 2012).

There is no consensus on how APOBEC2 achieves these effects. It has been hypothesized that APOBEC2 may be involved in DNA demethylation (Powell et al., 2012; Rai et al., 2008), RNA editing (Liao et al., 1999), or that it has lost its cytidine deaminase activity altogether and may act by a different mechanism (Etard et al., 2010; Sato et al., 2010; Vonica et al., 2011).

1.2.3 The APOBEC3 subfamily

Since its appearance in the placental lineage of mammals, the APOBEC3 subfamily of proteins have undergone a number of fusions and expansions, resulting in a single *Apobec3* gene in mice and 7 in humans (Conticello, 2008). The

rapid diversification of this subfamily in the primate lineage is not surprising given these enzymes' function: restriction of viruses and retrotransposons.

APOBEC3G has been the most intensely studied of the APOBEC3s because of its identification as the HIV restriction factor antagonized by the viral gene product Vif (Sheehy et al., 2002). This activity is achieved by hypermutation of retroviral (-)-strand cDNA (Harris et al., 2003; Lecossier et al., 2003; Mangeat et al., 2003; Zhang et al., 2003). Other APOBEC3 enzymes display similar antiviral activity against other viruses (Chen et al., 2006; Delebecque et al., 2006; Russell et al., 2005; Turelli et al., 2004), and restrict retrotransposons (Esnault et al., 2005; Muckenfuss et al., 2006), and foreign DNA (Stenglein et al., 2010).

The potent antiviral activities of APOBEC3s come at the cost of oncogenic mutations. APOBEC3s generally (Alexandrov et al., 2013; Nik-Zainal et al., 2012; Roberts et al., 2013) and APOBEC3B specifically (Burns et al., 2013a; 2013b) have been linked to a variety of human cancers by their characteristic mutational spectra.

1.2.4 Activation-induced cytidine deaminase (AID)

Activation-induced cytidine deaminase (AID) was initially identified in 1999 as a factor required for class-switch recombination (CSR) and immunoglobulin somatic hypermutation (SHM) (Muramatsu et al., 2000; 1999; Revy et al., 2000). The 24kD protein, encoded by the *Aicda* gene, is conserved among jawed vertebrates and appears to be the basal member of the APOBEC family (Conticello, 2004). Although it was initially hypothesized to be an RNA editor due to its similarity to APOBEC1 (Muramatsu et al., 1999), it has since been established that AID is able to convert cytosine to uracil in single-stranded

DNA, as demonstrated in *E. coli* (Petersen-Mahrt et al., 2002) and *in vitro* (Bransteitter et al., 2003; Chaudhuri et al., 2003; Dickerson et al., 2003).

1.2.4.1 Class-switch recombination and immunoglobulin somatic hypermutation

In the decade since AID's discovery, a broadly accepted model for its roles in antibody diversification has emerged (Delker et al., 2009; Di Noia and Neuberger, 2007). In this model, AID initiates CSR and SHM by conversion of cytosine to uracil in different regions of the immunoglobulin (Ig) loci. CSR occurs as a result of the double-stranded breaks frequently produced in the course of repair of such lesions in the S regions of the IgH locus. Joining of breaks in different S regions results in a different constant region immediately downstream of the transcribed V(D)J and consequently to antibodies of a different isotype (Xu et al., 2012). SHM is initiated by AID deamination within the V(D)J region of Ig loci. Repair of these lesions proceeds with an unusually high error rate, leading to mutations and thus altering the affinity of the encoded antibody (Di Noia and Neuberger, 2007). Selection of cells bearing these mutated immunoglobulins leads to affinity maturation. AID is the sole initiator of these processes: *Aicda*$^{-/-}$ mice exhibit a complete lack of secondary Ig isotypes and no mutations in Ig variable regions during an immune response (Muramatsu et al., 2000). Mutations in the *AICDA* gene in humans result in a similar condition known as hyper-IgM syndrome type 2 (Revy et al., 2000).

1.2.4.2 Off-target effects of AID in B cells

Although AID displays striking specificity in its action on the Ig loci, it can act at other points in the genome. A large number of loci are mutated at an elevated rate in the presence of AID, including *Bcl6* and *Fas* (Muschen et al., 2000; Pasqualucci et al., 1998; Shen et al., 1998). In some cases, notably mutations at *Myc*, these AID-catalyzed "non-Ig somatic hypermutations" have been shown to contribute to B cell tumors in mice (Pasqualucci et al., 2001). AID appears to act at a much larger number of loci, and that fidelity of repair following damage varies greatly at different sites (Liu et al., 2008). In addition, AID can catalyze the formation of DNA breaks, which can lead to translocations. The most notable of these is the *Myc*/*Igh* translocation that is characteristic of Burkitt's lymphoma (Pasqualucci et al., 2007; Robbiani et al., 2008), but it has become clear that AID-dependent DNA breaks are distributed broadly throughout the genome (Chiarle et al., 2011; Klein et al., 2011).

1.2.4.3 AID and DNA demethylation

Despite the lack of an obvious non-immune phenotype in AID-deficient mice, there are signs that AID has additional functions outside of antibody diversification. Notably, it is expressed in many cell types other than B cells, namely oocytes, PGCs, ES cells (Morgan et al., 2004), breast tissue (Pauklin et al., 2009), and prostate epithelial cells (Lin et al., 2009), although its presence in PGCs has recently been challenged (Hajkova et al., 2010). As AID is a DNA mutator, its expression outside of B cells would likely have been strongly selected against if it had no function in these tissues. Further suggestions of functions for AID beyond the immune system come from studies of lower vertebrates. As in mice, AID

expression is found during early development in *D. rerio* (Rai et al., 2008), *Xenopus* (Marr et al., 2007), and the newt *P. waltl* (Bascove and Frippiat, 2010). The broad conservation of AID expression in early development strongly suggests a function at that stage.

The first direct evidence that AID might have functions beyond the standard model of antibody diversification came in 2004, when it was shown by Petersen-Mahrt and colleagues that AID, along with the related cytidine deaminase APOBEC1, can convert 5mC in single-stranded DNA to thymidine *in vitro* (Morgan et al., 2004). This observation led to the proposal that these enzymes could function in DNA demethylation.

AID could initiate demethylation by a damage-and-repair mechanism similar to that used in SHM (Figure 1.3). The deamination of 5mC by AID yields thymidine. This T would then be removed by a T-G mismatch-specific glycosylase, of which two, TDG and MBD4, are known in mammals (Hardeland et al., 2003; Millar et al., 2002). The resulting abasic site would then be replaced by an unmethylated cytidine via base excision repair (BER) processes, yielding the net removal of methylation without alteration of sequence. BER could also proceed through either short or long patch repair, potentially yielding demethylation of multiple neighboring cytosines in the latter case. This could give rise to the appearance of processive demethylation, despite originating from a single deamination event.

The AID model of demethylation received a measure of *in vivo* validation from work in *D. rerio* early embryos (Rai et al., 2008). Introduction of a methylated DNA fragment into single-cell embryos induced expression of AID along with related putative cytidine deaminases Apobec2a and Apobec2b.

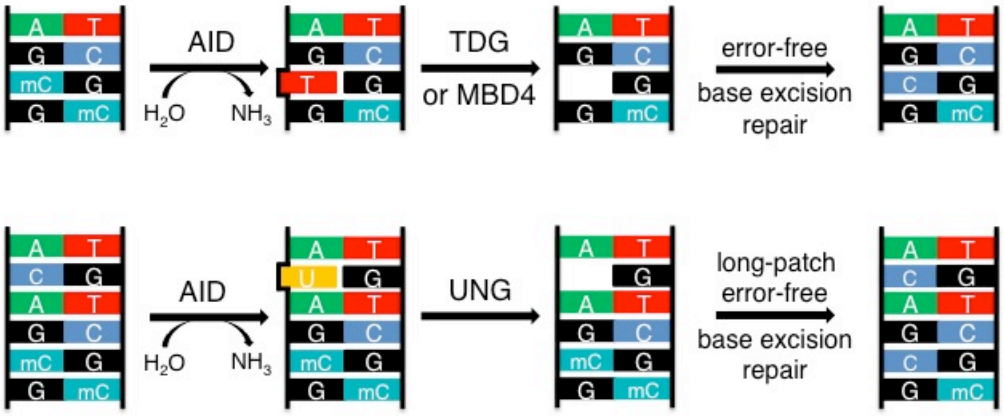


Figure 1.3. Proposed mechanisms of AID-dependent DNA demethylation. Deamination of 5mC by AID yields thymine. Excision of thymine by either of the mammalian T-G-specific glycosylases (TDG or MBD4) and subsequent error-free replacement with cytosine would yield loss of methylation at that base on one strand. Deamination of cytosine followed by excision of uracil and error-free long-patch BER could lead to replacement and net demethylation of neighboring 5mCs.

Additionally, overexpression of AID along with the T-G glycosylase MBD4 leads to efficient demethylation of a methylated DNA fragment, with observable conversion of 5mC to T. Morpholino knockdown of AID or of MBD4 in *D. rerio* single-cell stage embryos results in hypermethylation of the promoter of the neurogenesis-related transcription factor *neurod2* at 80% epiboly. Furthermore, both AID and MBD4 were also detectable at the locus by chromatin immunoprecipitation (ChIP), suggesting that the promoter's methylation state was dynamic through development and not just the result of transmission of altered methylation state at the single-cell stage. Consequently, AID morphants displayed severe defects in neurogenesis consistent with decreased *neurod2* expression. The tissue- and locus-specific phenotype suggests that active demethylation may serve as a mechanism for regulating specific lineage decisions, as opposed to only resetting methylation state grossly in the zygote and germline.

Intriguingly, Rai and colleagues also found that the Gadd45 family of proteins, previously implicated in active demethylation in *Xenopus* (Barreto et al., 2007), were involved in AID-dependent demethylation. Combined knockdown of four of the six Gadd45 family members found in zebrafish sharply reduced demethylation of a reporter and increased methylation of the genome as a whole. It was thus suggested that Gadd45 serves as a scaffold to couple AID and MBD4, supported by the facts that Gadd45α promotes their association with a methylated reporter plasmid and that the three proteins also co-immunoprecipitate. The apparent physical interaction of AID and Mbd4 afforded by Gadd45 provides a mechanism for the tight coupling of deamination

to repair that would be necessary for demethylation without attendant widespread mutation, as has been noted (Law and Jacobsen, 2010).

Although there is no direct evidence for deaminase-dependent effects on cytosine methylation in mammalian somatic tissue, a recent study has implicated AID in establishment of the hypomethylated state of mouse primordial germ cells (PGCs) (Popp et al., 2010). In this study, genomic DNA from sperm, total fetus, placenta, and male and female E13.5 PGCs from wild-type and *Aicda*$^{-/-}$ mice was bisulfite-converted and sequenced using the Illumina ultra-high throughput platform, a method termed BS-Seq or MethylC-Seq (Cokus et al., 2008; Lister et al., 2008). The resulting data provides a genome-wide map of cytosine methylation at single-nucleotide resolution. While the lack of sequencing depth precluded quantitative measures of methylation for every genomic cytosine, the coverage was more than sufficient to detect significant increase in methylation in *Aicda*$^{-/-}$ PGCs. This difference was more pronounced for female than for male PGCs, and was roughly homogeneous throughout the genome. The broad nature of the methylation increase implies that AID functions without regard to specific loci, and thus is a plausible component of germline methylation erasure. This report marks the first description of a non-immune phenotype in *Aicda*$^{-/-}$ mice.

Further evidence of AID's capacity to demethylate DNA in mammalian cells came in a recent study of reprogramming during interspecies heterokaryon formation (Bhutani et al., 2010). This system uses polyethylene glycol fusion of mouse ES cells with human fibroblasts to induce reprogramming of the human nucleus to an ES-like state with high frequency. Because reprogramming is known to involve demethylation, heterokaryon formation allows for study of demethylation in mammalian cells, albeit in a non-physiological context.

Knockdown of AID was found to significantly inhibit reprogramming as measured by transcript levels of the pluripotency markers Oct4 and Nanog. Methylation of the promoters of these genes was also significantly increased as a result of reduced levels of AID. As heterokaryons are non-dividing, AID-dependent demethylation in this system is necessarily active. Similar deficiencies for AID-deficient cells have been noted for reprogramming by Yamanaka factors as well (Kumar et al., 2013; Sabag et al., 2014).

While the evidence that AID has a role in demethylation is suggestive, there are still caveats. It has been found for all tested members of the family that deaminase activity is higher on C than on 5mC (Nabel et al., 2012). Additionally, a recent report detected no AID expression in mouse PGCs at E11.5, although a small level of AID expression was observed at E12.5 (Hajkova et al., 2010). As the epigenetic reprogramming of PGCs begins at E11.5, AID cannot be the sole agent of demethylation. This idea is consistent with the occurrence of significant, albeit reduced, demethylation in the PGCs of *Aicda*$^{-/-}$ mice compared to somatic cells. Whether AID-independent demethylation in these cells is a result of 5mC deamination by another member of the AID/APOBEC family or due to any of the mechanisms mentioned above is an open question. It is also possible that the AID/APOBEC family of deaminases have differing target gene preferences for demethylation, and thus play complementary roles in the genome-wide removal of cytosine methylation.

Another issue raised by these results is the functional importance of the observed PGC hypermethylation, as *Aicda*$^{-/-}$ mice are fully viable. Even if AID were the sole agent of active demethylation, it is not clear that loss of this mechanism would result in a drastic phenotype. This view is supported by

evidence from *A. thaliana*, in which loss of all three 5mC-removing glycosylases in vegetative tissue leads to viable plants displaying increased methylation only at certain loci, and no genome-wide increase (Penterman et al., 2007). As DNA methylation is essential for parental imprinting, Popp and coworkers suggested that the consequences of PGC AID deficiency may lie in retention of a parental-like epigenetic state. Data supporting this view lies in a small but significant difference that exists between wild-type and *Aicda*$^{-/-}$ mice in the relationship between litter size and birth weight (Popp et al., 2010). In normal mice, pups that are part of large litters tend to have lower birth weights, while Popp and colleagues report that in *Aicda*$^{-/-}$ mice this compensation is absent. Hypermethylated elements responsible for this phenotype have not been identified.

1.3 Statement of the problem

Despite over a decade of intense study, a number of open issues remain surrounding AID. The divergent findings regarding AID's role in DNA demethylation raises the question of the scope of its involvement. Additionally, while it is clear that AID-mediated damage occurs, the effects of this damage on the average B cell have not been characterized. Finally, the issue of whether AID is able to edit RNA *in vivo* has never been rigorously addressed in the literature.

In each of these cases, the advent of high-throughput sequencing provides methods for genome-wide characterization of AID's effects. This thesis presents the application of a number of genome-scale, sequencing-based methods to characterize the consequences of AID deficiency and overexpression on the activated B cell: mRNA-Seq and miRNA-Seq allow for measurements of RNA

expression and editing, while reduced-representation bisulfite sequencing assays DNA methylation. These analyses confirmed AID's known role in immunoglobulin isotype switching, while also demonstrating that it has little other effect on gene expression. Additionally, no evidence of AID-dependent mRNA or miRNA editing could be detected. Finally, RRBS data failed to support a role for AID in the regulation of DNA methylation. Thus, despite evidence of its additional activities in other systems, antibody diversification appears to be AID's sole physiological function in activated B cells.

Following the conclusion of my studies of AID's effects in B cells, I applied similar genomics tools to two amenable topics in nucleic acid modifications. First, I used mRNA-Seq to attempt to determine the substrate of the orphan cytidine deaminase Apolipoprotein B mRNA-editing enzyme, catalytic polypeptide 2 (APOBEC2). Next, I used whole-genome bisulfite sequencing to explore the distribution of 5-methylcytosine in *Trypanosoma brucei*. In both of these cases, results were inconclusive but suggest future directions for investigation.

Chapter 2. Systematic characterization of the effects of AID on the B cell transcriptome and DNA methylome[†]

2.1 Assaying AID-dependent changes in the activated B cell

While AID-dependent effects have been reported in a variety of systems, I chose to investigate activated murine B cells for a number of reasons. First, B cells are the physiological setting for the highest levels of AID. I hypothesized that AID-dependent effects would be more apparent as well as more likely to be physiologically meaningful in the cell type with the highest levels of AID. Additionally, there were a number of practical concerns that made *ex vivo* stimulated B cells an attractive system. They are simple to derive in large quantities, with a single spleen supplying enough material for several types of experiments. Moreover, primary cells from the mouse that are cultured *ex vivo* closely approximate the physiological conditions of CSR, while providing far more uniformity than possible for any true *in vivo* system. Finally two useful mouse strains have already been generated: *Aicda*$^{-/-}$ (Muramatsu et al., 2000) and AID–miR-155T, which has a transgene containing a C-terminal AID-GFP-fusion and a mutation in the miR-155 target site, leading to overexpression of AID-GFP (Teng et al., 2008). In combination with the wild-type, these strains allowed investigation of dose-response relationships for any observed effects.

The investigation began with high throughput sequencing of poly-A$^+$ RNA, or mRNA-Seq. This technique was an attractive starting point for a number of reasons. First, it is high-throughput, with a single experiment yielding

[†] The work described in this chapter was published in (Fritz et al., 2013)

expression data for every gene in the genome. Additionally, mRNA-Seq can provide useful data beyond simple expression profiles, most saliently here evidence of RNA editing. Finally, because most possible effects of a nucleic acid editing protein like AID would be manifested either directly or indirectly on the transcriptome, it is likely that mRNA-Seq would allow for detection of any of AID-dependent DNA demethylation, RNA editing, or damage responses to off-target AID activity.

2.1.1 Validation of the activated B cell culture system

To ensure that *ex vivo* stimulation was occurring as desired and that each of the genotypes behaved as expected, I sought to validate the cell isolation and culture conditions. The criteria for stimulation conditions were as follows: (1) that the cells isolated were in fact naïve B cells and (2) that AID was efficiently induced. Both of these criteria were assessed by flow cytometry following negative selection of $CD43^+$ splenic cells and stimulation in culture for 3d with anti-CD40, LPS, and IL-4. This suite of factors, which induce switching to IgG1, IgG3, and IgE, were used because they are known to produce high levels of CSR.

Following stimulation, cells displayed forward and side scatter consistent with blasting B cells. Additionally, over 98% of cells displayed the surface marker B220, indicating that the dissection and cell purification were performed correctly (data not shown). In order to determine whether AID was efficiently induced, the cells were also stained with anti-IgG1 antibodies. Because CSR is absolutely dependent on AID, surface expression of secondary Ig isotypes such as IgG1 functions as a proxy for AID expression. Flow cytometry demonstrated the expected pattern of $IgG1^+$ cells by genotype, with background levels for

Aicda$^{-/-}$, 14% for WT, and 31% for AID–miR-155T (Figure 2.1). Because AID level is known to be positively correlated with rates of CSR, this pattern also demonstrates that AID is in fact overexpressed in AID–miR-155T cells, and that the AID-GFP fusion is CSR-competent.

Additionally, AID expression itself could be detected by flow cytometry in the case of the AID–miR-155T mouse due to its C-terminal GFP fusion. As expected, no GFP expression was detectable for the *Aicda*$^{-/-}$ or WT mice. In the case of the AID–miR-155T mouse, over 85% of cells were GFP^{+}, demonstrating that the culture conditions robustly induce AID expression. Because there are no alterations in the transgene contained by AID–miR-155T cells that should affect transcription, this figure is also a fair estimate of the fraction of WT cells that express AID under these conditions. For all subsequent B cell experiments (except for the noted exceptions), all material was derived from the same single B cell culture per genotype assayed in Figure 2.1.

2.1.2 Generation, mapping, and validation of mRNA-Seq data

RNA extracted from *Aicda*$^{-/-}$, WT, and AID–miR-155T activated B cells was used to prepare mRNA-Seq libraries using a standard protocol with the addition of two sets of "spike in" control transcripts after poly-A^{+} selection (Figure 2.2). The first was the commercially available ERCC panel of precisely quantified RNAs (Jiang et al., 2011a) which allow determination of the lower limit of detection for gene expression. The second was 5 sets of "pre-edited" RNAs derived from *Trypanosoma brucei* variant surface glycoprotein (VSG) genes with a single C-to-T change introduced at a frequency of 50%, which serve as a positive control for the detection of RNA editing.

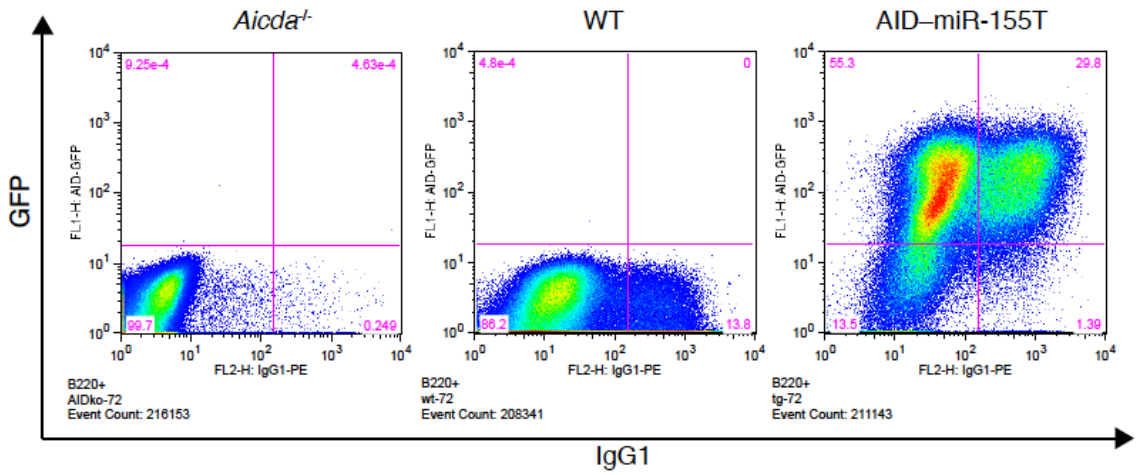


Figure 2.1. Quantification of $IgG1^+$ and GFP^+ populations of *ex vivo* stimulated naïve splenic B cells used for subsequent experiments as determined by flow cytometry. Plotted populations are $7AAD^-$, $B220^+$.

Following purification of poly-A^+ RNA and addition of exogenous spikes, the RNA-Seq library was prepared via modifications of a standard protocol (Rosenberg et al., 2011). Briefly, the RNA was chemically fragmented with Mg^{2+} and high heat, then reverse transcribed with random hexamer priming. The resulting cDNA was made double stranded with a cocktail of DNA polymerases, and following end repair and generation of A-overhangs, this cDNA was ligated to Illumina adaptors to facilitate sequencing. The adaptor-ligated product was size selected by gel electrophoresis, enriched by PCR, and size selected once more before 100-cycle, single-end sequencing on the Illumina HiSeq 2000.

After verifying that the resulting sequencing reads passed basic quality control measures, they were aligned to the reference genome in a splice-junction-conscious manner, and with the exogenous spike sequences added as extra "chromosomes." The resulting mapped reads were then compared to the Ensembl reference gene annotation (with certain alterations, as noted later) using the program Cufflinks in order to generate relative gene and isoform expression values. The resulting expression values have units of fragments per kilobase of transcript per million reads mapped, or FPKM. Because it is normalized to both length of transcript and number of reads mapped, FPKM provides a count-like unit that allows for comparison between samples for a given transcript, as well as between transcripts for a given sample.

To determine the limits of detection for this measurement of gene expression, the resulting values were first compared for the ERCC controls alone. The ERCC panel consists of 92 RNAs with no similarity to any sequences in the mouse genome, which allows for fully exogenous measurement of the limit of detection of abundance for transcripts (Jiang et al., 2011b). The resulting

estimates of abundance were well correlated for each pair of samples for transcripts with > 5 FPKM, and were within a 2-fold difference as low as 0.1 FPKM (Figure 2.2). Since an FPKM of 5 corresponds to the 16th percentile of expressed genes for this data set, this analysis demonstrates that the data presented should be sensitive to relatively subtle changes in most expressed genes and large changes even in poorly expressed transcripts.

2.1.3 AID-dependent differences in immunoglobulin isotype abundance

After verifying that the exogenous controls behaved as expected with respect to transcript quantification, the next validation step was to detect the expected difference between the samples sequenced: transcription of productive secondary Ig isoforms.

Conceptually, this was no different than detecting differences in isoform-level expression differences in any other gene, as all of the J_H-C spliced transcripts should be derived from productive transcripts. However, because the existing standard gene annotations do not include J_H segments, a standard gene- or isoform-expression analysis would give no information about the levels of CSR in the source cell population. I first found the correct coordinates for each Ig segment by mapping the sequences for each as listed in IMGT (Lefranc et al., 2009) to the genome. I then manually generated annotations for each of the 5 theoretically possible "isoforms" for each isotype (each of the germline promoter and $J_{H,1\text{-}4}$ spliced to C_1, followed by the remaining C exons) and used this set to replace the existing Ensembl IgH locus annotation. Quantifying expression at the isoform level and then summing the FPKM values for each of the J-containing IgH isoforms for each isotype then allowed for relative measures of CSR.

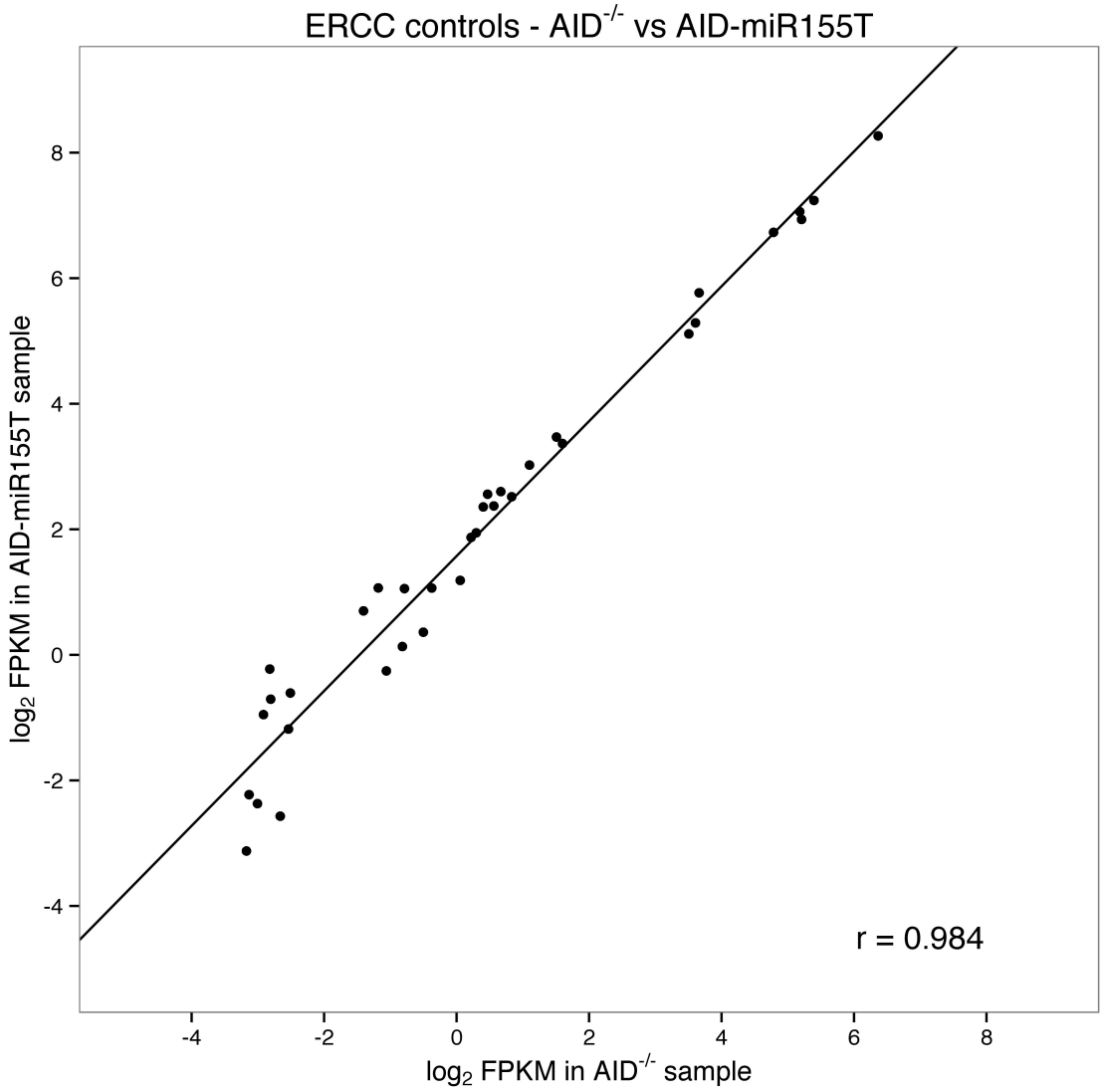


Figure 2.2 Quantification of ERCC controls for RNA-Seq. Pairwise comparisons of ERCC RNA spike levels for (a) *Aicda*$^{-/-}$ and AID–miR-155T samples (r = Pearson correlation coefficient)

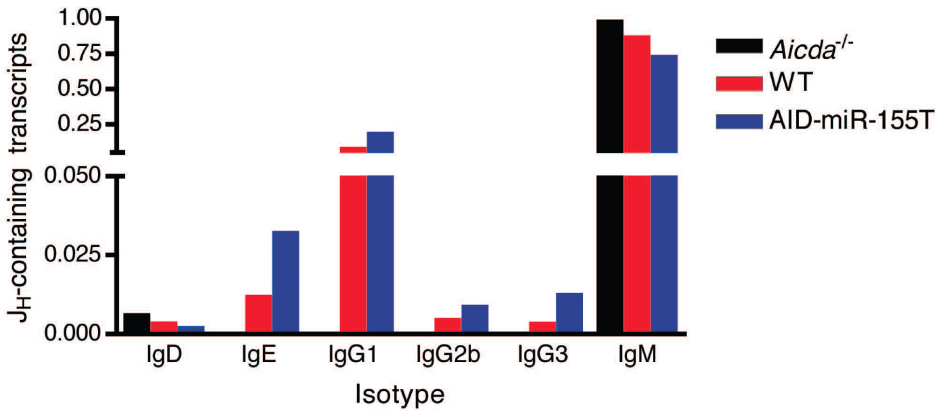
As expected, nearly all J_H-C spliced transcripts for the *Aicda*$^{-/-}$ sample were IgM or IgD, while the wild-type and AID–miR-155T samples both displayed appreciable amounts of IgE, IgG1, IgG2a, and IgG3-derived transcripts (Figure 2.3 A). Furthermore, apparent CSR frequencies were higher for AID–miR-155T than for wild-type for each secondary isotype, consistent with previous reports (Teng et al., 2008). This result demonstrates that the RNA-Seq data generated are of sufficient depth and quality to detect large differences in isoform abundance, and also indicates that the B cells used display the expected AID-dependent differences at the RNA level.

One caveat of measuring CSR by RNA-Seq is that it provides only a relative measurement. To determine how CSR frequencies quantified by RNA-Seq analysis compare to standard measurements, and thus to generate a fixed peg that would allow conversion to absolute values, the fraction of IgG1$^+$ cells as determined by flow cytometry were compared to the abundance of J_H-$C_{\gamma 1}$ transcripts as a fraction of all J_H-C transcripts (Figure 2.3 B). A clear linear relationship was observed, albeit with slope not equal to 1, which likely represents differing per-cell levels and/or sequencing efficiencies of different isotype transcripts. Determining these correction factors for each isotype should allow RNA-Seq to be used as quantitative tool for assaying absolute frequencies of CSR.

2.1.4 AID has little effect on non-immunoglobulin gene expression

With the sensitivity of the mRNA-Seq system thoroughly validated, effective comparisons of gene expression could be made between samples.

A.



B.

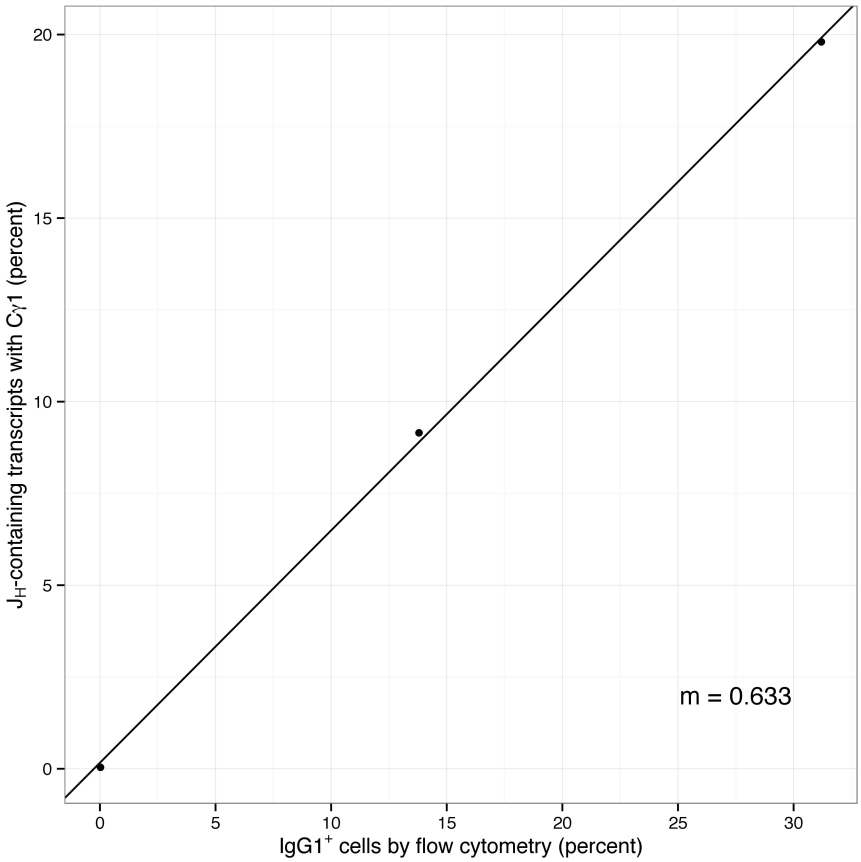


Figure 2.3. Quantification of immunoglobulin isotype abundance. (A) Relative Ig isotype abundance by AID level as calculated by RNA-Seq. (B) Comparison of CSR to IgG1 as calculated by RNA-Seq and flow cytometry for 3 analyzed B cell genotypes. (m = slope of line of best fit)

Overall, the gene expression profiles were similar for the three samples, with a Pearson correlation coefficient > 0.995 for each pairwise comparison (Figure 2.4). Within these expression profiles, AID itself was clearly expressed at the expected level for each dataset. But besides AID, the short list of genes that with adjusted p-value < 0.05, > 2 fold-change differences between conditions, and FPKM of at least 5 for at least 1 condition (Table 2.1) was composed of elements predominantly annotated as pseudogenes in other references, or elements with RNA-Seq coverage characteristic of mismapped reads derived from paralogous transcripts. In addition, few transcripts that displayed a > 2 fold difference did so for more than one of the binary comparisons, suggesting that these apparent differences were the result of noise rather than authentic AID-dependent effects. These findings are concordant with those from a lower-depth RNA-Seq comparison of *Aicda*$^{-/-}$ and wild-type under slightly different conditions (36 nt paired-end sequencing, and IL-4 plus anti-CD-40 stimulation, data not shown).

I also used the RNA-Seq data to estimate the abundance of different transcript isoforms. Expression analysis at the isoform level again shows a high degree of similarity between *Aicda*$^{-/-}$, wild-type and AID–miR-155T samples, with the previously discussed exception of IgH transcripts (Figure 2.5). The lower degree of correlation for the isoform-level comparison as compared to the gene-level is expected, due to the uncertainty inherent in assigning ambiguous reads to one of several isoforms.

Overall, the analysis failed to detect any clear difference in gene expression outside of the IgH locus. While these results do not exclude the

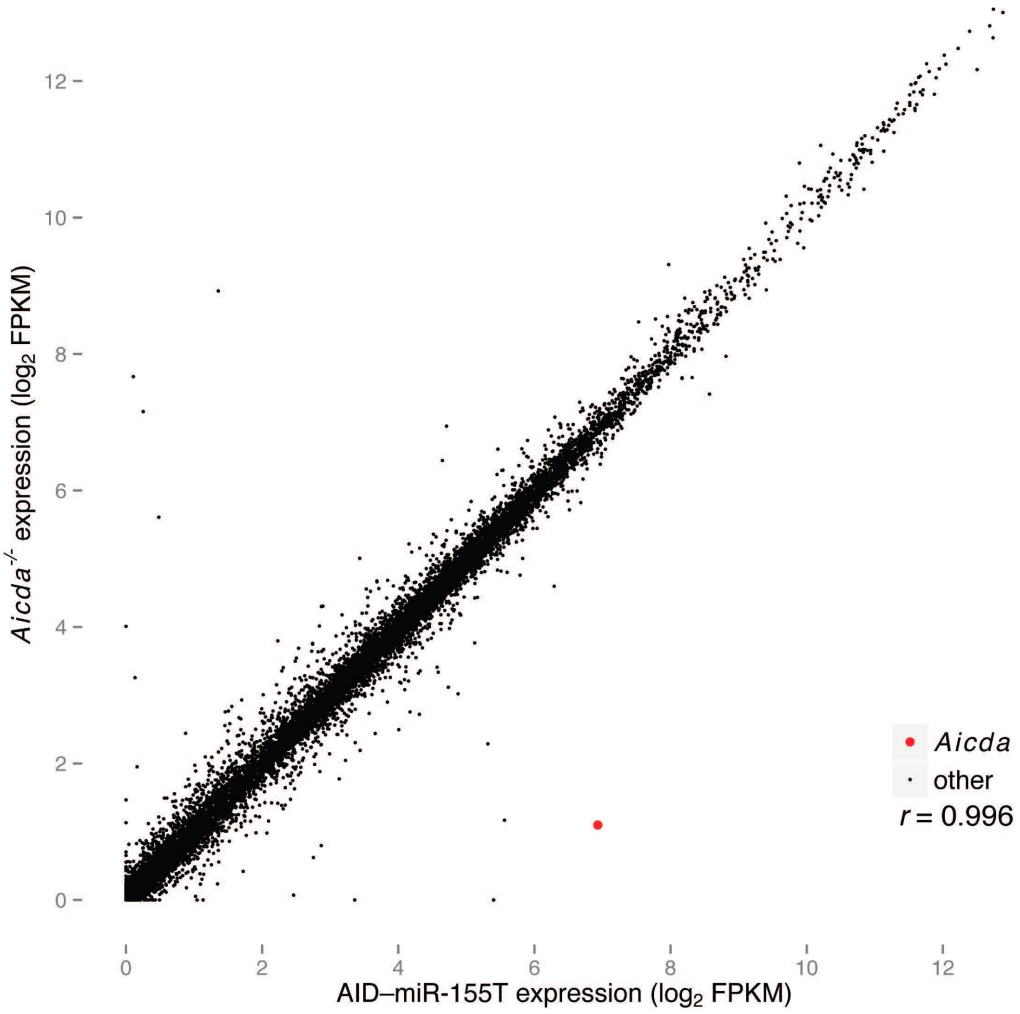


Figure 2.4 Gene expression comparison for *Aicda*$^{-/-}$ and AID–miR-155T B cells. (r = Pearson correlation coefficient)

Table 2.1. Genes with at least 2-fold difference in expression, FPKM of at least 5 for one condition, and adjusted p-value < 0.05 for *Aicda*$^{-/-}$ (KO) and AID–miR-155T (TG) samples, from RNA-Seq data as determined by Cuffdiff.

gene	KO_FPKM	TG_FPKM	log2(fold_change)	q_value
Aicda	1.14	120.99	6.73016	0.00E+00
Rpl7a-ps5	1.25	46.18	5.2117	0.00E+00
Mfap5	0.54	5.74	3.4151	4.20E-09
Gm14431	0.74	6.30	3.09394	3.71E-02
Gm4245	0.74	6.30	3.09394	3.71E-02
Havcr1	4.64	15.08	1.7006	3.91E-08
Sgip1	5.75	17.04	1.5663	6.11E-09
Cd9	3.57	9.84	1.46525	2.23E-03
Gm3839	12.60	33.83	1.42458	3.89E-06
Gm10327	3.13	8.35	1.41717	5.53E-03
Gm10293	9.64	22.76	1.24041	3.14E-04
Gm10709	169.32	379.70	1.16513	8.12E-12
Ccr4	7.46	16.09	1.10807	8.60E-05
Gnb4	8.24	16.75	1.0237	3.17E-03
Pld4	17.09	8.03	-1.09064	4.12E-04
Fcrl5	7.90	3.63	-1.12079	1.85E-02
Oas1g	12.54	5.68	-1.14153	2.38E-02
Ifitm3	96.35	43.11	-1.16026	1.63E-05
BC094916	10.29	4.60	-1.16053	1.45E-02
Oas2	9.24	3.89	-1.24865	1.07E-02
Gm5431	5.29	2.22	-1.25388	1.45E-02
Pydc3	9.58	3.96	-1.27535	3.76E-04
Sla	15.16	6.17	-1.29709	1.04E-04
Ccl5	634.25	250.14	-1.34233	0.00E+00
Ifit3	18.77	6.42	-1.54769	8.43E-08
AI607873	5.81	1.84	-1.66161	3.34E-04
Serpinb1a	6.14	1.94	-1.66476	5.73E-05
Pydc4	31.13	9.80	-1.66687	4.28E-12
Plac8	85.76	24.07	-1.83303	1.13E-10
Hist1h2af	122.00	25.14	-2.27855	0.00E+00
C530028O21Rik	8.56	0.10	-6.4739	1.88E-03
Alox5ap	47.78	0.40	-6.91307	1.95E-10
Gm9493	484.50	1.56	-8.28029	0.00E+00
Gm2606	202.40	0.08	-11.3459	0.00E+00

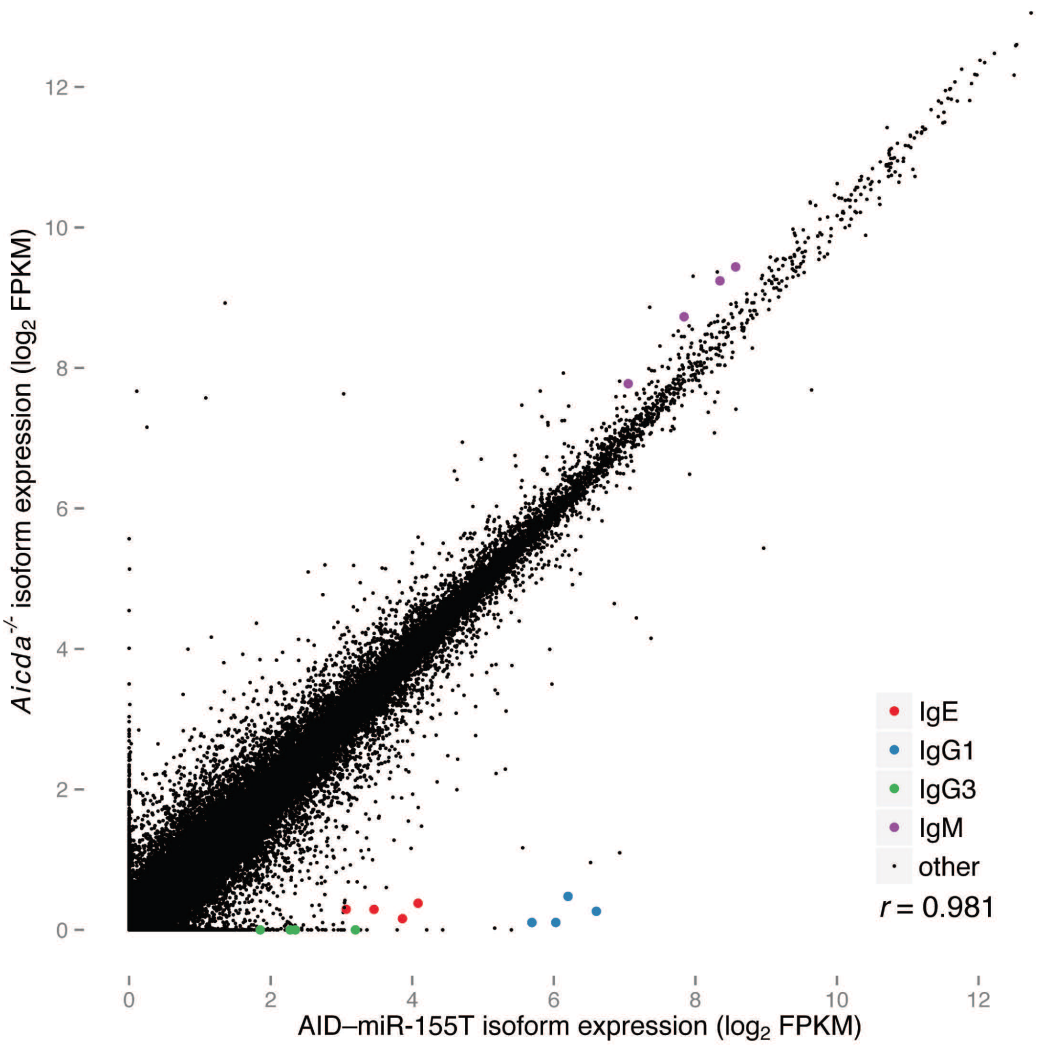


Figure 2.5 Isoform expression comparison for *Aicda*$^{-/-}$ and AID–miR-155T B cells. (r = Pearson correlation coefficient)

possibility of AID-dependent changes in gene expression in B cells, they demonstrate that if such changes exist outside the *Igh* locus, they are likely too small to be physiologically relevant. While these findings are derived from analysis of a single sample in each case, a lower-depth RNA-Seq comparison of *Aicda*$^{-/-}$ and wild-type under slightly different conditions (36 nt paired-end sequencing, and IL-4 plus anti-CD-40 stimulation) also found high concordance between samples by AID level.

2.1.5 AID has no effect on V_H segment usage in naïve B cells

AID deficiency has been associated with autoimmune disease and a skewed usage pattern of V_H segments in both humans (Meyers et al., 2011) and mice (Kuraoka et al., 2011). Because our data were derived from naïve B cells non-specifically stimulated to undergo CSR *ex vivo*, any differences in V_H gene usage between samples ought to mirror the *in vivo* repertoire prior to affinity maturation. Thus the mRNA-Seq data set generated should allow for determination of whether this previously observed effect occurred prior to affinity maturation.

FPKM values for the entire set of V_H segments were derived by using a separate annotation consisting of only Ig V segments, defined in relation to the IMGT sequences as previously described for J segments. Because only one V_H is transcribed in a mature B cell, the measurements correspond to relative segment usage between samples. This transcript abundance analysis revealed only minor differences in V_H transcript abundance between *Aicda*$^{-/-}$, wild-type and AID–miR-155T samples. Importantly, none of these exhibited a clear relationship with AID expression (Figure 2.6), strongly suggesting that the pattern of V_H usage was

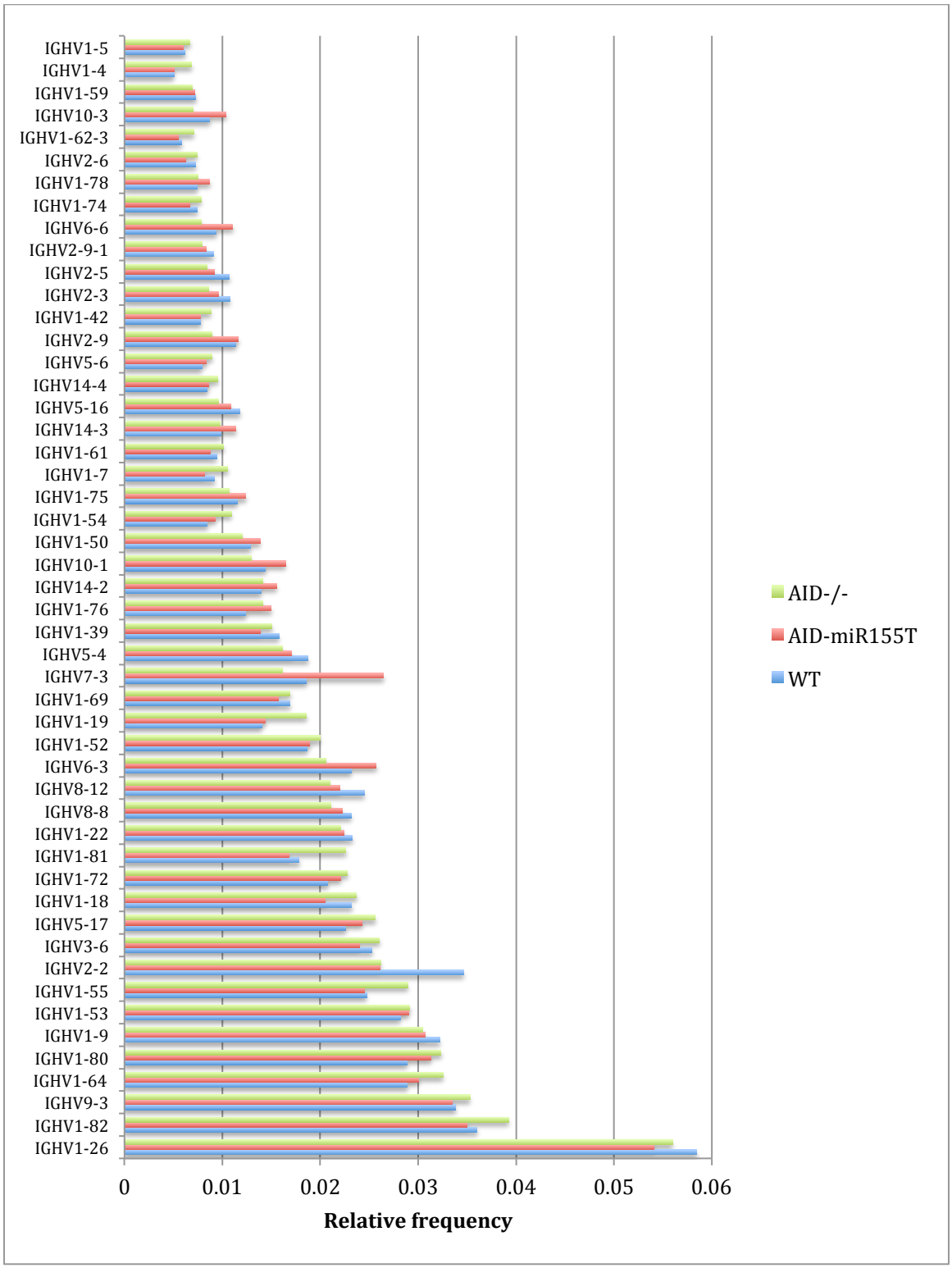


Figure 2.6. Relative frequencies of V_H segment usage by AID level, as calculated by mRNA-Seq. Values shown are the top 50 V_H segments for the WT sample.

unaffected by AID. This result is in contrast to the pattern of V_H usage in newly emigrant B cells in AID-deficient humans (Meyers et al., 2011; Yamane et al., 2010), suggesting that the influence of AID over V_H usage occurs after establishment of the primary repertoire and probably reflects the dynamics between AID-mediated affinity maturation and B cell survival, rather than a role in early B cell development.

2.2 Assaying AID-dependent mRNA editing

While it has become clear that AID's roles in initiating CSR and SHM proceed via a DNA-editing mechanism (Di Noia and Neuberger, 2007), the question of whether AID can edit RNA in a cellular context remains open. Indeed, a number of reports have hypothesized AID-dependent RNA editing activity in B cells (Kobayashi et al., 2009; Muramatsu et al., 2000; Nonaka et al., 2009). The RNA-Seq reads described in the previous sections were precisely the dataset required for answering this question in a rigorous manner.

2.2.1 Refinement and validation of a comparative mRNA-Seq RNA editing-detection pipeline

While RNA-Seq is often treated as simply a method for gene expression profiling, the sequences it generates contain far more information than just RNA abundance. In particular, comparing mapped reads to reference to identify recurrent mismatches is an effective strategy for identifying RNA editing events, as has been demonstrated in a number of reports (Bazak et al., 2014; Eisenberg et al., 2010; Ramaswami et al., 2013). Previous work in the laboratory demonstrated the particular utility of paired wild-type and specific deaminase-deficient

comparative RNA-Seq for identifying RNA editing events (Rosenberg et al., 2011). By using the sequences from near-congenic, deaminase-deficient samples as a final filter, false positives arising from mismapping and genomic differences can be minimized.

The pipeline applied to the sequencing data is described in Figure 2.7. Briefly, single nucleotide variants (SNVs) were considered candidate editing sites if they conformed to the following criteria: (1) had greater than 30x read coverage, (2) at least 20% apparent C-to-T editing, (3) were a minimum distance from a non-C-to-T SNV (1 kb if using reference, 10 kb if not), (4) were not significantly strand-biased, as determined by Fisher's exact test, (5) were not located in regions that were not isogenic between the mice used, as determined by the frequency of non-C-to-T SNVs, and (6) did not occur in the *Aicda*$^{-/-}$ sample. These cutoff values were determined to minimize the false positive rate for a comparison of RNA-Seq data from *Apobec1*$^{-/-}$ and wild-type macrophages. Two parallel analyses were performed for positions within reference exons (to achieve the lowest possible background) and for all positions (to include positions in transcripts not found in the reference annotation).

2.2.2 Validation of the RNA editing detection pipeline

In order to determine the limits of detection of this pipeline, three analyses were undertaken. First, the fraction of the transcriptome covered under the parameters described was quantified. Second, apparent editing was analyzed for the "pre-edited" RNA spikes. Finally, the pipeline was applied to analogous RNA-Seq data for the bona fide RNA editor APOBEC1 (C. Hamilton and F.N. Papavasiliou, unpublished data).

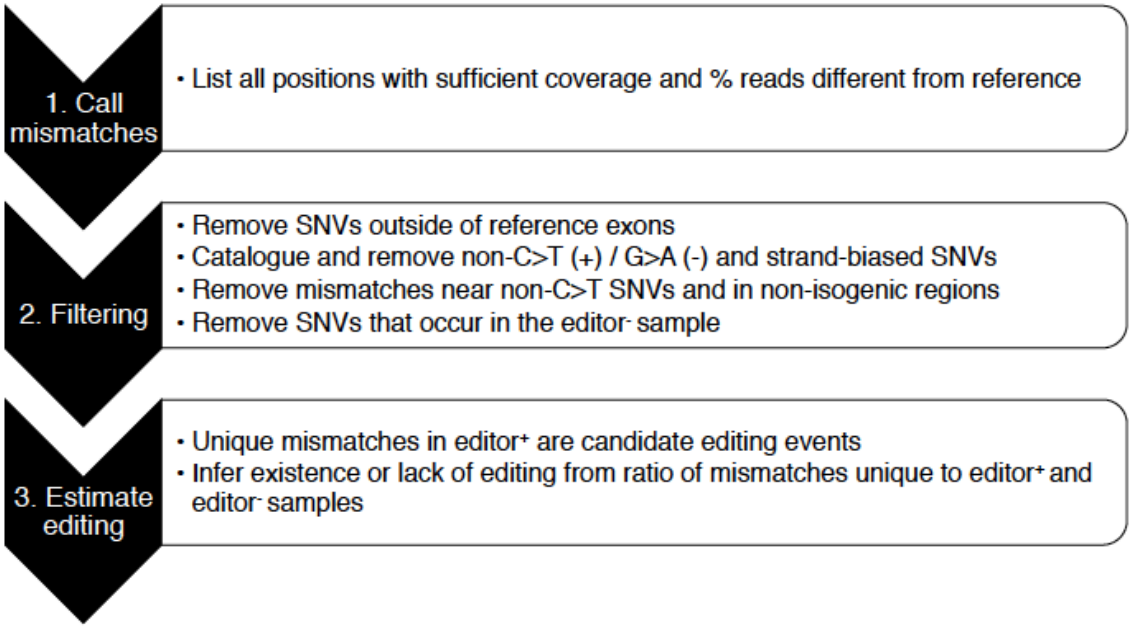


Figure 2.7. Schematic of RNA editing detection pipeline.

To estimate the fraction of the transcriptome covered at least 30x and thus visible to this pipeline, the amount of coverage by base was computed for ERCC spikes with various FPKM values (Figure 2.8). More than 90% of bases of the transcripts with FPKM ≥ 10 had at least 30x coverage, while those with FPKM = 8 had about 60% of bases with sufficient depth to be interrogated for editing by this strategy. For the B cell RNA-Seq datasets, roughly the top 6600 transcripts have FPKM ≥ 10 and are thus well covered by the RNA editing analysis. Thus this combination of analysis strategy and dataset is appropriate for identifying editing events in moderately to highly expressed transcripts.

Next, editing was determined for the "pre-edited" controls added during preparation of the library. These RNAs were derived from 5 *Typanosoma brucei* variant surface glycoprotein (VSG) genes, which had a single C-to-T mutation through site-directed mutagenesis. For each VSG the wild-type and "pre-edited" varieties were transcribed *in vitro*, purified, quantified, and then mixed in a 1:1 ratio, resulting in a population of transcripts approximating a 50% editing ratio at a single site. These pairs were added at a range of concentrations to the B cell RNA following poly-A^+ selection.

Application of the described pipeline allowed detection of the editing event in the exogenous "pre-edited" VSG RNA with FPKM = 23, but not for the VSG with FPKM = 7. Therefore this more stringent measure of the limit of detection for editing yields a similar answer to depth of coverage alone. Both show that, for this pipeline and this depth of sequencing, medium- to highly-expressed transcripts are thoroughly interrogated for editing. However, editing events in poorly expressed transcripts are below the limit of detection.

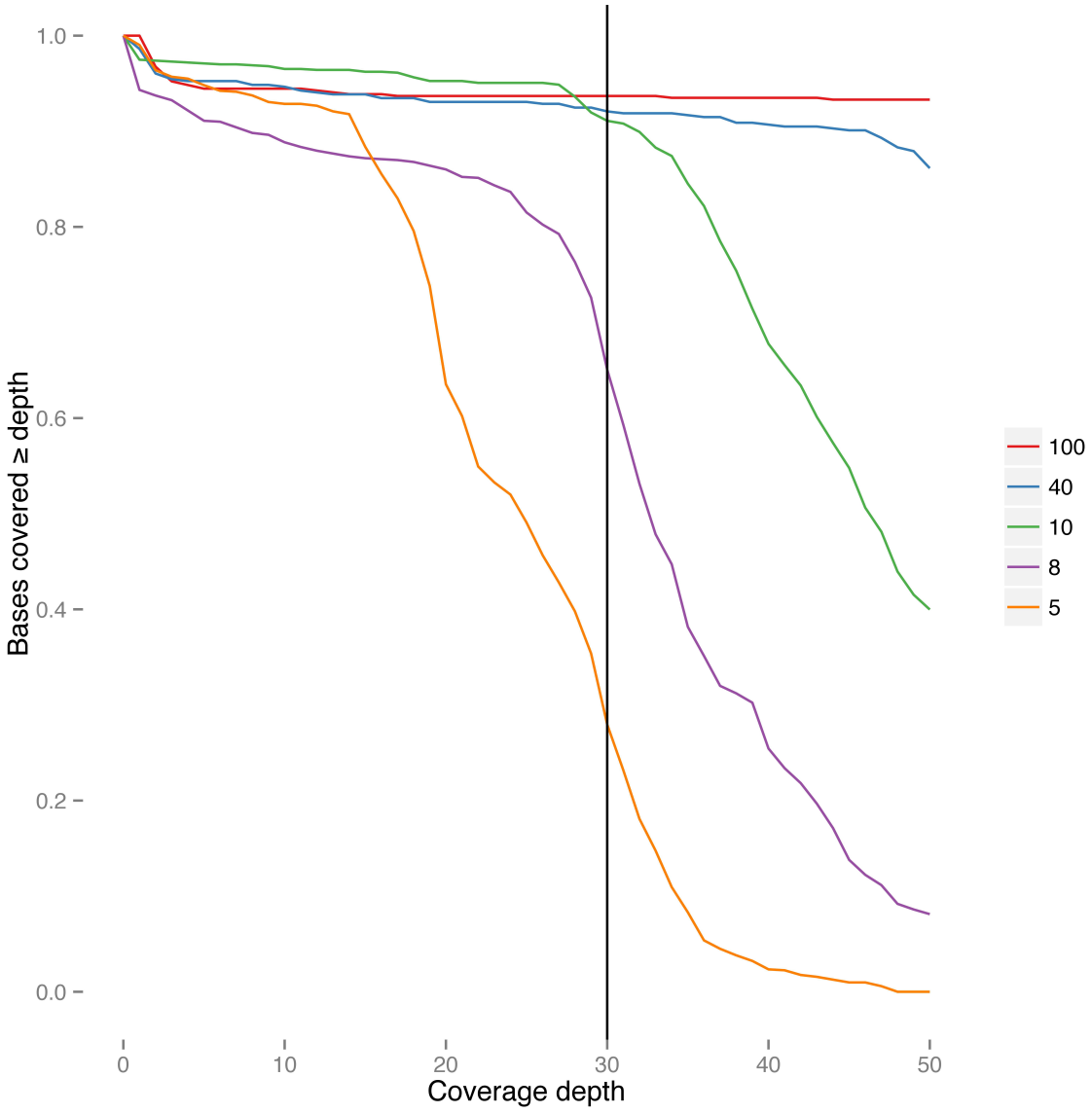


Figure 2.8. Cumulative per-base coverage depth for ERCC spike transcripts with various expression values (FPKM), illustrating the portion of the transcriptome accessible to the RNA editing analysis workflow for the AID-overexpressing 100bp mRNA-Seq.

As a final validation, the pipeline was applied to RNA-Seq data for wild-type and *Apobec1*$^{-/-}$ macrophages that had been generated for other purposes (C. Hamilton, unpublished data). A large number (> 40) of APOBEC1-dependent RNA editing events were identified by this method (Figure 2.9). Importantly, the reciprocal comparison was also performed; that is, the number of apparent editing events that were present in the *Apobec1*$^{-/-}$ sample but not in the wild-type. This count provides a measurement of the background noise of the technique (due to mismapping or genomic sequence differences). It also allows for computation of an implied false positive rate (IFPR), as # of events in the deaminase-deficient sample / # of events in the wild-type. For APOBEC1-dependent editing in macrophages, the IFPR was 7-8%. This demonstrates that the pipeline is highly specific for detecting true editing events in mRNA, at least if the events have APOBEC1-like properties.

2.2.3 No AID-dependent RNA editing events can be detected in mRNA

With the pipeline thoroughly validated, it was then applied to the B cell RNA-Seq data discussed in section 2.1. This analysis revealed less than 10 candidate editing sites in the AID–miR-155T and wild-type samples that were absent in the *Aicda*$^{-/-}$ sample (Figure 2.9). Approximately equal numbers of candidate sites were found in the reciprocal comparison, resulting in IFPRs of > 75% for each condition. In contrast to APOBEC1, this strongly suggests that AID does not edit a large number of RNAs. Because the data presented is derived from a single sample for each genotype, estimates of the variance in apparent candidate editing events was not possible. However, the small number of such

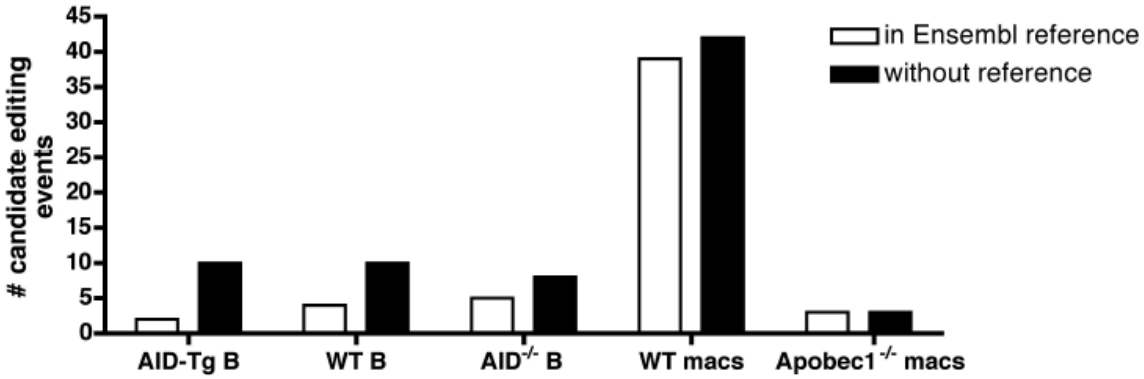


Figure 2.9. Candidate editing event counts derived from samples from AID overexpressing, wild-type, and *Aicda*$^{-/-}$, B cells, and wild-type and *Apobec1*$^{-/-}$ macrophages, for analysis pipelines incorporating the Ensembl reference gene model and using no outside annotation.

events enabled manual inspection of these candidates to determine with greater certainty whether AID-dependent RNA editing was occurring in these samples.

To determine if there was a highly specific AID-dependent RNA editing event that was present in the sequencing data, reads for the few candidate AID-dependent editing events were visually inspected. In each case these events were adjudged to be false positives because of one of the following criteria: near-threshold distance from non-C-to-T SNVs, C-to-T mismatches also occurring in the *Aicda*$^{-/-}$ sample, or complete absence of apparent editing in the AID–miR-155T sample. Thus, the method fails to detect AID-dependent editing of moderately or highly expressed polyadenylated RNAs. While these results do not exclude the possibility of AID-dependent editing of low-expressed or non-polyadenylated RNAs, it is highly unlikely that editing of a highly expressed protein-coding transcripts takes place in B cells.

2.3 Assaying AID-dependent changes in DNA methylation

Although the mRNA-Seq analyses suggest that AID does not have large effects in B cells beyond its known roles in SHM and CSR, they did not exclude the possibility of AID-dependent DNA methylation occurring in a way that does not dramatically alter gene expression. Because AID-dependent changes in DNA methylation has been an area of intense interest (Bhutani et al., 2013; 2010; Cortellino et al., 2011; Kumar et al., 2013; Popp et al., 2010; Rai et al., 2008; 2010; Sabag et al., 2014), I decided to compare the methylomes of *Aicda*$^{-/-}$, wild-type and AID–miR-155T B cells to determine whether AID demethylates DNA in this cell type.

2.3.1 Generation of genome-scale methylation data by the reduced-representation bisulfite sequencing (RRBS) method

A number of techniques exist for generating genome-scale DNA methylation data (Harris et al., 2010). The most comprehensive of these is whole-genome bisulfite sequencing (WGBS), in which genomic DNA undergoes bisulfite conversion (converting all cytosines to thymines, but leaving 5-methylcytosines unchanged). The converted DNA is then sequenced and mapped to a reference genome. This allows the fraction of cytosines methylated at a given site to be calculated as the ratio of reads containing C to total number of reads covering that site. The advantages of this technique are its single-base resolution, direct and quantitative readout of methylation, and near-complete genomic coverage. However this superior coverage comes at a significant cost: because high depth is required for methylation ratios to be quantitative, it requires very deep sequencing per sample. For example, in a recent report the equivalent of 5 Illumina Hi-Seq lanes was required to achieve just 13x average coverage for a human sample (Heyn et al., 2012).

To retain the advantages of bisulfite sequencing without the issues that come along with obtaining its extensive coverage, I chose to use reduced-representation bisulfite sequencing (RRBS) to generate genome-scale methylation data (Meissner et al., 2008). This technique differs from WGBS by including digestion with the restriction enzyme MspI (which has restriction site C^CGG, and is insensitive to methylation) and size selection prior to bisulfite conversion. This strategy eliminates the completely uninformative reads that are frequent in WGBS, because each read must begin with a CpG dinucleotide. Because CpGs are very non-uniformly distributed in mammalian genomes (Illingworth and

Bird, 2009), using MspI sites as the ends of each sequencing insert has the added advantage of focusing sequencing on the CpG-rich, genic portion of the genome, as well as yielding more insert-internal CpGs, and thus more informative sequencing. The drawback of this focused coverage is that large CpG-poor portions of the genome will be entirely excluded by this technique.

2.3.2 Validation and coverage analysis of RRBS data

Genomic DNA from the same three B cell samples described in section 2.1.1 was used to prepare RRBS libraries by a standard protocol (Gu et al., 2011a). Two lanes of multiplexed 50-cycle sequencing yielded 47-54 million reads per sample. Following removal of 3' adapter sequence, the reads were mapped using the program Bismark (Krueger and Andrews, 2011). This program accounts for the complications of the non-complementarity of the two strands of the genome following bisulfite conversion and the expected incomplete conversion due to the presence of 5mC by first temporarily converting all C's to T's in the read and mapping against a C-to-T converted version of the genome, then doing the same for the G-to-A conversion, and finally reporting the best single result for each read. This technique resulted in 68-70% mapping efficiency for trimmed reads.

Because 5mC is extremely rare in non-CpG contexts in B cells (Ziller et al., 2011), the apparent level of methylation at these sites can be used to approximate the background error of RRBS-derived methylation measurements, which are due to incomplete bisulfite conversion and mismapping. All 3 samples displayed apparent overall methylation levels of < 1.5% for non-CpG sites, demonstrating that non-conversion error is not a factor in interpretation of this dataset.

In order to determine the extent of genomic coverage in the dataset, the number of informative basecalls was quantified for a number of types of features. This analysis demonstrated the excellent coverage of CpG islands (85% with ≥ 100 individual CpG measurements) and promoters (64%) (Figure 2.10). This coverage was achieved despite low overall genomic coverage (2% of 1 kb windows with ≥ 100 individual CpG measurements), demonstrating the efficiency of targeting achieved by RRBS. To be included in subsequent analyses it was required that a CpG be covered at least 10x in all 3 samples, which yielded nearly 950,000 sites.

2.3.3 RRBS fails to detect AID-dependent differences in DNA methylation

To determine whether AID has a gross effect on the B cell DNA methylome, the distribution of methylation frequency for various genomic features was compared for the three AID genotypes. For each genomic feature type analyzed (1 kb windows, individual CpGs, CpG islands and promoters), there was no apparent difference in DNA methylation distributions associated with AID expression (Figure 2.11). The mean methylation frequency for each set of features was highly similar for each genotype. For each feature type, the expected bimodal distribution of methylation was observed, with proportionally more 1 kb windows near-fully methylated than the other feature types.

To assess more subtle differences in DNA methylation by AID expression, methylation frequencies of individual features for each pair of samples were compared. For each feature set analyzed, methylation frequencies between samples were very strongly correlated (Figure 2.12). For 1 kb windows, a Pearson $r = 0.997\text{-}0.998$ was observed for each pair; in comparison, the maximum

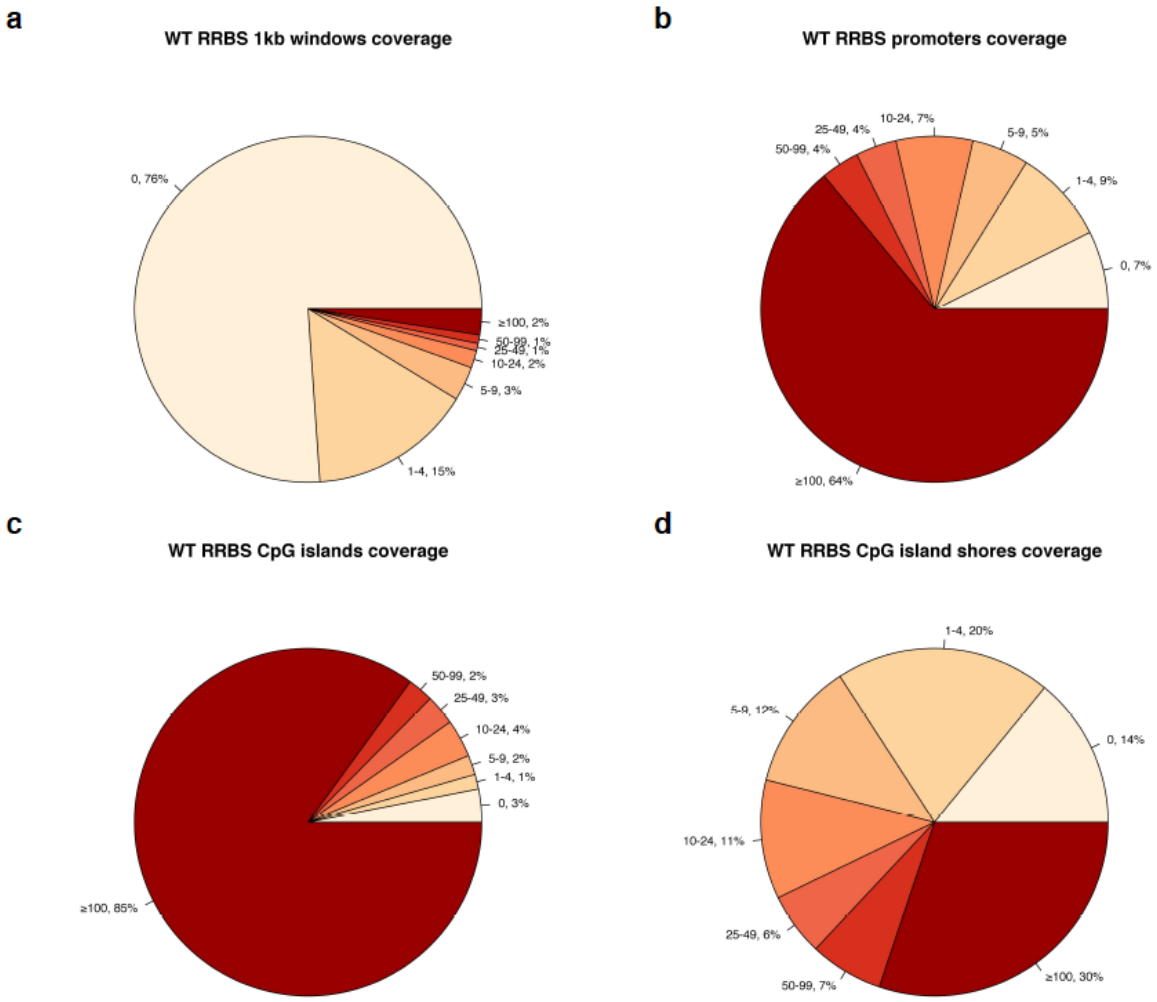


Figure 2.10. RRBS coverage of (a) 1kb genomic windows, (b) gene promoters, (c) CpG islands, and (d) CpG island shores for the WT sample. Promoters were defined as -5kb to +1kb from the TSS in Ensembl annotation, CpG islands were taken from the cpgIslandExt track of the UCSC table browser, and island shores were defined as 2kb up or downstream of a CpG island.

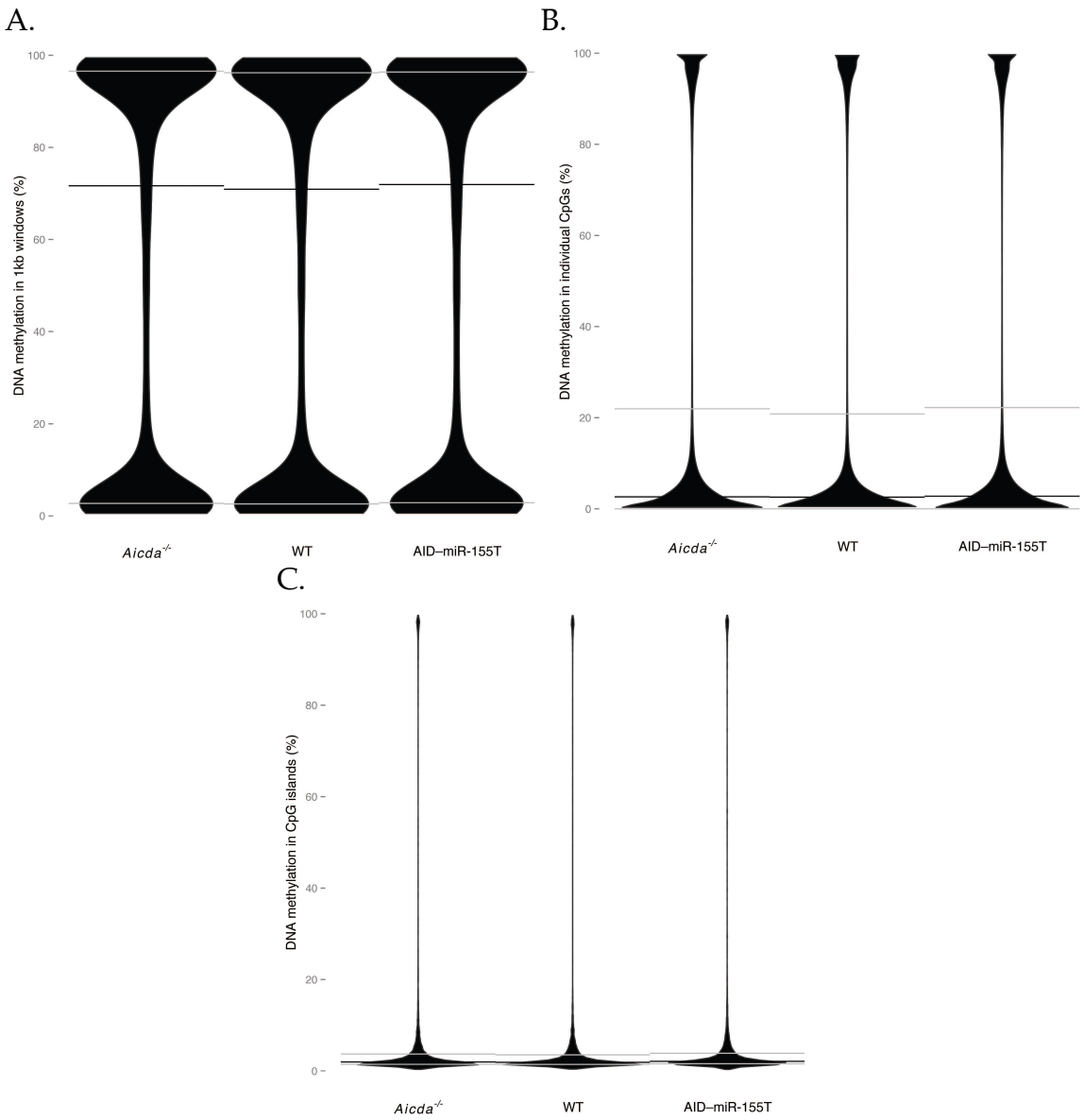
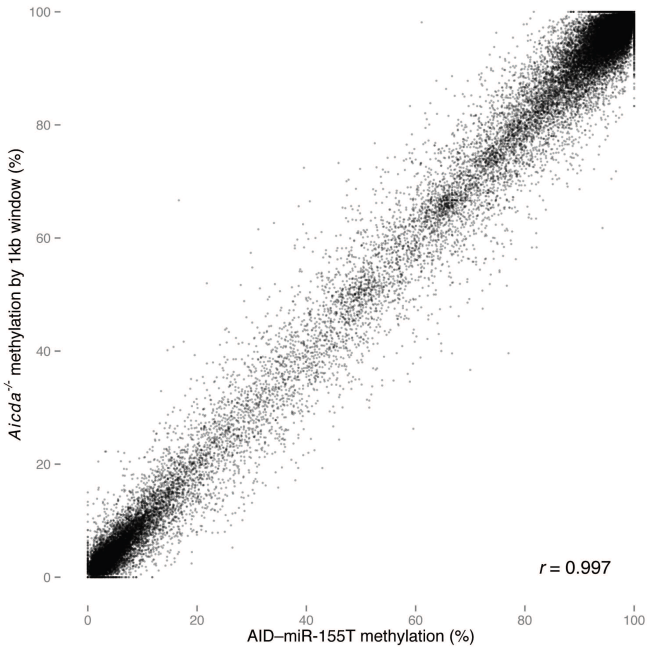


Figure 2.11. Distribution of DNA methylation frequencies in activated B cells by AID expression as determined by RRBS for (**a**) 1 kb windows, (**b**) individual CpGs, and (**c**) CpG islands. Width along *x*-axis denotes relative frequency of features with given level of methylation. Black horizontal line is sample median; gray horizontal lines are first and third quartiles.

A.



B.

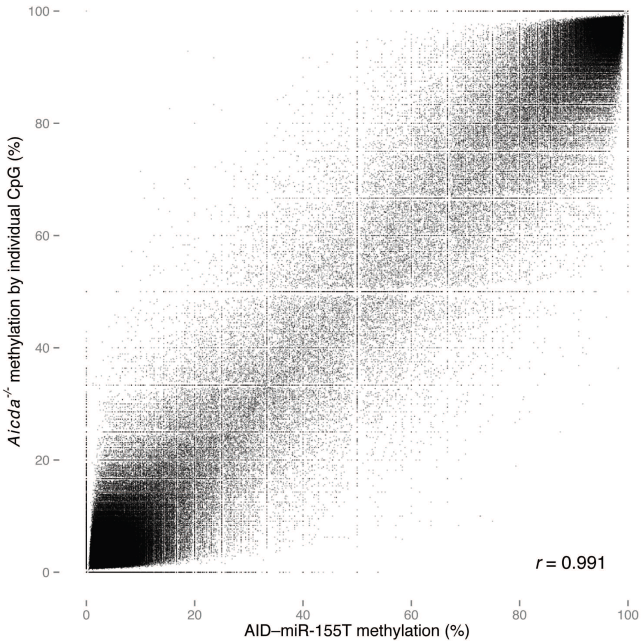


Figure 2.12. Pairwise comparisons of methylation frequency in AID-overexpressing and *Aicda*$^{-/-}$ B cells for (**a**) 1 kb windows and (**b**) all CpGs, as determined by RRBS. (*r* = Pearson's correlation coefficient)

reported *r* for methylation values of 1 kb windows for biological replicates of cells in the hematopoietic lineage is 0.997 (Bock et al., 2012). This suggests that any overall changes in DNA methylation attributable to AID are much smaller in magnitude than can be detected by this technique.

2.3.4 Attempted validation of AID-dependent differentially methylated regions

The high degree of correlation of methylation values between samples did not exclude the possibility that sampling noise could mask small numbers of true AID-dependent changes in methylation. To determine if this was the case, I sought to independently verify the largest apparent AID-dependent decreases in DNA methylation identified by RRBS. To this end, the Sequenom Epityper system was used to assay the most promising candidate regions for AID-dependent demethylation. This technique involves bisulfite conversion of DNA followed by PCR amplification, in vitro transcription, base-specific RNA cleavage, and mass spectrometry of the resulting product. The relative masses of the resulting spectral products can be used to determine absolute levels of methylation at each cytosine in the original genomic DNA. Because it is targeted to a known genomic region and samples large numbers of DNA molecules in one measurement, it reduces the error resulting from mapping and sampling noise that are inherent in RRBS.

The regions selected for analysis were those that had > 20% higher methylation in the *Aicda*$^{-/-}$ sample than in the wild-type as measured by RRBS and that also were easily accessible in terms of primer design. Randomly selected regions with < 10% difference in methylation between samples were also assayed

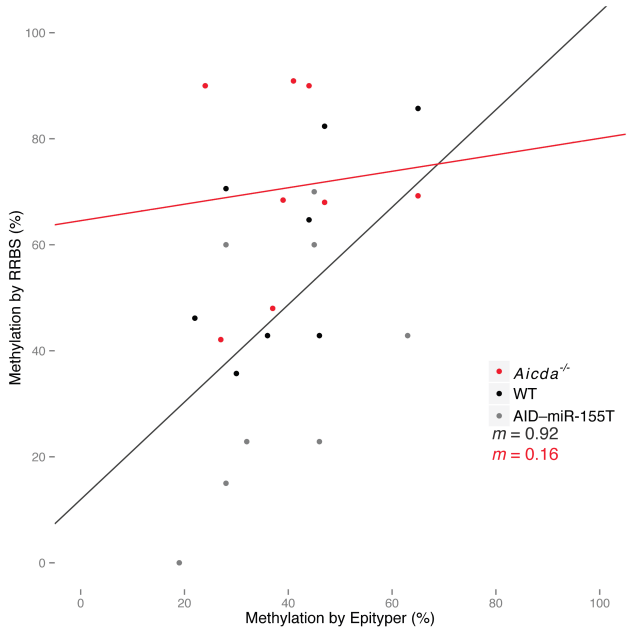
to validate the technique. High quality data for all 3 genotypes was obtained for 8 apparently AID-dependent hypermethylated CpGs and 18 similarly methylated CpGs that were also well covered by RRBS.

For the candidate differentially methylated CpGs, the methylation frequencies for the *Aicda*$^{-/-}$ sample as determined by RRBS were generally much higher than as determined by Epityper (Figure 2.13 A). In contrast, these methods gave similar values for the wild-type and AID–miR-155T samples. The fact that the apparent AID-dependent hypermethylated CpGs are not reproducible and yield uniformly lower methylation values when assayed by an independent technique strongly suggests that the RRBS-derived values for these CpGs represent overestimates of the true population mean methylation frequency, due to noise or an artifact of the method. Additionally, the set of CpGs with similar methylation values between samples by RRBS displayed excellent agreement in methylation frequencies as determined by the two methods (Figure 2.13 B), demonstrating that RRBS as performed here yields accurate methylation values for well-covered CpGs. Taken together, these results suggest that the most extremely hypermethylated CpGs in the *Aicda*$^{-/-}$ sample were not a result of an authentic AID-dependent process.

2.3.5 AID-dependent changes in DNA methylation and mRNA abundance do not suggest function

Finally, in an attempt to locate any subtle but biologically meaningful AID-dependent changes in DNA methylation, the fold-changes in gene expression by RNA-Seq were compared to the changes in promoter methylation

A.



B.

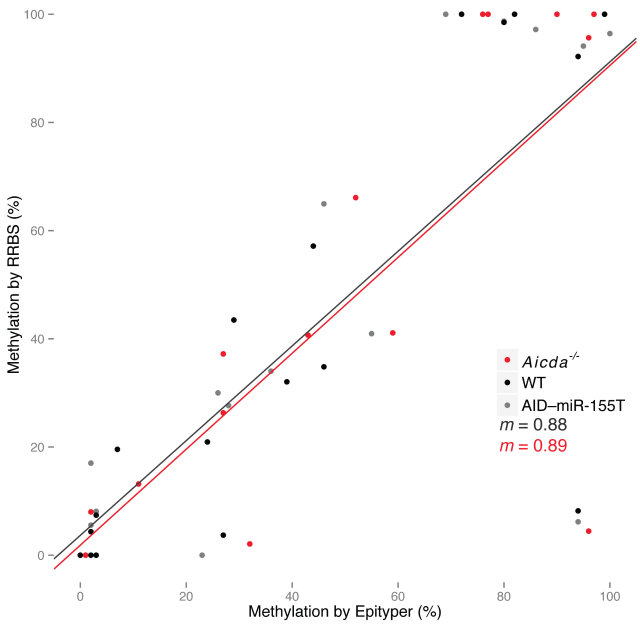


Figure 2.13 Comparison of DNA methylation frequencies as determined by RRBS and Epityper for a random subset of CpGs with (A) > 20% greater methylation in *Aicda*$^{-/-}$ than in WT and (b) CpGs with < 10% difference between *Aicda*$^{-/-}$ and WT. Lines are best linear fit for *Aicda*$^{-/-}$ or pooled WT and AID–miR-155T data. (m = slope of line of best fit)

for each pair of samples (Figure 2.14). Each pairwise comparison between genotypes showed that these variables were uncorrelated ($|r| < 0.03$ in all cases), suggesting that the observed modest differences in DNA methylation were not associated with changes in gene expression and therefore were unlikely to be physiologically relevant. As a whole, these results demonstrate that no candidate for consistent AID-dependent loss of methylation in B cells can be identified by RRBS.

2.4 Assaying AID-dependent changes in miRNA abundance and sequence

Although it appears that none of mRNA expression levels, RNA editing, or DNA methylation are affected by AID outside of the Ig loci, there of course remain other classes of nucleic acids that the previously described analyses do not address, such as miRNAs. Because AID has been implicated in the mutation of miRNA genes (Robbiani et al., 2009) as well as the editing of miRNAs themselves (Kobayashi et al., 2009; 2011), I sought to characterize the effects of AID deficiency on the miRNAome in terms of both abundance and sequence.

2.4.1 Generation and mapping of miRNA-Seq data

In order to minimize the differences in genomic sequence between the samples used and maximize the AID levels in the B cell, a retroviral complementation strategy was used to generate material for miRNA-Seq. *Aicda*$^{-/-}$ B cells were stimulated in culture with LPS and IL-4 and infected with retroviruses to induce stable overexpression of either AID and GFP, or GFP alone. At 24h and 48h after infection, GFP^{+} cells were sorted by flow cytometry to yield pure populations of infected cells. These flow cytometric measurements

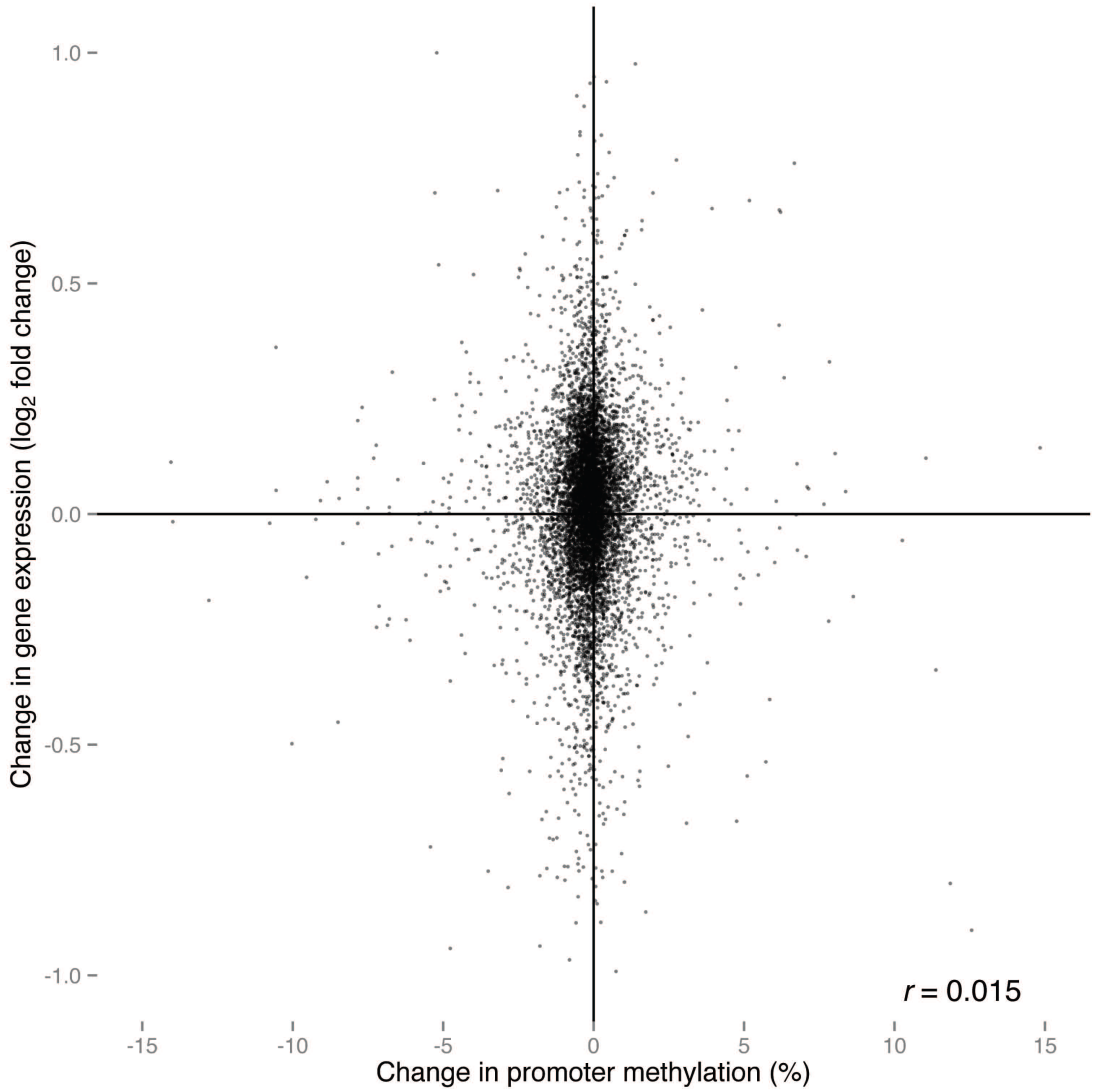


Figure 2.14 Comparison of differences in gene expression and methylation in the associated promoters for AID–miR-155T and *Aicda*$^{-/-}$ B cells. (r = Pearson's correlation coefficient)

indicated that GFP was induced in both populations, and that switching to IgG1 occurred only in the AID-overexpressing sample.

After total RNA was extracted from the sorted B cells, miRNA-Seq libraries were prepared according to a standard protocol (Hafner et al., 2012). First, an RNA linker containing a sample-specific barcode was ligated to the 3′ end of all RNAs. This was followed by PAGE size selection to yield only miRNA-sized products. Next, another RNA linker was ligated to the 5′ end, ensuring that only 5′-phosphorylated small RNAs would be sequenced. These species were then reverse transcribed, amplified by PCR, and sequenced on an Illumina HiSeq 2000. The resulting sequencing data was first trimmed to remove adapter sequence and then mapped against the genome, allowing unique mappings only.

2.4.2 AID has little effect on miRNA abundance

Reads overlapping miRNA sequences as annotated in miRBase were counted and normalized to million reads mapped for each of the four samples. These counts were compared for each pair of samples. For both time points, miRNA abundance was well correlated between the samples, with Pearson correlation coefficients exceeding 0.95 in every case (Figure 2.15). While this does not exclude the possibility that AID may have some subtle effect on miRNA abundance, it suggests that any such effect is unlikely to be large enough to be physiologically significant.

2.4.3 No AID-dependent RNA editing events can be detected in miRNA

To determine whether AID acts as an editor of miRNAs, an RNA editing detection strategy similar to that used for mRNAs was employed. Briefly,

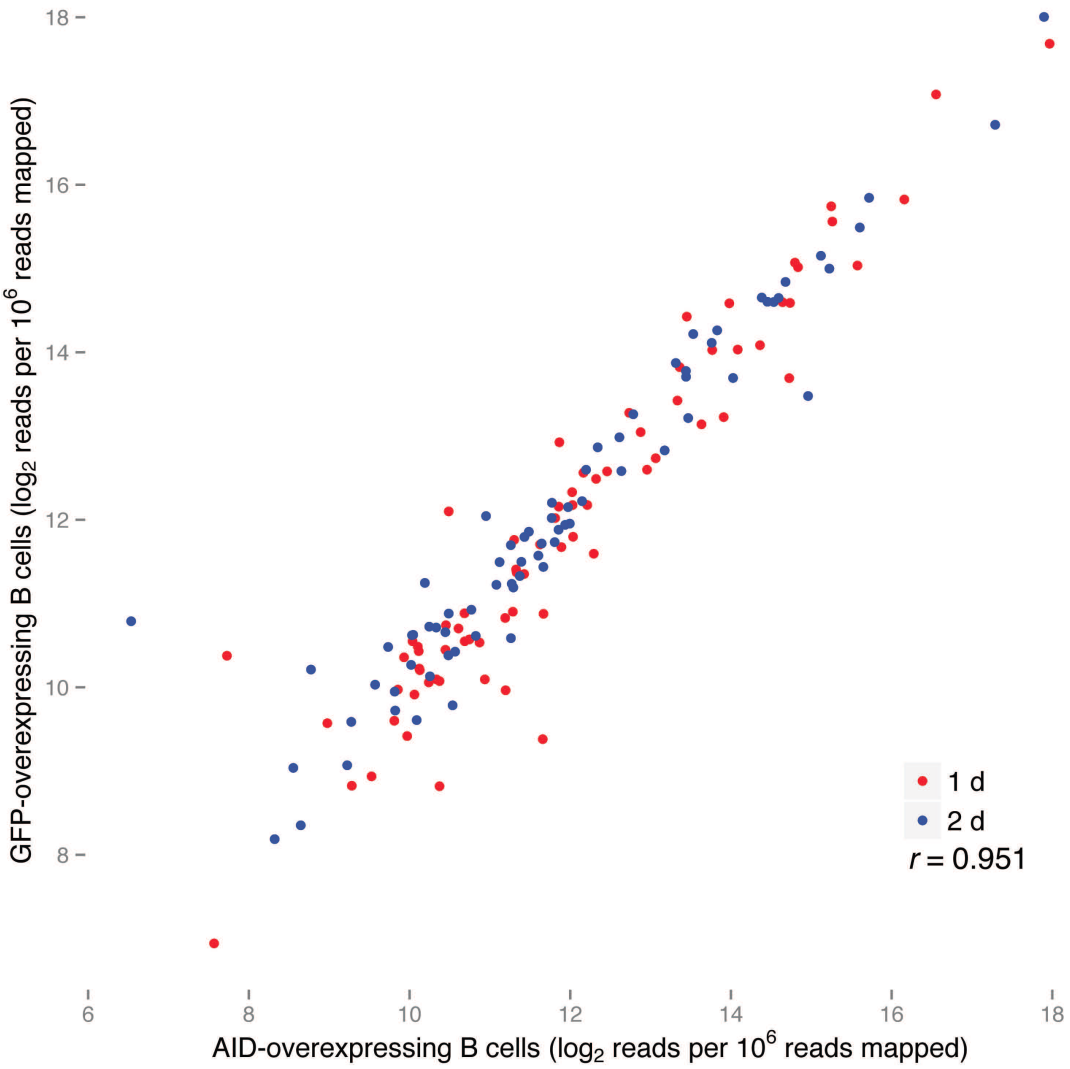
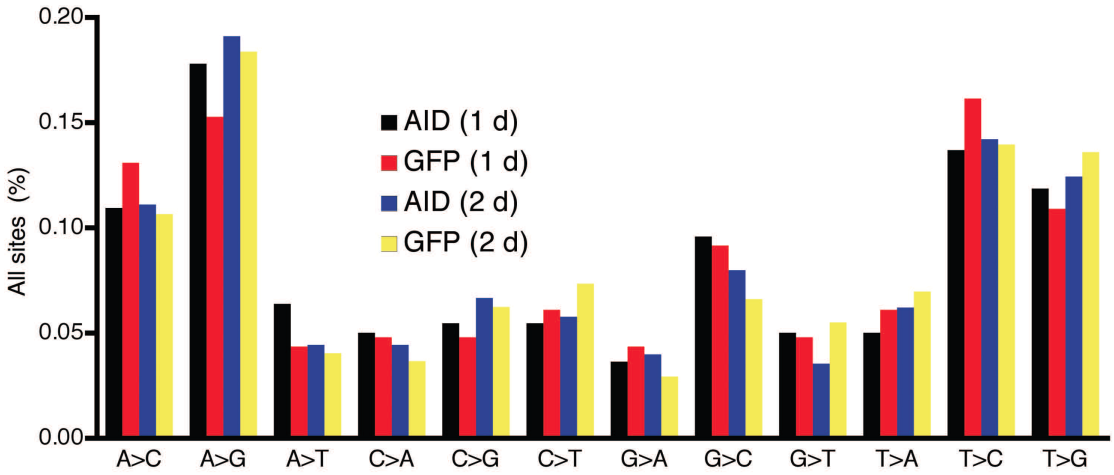


Figure 2.15. Abundance of small RNA-Seq reads overlapping annotated miRNAs that make up at least 0.1% of any sample for B cells overexpressing AID or GFP for 1 d or 2 d, normalized to million reads mapped. Points are colored according to time.

positions with at least 10% mismatches and at least 10x depth were identified and classified by type of base change. Because there is no known example of C-to-T editing in miRNAs, the observed frequencies of each possible base change type were quantified to get an idea of whether AID-dependent miRNA editing could be widespread. However, C-to-T changes were among the least common base alterations observed (Figure 2.16 A). For each base change type, including C-to-T, the frequency observed was approximately the same for each sample. These likely represent sequencing error, mismapping, or other types of RNA editing that do not depend on AID. For instance, A-to-G alterations constitute the most common change observed for all samples, and are possibly the results of ADAR-catalyzed editing (Alon et al., 2012; Kawahara et al., 2007; Vesely et al., 2012; Yang et al., 2005).

To look more specifically at potential AID-dependent editing, unique mismatched sites were analyzed. These were defined as events that did not occur in the corresponding sample at the same time point while using a lower depth threshold. Again, C-to-T sites were among the least frequent (Figure 2.16 B). Only one C-to-T site for each of the AID-overexpressing samples fulfilled these criteria. These sites do not appear to be authentic RNA editing as they have insufficient depth in the GFP-only overexpressing sample to compare, and they occur at nearly 100%, suggesting a genomic SNV or mismapping. Thus, if AID-catalyzed editing of miRNAs occurs, it does so at a rate that is well below the background of the sequencing protocol, and is therefore unlikely to be physiologically significant.

A.



B.

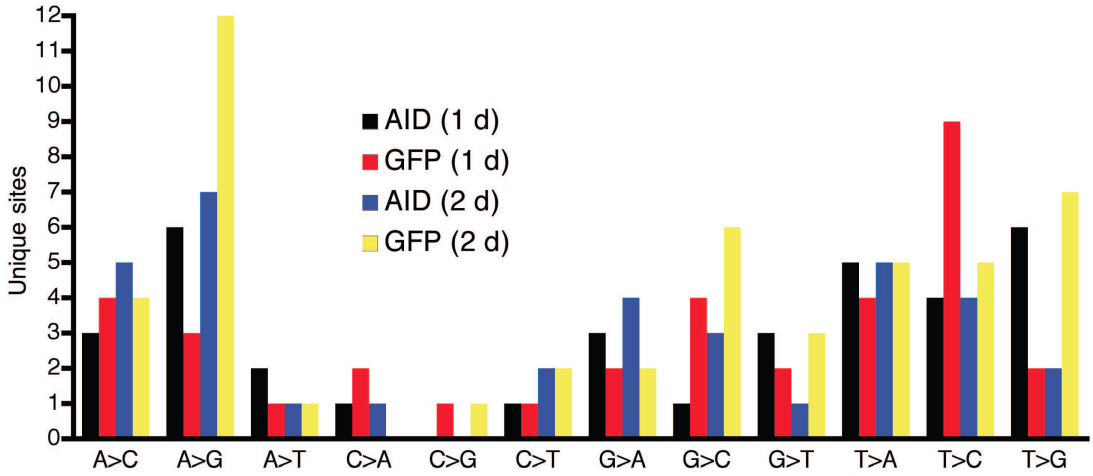


Figure 2.16. Number of (A) total and (B) unique observed mismatches for each base change type in aligned small RNA-Seq reads. (r = Pearson's correlation coefficient)

Chapter 3. Progress towards identification of the substrate of APOBEC2

3.1 Motivation

Despite the existence of a crystal structure (Prochnow et al., 2006), knockout mouse (Mikl et al., 2005), and its long evolutionary conservation (Conticello, 2004), very little is known about the biological function of the putative cytosine deaminase APOBEC2. While the *Apobec2*$^{-/-}$ mouse displays shift in muscle fiber type balance (Sato et al., 2010) and knockdown in zebrafish leads to muscular dystrophy (Etard et al., 2010), precisely how this is accomplished is unclear. More fundamentally, no deaminase activity for APOBEC2 has been demonstrated on any substrate, DNA or RNA. Because this problem seemed amenable to the genomics techniques that I had become familiar with in the work described in chapter 2, I decided to undertake studies with the primary goal of identifying the substrate of APOBEC2.

3.2 Analysis of APOBEC2 activity in primary myoblasts

I chose to use primary myoblasts differentiated in culture as a system for investigating the role of APOBEC2 because they fulfilled a number of key criteria. First, these cells are a commonly used system for recapitulating muscle differentiation in culture. Additionally, it had previously been demonstrated that APOBEC2 is induced upon differentiation in a myoblastic cell line (Vonica et al., 2011). Because the choice of fiber-type is thought to be set before fiber formation (Braun and Gautel, 2011), the differentiation process seems like the mostly likely setting for APOBEC2 to be exerting its physiological effects.

3.3 Establishment and validation of the primary myoblast culture system

Myoblasts are naturally resident in the muscles of neonatal mice, and protocols for their culture are well established (Danoviz and Yablonka-Reuveni, 2011). Following thorough dissection of the leg muscles, I isolated myoblasts by taking advantage of their resistance to adhesion. Rounds of pre-plating on standard tissue culture dishes cause the contaminating fibroblasts to adhere, while leaving myoblasts in the supernatant. Re-plating on collagen-coated dishes in the presence of basic fibroblast growth factor, hepatocyte growth factor, and high levels of FBS (20%) allows the myoblasts to adhere and divide for periods of weeks. Once myoblasts grow to roughly 70% confluence, they can be induced to differentiate by withdrawal of growth factors and changing from 20% FBS to 5% horse serum.

To determine if the cultured cells were in fact myoblasts, immunofluorescence microscopy was performed to detect the muscle lineage marker MyoD and the myotube differentiation marker MyHC. The images from undifferentiated cells show that nearly all cells present express MyoD, and thus are not contaminating fibroblasts (Figure 3.1 A). At this time point, very few cells express MyHC, and no cell fusion is apparent. However at 24h after serum withdrawal, nearly all cells express MyHC, and many fusion events are observed (Figure 3.1 B). By 48h after serum withdrawal, fusion has progressed to the point that nearly all nuclei are part of a continuous unicellular network (Figure 3.1 C). Unfortunately, attempts to stain for APOBEC2 using a Western-competent antibody failed, with similar levels of signal observed in the *Apobec2*$^{-/-}$ sample.

To verify that APOBEC2 induction in primary myoblast system had similar kinetics to that previously reported for a myoblastic cell line

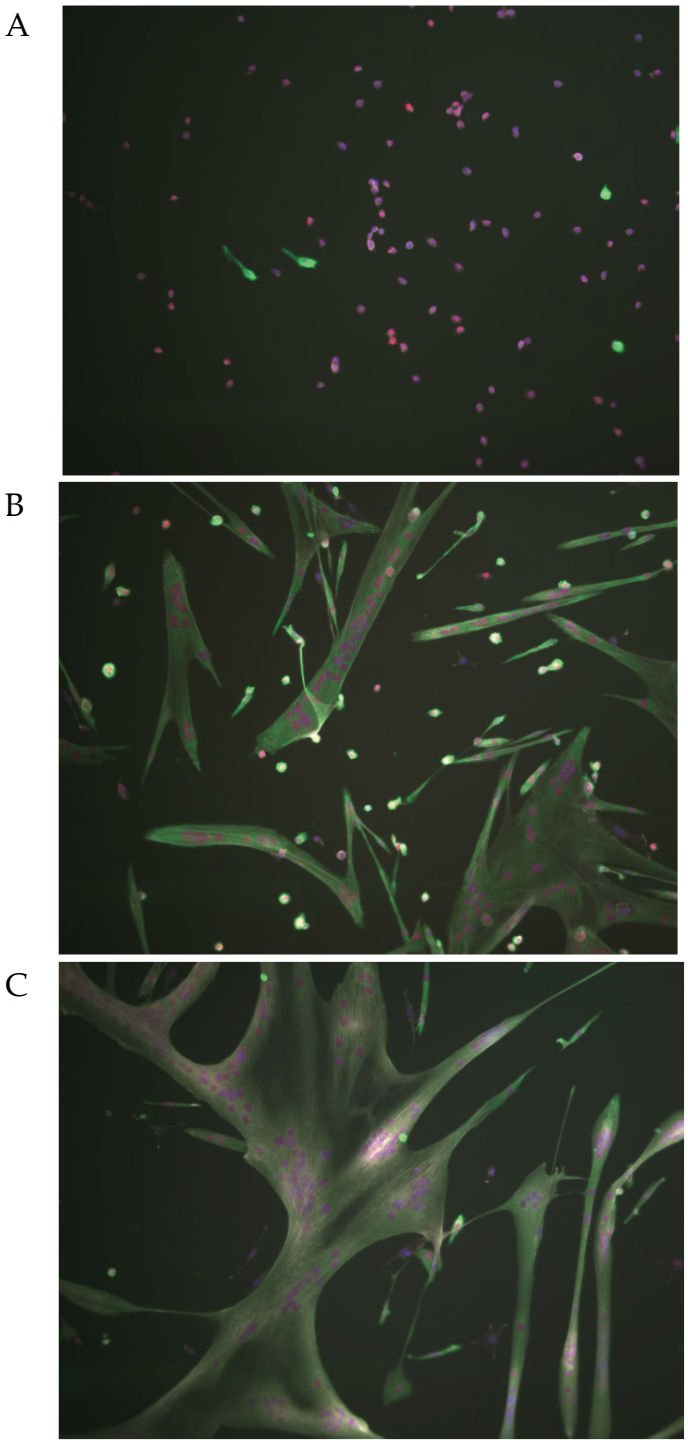


Figure 3.1 Immunofluorescence microscopy images of myoblasts at 10x magnification (A) prior to differentiation, (B) 24h after differentiation and (C) 48h after differentiation. (Hoescht in blue, MyoD in red, MyHC in green, and actin in grayscale)

(Vonica et al., 2011), *Apobec2* RNA levels were assayed over the course of differentiation. Two sets of wild-type myoblasts were lysed for RNA immediately before differentiation and for each of the next 5 days. Following conversion to cDNA, *Apobec2* levels were measured by quantitative RT-PCR. This analysis revealed that *Apobec2* transcript levels increase approximately 8-fold 1 day after serum withdrawal, and remain at roughly that level throughout differentiation (Figure 3.2).

To determine the kinetics of APOBEC2 protein expression in the system, protein lysates were harvested from both wild-type and *Apobec2*$^{-/-}$ cultures prior to differentiation and 1 and 2 days after. Western blotting revealed that, in contrast to RNA levels, protein levels of APOBEC2 are still rising at 1 day after differentiation, and reach high levels at 2 days after (Figure 3.3). For this reason, days 1 and 2 after differentiation were used as time points for subsequent studies. Additionally, no APOBEC2 protein could be detected in the *Apobec2*$^{-/-}$ samples, despite similar reductions in levels of MYOD1 and overall cell morphology.

3.4 APOBEC2-dependent changes in gene expression

To identify APOBEC2-dependent changes in gene expression or RNA sequence that could lead to identification of a substrate, an mRNA-Seq dataset was generated from differentiating myoblasts. Myoblasts were prepared from a pair of *Apobec2*$^{-/-}$ and wild-type 10-day old littermates and grown to 70% confluency. RNA was harvested prior to differentiation and at 24 and 48 hours afterwards, and this RNA was used to prepare mRNA-Seq libraries by a protocol similar to that described in 2, but lacking exogenous spikes.

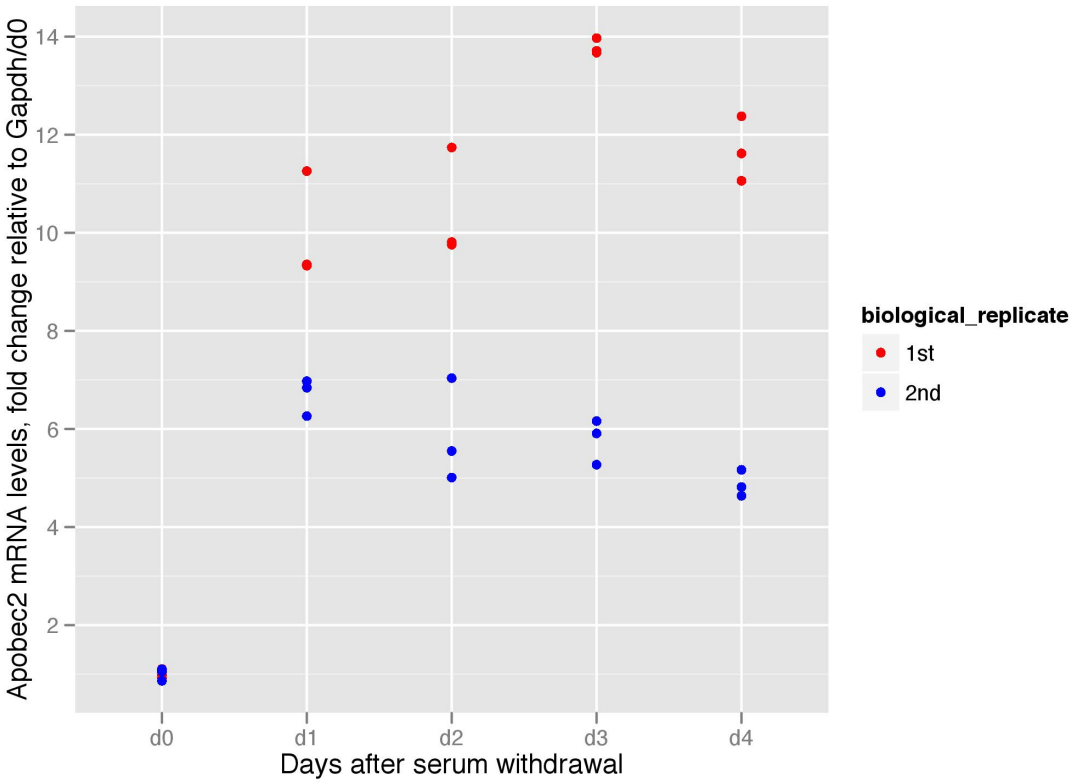


Figure 3.2. Apobec2 RNA levels after differentiation for primary myoblasts, as determined by quantitative RT-PCR. Values are relative to Gapdh and to Apobec2 levels for the undifferentiated sample. For each of the 5 timepoints, each measurement was taken in triplicate. Colors correspond to replicates.

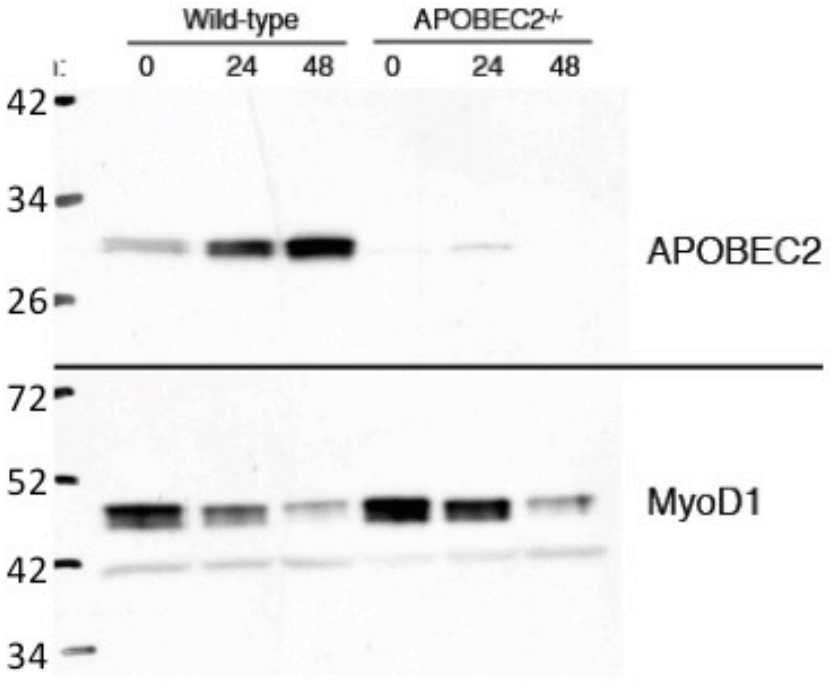


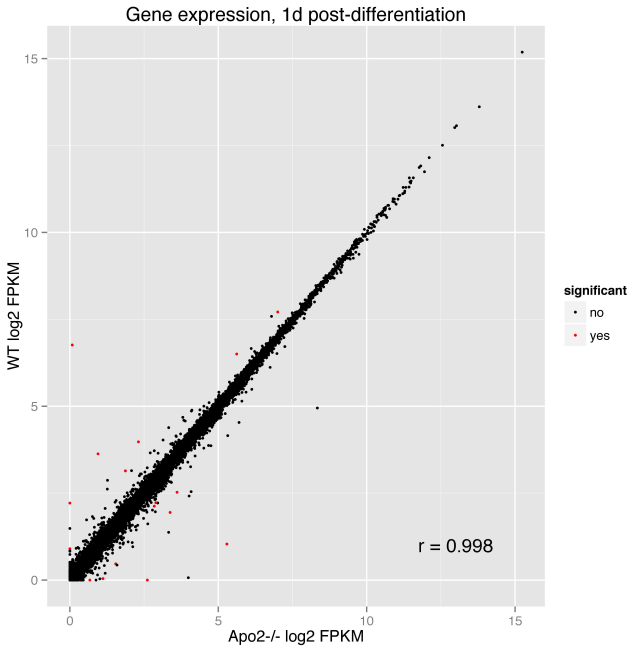
Figure 3.3. Apobec2 and MyoD protein levels after differentiation for primary myoblasts, as determined by Western blot.

For both the 24h and 48h pairs of samples, the correlation between the resulting gene expression values was very high (Figure 3.4). However in contrast to the results observed in the B cell mRNA-Seq, a number of genes were well-covered and had significantly different expression levels between the samples (Table 3.1). Besides *Apobec2* itself, this list of genes notably includes genes from two imprinted loci: the miRNA genes *Rian* and *Meg3*, which are derived from the Gtl2 locus, and *Cdkn1c*, which is derived from the Kcnq1 locus, were all upregulated approximately 2-fold in both wild-type samples. Given APOBEC2's reported involvement in DNA demethylation (Rai et al., 2008; 2010) and the importance of DNA methylation in the regulation of imprinted genes (Bartolomei and Ferguson-Smith, 2011), these loci appeared to be prime candidates as targets of APOBEC2.

3.5 APOBEC2-dependent changes in miRNA abundance

Because APOBEC2-dependent changes appear to be present in 2 miRNA primary transcripts in myoblasts, the RNA samples described in the previous section as well as another biological replicate similarly prepared were also used to prepare miRNA-Seq libraries. Quantification revealed approximately 2-fold increases in the wild-type samples for a large number of miRNAs derived from the *Gtl2* locus, in concordance with the mRNA-Seq findings (Table 3.2). This further suggested that APOBEC2 may have a role in the regulation of the *Gtl2* locus during myoblastic differentiation.

A.



B.

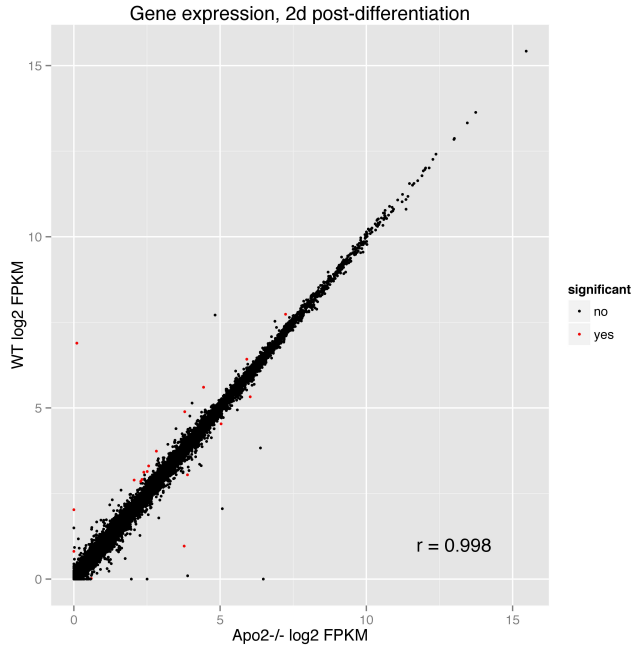


Figure 3.4 Gene expression values for wild-type and *Apobec2*$^{-/-}$ samples at (A) 1d and (B) 2d after differentiation, from mRNA-Seq. Points in red are significantly different, as determined by Cufflinks. (r = Pearson correlation coefficient)

Table 3.1 Genes significantly different in myoblast for at least one comparison and with FPKM > 1 in at least 1 sample. (W_d1 = wild-type FPKM at 1 day post-differentiation, K_d1 = Apobec2$^{-/-}$ FPKM at 1 day post-differentiation, W_d2 = wild-type FPKM at 2 days post-differentiation, K_d2 = Apobec2$^{-/-}$ FPKM at 1 day post-differentiation, d1_q = adjusted p-value for day 1 comparison, d2_q = adjusted p-value for day 2 comparison)

gene	W_d1	K_d1	d1_q	W_d2	K_d2	d2_q
Gm7325	379.45	353.53	1.00E+00	212.27	149.90	1.30E-03
Cdkn1c	208.59	127.35	4.61E-05	84.76	59.23	5.86E-05
Meg3	191.18	109.97	1.30E-01	47.65	20.62	9.75E-04
Apobec2	107.58	0.06	0.00E+00	117.76	0.07	0.00E+00
Rian	89.68	48.22	1.82E-06	28.62	12.84	4.61E-12
Eif2s3y	34.44	54.52	1.00E+00	39.07	64.34	7.83E-03
2410018M08Rik	1.05	38.13	3.25E-07	42.15	36.28	1.00E+00
Ptprs	13.81	19.55	8.34E-01	22.18	31.76	2.67E-03
2810453I06Rik	14.74	3.96	1.95E-06	20.39	21.32	1.00E+00
Daam2	4.76	11.22	1.95E-09	7.26	13.76	3.07E-12
Gm17517	11.38	0.93	0.00E+00	0.95	12.65	0.00E+00
Rftn1	7.82	2.66	4.60E-05	12.34	6.07	5.85E-05
Gm10269	2.85	9.39	2.13E-02	3.08	3.37	1.00E+00
Polh	7.62	4.77	1.00E+00	8.88	4.88	3.78E-03
Mafb	2.79	1.79	1.00E+00	7.83	4.69	5.35E-03
Snapc1	7.29	4.31	1.00E+00	7.71	4.25	2.96E-02
Gm17646	3.68	6.44	3.20E-03	6.52	4.02	5.46E-05
Ifi202b	3.72	1.88	8.53E-01	6.42	3.17	3.05E-02
Gm17492	3.36	6.18	1.99E-03	6.23	3.88	2.55E-04
Rbpsuh-rs3	0.00	5.11	3.57E-09	3.07	0.00	1.64E-08
Usp9x	3.64	0.00	5.95E-23	0.01	0.01	1.00E+00
H2-Gs10	0.38	1.92	3.70E-05	0.49	0.45	1.00E+00
Spna1	0.03	1.17	2.45E-11	0.01	0.51	1.31E-06

Table 3.2 miRNAs with greater than 1.5-fold average increases for the wild-type as compared to *Apobec2*$^{-/-}$ in both 1d and 2d samples. miRNAs derived from the *Gtl2* locus are highlighted in yellow.

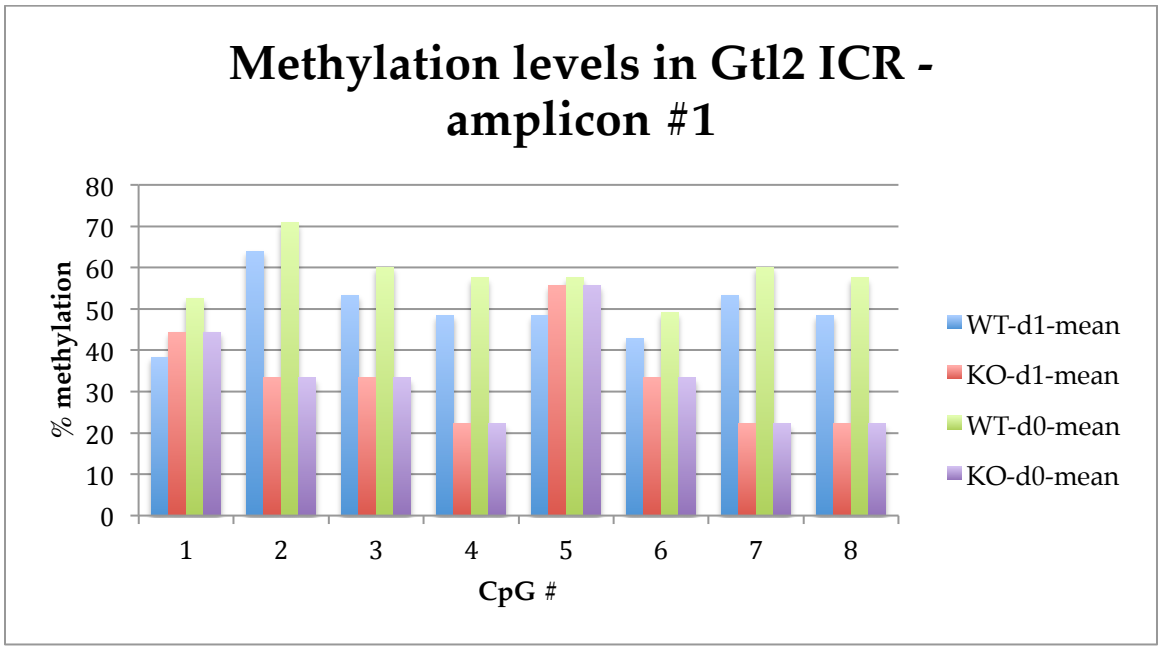
name	chr	start	end	avg_fc_d1	avg_fc_d2
Mir669l	2	10390598	10390695	2.3274	2.6468
Mir873	4	36615543	36615619	3.0290	3.0820
Mir592	6	27886655	27886750	1.5432	3.5210
Mir196a-1	11	96126478	96126579	1.9639	2.5010
Mir338	11	119876079	119876176	2.5706	2.0878
Mir673	12	110810200	110810290	1.5325	2.0765
Mir493	12	110818443	110818525	1.5693	2.2964
Mir433	12	110829925	110830048	1.5672	1.7134
Mir379	12	110947235	110947357	2.2052	1.5072
AC121784.4	12	110948819	110948935	2.1054	1.7742
Mir666	12	110955295	110955393	1.5174	1.5725
Mir543	12	110955466	110955545	1.8647	2.0442
Mir495	12	110956957	110957036	1.7691	1.6591
Mir381	12	110965032	110965106	1.9228	1.6067
Mir487b	12	110965543	110965624	1.5535	2.1474
Mir154	12	110976634	110976717	2.4047	1.5996
Mir412	12	110981499	110981578	1.6868	1.8079
Mir369	12	110981628	110981706	1.7411	1.8111
Mir410	12	110981925	110982005	1.7402	1.5189
Mir203	12	113369091	113369166	1.5883	2.2689

3.6 Methylation analysis of candidate imprinted loci

The APOBEC2-dependent changes in transcript and miRNAs levels derived from the *Kcnq1* and *Gtl2* loci suggested that APOBEC2 might regulate these imprinted loci via DNA demethylation. To investigate this further, the methylation status of the imprinted control regions (ICRs) of these loci was assayed by targeted bisulfite sequencing using genomic DNA from the same myoblast samples. Following bisulfite conversion of the DNA from the undifferentiated and d1-differentiated samples, amplified products from the ICRs were subcloned into a sequencing vector and at least 10 clones were sequenced from each sample for each amplicon.

For the only amplicon tested for *Kcnq1* and for 3 of the 5 amplicons tested for *Gtl2*, there were no differences between the samples. For the remaining 2 Gtl2 amplicons, methylation was consistently 20-30% higher in the wild-type than in the *Apobec2*$^{-/-}$ sample (Figure 3.5). This was not the expected result, as methylation of this ICR has previously been shown to be inhibitory to expression of *Meg3* and *Rian* (Ferguson-Smith, 2011). To determine whether this difference was due to APOBEC2, the undifferentiated samples of each genotype were also sequenced. This revealed that there was almost no change in methylation over the course of differentiation, and thus the observed differences were not APOBEC2-dependent. It is possible that the observed differences in methylation were due to heterogeneity in preparation of the myoblasts or strain differences in the mice used.

A.



B.

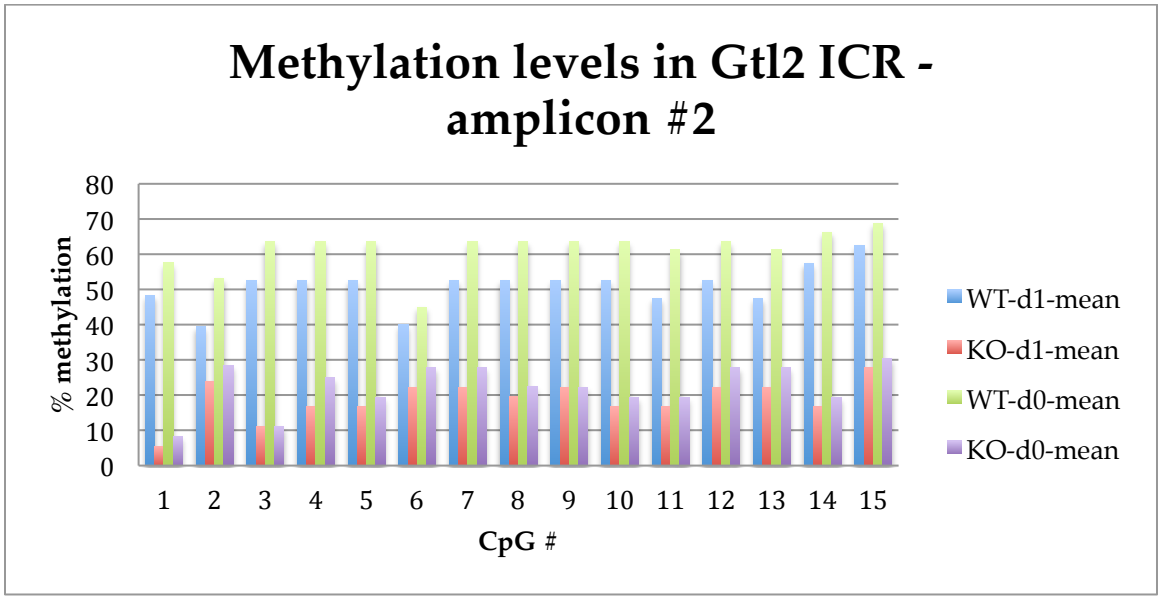


Figure 3.5 Methylation levels for 2 regions of the *Gtl2* ICR, for wild-type and *Apobec2*$^{-/-}$ samples prior to differentiation and 1 day post-induction. Each column represents a single CpG within the amplicon.

3.7 No APOBEC2-dependent editing is apparent in cultured myoblast mRNA

To check whether APOBEC2-dependent RNA editing was occurring in the differentiating myoblasts, an RNA editing detection pipeline similar to that described in chapter 2 was applied to the mRNA-Seq data. Again, subject to cutoffs for minimum coverage depth and percentage apparent editing, C-to-T changes that were present in the wild-type but not the *Apobec2*$^{-/-}$ sample were identified. In this case another filter was added, in which reads overlapping candidate editing sites were realigned with the program BLAT (Kent, 2002) and then the fraction of mismatches at the site recounted. This strategy has been previously shown to reduce false positives that are the result of mismapping (Peng et al., 2012). While VSG spikes were not used in the preparation of this library, the numbers of aligned reads were similar to those in the B cell samples, so a similar portion of the transcriptome was likely covered.

Application of the pipeline to the 1d-differentiated sequencing data resulted in a greater number of apparent candidate editing events (81) in the *Apobec2*$^{-/-}$ sample as in the wild-type sample (59), for an IFPR of > 100%. This strongly suggests that APOBEC2-dependent RNA editing, if it occurs, does not occur at a high frequency. To determine whether such editing occurs at all, the supporting RNA-Seq reads for each of the candidate editing events were visually inspected. All but two were rejected because of distance to a non-C-to-T SNV or distance from a splice junction. To determine whether these two candidate events were in fact true RNA editing, the sequences were amplified from the genomic DNA of the wild-type myoblasts by PCR and analyzed by Sanger sequencing. These sequences revealed that the mismatches were a result of

heterozygous SNVs at the sites in question in the wild-type sample. Thus, no candidate APOBEC2-dependent RNA editing events could be identified in the myoblast-derived mRNA-Seq data.

Because miRNA-Seq data was also available, a similar comparison was made to check for APOBEC2-dependent editing in small RNAs. Although for some samples more unique C-to-T events were apparent in the wild-type sample, these were not consistent between replicates and appeared to be artifactual upon visual inspection (Table 3.3).

3.8 Analysis of APOBEC2 RNA editing activity in adult murine muscle

Experiments in the myoblast system failed to reveal a function for APOBEC2. Plausible issues with the myoblast experiments include the facts that the system is not physiological, or that APOBEC2 may act later than the induction of differentiation. Alternatively, the issue may have been with the experimental techniques and not the system itself. Examples of functions that could have been missed by the techniques used include editing of a long non-coding RNA, editing of a low-expressed RNA, or a scaffold role for APOBEC2 in a complex that does not make use of its putative catalytic activity.

In order to comprehensively characterize any possible APOBEC2-dependent RNA editing in physiological settings, directional, ribo-minus mRNA-Seq libraries were prepared from muscles from a pair of wild-type and *Apobec2*$^{-/-}$ adults. The three muscles used were chosen because they are classic examples of the easily accessible muscle types: the heart (cardiac), the extensor digitorum longus (fast-twitch skeletal), and the soleus (slow-twitch skeletal). The directional and ribo-minus library preparation conditions were chosen to

Table 3.3 Counts of apparent unique C-to-T SNVs (candidate editing events) in miRNA-Seq from wild-type and *Apobec2*$^{-/-}$ myoblasts. (IFPR = implied false positive rate)

Comparison	WT sites	Apo2$^{-/-}$ sites	IFPR
d0 replicate 1	15	15	100%
d0 replicate 2	15	15	100%
d1 replicate 1	12	7	58%
d1 replicate 2	15	14	93%
d2 replicate 1	18	18	100%
d2 replicate 2	24	6	25%

decrease the noise in identifying editing events due to incomplete annotations and to analyze non-polyadenlyated RNAs, respectively.

Application of a similar RNA editing detection pipeline to that used on the myoblast samples again failed to identify any RNA editing events. No candidate events were identified in the wild-type soleus or EDL (as well as in the *Apobec2*$^{-/-}$ reciprocal comparison). For the heart, 59 candidate events were found in the wild-type compared to only 19 for the *Apobec2*$^{-/-}$ sample. While the counts alone were consistent with the existence of APOBEC2-dependent editing, nearly all of the events were located in regions of apparent non-isogenicity. The few that were from other genomic locations were found to be the result of mismapping upon visual inspection of the supporting reads.

Chapter 4. Investigation of 5-methylcytosine content of *T. brucei* DNA

4.1 Motivation and preliminary findings

While DNA 5-cytosine methylation is best known as a silencing modification in mammals, the mark is used for a wide variety of other purposes in other organisms. For example, it is involved in regulating life cycle stages in a parasitic worm (Gao et al., 2012), in marking DNA for degradation in a ciliated protist (Bracht et al., 2012), and in setting caste identity in the honeybee (Elango et al., 2009; Wang et al., 2006).

In the unicellular eukaryotic parasite *Trypanosoma brucei*, 5-cytosine methylation is present in nuclear DNA, albeit at a lower level than in mammals (Militello et al., 2008). However its function in this organism, as well as the pattern of methylation within its genomic DNA, are unknown. As part of a study focused on another epigenetic mark in *T. brucei*, a colleague, Hee-sook Kim, had previously generated strains deficient in each of the 3 closest homologues of known 5-cytosine methyltransferase enzymes (henceforth CMT1, CMT2, and CMT3). Although none of the mutants displayed decreased methylation in bulk as assayed by dot blot, this experiment verified that 5-methylcytosine was present in DNA at levels well above background. Because of my interest in cytosine methylation, I decided to undertake whole-genome bisulfite sequencing of *T. brucei* with the goal of producing the first map of methylation in this organism and generating hypotheses for its function.

4.2 Identification of candidate methylated sites by whole-genome bisulfite sequencing

Because the genome of *T. brucei* is small (25 Mb), whole-genome bisulfite sequencing was the obvious choice for assaying methylation. To determine whether whole-genome bisulfite sequencing was possible for the *T. brucei* genome, as well as to validate the alignment strategy, simulated bisulfite-converted reads were generated from a standard genome (427 strain) and then aligned back with Bismark. This resulted in 77.6% of simulated 50bp reads aligning uniquely, with only 15.6% of sites left completely uncovered. This was taken as evidence that the T. brucei genome was sufficiently complex to allow for efficient mapping of 50bp bisulfite-converted reads.

Libraries were prepared from wild-type genomic DNA from both the bloodstream form (BF) and procyclic form (PF) of wild-type 427-strain T. brucei, and the BF of CMT1$^{-/-}$ and CMT2 and CMT3 knockdown strains. Although none of the mutants had displayed decreased methylation in bulk, libraries were still prepared from both in case either was responsible for methylation of a particular subset of cytosines, as is the case in *Arabidopsis* (Lister et al., 2008). As a control for library preparation and mapping, an unconverted wild-type BF sample was run in parallel.

Unmethylated lambda phage DNA was added to the genomic DNA prior to library preparation to serve as a positive control for bisulfite conversion. Genomic DNA was fragmented by sonication, repaired, ligated to sequencing adapters, bisulfite converted twice, and amplified by PCR.

Once the sequencing data was in hand, test alignments were performed to determine the optimal genome configuration in terms of maximum number of

reads aligning uniquely. The BF sample was mapped to each permutation of the 927-strain genome (for which genomic sequence is more complete, but belonging to a related strain and thus containing a number of sequence differences) or 427-strain (which is less complete but corresponds to the genomic DNA actually used here) with or without a supplementary file containing known VSGs and repetitive elements. While all 4 of these options performed acceptably well, the best performing was 427 with VSGs, which yielded 43.4% of reads aligning uniquely. While this is much lower than the mapping efficiency for the simulated reads (77.6%), this is expected due to the repetitive sequences, and polymerase and sequencing errors that the simulation does not account for.

After aligning all of the reads to the 427+VSG genome, apparent methylation levels were checked for the phage genome to verify that bisulfite conversion proceeded efficiently. For the converted libraries, the fraction of phage cytosines uncoverted ranged from 0.1-0.6%, while for the unconverted library it was 99.8%. Thus the background error due to incomplete bisulfite conversion, mismapping, and incorrect demultiplexing is less than 0.6%.

Genomic coverage was then visually inspected with IGV. This revealed that for the bisulfite converted libraries coverage was "lumpy," with sharp peaks as well as large stretches of zero coverage present. In contrast, the unconverted control displayed much more uniform coverage. Quantification of the theoretically mappable but uncovered ("anomalously uncovered") regions of the genome revealed that it was much higher for the bisulfite-converted samples (13.7-13.9%) than for the uncoverted sample (2.1%) as well as the ideal Poisson-distributed case (<0.1%). Comparison of the anomalously unmapped regions for the BF and PF samples revealed that they frequently overlapped (61% overlap

observed, with 14% expected if these bases were randomly distributed). Additionally, these anomalously uncovered regions had significantly lower CG-content than covered regions (41.7%, compared to 47.4% for covered regions). Taken together, this suggests that the coverage pattern displayed in the bisulfite converted samples was a result of the properties of the bisulfite converted sequences themselves, perhaps as a result of preferential amplification of CG-rich sequences, and not a failure of the mapping strategy or insufficient number of raw seqeunces. This shortcoming could likely be fixed in the future by use of a PCR-free library preparation strategy.

Bulk-level analysis of the methylation levels was then performed on the mapped reads. For each of the bisulfite-converted samples, the apparent 5mC/C ratio was 0.2-0.4%. These values are well within the range of background, and are much lower than was expected based on the initial 5mC dotblots.

Despite the imperfect coverage and the surprisingly low levels of overall cytosine methylation, the sequencing data were analyzed to identify methylated cytosines. For each site, the binomial test was performed with the probability of success equal to the background uncovered percentage for that sample as determined by analysis of the unmethylated phage control. The resulting p values were then adjusted with the Benjamini-Hochberg method, and sites with q values < 0.05 considered candidate methylated sites. These candidate sites were few in number and were mostly derived from a few discrete genomic locations (Table 4.1). Visual inspection of the candidates revealed that most of these locations displayed hallmarks of mismapping, such as non-bisulfite-type mismatches nearby, or all of the methylation-supporting bases occurring at the

Table 4.1. Sites with significant levels of unconverted cytosine, as determined by the binomial test. Values for each library are apparent percentage methylated. Sites in orange are located in the 177-base pair repeat.

site	BF	PF	CMT1	CMT2	CMT3	BF_unconv
Tb427VSG-999904:11	5%	0%	1%	1%	3%	99%
Tb427VSG-999904:81	4%	2%	2%	3%	5%	100%
Tb427VSG-999904:142	14%	29%	10%	19%	0%	97%
Tb427VSG-999904:188	5%	2%	2%	4%	5%	99%
Tb427_01_v4:93	19%	23%	19%	19%	19%	35%
Tb427_01_v4:623754	0%	0%	33%	0%	0%	100%
Tb427_01_v4:1061477	29%	63%	33%	61%	67%	45%
Tb427_01_v4:1061479	100%	100%	100%	100%	100%	100%
Tb427_01_v4:1061485	100%	100%	100%	100%	100%	100%
Tb427_01_v4:1061486	100%	100%	99%	100%	100%	100%
Tb427_01_v4:1061489	100%	100%	99%	100%	100%	100%
Tb427_01_v4:1061492	100%	100%	100%	98%	100%	100%
Tb427_01_v4:1061495	100%	100%	100%	100%	100%	100%
Tb427_05_v4:1470913	18%	36%	0%	50%	0%	100%
Tb427_08_v4:26469	33%	12%	29%	31%	28%	0%
Tb427_09_v4:2240282	25%	15%	28%	14%	9%	100%
Tb427_09_v4:2240286	25%	15%	26%	13%	9%	100%
Tb427_09_v4:2240416	26%	6%	23%	11%	29%	100%
Tb427_09_v4:2240417	26%	6%	23%	11%	29%	100%
Tb427_09_v4:2240418	26%	7%	25%	11%	25%	100%
Tb427_09_v4:2240426	25%	13%	33%	13%	28%	100%
Tb427_09_v4:2240429	29%	13%	38%	15%	28%	100%
Tb427_09_v4:2240430	29%	12%	43%	17%	28%	100%
Tb427_09_v4:2240433	31%	13%	55%	22%	25%	100%
Tb427_09_v4:2240435	36%	13%	55%	22%	31%	100%
Tb427_09_v4:2240436	36%	14%	55%	13%	31%	100%
Tb427_09_v4:2240440	36%	20%	67%	25%	31%	100%
Tb427_09_v4:2240442	31%	21%	67%	25%	31%	100%
Tb427_09_v4:2240450	33%	18%	71%	20%	40%	100%
Tb427_09_v4:2240451	25%	20%	67%	0%	33%	76%
Tb427_10_v5:449	4%	0%	1%	3%	18%	0%
Tb427_10_v5:125988	4%	9%	6%	5%	6%	100%
Tb427_10_v5:4057088	83%	86%	83%	88%	87%	91%
Tb427_10_v5:4057096	88%	93%	94%	91%	91%	100%
Tb427_10_v5:4057098	67%	81%	76%	79%	82%	89%
Tb427_11_01_v4:640	50%	54%	49%	50%	50%	58%

site	BF	PF	CMT1	CMT2	CMT3	BF_unconv
Tb427_11_01_v4:2960	16%	25%	50%	29%	19%	0%
Tb427_11_01_v4:530862	7%	16%	25%	15%	0%	100%
Tb427_11_01_v4:1127249	8%	11%	7%	8%	5%	100%
Tb427_11_01_v4:4125370	5%	5%	3%	10%	3%	100%
Tb427_11_01_v4:4633063	89%	89%	91%	87%	89%	94%
Tb427VSG-1494:687	0%	18%	14%	36%	14%	100%

end of reads. The exception to this was sites within the 177-base pair repeat, an element found in the minichromosomes of *T. brucei* (Sloof et al., 1983).

4.3 Validation of candidate methylated sites by methylated DNA immunoprecipitation

To determine whether the 177-base pair repeat was truly methylated, I attempted to validate the bisulfite result by an orthogonal technique: methylated DNA immunoprecipitation. In this technique, DNA is immunoprecipitated with an antibody specific for 5mC and target regions are compared to input or to an isotype control by quantitative PCR.

Genomic DNA was sonicated to an average size of 500bp and immunoprecipitated with a monoclonal anti-5mC antibody or an isotype control. As a control, mouse tail-tip DNA was immunoprecipitated in parallel.
Two sets of primers to specifically amplify the 177-base pair repeat along with two sets of primers to amplify separate well-covered areas that had no methylated basecalls were designed and successfully tested for specificity on the sonicated DNA. Mouse DNA was analyzed at known methylated (H19) and unmethylated (Actb) loci. These primers were used for SYBR Green real-time PCR on the immunoprecipitated DNA.

In the mouse positive control, the methylated sequence was clearly enriched in the IP sample, with an enrichment of about 50x relative to the unmethylated amplicon (Figure 4.1). There was an even more dramatic enrichment in comparison to the isotype control, with so little DNA immunoprecipitated that real-time amplification failed. However for both *T. brucei* samples, an apparent disenrichment of the 177-bp sequences was observed

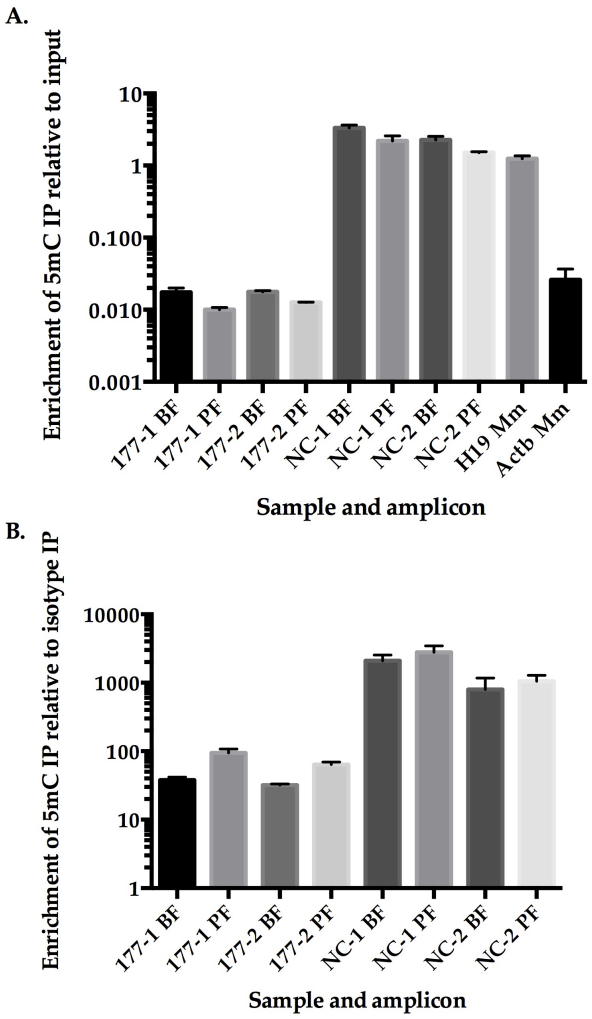


Figure 4.1. Validation of apparently methylated sites by MeDIP-qPCR. Enrichment was calculated relative to known unmethylated regions and (A) input or (B) isotype control IP. Values are averages for 2 sets of primers each for 177-bp repeat and unmethylated regions. (177-1 and -2 = amplicons spanning the 177-bp repeat, NC-1 and -2 = amplicons spanning unmethylated negative control regions, BF = bloodstream form, PF = procyclic form, Mm = *Mus musculus*)

in the IP relative to both input and isotype control. Relative to input as well as to isotype IP, the 177-bp repeat sequences appeared to be at least 10-fold less abundant than the sequences used as negative controls. A simple lack of enrichment would be interpretable as a lack of methylation at the 177-base pair repeat, but the magnitude of the apparent disenrichment suggests that the situation is more complex. It is possible that the sequences used as negative controls contain modified bases that cross-react with the 5mC antibody.

Chapter 5. Discussion[‡]

5.1 Refined model of AID activity in B cells

The genome-wide analyses of AID-dependent changes in gene expression and DNA methylation allow for a reexamination of the various activities that have been ascribed to AID in the B cell. The most obvious of these is the requirement of AID for CSR, which was made apparent by the presence of secondary-isotype-derived productive transcripts in the AID–miR-155T and wild-type samples and their absence in the *Aicda*$^{-/-}$ sample. Additionally, the increase in secondary isotype transcript levels in the AID–miR-155T sample compared to wild-type reconfirms the positive correlation between AID levels and levels of CSR (Teng et al., 2008). As has been previously discussed (Wang et al., 2009), this suggests that under physiological conditions the AID system is not optimized for maximum CSR, and so there is likely a tradeoff between CSR and off-target AID activity that maintains an intermediate level of the protein.

The second known physiological role for AID, initiating somatic hypermutation (SHM), is unfortunately not amenable to the analyses described here. This is due to two factors: cell type and sequencing technology. Attempts to detect elevated levels in mismatch rate at V-segments in the wild-type and AID–miR-155T samples failed. This was not surprising, given the low levels of SHM in *ex vivo* stimulated naïve B cells and the sequencing error rates for the Illumina HiSeq platform. The likely bias towards correctly mapping unmutated V(D)J-derived reads more frequently than mutated ones would further confound any conversion of mRNA-Seq data into a true SHM rate. It should however be

[‡] Portions of this chapter were published in (Fritz and Papavasiliou, 2010) and (Fritz et al., 2013)

possible to directly examine levels of SHM by mRNA-Seq given a system in which these levels are higher (for example, germinal center B cells, with a transgenic BCR, after immunization), and perhaps greater sequencing depth and/or accuracy.

Although the mRNA-Seq results were consistent with the known "classical" functions of AID, the analyses failed to detect any signs of the known off-target effects of AID in B cells. For the well established AID-catalyzed hypermutation of certain non-Ig loci (Liu et al., 2008; Pasqualucci et al., 1998; Shen et al., 1998; Yamane et al., 2010), this lack of evidence was not surprising. The challenges of using mRNA-Seq to quantify SHM would present an even greater problem for detecting mutation of non-Ig targets, given their much lower mutation rate.

However, it is somewhat surprising that the DNA damage caused by AID does not display greater effects on the transcriptome or DNA methylome. Despite the known broad occupancy of AID in the genome (Liu et al., 2008; Yamane et al., 2010) and occurrence of AID-dependent translocations (Klein et al., 2011; Ramiro et al., 2004; Robbiani et al., 2009), there was no clear upregulation or promoter demethylation of DNA repair genes in the AID-containing samples. There are a number of non-exclusive explanations for these observations. First, it is likely that off-target AID-dependent damage is rare in absolute terms. Previous reports have estimated the frequencies of AID deamination of its most frequent non-Ig target at 10^{-4} per basepair, and at of one of the most common AID-catalyzed translocations at 4×10^{-7} per wild-type cell (Robbiani et al., 2009). While this level of damage can have clear effects at the organismic level over time, as in the c-myc/IgH translocation in Burkitt's

lymphoma, this sort of selection-dependent effect would not be apparent in the naïve culture system used here.

Another possibility is that the repair of AID-dependent DNA damage does not require a large transcriptional response. The fact that AID-dependent abundance differences are apparent for Ig isoforms, but not for the other genes known to play a role in CSR such as RPA (Chaudhuri et al., 2004) and Spt5 (Pavri et al., 2010) indicates that this is at least partly the case. It is possible that B cells are primed to deal with such widespread DNA damage without requiring a dynamic transcriptional response. The view that B cells are at least somewhat primed to perform error-free DNA repair is supported by the finding that constitutive AID expression leads to tumors in non-B cell types (Okazaki et al., 2003).

Finally, it is possible that AID-dependent damage does elicit an appreciable population-level transcriptional response, but in a narrower population than the one sequenced here. It is possible that sequencing of a more uniform population in terms of AID induction or cell cycle phase may reveal such an AID-dependent upregulation of DNA damage response genes.

A clear limitation of the work presented here is that none of the analyses are direct measurements of the frequency of off-target AID activity or AID occupancy in the genome *per se*; use of a more physiological system, such as germinal center B cells after immunization, may be able to yield more quantitative answers about the frequency and consequences of off-target AID activity. However it is clear that for the subset of cells studied here the only observable transcriptional consequences of AID at the population level are at the Ig loci despite its ability to act at many places in the genome.

5.2 Reconciliation of AID activity in B cells and other systems

Despite previous data showing AID occupancy throughout the genome (Yamane et al., 2010) and DNA demethylation concurrent with B cell activation (Shaknovich et al., 2011), no AID-dependent demethylation events were observed in activated B cells, even when AID is overexpressed. This is in contrast to observations of AID-dependent demethylation in iPS cells (Bhutani et al., 2013; Kumar et al., 2013), primordial germ cells (Popp et al., 2010), heterokaryons (Bhutani et al., 2010), neurons (Guo et al., 2011), and zebrafish embryos (Rai et al., 2008). Three classes of explanations are possible to explain this difference: (1) that AID-dependent DNA methylation is cell-type specific and does not occur in B cells, (2) that AID-dependent DNA methylation occurs in B cells, but was not detected by the assays used here, and (3) that AID-dependent DNA methylation does not occur at all.

Perhaps the easiest explanation for the gap between the dramatic effects attributed to AID in more exotic systems and the limited changes observed in B cells is that the AID-dependent demethylation pathway is cell-type specific. It is possible that AID does not catalyze DNA demethylation at all in B cells because cofactors required for AID-dependent DNA demethylation are not present in B cells, or alternatively B cells could possess some other factor that prevents deleterious demethylation in the presence of AID. The latter seems more likely than the former, given the expression in B cells of the two alleged cofactors of AID required for DNA demethylation, Gadd45 and TDG (Rai et al., 2008) and the known deleterious effects of AID expression in other cell types (Okazaki et al., 2003). While no candidate that could act as such a cofactor has been identified,

there is precedent for altered treatment of AID-induced damage, which appears to be responsible for the varying error rates of repair at different loci (Liu et al., 2008).

There are also a number of circumstances that would lead to AID-dependent DNA methylation that is undetectable by RRBS. For example, it is possible that AID-dependent DNA demethylation occurs at a subset of sites that are not covered by RRBS, such as CpG-poor regions of the genome. This possibility could be investigated by use of WGBS on the existing samples. Alternatively, AID-dependent demethylation in B cells could be obscured by rapid re-methylation, for example by DNMT1. However if this is the case, it is unlikely to be physiologically meaningful even if it were to be observable. Finally, it is possible that the differences are so small in number and magnitude that they are simply below the limit of detection of the RRBS technique. While it is clear that there is no difference in bulk levels of methylation between the samples, it is possible that a subset of the observed differences that are consistent with noise are biologically meaningful. Increased sequencing depth and sequencing of replicates would provide information about this possibility, but it is impossible to eliminate completely.

Finally, there are reasons to doubt the existing reports of AID-dependent demethylation in mouse cells. One particularly concerning issue regarding publications that have found differences in methylation between cells from *Aicda*$^{-/-}$ and wild-type mice in reprogramming contexts (Kumar et al., 2013; Popp et al., 2010) is that they make use of an *Aicda*$^{-}$ allele that is on a C57BL/6 x CBA F1 background (Muramatsu et al., 2000) that has been backcrossed for many generations to C57BL/6. This is a concern here because the AID locus is in close

proximity to the locus of *Nanog*, a key factor in reprogramming to pluripotency. It is clear from both the distribution of SNVs in the mRNA-Seq data reported here and the observations of others (Hogenbirk et al., 2013) that the CBA-like region encompasses the Nanog locus. Reprogramming efficiency is known to be highly strain-dependent (Schnabel et al., 2012), and it is possible that the differences attributed to AID-deficiency were actually the result of differences between $Nanog^{C57BL/6}$ and $Nanog^{CBA}$.

Furthermore, there are aspects of the reports of the AID's influence on reprogramming that call their authors' conclusions into question. Other laboratories have observed no effect of AID deficiency for reprogramming experiments using both Yamanaka factors (Habib et al., 2014) and heterokaryon formation (Foshay et al., 2012). While these differences do not invalidate the original reports, they do suggest that the effects attributed to AID are not robust.

There are also reasons to suspect that the evidence for AID-dependent demethylation from lower vertebrates is not generalizable to mammals. In zebrafish, identical methylation phenotypes were observed when *Aid*, *Apobec2a*, or *Apobec2b* were knocked down (Rai et al., 2008). Because mammalian AID and APOBEC2 clearly behave very differently in terms of *in vitro* activity (Bransteitter et al., 2003; Chaudhuri et al., 2003; Dickerson et al., 2003; Mikl et al., 2005) and knockout phenotype (Muramatsu et al., 2000; Sato et al., 2010), it is difficult to imagine a situation where they would be interchangeable, as appears to be the case in zebrafish.

Most significantly, the viability of the $Aicda^{-/-}$ mouse suggests that any possible role that AID plays in the regulation of DNA methylation is relatively modest. In contrast to the embryonic lethal phenotypes for the knockouts of bona

fide regulators of DNA methylation such as *Dnmt1* (Li et al., 1992), *Dnmt3a* (Kaneda et al., 2004), *Tet1* (Yamaguchi et al., 2013), and *Tet3* (Gu et al., 2011b), *Aicda*$^{-/-}$ mice have no apparent defects in fertility, size, or health outside of the expected immune deficiencies. The similar phenotype of humans lacking AID (Revy et al., 2000) suggests that this is generally true of AID's role in mammals.

A consolidated explanation for the observations relating to AID and DNA demethylation would be as follows: in B cells, even though AID activity is widespread, AID-induced DNA damage does not lead to DNA demethylation because B cells possess a DNA repair system that leads to faithful transmission of methylation state. In reprogramming by iPS and heterokaryon formation, AID-induced DNA damage can rarely lead to DNA demethylation due to differences in cofactors or available substrates. This demethylation is more likely to occur by deamination of nearby C and long-patch-type repair than by direct deamination of 5mC because of AID's strong preference for the unmodified base (Larijani et al., 2005; Morgan et al., 2004; Nabel et al., 2012; Wijesinghe and Bhagwat, 2012).

5.3 RNA-Seq strategies for characterizing B cell populations

In addition to confirming a narrow role for AID action in B cells, the analyses presented here also demonstrate the broad capabilities of high throughput RNA sequencing. In addition to its standard usage as a gene expression assay, its utility in detecting RNA editing, CSR frequency and V_H segment usage demonstrate its flexibility as a tool for characterizing populations of B cells.

Although not performed on the data here, mRNA-Seq has the capacity to assay mutation rates. With improved depth, and when used on a system with

higher mutation rates, RNA-Seq may be a useful tool for assessing AID-dependent mutation genome-wide. As the most comprehensive data to date regarding AID-catalyzed non-Ig hypermutation still only cover about a hundred loci (Liu et al., 2008), use of very deep RNA-Seq on B cells from immunized mice could lead to a much more complete picture of AID's off-target activities. This same sequencing data could be useful in identifying AID-dependent translocations in a truly wild-type context by finding recurrent fusion-derived reads (Kim and Salzberg, 2011).

5.4 Potential roles for APOBEC2

The clearest conclusion that can be drawn from the data presented here is that APOBEC2 is not likely to be an RNA editor. In both differentiating myoblasts and mature muscle tissue, no such activity could be observed. While I cannot exclude the possibility that it edits a very rare transcript or a common transcript at a very low level, it is also unlikely that such activity, should it exist, has physiological significance.

While the data presented do not demonstrate a role for APOBEC2, they do help to clarify its possible activities. It remains possible that APOBEC2 has a function in DNA demethylation during muscle development. The observations of expression differences at known methylation-regulated genes between wild-type and *Apobec2*$^{-/-}$ samples supports this, as do the impaired DNA demethylation phenotype observed in *Apobec2a/b* morphant zebrafish (Rai et al., 2008). However, APOBEC2's lack of enzymatic activity on DNA *in vitro* (Mikl et al., 2005; Nabel et al., 2012), and cytoplasmic localization (Etard et al., 2010) argue

against APOBEC2-mediated DNA demethylation. If APOBEC2 does function as a DNA demethylase in mammals, this activity must not be essential.

Alternatively, it is possible that mammalian APOBEC2 is not a cytidine deaminase at all, despite the clear conservation of its cytidine deaminase catalytic domain throughout evolution (Conticello, 2004). The lack of enzymatic activity *in vitro* and failure to find a substrate *in vivo* has led previous researchers in the field to suggest this possibility (Sato et al., 2010; Vonica et al., 2011), and the data presented here gives little reason to argue. APOBEC2 could have a structural function, or perhaps could catalyze deamination of a non-C substrate. Its enzymatic activity may also differ between species. While there is evidence that it serves a deamination-dependent function in lower vertebrates, it may have lost this activity in the mammalian lineage, which would explain the divergent phenotypes of APOBEC2-deficiency for zebrafish and mouse. Future studies of the proteins and/or nucleic acids that directly interact with APOBEC2 will likely be helpful in illuminating the function of the most obscure of the APOBECs.

5.5 DNA cytosine methylation in *T. brucei*

The failure of whole-genome bisulfite sequencing to identify methylated sites in the *T. brucei* genome can be explained in two nonexclusive ways: that flaws in the experimental or analytical techniques prevented detection of methylation, or that this methylation does not exist.

One clear issue with the WGBS data described here is the poor genomic coverage of AT-rich regions. It is possible that all of the methylated sites in the *T. brucei* genome are located in AT-rich areas, and thus were not identified because they were not covered. Improved coverage could be achieved in future studies

by use of PCR-free library preparation protocols. Biases inherent in PCR have been shown to be the driving force in AT-content bias in library preparation (Aird et al., 2011), and PCR-free library preparation strategies have been shown to dramatically correct this bias (Kozarewa et al., 2009; Miura et al., 2012).

Even if ideally distributed coverage were attained, the ability to discriminate between methylated and non-methylated sites depends on the depth of coverage. If methylation occurs at low frequency, extremely high levels of coverage would be required to identify the methylated sites. While 10x average coverage would be more than sufficient for identification of methylated sites for a mammalian-like methylation landscape, it is impossible to know what levels of coverage are required for *T. brucei* without some prior knowledge of the distribution of methylated sites.

Another complicating factor in the interpretation of WGBS data in *T. brucei* is the status of the genome. If methylated sites are located in regions that are difficult to assemble (such as repetitive sequences) or that differ between strains, these sites would be invisible to the strategy employed here. However, the extremely low frequency of C in reads that fail to map suggests that discordance between reference genome and genome of the sample is not the driving factor that prevented detection of methylation.

Finally, bisulfite sequencing in *T. brucei* is further complicated by the presence of beta-d-glucopyranosyloxymethyluracil, or base J, in genomic DNA. While there appears to be no study of the reactivity of bisulfite towards J, there is precedent for 5-oxidized pyrimidines forming stable adducts with bisulfite that are poor substrates for PCR (Huang et al., 2010). If J behaves similarly to 5hmC in

terms of bisulfite reactivity, coverage of J-rich regions may be impossible by bisulfite sequencing.

While there are many reasons why WGBS as deployed here could have failed to identify methylation, it is also possible that such methylation does not exist at all. The evidence that 5mC occurs in *T. brucei* DNA comes from two dot-blot experiments, which both found that trypanosome genomic DNA bound anti-5mC antibody more than *Dcm⁻ E. coli* genomic DNA (Militello et al., 2008) and (H.-S. Kim, unpublished data). While the 5mC antibody is clearly not reactive towards C or any other base found in mammalian DNA, it remains possible that it crossreacts with some metabolite or other base found in *T. brucei*. It is also known that the antibody used is reactive towards 5mC in RNA. Although an RNAse A-treated control was performed with indistinguishable levels of signal, it is possible that the treatment conditions were insufficient to remove contaminating methylated RNA.

Although no 5mC signal was observed in *Dcm⁻ E. coli* DNA, it is possible that the differences in DNA extraction procedures or chromosome sizes between the *E. coli* and *T. brucei* samples led to differences in background that were interpreted as positive signal. One such source of background could be the presence of serum. Although almost none of the reads that fail to map to the *T. brucei* genome map to the *B. taurus* genome, if such fragments are very small they may have been excluded from the library preparation process.

If it is the case that cytosine methylation is absent or very rare in *T. brucei*, it raises the question of what the identified putative CMT have as substrates. Because *T. brucei* has appreciable levels of 5mC in tRNA (Militello et al., 2014), it is likely that at least one of these CMTs acts on RNA. Techniques such as Aza-IP

(Khoddami and Cairns, 2013) could be a fruitful way to match these CMTs to their RNA substrates.

5.6 Closing remarks

Although the work presented here provides a number of directions for future study. Perhaps the most obvious space for such work is the gap between AID's apparent behavior in mouse B cells and iPS. Paired wild-type/knockout genome-scale methylation studies in cells undergoing reprogramming could be useful in assessing precisely how AID produces its methylation phenotype. *In vitro* characterization of AID from different species with respect to activity towards modified cytosines may also clarify the divergent phenotypes of AID deficiency in mammals and fish.

With respect to APOBEC2, the failure of pure sequencing methods to find a substrate suggests that looking first for interacting molecules may be a more fruitful approach. Mass spectrometry should allow clarification of APOBEC2's interacting proteins, and ChIP or CLIP would allow identification of interacting nucleic acids even if they are not themselves edited. *In vitro* experiments using Apobec2 from zebrafish and mice may reveal the reason for the large difference in its importance between species.

The *T. brucei* results reported here suggest a clear way forward for characterizing this organism's methylome: PCR-free WGBS. By solving the issue of uneven coverage, this technique should allow for a high-resolution map of DNA methylation, or make it clear that it does not exist. Once this map is established, the details of how methylation is regulated and what role it plays in *T. brucei* could be a rich field of study.

Chapter 6. Materials and methods

6.1 Mice

All C57BL/6 wild-type, *Aicda*$^{-/-}$ (Muramatsu et al., 2000), and AID-miR155T (Teng et al., 2008) were used at 6-8 weeks of age for B cell experiments. C57BL/6 wild-type and *Apobec2*$^{-/-}$ (provided by L. Chan, Baylor College of Medicine) were used at 10-12 days of age for primary myoblasts. All mice were bred and maintained under specific pathogen-free conditions at the Rockefeller University Animal Care Facility and all procedures involving mice were approved by The Rockefeller University Institutional Animal Care and Use Committee.

6.2 Cell culture

6.2.1 B cells

Mice were euthanized by cervical dislocation and spleens removed. The spleens were crushed with a syringe plunger in a cell strainer in a 6cm dish with 5 mL cold 1x PBS, and the flowthrough pipetted vigorously and spun at 1500 rpm for 4′ at 4ºC. The resulting pellet was resuspended in 1 mL 0.16M NH_4Cl in water and incubated 5′ at RT to lyse red blood cells. This solution was underlaid with 1 mL FBS and spun 1500 rpm for 5′ at 4C. The resulting pellet was washed twice with 1x PBS with 10% FBS, resuspended in 2.5 mL PBS with 10% FBS, and counted with a hemocytometer. Cells were then resuspended in 90 μL PBS with 0.5% BSA and 10ul anti-CD43 MACS beads (Miltenyi Biotec) per 10^7 cells. This mixture was incubated 15′ in refrigerator, then applied to an LS magnetic column (Miltenyi Biotec) that had been prewashed with 3mL 1x PBS with 0.5% BSA and

2mM EDTA. The column was eluted with 3.5 mL 1x PBS with 0.5% BSA and 2mM EDTA, and the cell count in the eluent determined with a hemocytometer. The cells were then centrifuged and resuspended at 10^6 cells per mL in pre-heated RPMI with glutamine (Gibco) supplemented with 1x Pen/strep (Life Technologies), 2 mM L-glutamine (Life Technologies) 10% FBS, 50 uM 2-mercaptoethanol, 5 ng/mL IL-4 (Sigma), 1ul/mL monoclonal IgM a-CD40 (Ebiosciences, clone HM40-3), and 25 μg/mL LPS (Sigma). The cells were cultured in 6-well plates at 37C and 5% CO_2. Prior harvesting of nucleic acids for mRNA-Seq and RRBS, dead cells were removed by resuspending 10^7 cells in 100 μL dead cell removal beads (Miltenyi), incubating 15' at RT, adding 1 mL dead cell binding buffer, then applying to an LS column (Miltenyi) that had been prewashed with 3 mL dead cell binding buffer. The columns were then washed with 12 mL binding buffer and split into thirds for flow cytometry or lysis for DNA or RNA.

6.2.2 Primary myoblasts

Congenic 10-12 day-old C56BL/6 wild-type and *Apobec2*$^{-/-}$ pairs were euthanized by decapitation and muscles of the hind leg dissected. The muscles were pulped with scalpels in 1.25 mL of a collagenase/dispase solution (2.4 U/mL dispase II [Roche], 1% collagenase B [Roche], 2.5 mM $CaCl_2$) in a 35mm dish. This suspension was incubated 24' at 37C and 5% CO_2, pipetting thoroughly every 12'. PBS (2 mL) was added to the resulting slurry and filtered through a cell strainer inside of a 50 mL conical flask that had been pre-wet with 1 mL 1x PBS. A further 7 mL PBS was used to wash the plate and strainer, and

the flowthrough moved to a 15 mL conical. This suspension was then pelleted (5', 2000 rpm, 4C) and resuspended in 3 mL growth media (Ham's/F10 media [Gibco] supplemented with 20% FBS, 2x Pen/strep [Life Technologies], 2.5 ng/mL bFGF, 10 ng/mL HGF, and 5 ng/mL heparin sulfate [Sigma]). The cells were then plated onto an uncoated 60mm tissue culture dish and incubated at 37C and 5% CO2 for 3h to allow fibroblasts to settle. The supernatant was then moved to a collagen-coated 60mm dish, and the replating procedure repeated the following day. Cells were expanded for 3 weeks and differentiated at 70% confluency by changing to differentiation media (DMEM [Gibco] supplemented with 1x Pen/strep and 5% horse serum [Gibco]).

6.3 Flow cytometry

For flow cytometric analysis of B cells, cells were pelleted (5000 rcm, 4C, 30"), resuspended in 3 μL 7-AAD solution (BD), and incubated on ice 5'. The cells were then stained with 100 μL of FACS buffer (1x PBS with 1% FBS) along with 1:200 IgG1-PE (BD) and 1:400 B220-APC (BD). The cells were incubated 5' on ice, washed twice with 500 μL cold FACS buffer, and resuspended in 300 μL FACS buffer. For wild-type, *Aicda*$^{-/-}$, and AID-miR155T B cells used for mRNA-Seq and RRBS, data was acquired on a FACSCalibur flow cytometer (BD). For retrovirally-complemented *Aicda*$^{-/-}$ B cells used for miRNA-Seq, data was acquired on a FACSAria cell sorter (BD), and singlet, 7-AAD^{-}, B220^{+}, GFP^{+} cells were sorted into culture tubes coated with FBS.

6.4 Retroviral infection

For preparation of retroviral particles, 12 μg pCL-Eco and 12 μg pMX-GFP or pMX-AID-IRES-eGFP in 1.4 mL OptiMEM was mixed with 60 μL lipofectamine 2000 (Life Technologies) in 1.6 mL OptiMEM and incubated 20' at RT. This mixture was added dropwise to a 10cm dish of 90% confluent 293T cells in DMEM (Gibco) with 10% FBS and 1x Pen/strep, and the cells then moved to an incubator. After 6h, media was changed to virus collection media (IMDM (Gibco) with 5% FBS and 1x Pen/strep). Media was changed and discarded at 24h after transfection, changed and collected at 48h, and collected at 72h. The collected supernatant was spun 2' at 1000 rcm to remove debris, filtered through a 45 um syringe filter, and stored at -80C until use.

Aicda$^{-/-}$ B cells in culture for 1d were infected with retroviral supernatant supplemented with 8 μg /mL polybrene (1 mL retroviral supernatant per 10^6 initial B cells), and spun 800 rcm, 10C, 2h. Cells were moved back to the incubator without changing media, and GFP expression was monitored by microscopy.

6.5 Generation of RNA spikes

ERCC spikes were the gift of C. Mason (MSKCC). VSG spikes were generated by K. Lay from pGEM-VSG constructs (contributed by G. Hovel-Miner), and VSG* constructs, in which single C>T changes were introduced by mutagenesis PCR using Phusion polymerase. These constructs were linearized with SalI, gel-purified, and transcribed using the Mega T7 *In Vitro* Transcription

kit (Ambion). The resulting transcripts were precipitated, resuspended, and quantified using a Bioanalyzer RNA chip (Agilent).

6.6 Generation of sequencing libraries

6.6.1 mRNA-Seq

For preparation of libraries from B cells and primary myoblasts, RNA was harvested by lysing cells in Trizol (Invitrogen) according to the manufacturer's instructions, using 1 mL for 2 x 10^6 B cells or a 6 cm dish of myoblasts. Libraries were then prepared according to a protocol adapted from (Rosenberg et al., 2011). Total RNA (10 μg) was selected twice with Sera-Mag oligo(dT) beads, and for the B cell libraries, ERCC (0.5 μL of 1:100 dilution per sample) and VSG spikes (50-300 pg/sample) were added. RNA was then chemically fragmented to a mode size of about 200 nt (4' at 94C in 30 mM Mg^{2+}), and precipitated. The fragmented RNA was then reverse-transcribed with Superscript III (Invitrogen) and random hexamer primers, and made double stranded with a mix of *E. coli* DNA polymerase I, *E. coli* DNA ligase, and RNAse H (all NEB). The ends of the double-stranded cDNA were then repaired with T4 DNA polymerase and T4 polynucloetide kinase and adenylated with Klenow exo^- (all NEB). These products were then ligated to Illumina TruSeq DNA adapters with T4 DNA Quickligase (NEB), and size-selected on a 2% agarose gel, taking fragments in the range 300-350 bp. These fragments were then enriched by PCR using Phusion DNA polymerase (NEB) for 15 cycles. Following validation using a Bioanalyzer Pico chip (Agilent), the libraries were sequenced (100-cycle, single-end) on an Illumina HiSeq 2000.

For preparation of libraries from muscle tissue, muscles were dissected and put immediately into 1.5 mL Trizol, then homogenized with a 5 mm stainless steel bead in a Qiagen TissueLyser LT at 50 Hz for 10'. Ribosomal RNA was depleted with the Ribo-Zero Gold Magnetic Human/Mouse/Rat kit (Epicentre) according to the manufacturer's protocol. Directional RNA-Seq libraries were then prepared according to the manufacturer's protocol using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina, along with Illumina TruSeq DNA Adapter oligos, and sequenced on an Illumina HiSeq 2000 (100-cycle, single-end).

6.6.2 miRNA-Seq

Libraries were prepared by A. Mihailovic according to a published protocol (Hafner et al., 2012). Sequencing was performed on an Illumina HiSeq 2000 sequencer (50-cycle, single-end).

6.6.3 Reduced-representation bisulfite sequencing

Libraries were prepared according to a protocol adapted from (Gu et al., 2011a). DNA was harvested from 5 x 10^6 B cells per genotype with the DNeasy Blood and Tissue kit (Qiagen), according to the manufacturer's instructions. The resulting genomic DNA (500 ng per sample) was digested with 2 µL MspI (NEB) for 18h and purified by phenol/chloroform extraction. Following end repair and adenylation, Illumina methylated adapter oligos were ligated to the products. Gel purification for products of size 200-350 bp was followed by two rounds of bisulfite conversion using the Qiagen Epitect kit. The resulting fragments were

enriched by PCR with Pfu Cx Hotstart polymerase (Agilent) and sequenced at low density on an Illumina HiSeq 2000 (50-cycle, single end).

6.6.4 Whole genome bisulfite sequencing

Whole genome bisulfite sequencing libraries were prepared according to a protocol adapted from (Johnson et al., 2012). Genomic DNA (3.2 μg per sample) was spiked 1:1000 with unmethylated phage DNA (Promega) and fragmented to an average size of 300 bp using a Covaris S2 sonicator. The fragmented DNA was then end-repaired, adenylated, and quantified using the Quant-iT reagents (Life Technologies) and a FLUOstar Omega fluorescence plate reader. Illumina TruSeq DNA Adapters were then ligated at a 10:1 molar ratio, and 2 rounds of bisulfite conversion were performed using the Qiagen Epitect kit. Resulting fragments were amplified with Pfu Turbo Cx Hotstart polymerase and size selected on a 2% agarose gel to yield fragments of size 375-475 bp. Resulting products were quantified with Quant-iT reagents and sequenced on an Illumina HiSeq 2000 (50 cycle, single-end).

6.7 Sequencing data analysis

6.7.1 mRNA-Seq

Following quality control inspection with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), reads were aligned to the Ensembl reference transcriptome (release 63) supplemented with *Igh* transcripts with $J_{1\text{-}4}$-C and I-C junctions explicitly added for each isotype, and then to the NCBI37 reference genome supplemented with ERCC and VSG

sequences. This was done using Tophat v2.0.3 (Kim et al., 2009a) and with the parameters "--b2-sensitive --no-novel-juncs". Values for gene expression were calculated with Cuffdiff v.2.0.2 (Trapnell et al., 2010), using the Ensembl gene set supplemented with *Igh* transcripts and a masking file with annotated mitochondrial genes, rRNA and tRNA genes and pseudogenes. Values for Ig isotype abundance were calculated by dividing the sum of the FPKM values for $J_{1\text{-}4}$-C_x for a given isotype by the sum of the FPKM values for all $J_{1\text{-}4}$-C isoforms.

For V_H segment usage analysis, all mouse Ig segments listed in IMGT (Lefranc et al., 2009) were first aligned to chromosome 12 with Bowtie v0.12.7 (Langmead et al., 2009). These alignments were then curated by manually inspected with IGV (Robinson et al., 2011).

6.7.2 miRNA-Seq

Adapters were removed from sequence data using the script fastx_barcode_splitter.pl and two bases were trimmed from the 3′ end using fastx_trimmer, both from the FASTX-toolkit suite (http://hannonlab.cshl.edu/fastx_toolkit/). Trimmed reads were aligned with Bowtie v0.12.7 against NCBI37 using the parameters "-l 15 -v 2 --best --strata -m 1". The resulting hits that overlapped with sequences annotated in miRbase (Kozomara and Griffiths-Jones, 2014) release 18 were then quantified using the tool Seqmonk (http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/).

6.7.3 RRBS

Following quality control inspection with FastQC, adaptors were sequentially trimmed from raw reads by using the program TrimGalore v0.2.2 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with the parameters "-a AGATCGGAAGAGC", then "-s 8 -a CGGTTCAG", then "-s 8 AGCAGGAA". Trimmed reads were aligned to the mouse genome (NCBI37) using Bismark v0.7.4 (Krueger and Andrews, 2011) with Bowtie v0.12.7 with parameters "-l 20". Initial methylation counts were found using Seqmonk, and statistical analyses were performed using R. Each position was used for analysis if it was covered at least 10x for all of the samples, and features were used if they contained at least 3 such positions. Features were defined using the Ensembl reference. Promoters were defined as -5 kb to +1 kb from the reference TSS.

6.7.4 WGBS

Reads were aligned using Bismark v0.7.7 and Bowtie v0.12.7 with the parameters "-q -n 1 -k 2 --best --chunkmbs 512". A custom genome was used for alignment, consisting of the following sequences: Tb427 v4 for chromosomes 1-9 and 11, Tb427 v5 for chromosome 10, a composite BES site ("chromosome 12"), a curated set of VSG and repetitive sequences (G. Cross), the phage strain J02459_1 genome, the phiX174 genome used for Illumina quality control, and the Illumina adaptor sequences. The number of C and T basecalls at every genomic C was then counted using Seqmonk, and bisulfite conversion efficiency calculated as the percentage of apparently methylated basecalls mapping to the phage genome. Candidate methylated positions were then identified by applying a

binomial test with chance of success set at the bisulfite conversion efficiency for that sample (in R, with the function binom.test). The resulting p values were adjusted for multiple testing using the Benjamini-Hochberg method (in R, with the function p.adjust with method="BH"), and sites with $q<0.05$ in any sample were taken as candidates.

6.8 Epityper assays

Primers for use in the Epityper assay were designed using the Epidesigner tool (www.epidesigner.com). The assays were performed by the Weill Cornell Medical College Epigenomics Core.

6.9 Targeted bisulfite sequencing

Genomic DNA from primary myoblasts was bisulfite converted with the Qiagen Epitect kit, then amplified with Hotstar Taq polymerase (Qiagen). Aliquots of each reaction were electrophoresed on an agarose gel to verify amplification, and then cloned using the Strataclone PCR cloning kit (Agilent) according to the manufacturer's protocol. The ligation reaction was transformed into Strataclone Solopack chemically competent *E. coli* (Agilent) and plated on LB-agarose with ampicillin and X-gal. White colonies (10 per plate) were sequenced with the T3 primer by Genewiz. Resulting sequences were analyzed using the QuMA tool (Kumaki et al., 2008).

6.10 Immunofluorescence imaging

Primary myoblasts in 35 mm dishes were fixed in PBS with 4% formaldehyde and blocked in PBS with 1% BSA, 1% gelatin, 5% normal goat

serum, and 0.1% Triton-X. Cells were then stained in blocking buffer with 1:10 mouse monoclonal a-MyHC (clone MF-20, DSHB, supernatant) and 1:50 rabbit polyclonal a-MyoD (M-318, SC760, Santa Cruz). After washing 3 times with PBS, secondary stain was performed in blocking buffer with 1:1000 goat a-rabbit Alexa Fluor 546, 1:1000 goat a-mouse Alexa Fluor 488, and 1:200 phalloidin-Alexa Fluor 657. After 3 washes, cells were incubated with 1:1000 Hoescht in PBS, and a cover slip was mounted using Prolong Gold (Life). Images were acquired using a Zeiss Wide-field fluorescence/brightfield/DIC microscope.

6.11 Western blotting

Western blotting of primary myoblasts was performed by B. Rosenberg. One 60mm dish per condition was rinsed in 1x PBS, trypsinized, quenched, and pelleted before flash-freezing and storage at -80C. These pellets were lysed in 60 μL RIPA buffer with cOmplete mini protease inhibitor (Roche) and PMSF for 20 minutes at 4C and centrifuged at 12 krpm for 20 minutes at 4C. Following quantification of the supernatant using the Bradford assay, for each of two blots 10 μg protein was diluted 1:1 with 2x Laemmli buffer with 200 mM DTT and incubated for 5 minutes at 95C. These samples were then electrophoresed on a Criterion 12.5% Tris-HCl gel (Bio-Rad) at 150V. Following wet transfer to PVDF membrane (100V, 30 minutes) and blocking (PBS with 0.1% Tween and 5% milk, 2.5 hours at 4C), the following primary incubations were performed: 1:1000 of polyclonal rabbit anti-APOBEC2 (gift of Alin Vonica) in PBS with 0.1% Tween and 5% milk overnight at 4C; 1:250 of polyclonal rabbit anti-MyoD1 (Santa Cruz, M-318) 1 hour at room temperature. Following 4 washes (5 minutes, PBS with

0.1% Tween, room temperature), secondary incubation was performed with RG16 anti-rabbit HRP at 1:20000 in PBS with 0.1% Tween and 1% milk for 1 hour at room temperature. The membrane was then washed twice in PBS and visualized by ECL (1 minute exposure for APOBEC2, 4 minute exposure for MyoD1).

6.12 Methylated DNA immunoprecipitation

Immunoprecipitation of methylated DNA was performed according to a protocol adapted from (Mohn et al., 2009). Genomic DNA was fragmented to an average size of 500 bp using a Covaris S2 sonicator, then precipitated. After setting aside input samples, DNA was immunoprecipitated using Dynabeads sheep a-mouse IgG magnetic beads and mouse monoclonal a-5mC (Eurogentec clone 33D3) or mouse IgG1 isotype control antibody (Cell Signaling). DNA was digested from the beads with proteinase K and phenol/chloroform extracted. Quantitative real-time PCR was then performed using Sybr Green master mix (Life Technologies) and an ABI 7900 thermocycler.

6.13 Primer sequences

Below are the sequences of the primers used in the experiments described.

Illumina amplification primers (asterisk denotes a phosporothioate linkage):
TS1.0: AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACG*A
TS2.0: CAAGCAGAAGACGGCATACGAGA*T

Targeted bisulfite sequencing primers (Gtl2 primers from (Sato et al., 2011), Kcnq1 primers from (Rivera et al., 2008)):

Gtl2-1F: TGTGTTGTGGATTTAGGTTGTAGTTTA

Gtl2-1R: TAATCCCATTCCCAATCTATAAAAATA

Gtl2-2F: CCAAAACAAACCCAATAAATTCTAA

Gtl2-2R: TGGTGAGTTTTGGTTAGAAAAGTGT

Gtl2-3F: CCCCCAATAACTTATAAACCATAATACT

Gtl2-3R: GGATGGTAGTAGATAATTTGTTGTTTGA

Gtl2-4F: AAATCAAAATCCTTTTACCTCAACAATA

Gtl2-4R: GGAAATAATTTTAATTGGTGATTGTTTT

Gtl2-5F: AAATTTTGTAAGGAAAAGAATTTTTAGG

Gtl2-5R: TTCAAAATTACTAATCAACATAAACCTC

Kcnq1-outer-F: AGTGTTTGTTTTGAGTTTAGAT

Kcnq1-outer-R: CCTCAAAACCACCCCTACTTC

Kcnq1-inner-F: GATTTTTATGGTGAGGTTTTA

Kcnq1-inner-R: CAAAACCACCCCTACTTCTAT

MeDIP primers (from (Weber et al., 2005)):

177-1 F: GCGCAGTTAACGCTATTATAC

177-1 R: CTTTGTTGCACACATTAAACAC

177-4 F: GTGCAACAAAGCTAATAAATGGTTC

177-4 R: CACACTTGTATTTAATGTTGCACACTTG

Nc-1 F: TGAAATACTTAGGGTGACGGATG

Nc-1 R: ATCCCTCTCCTCAACACAAATC

Nc-2 F: TGTACGTGTCTGCTCGTTTG

Nc-2 R: TTCGGGTGGAGTCGGAA

H19 F: GCATGGTCCTCAAATTCTGCA

H19 R: GCATCTGAACGCCCCAATTA

Actb F: AGCCAACTTTACGCCTAGCGT

Actb R: TCTCAAGATGGACCTAATACGGC

References

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. *12*, R18.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Alon, S., Mor, E., Vigneault, F., Church, G., Locatelli, F., Galeano, F., Gallo, A., Shomron, N., and Eisenberg, E. (2012). Systematic identification of edited microRNAs in the human brain. Genome Res. *22*, 1533–1540.

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.

Barras, F., and Marinus, M.G. (1989). The great GATC: DNA methylation in E. coli. Trends Genet. *5*, 139–143.

Barreto, G., Schäfer, A., Marhold, J., Stach, D., Swaminathan, S.K., Handa, V., Döderlein, G., Maltry, N., Wu, W., Lyko, F., et al. (2007). Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation. Nature *445*, 671–675.

Bartolomei, M.S., and Ferguson-Smith, A.C. (2011). Mammalian Genomic Imprinting. Cold Spring Harb. Perspect. Biol. *3*, a002592–a002592.

Bascove, M., and Frippiat, J.-P. (2010). Molecular characterization of Pleurodeles waltl activation-induced cytidine deaminase☆. Mol. Immunol. *47*, 1640–1649.

Bazak, L., Haviv, A., Barak, M., Jacob-Hirsch, J., Deng, P., Zhang, R., Isaacs, F.J., Rechavi, G., Li, J.B., Eisenberg, E., et al. (2014). A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. Genome Res. *24*, 365–376.

Benne, R., Van Den Burg, J., Brakenhoff, J.P.J., Sloof, P., Van Boom, J.H., and Tromp, M.C. (1986). Major transcript of the frameshifted coxll gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell *46*, 819–826.

Bhattacharya, S.K., Ramchandani, S., Cervoni, N., and Szyf, M. (1999). A mammalian protein with specific demethylase activity for mCpG DNA. Nature *397*, 579–583.

Bhutani, N., Decker, M.N., Brady, J.J., Bussat, R.T., Burns, D.M., Corbel, S.Y., and Blau, H.M. (2013). A critical role for AID in the initiation of reprogramming to

induced pluripotent stem cells. Faseb J. *27*, 1107–1113.

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.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. *16*, 6–21.

Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. Nucl. Acids Res. *8*, 1499–1504.

Bock, C., Beerman, I., Lien, W.-H., Smith, Z.D., Gu, H., Boyle, P., Gnirke, A., Fuchs, E., Rossi, D.J., and Meissner, A. (2012). DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. Mol. Cell *47*, 633–647.

Bracht, J.R., Perlman, D.H., and Landweber, L.F. (2012). Cytosine methylation and hydroxymethylation mark DNA for elimination in Oxytricha trifallax. Genome Biol. *13*, R99.

Bransteitter, R., Pham, P., Scharff, M.D., and Goodman, M.F. (2003). Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. Proc. Natl. Acad. Sci. USA *100*, 4102–4107.

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

Burns, C.M., Chu, H., Rueter, S.M., Hutchinson, L.K., Canton, H., Sanders-Bush, E., and Emeson, R.B. (1997). Regulation of serotonin-2C receptor G-protein coupling by RNA editing. Nature *387*, 303–308.

Burns, M.B., Lackey, L., Carpenter, M.A., Rathore, A., Land, A.M., Leonard, B., Refsland, E.W., Kotandeniya, D., Tretyakova, N., Nikas, J.B., et al. (2013a). APOBEC3B is an enzymatic source of mutation in breast cancer. Nature *494*, 366–370.

Burns, M.B., Temiz, N.A., and Harris, R.S. (2013b). Evidence for APOBEC3B mutagenesis in multiple human cancers. Nat. Genet. *45*, 977–983.

Cedar, H., and Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. Nat. Rev. Genet. *10*, 295–304.

Chaudhuri, J., Khuong, C., and Alt, F.W. (2004). Replication protein A interacts with AID to promote deamination of somatic hypermutation targets. Nature *430*, 992–998.

Chaudhuri, J., Tian, M., Khuong, C., Chua, K., Pinaud, E., and Alt, F.W. (2003). Transcription-targeted DNA deamination by the AID antibody diversification enzyme. Nature *422*, 726–730.

Chen, H., Lilley, C.E., Yu, Q., Lee, D.V., Chou, J., Narvaiza, I., Landau, N.R., and Weitzman, M.D. (2006). APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons. Curr. Biol. *16*, 480–485.

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.

Chiarle, R., Zhang, Y., Frock, R.L., Lewis, S.M., Molinie, B., Ho, Y.-J., Myers, D.R., Choi, V.W., Compagno, M., Malkin, D.J., et al. (2011). Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. Cell *147*, 107–119.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature *452*, 215–219.

Conticello, S.G. (2004). Evolution of the AID/APOBEC Family of Polynucleotide (Deoxy)cytidine Deaminases. Mol. Biol. Evol. *22*, 367–377.

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

Cooke, M.S., Evans, M.D., Dizdaroglu, M., and Lunec, J. (2003). Oxidative DNA damage: mechanisms, mutation, and disease. Faseb J. *17*, 1195–1214.

Cortellino, S., Xu, J., Sannai, M., Moore, R., Caretti, E., Cigliano, A., Le Coz, M., Devarajan, K., Wessels, A., Soprano, D., et al. (2011). Thymine DNA Glycosylase Is Essential for Active DNA Demethylation by Linked Deamination-Base Excision Repair. Cell *146*, 67–79.

Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561–563.

Danoviz, M.E., and Yablonka-Reuveni, Z. (2011). Skeletal Muscle Satellite Cells: Background and Methods for Isolation and Analysis in a Primary Culture System. (Totowa, NJ: Myogenesis: Methods and Protocols), pp. 21–52.

Decatur, W.A., and Fournier, M.J. (2002). rRNA modifications and ribosome function. Trends Biochem. Sci. *27*, 344–351.

Delebecque, F., Suspène, R., Calattini, S., Casartelli, N., Saïb, A., Froment, A., Wain-Hobson, S., Gessain, A., Vartanian, J.-P., and Schwartz, O. (2006). Restriction of foamy viruses by APOBEC cytidine deaminases. J. Virol. *80*, 605–614.

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. Annu. Rev. Biochem. *76*, 1–22.

Dickerson, S.K., Market, E., Besmer, E., and Papavasiliou, F.N. (2003). AID mediates hypermutation by deaminating single stranded DNA. J. Exp. Med. *197*, 1291–1296.

Ehrlich, M., Gama-Sosa, M.A., Huang, L.-H., Midgett, R.M., Kuo, K.C., McCune, R.A., and Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. Nucl. Acids Res. *10*, 2709–2721.

Eisenberg, E., Li, J.B., and Levanon, E.Y. (2010). Sequence based identification of RNA editing sites. RNA Biol. *7*, 248–252.

Elango, N., Hunt, B.G., Goodisman, M.A.D., and Yi, S.V. (2009). DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, Apis mellifera. Proc. Natl. Acad. Sci. USA *106*, 11206–11211.

Esnault, C., Heidmann, O., Delebecque, F., Dewannieux, M., Ribet, D., Hance, A.J., Heidmann, T., and Schwartz, O. (2005). APOBEC3G cytidine deaminase inhibits retrotransposition of endogenous retroviruses. Nature *433*, 430–433.

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

Falnes, P.Ø., Johansen, R.F., and Seeberg, E. (2002). AlkB-mediated oxidative demethylation reverses DNA damage in Escherichia coli. Nature *419*, 178–182.

Feldman, N., Gerson, A., Fang, J., Li, E., Zhang, Y., Shinkai, Y., Cedar, H., and Bergman, Y. (2006). G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. Nat. Cell Biol. *8*, 188–194.

Ferguson-Smith, A.C. (2011). Genomic imprinting: the emergence of an epigenetic paradigm. Nat. Rev. Genet. *12*, 565–575.

Foshay, K.M., Looney, T.J., Chari, S., Mao, F.F., Lee, J.H., Zhang, L., Fernandes, C.J., Baker, S.W., Clift, K.L., Gaetz, J., et al. (2012). Embryonic stem cells induce pluripotency in somatic cell fusion through biphasic reprogramming. Mol. Cell *46*, 159–170.

Fritz, E.L., and Papavasiliou, F.N. (2010). Cytidine deaminases: AIDing DNA demethylation? Genes Dev. *24*, 2107–2114.

Fritz, E.L., Rosenberg, B.R., Lay, K., Mihailovic, 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.

Gao, F., Liu, X., Wu, X.-P., Wang, X.-L., Gong, D., Lu, H., Xia, Y., Song, Y., Wang, J., Du, J., et al. (2012). Differential DNA methylation in discrete developmental stages of the parasitic nematode Trichinella spiralis. Genome Biol. *13*, R100.

Garrett, S., and Rosenthal, J.J.C. (2012). RNA Editing Underlies Temperature Adaptation in K+ Channels from Polar Octopuses. Science *335*, 848–851.

Grosjean, H. (2009). DNA and RNA modification enzymes (Austin, Texas: Landes Bioscience).

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

Gu, T.-P., Guo, F., Yang, H., Wu, H.-P., Xu, G.-F., Liu, W., Xie, Z.-G., Shi, L., He, X., Jin, S.-G., et al. (2011b). The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. Nature *477*, 606–610.

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

Habib, O., Habib, G., Do, J.T., Moon, S.-H., and Chung, H.-M. (2014). Activation-Induced Deaminase-Coupled DNA Demethylation Is Not Crucial for the Generation of Induced Pluripotent Stem Cells. Stem Cells Dev. *23*, 209–218.

Hackett, J.A., Sengupta, R., Zylicz, J.J., Murakami, K., Lee, C., Down, T.A., and Surani, M.A. (2013). Germline DNA Demethylation Dynamics and Imprint Erasure Through 5-Hydroxymethylcytosine. Science *339*, 448–452.

Hafner, M., Renwick, N., Farazi, T.A., Mihailovic, A., Pena, J.T.G., and Tuschl, T. (2012). Barcoded cDNA library preparation for small RNA profiling by next-generation sequencing. Methods *58*, 164–170.

Hajkova, P., Jeffries, S.J., Lee, C., Miller, N., Jackson, S.P., and Surani, M.A. (2010). Genome-Wide Reprogramming in the Mouse Germ Line Entails the Base Excision Repair Pathway. Science *329*, 78–82.

Hajkova, P., Erhardt, S., Lane, N., Haaf, T., El-Maarri, O., Reik, W., Walter, J., and Surani, M.A. (2002). Epigenetic reprogramming in mouse primordial germ cells. Mech. Dev. *117*, 15–23.

Hamilton, C.E., Papavasiliou, F.N., and Rosenberg, B.R. (2010). Diverse functions for DNA and RNA editing in the immune system. RNA Biol. *7*, 220–228.

Hardeland, U., Bentele, M., Jiricny, J., and Schär, P. (2003). The versatile thymine DNA-glycosylase: a comparative characterization of the human, Drosophila and fission yeast orthologs. Nucl. Acids Res. *31*, 2261–2271.

Harris, R.A., Wang, T., Coarfa, C., Nagarajan, R.P., Hong, C., Downey, S.L.,

Johnson, B.E., Fouse, S.D., Delaney, A., Zhao, Y., et al. (2010). Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat. Biotechnol. 1–12.

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.

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. Mol. Cell *10*, 1247–1253.

Hartner, J.C., Walkley, C.R., Lu, J., and Orkin, S.H. (2009). ADAR1 is essential for the maintenance of hematopoiesis and suppression of interferon signaling. Nat. Immunol. *10*, 109–115.

Hattori, N., Nishino, K., Ko, Y.-G., Hattori, N., Ohgane, J., Tanaka, S., and Shiota, K. (2004). Epigenetic control of mouse Oct-4 gene expression in embryonic stem cells and trophoblast stem cells. J. Biol. Chem. *279*, 17063–17069.

Heyn, H., Li, N., Ferreira, H.J., Moran, S., Pisano, D.G., Gomez, A., Diez, J., Sanchez-Mut, J.V., Setien, F., Carmona, F.J., et al. (2012). Distinct DNA methylomes of newborns and centenarians. Proc. Natl. Acad. Sci. USA *109*, 10522–10527.

Higuchi, M., Single, F.N., Köhler, M., Sommer, B., Sprengel, R., and Seeburg, P.H. (1993). RNA editing of AMPA receptor subunit GluR-B: A base-paired intron-exon structure determines position and efficiency. Cell *75*, 1361–1370.

Hogenbirk, M.A., Heideman, M.R., Velds, A., van den Berk, P.C., Kerkhoven, R.M., van Steensel, B., and Jacobs, H. (2013). Differential Programming of B Cells in AID Deficient Mice. PLoS ONE *8*, e69815.

Holliday, R., and Pugh, J.E. (1975). DNA modification mechanisms and gene activity during development. Science *187*, 226–232.

Hoopengardner, B., Bhalla, T., Staber, C., and Reenan, R. (2003). Nervous system targets of RNA editing identified by comparative genomics. Science *301*, 832–836.

Hotchkiss, R.D. (1948). THE QUANTITATIVE SEPARATION OF PURINES, PYRIMIDINES, AND NUCLEOSIDES BY PAPER CHROMATOGRAPHY. J. Biol. Chem. 315–332.

Howell, C.Y., Bestor, T.H., Ding, F., Latham, K.E., Mertineit, C., Trasler, J.M., and Chaillet, J.R. (2001). Genomic Imprinting Disrupted by a Maternal Effect Mutation in the Dnmt1 Gene. Cell *104*, 829–838.

Howlett, S.K., and Reik, W. (1991). Methylation levels of maternal and paternal genomes during preimplantation development. Development *113*, 119–127.

Hsieh, C.-L. (2000). Dynamics of DNA methylation pattern. Curr. Opin. Gen. Dev. *10*, 224–228.

Huang, Y., Pastor, W.A., Shen, Y., Tahiliani, M., Liu, D.R., and Rao, A. (2010). The Behaviour of 5-Hydroxymethylcytosine in Bisulfite Sequencing. PLoS ONE *5*, e8888.

Illingworth, R.S., and Bird, A.P. (2009). CpG islands – "A rough guide." FEBS Lett. *583*, 1713–1720.

Inoue, A., and Zhang, Y. (2011). Replication-Dependent Loss of 5-Hydroxymethylcytosine in Mouse Preimplantation Embryos. Science *334*, 194–194.

Iurlaro, M., Ficz, G., Oxley, D., Raiber, E.-A., Bachman, M., Booth, M.J., Andrews, S., Balasubramanian, S., and Reik, W. (2013). A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. Genome Biol. *14*, R119.

Jackman, J.E., and Alfonzo, J.D. (2012). Transfer RNA modifications: nature's combinatorial chemistry playground. WIREs RNA *4*, 35–48.

Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., and Oliver, B. (2011a). Synthetic spike-in standards for RNA-seq experiments. Genome Res. *21*, 1543–1551.

Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., and Oliver, B. (2011b). Synthetic spike-in standards for RNA-seq experiments. Genome Res. *21*, 1543–1551.

Johnson, M.D., Mueller, M., Game, L., and Aitman, T.J. (2012). Single Nucleotide Analysis of Cytosine Methylation by Whole-Genome Shotgun Bisulfite Sequencing (Hoboken, NJ, USA: Current Protocols in Molecular Biology).

Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. *13*, 484–492.

Jost, J.P. (1993). Nuclear extracts of chicken embryos promote an active demethylation of DNA by excision repair of 5-methyldeoxycytidine. Proc. Natl. Acad. Sci. USA *90*, 4684–4688.

Kahramanoglou, C., Prieto, A.I., Khedkar, S., Haase, B., Gupta, A., Benes, V., Fraser, G.M., Luscombe, N.M., and Seshasayee, A.S.N. (2012). Genomics of DNA cytosine methylation in Escherichia coli reveals its role in stationary phase transcription. Nat. Comms. *3*.

Kaneda, M., Okano, M., Hata, K., Sado, T., Tsujimoto, N., Li, E., and Sasaki, H. (2004). Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature *429*, 900–903.

Kangaspeska, S., Stride, B., Métivier, R., Polycarpou-Schwarz, M., Ibberson, D., Carmouche, R.P., Benes, V., Gannon, F., and Reid, G. (2008). Transient cyclical methylation of promoter DNA. Nature *452*, 112–115.

Karijolich, J., and Yu, Y.-T. (2010). Spliceosomal snRNA modifications and their function. RNA Biol. *7*, 192–204.

Kawahara, Y., Zinshteyn, B., Sethupathy, P., Iizasa, H., Hatzigeorgiou, A.G., and Nishikura, K. (2007). Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. Science *315*, 1137–1140.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res. *12*, 656–664.

Khoddami, V., and Cairns, B.R. (2013). Identification of direct targets and modified bases of RNA cytosine methyltransferases. Nat. Biotechnol. *31*, 458–464.

Kim, D., and Salzberg, S.L. (2011). TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol. *12*, R72.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2009a). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *10*, R25.

Kim, M.-S., Kondo, T., Takada, I., Youn, M.-Y., Yamamoto, Y., Takahashi, S., Matsumoto, T., Fujiyama, S., Shirode, Y., Yamaoka, I., et al. (2009b). DNA demethylation in hormone-induced transcriptional derepression. Nature *461*, 1007–1012.

Klein, I.A., Resch, W., Jankovic, M., Oliveira, T., Yamane, A., Nakahashi, H., Di Virgilio, M., Bothmer, A., Nussenzweig, A., Robbiani, D.F., et al. (2011). Translocation-Capture Sequencing Reveals the Extent and Nature of Chromosomal Rearrangements in B Lymphocytes. Cell *147*, 95–106.

Klug, M., Schmidhofer, S., Gebhard, C., Andreesen, R., and Rehli, M. (2013). 5-Hydroxymethylcytosine is an essential intermediate of active DNA demethylation processes in primary human monocytes. Genome Biol. *14*, R46.

Kobayashi, M., Aida, M., Nagaoka, H., Begum, N.A., Kitawaki, Y., Nakata, M., Stanlie, A., Doi, T., Kato, L., Okazaki, I.-M., et al. (2009). AID-induced decrease in topoisomerase 1 induces DNA structural alteration and DNA cleavage for class switch recombination. Proc. Natl. Acad. Sci. USA *106*, 22375–22380.

Kobayashi, M., Sabouri, Z., Sabouri, S., Kitawaki, Y., Pommier, Y., Abe, T., Kiyonari, H., and Honjo, T. (2011). Decrease in topoisomerase I is responsible for activation-induced cytidine deaminase (AID)-dependent somatic hypermutation. Proc. Natl. Acad. Sci. USA *108*, 19305–19310.

Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M., and Turner, D.J. (2009). Amplification-free Illumina sequencing-library preparation facilitates

improved mapping and assembly of (G+C)-biased genomes. Nat. Methods *6*, 291–295.

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucl. Acids Res. *42*, D68–D73.

Kress, C., Thomassin, H., and Grange, T. (2006). Active cytosine demethylation triggered by a nuclear receptor involves DNA strand breaks. Proc. Natl. Acad. Sci. USA *103*, 11112–11117.

Kriaucionis, S., and Heintz, N. (2009). The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. Science *324*, 929–930.

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

Kumaki, Y., Oda, M., and Okano, M. (2008). QUMA: quantification tool for methylation analysis. Nucl. Acids Res. *36*, W170–W175.

Kumar, R., DiMenna, L., Schrode, N., Liu, T.-C., Franck, P., Muñoz-Descalzo, S., Hadjantonakis, A.-K., Zarrin, A.A., Chaudhuri, J., Elemento, O., et al. (2013). AID stabilizes stem-cell phenotype by removing epigenetic memory of pluripotency genes. Nature *500*, 89–92.

Kuraoka, M., Holl, T.M., Liao, D., Womble, M., Cain, D.W., Reynolds, A.E., and Kelsoe, G. (2011). Activation-induced cytidine deaminase mediates central tolerance in B cells. Proc. Natl. Acad. Sci. USA *108*, 11560–11565.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Larijani, M., Frieder, D., Sonbuchner, T., Bransteitter, R., Goodman, M., Bouhassira, E., Scharff, M., and Martin, A. (2005). Methylation protects cytidines from AID-mediated deamination. Mol. Immunol. *42*, 599–604.

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C.T., Low, H.M., Sung, K.W.K., Rigoutsos, I., Loring, J., et al. (2010). Dynamic changes in the human methylome during differentiation. Genome Res. *20*, 320–331.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat. Rev. Genet. 1–18.

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.

Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., et al. (2009). IMGT, the

international ImMunoGeneTics information system. Nucl. Acids Res. *37*, D1006–D1012.

Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. Nature *366*, 362–365.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell *69*, 915–926.

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. Biochem. Biophys. Res. Commun. *260*, 398–404.

Limbach, P.A., Crain, P.F., and McCloskey, J.A. (1994). Summary: the modified nucleosides of RNA. Nucl. Acids Res. *22*, 2183–2196.

Lin, C., Yang, L., Tanasa, B., Hutt, K., Ju, B.-G., Ohgi, K.A., Zhang, J., Rose, D.W., Fu, X.-D., Glass, C.K., et al. (2009). Nuclear Receptor-Induced Chromosomal Proximity and DNA Breaks Underlie Specific Translocations in Cancer. Cell *139*, 1069–1083.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. Cell *133*, 523–536.

Liu, M., Duke, J.L., Richter, D.J., Vinuesa, C.G., Goodnow, C.C., Kleinstein, S.H., and Schatz, D.G. (2008). Two levels of protection for the B cell genome during somatic hypermutation. Nature *451*, 841–845.

Machnicka, M.A., Milanowska, K., Osman Oglou, O., Purta, E., Kurkowska, M., Olchowik, A., Januszewski, W., Kalinowski, S., Dunin-Horkawicz, S., Rother, K.M., et al. (2012). MODOMICS: a database of RNA modification pathways--2013 update. Nucl. Acids Res. *41*, D262–D267.

Macleod, D., Charlton, J., Mullins, J., and Bird, A.P. (1994). Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes Dev. *8*, 2282–2292.

Maiti, A., and Drohat, A.C. (2011). Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine: POTENTIAL IMPLICATIONS FOR ACTIVE DEMETHYLATION OF CpG SITES. J. Biol. Chem. *286*, 35334–35338.

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.

Marr, S., Morales, H., Bottaro, A., Cooper, M., Flajnik, M., and Robert, J. (2007). Localization and differential expression of activation-induced cytidine

deaminase in the amphibian Xenopus upon antigen stimulation and during early development. J. Immunol. *179*, 6783–6789.

Mayer, W., Niveleau, A., Walter, J., Fundele, R., and Haaf, T. (2000). Embryogenesis: Demethylation of the zygotic paternal genome. Nature *403*, 501–502.

Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B., et al. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature *454*, 766–770.

Mellén, M., Ayata, P., Dewell, S., Kriaucionis, S., and Heintz, N. (2012). MeCP2 Binds to 5hmC Enriched within Active Genes and Accessible Chromatin in the Nervous System. Cell *151*, 1417–1430.

Meyers, G., Ng, Y.-S., Bannock, J.M., Lavoie, A., Walter, J.E., Notarangelo, L.D., Kilic, S.S., Aksu, G., Debré, M., Rieux-Laucat, F., et al. (2011). Activation-induced cytidine deaminase (AID) is required for B-cell tolerance in humans. Proc. Natl. Acad. Sci. USA *108*, 11554–11559.

Métivier, R., Gallais, R., Tiffoche, C., Le Péron, C., Jurkowska, R.Z., Carmouche, R.P., Ibberson, D., Barath, P., Demay, F., Reid, G., et al. (2008). Cyclical DNA methylation of a transcriptionally active promoter. Nature *452*, 45–50.

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.

Militello, K.T., Chen, L.M., Ackerman, S.E., Mandarano, A.H., and Valentine, E.L. (2014). A map of 5-methylcytosine residues in Trypanosoma brucei tRNA revealed by sodium bisulfite sequencing. Mol. Biochem. Parasitol. *193*, 122–126.

Militello, K.T., Wang, P., Jayakar, S.K., Pietrasik, R.L., Dupont, C.D., Dodd, K., King, A.M., and Valenti, P.R. (2008). African trypanosomes contain 5-methylcytosine in nuclear DNA. Eukaryotic Cell *7*, 2012–2016.

Millar, C.B., Guy, J., Sansom, O.J., Selfridge, J., MacDougall, E., Hendrich, B., Keightley, P.D., Bishop, S.M., Clarke, A.R., and Bird, A. (2002). Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice. Science *297*, 403–405.

Miura, F., Enomoto, Y., Dairiki, R., and Ito, T. (2012). Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. Nucl. Acids Res. *40*, e136–e136.

Mohandas, T., Sparkes, R., and Shapiro, L. (1981). Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. Science *211*, 393–396.

Mohn, F., Weber, M., Schübeler, D., and Roloff, T.-C. (2009). Methylated DNA

immunoprecipitation (MeDIP). Methods Mol. Biol. *507*, 55–64.

Morgan, H.D., Dean, W., Coker, H.A., Reik, W., and Petersen-Mahrt, S.K. (2004). Activation-induced Cytidine Deaminase Deaminates 5-Methylcytosine in DNA and Is Expressed in Pluripotent Tissues: IMPLICATIONS FOR EPIGENETIC REPROGRAMMING. J. Biol. Chem. *279*, 52353–52360.

Muckenfuss, H., Hamdorf, M., Held, U., Perkovic, M., Löwer, J., Cichutek, K., Flory, E., Schumann, G.G., and Münk, C. (2006). APOBEC3 proteins inhibit human LINE-1 retrotransposition. J. Biol. Chem. *281*, 22161–22172.

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.

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.

Murphy, F.V., and Ramakrishnan, V. (2004). Structure of a purine-purine wobble base pair in the decoding center of the ribosome. Nat. Struct. Mol. Biol. *11*, 1251–1252.

Muschen, M., Re, D., Jungnickel, B., Diehl, V., Rajewsky, K., and Küppers, R. (2000). Somatic Mutation of the Cd95 Gene in Human B Cells as a Side-Effect of the Germinal Center Reaction. J. Exp. Med. *192*, 1833–1840.

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.

Nan, X., Ng, H.H., Johnson, C.A., Laherty, C.D., Turner, B.M., Eisenman, R.N., and Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature *393*, 386–389.

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., et al. (2012). Mutational processes molding the genomes of 21 breast cancers. Cell *149*, 979–993.

Nishikura, K., Yoo, C., Kim, U., Murray, J.M., Estes, P.A., Cash, F.E., and Liebhaber, S.A. (1991). Substrate specificity of the dsRNA unwinding/modifying activity. Embo J. *10*, 3523–3532.

Nonaka, T., Doi, T., Toyoshima, T., Muramatsu, M., Honjo, T., and Kinoshita, K. (2009). Carboxy-terminal domain of AID required for its mRNA complex formation in vivo. Proc. Natl. Acad. Sci. USA *106*, 2747–2751.

Okada, Y., Yamagata, K., Hong, K., Wakayama, T., and Zhang, Y. (2010). A role for the elongator complex in zygotic paternal genome demethylation. Nature *463*, 554–558.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247–257.

Okazaki, I.-M., Hiai, H., Kakazu, N., Yamada, S., Muramatsu, M., Kinoshita, K., and Honjo, T. (2003). Constitutive expression of AID leads to tumorigenesis. J. Exp. Med. *197*, 1173–1181.

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.

Pasqualucci, L., Migliazza, A., Fracchiolla, N., William, C., Neri, A., Baldini, L., Chaganti, R.S., Klein, U., Küppers, R., Rajewsky, K., et al. (1998). BCL-6 mutations in normal germinal center B cells: evidence of somatic hypermutation acting outside Ig loci. Proc. Natl. Acad. Sci. USA *95*, 11816–11821.

Pasqualucci, L., Neumeister, P., Goossens, T., Nanjangud, G., Chaganti, R.S., Küppers, R., and Dalla-Favera, R. (2001). Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. Nature *412*, 341–346.

Pasqualucci, L., Bhagat, G., Jankovic, M., Compagno, M., Smith, P., Muramatsu, M., Honjo, T., Morse, H.C., Nussenzweig, M.C., and Dalla-Favera, R. (2007). AID is required for germinal center–derived lymphomagenesis. Nat. Genet. *40*, 108–112.

Pastor, W.A., Pape, U.J., Huang, Y., Henderson, H.R., Lister, R., Ko, M., McLoughlin, E.M., Brudno, Y., Mahapatra, S., Kapranov, P., et al. (2011). Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature *473*, 394–397.

Pauklin, S., Sernandez, I.V., Bachmann, G., Ramiro, A.R., and Petersen-Mahrt, S.K. (2009). Estrogen directly activates AID transcription and function. J. Exp. Med. *206*, 99–111.

Pavri, R., Gazumyan, A., Jankovic, M., Di Virgilio, M., Klein, I., Ansarah-Sobrinho, C., Resch, W., Yamane, A., Reina-San-Martin, B., Barreto, V., et al. (2010). Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. Cell *143*, 122–133.

Peng, Z., Cheng, Y., Tan, B.C.-M., Kang, L., Tian, Z., Zhu, Y., Zhang, W., Liang, Y., Hu, X., Tan, X., et al. (2012). Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nat. Biotechnol. *30*, 253–260.

Penterman, J., Zilberman, D., Huh, J.H., Ballinger, T., Henikoff, S., and Fischer,

R.L. (2007). DNA demethylation in the Arabidopsis genome. Proc. Natl. Acad. Sci. USA *104*, 6752–6757.

Petersen-Mahrt, S.K., Harris, R.S., and Neuberger, M.S. (2002). AID mutates E. coli suggesting a DNA deamination mechanism for antibody diversification. Nature *418*, 99–104.

Popp, C., Dean, W., Feng, S., Cokus, S.J., Andrews, S., Pellegrini, M., Jacobsen, S.E., and Reik, W. (2010). Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature *463*, 1101–1105.

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

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. (2006). 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.

Rai, K., Sarkar, S., Broadbent, T.J., Voas, M., Grossmann, K.F., Nadauld, L.D., Dehghanizadeh, S., Hagos, F.T., Li, Y., Toth, R.K., et al. (2010). DNA Demethylase Activity Maintains Intestinal Cells in an Undifferentiated State Following Loss of APC. Cell *142*, 930–942.

Ramaswami, G., Lin, W., Piskol, R., Tan, M.H., Davis, C., and Li, J.B. (2012). Accurate identification of human Alu and non-Alu RNA editing sites. Nat. Methods *9*, 579–581.

Ramaswami, G., Zhang, R., Piskol, R., Keegan, L.P., Deng, P., O'Connell, M.A., and Li, J.B. (2013). Identifying RNA editing sites using RNA sequencing data alone. Nat. Methods *10*, 128–132.

Ramiro, A.R., Jankovic, M., Eisenreich, T., Difilippantonio, S., Chen-Kiang, S., Muramatsu, M., Honjo, T., Nussenzweig, A., and Nussenzweig, M.C. (2004). AID is required for c-myc/IgH chromosome translocations in vivo. Cell *118*, 431–438.

Ramsahoye, B.H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A.P., and Jaenisch, R. (2000). Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proc. Natl. Acad. Sci. USA *97*, 5237–5242.

Revy, P., Muto, T., Levy, Y., Geissmann, F., Plebani, A., Sanal, O., Catalan, N.,

Forveille, M., Dufourcq-Labelouse, R., Gennery, A., et al. (2000). Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). Cell *102*, 565–575.

Rice, G.I., Kasher, P.R., Forte, G.M.A., Mannion, N.M., Greenwood, S.M., Szynkiewicz, M., Dickerson, J.E., Bhaskar, S.S., Zampini, M., Briggs, T.A., et al. (2012). Mutations in ADAR1 cause Aicardi-Goutières syndrome associated with a type I interferon signature. Nat. Genet. *44*, 1243–1248.

Riggs, A.D. (1975). X inactivation, differentiation, and DNA methylation. Cytogenet. Cell Genet. *14*, 9–25.

Rivera, R.M., Stein, P., Weaver, J.R., Mager, J., Schultz, R.M., and Bartolomei, M.S. (2008). Manipulations of mouse embryos prior to implantation result in aberrant expression of imprinted genes on day 9.5 of development. Hum. Mol. Gen. *17*, 1–14.

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.

Robbiani, D.F., Bunting, S., Feldhahn, N., Bothmer, A., Camps, J., Deroubaix, S., McBride, K.M., Klein, I.A., Stone, G., Eisenreich, T.R., et al. (2009). AID produces DNA double-strand breaks in non-Ig genes and mature B cell lymphomas with reciprocal chromosome translocations. Mol. Cell *36*, 631–641.

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.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.

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.

Rountree, M.R., and Selker, E.U. (1997). DNA methylation inhibits elongation but not initiation of transcription in Neurospora crassa. Genes Dev. *11*, 2383–2395.

Russell, D.W., and Zinder, N.D. (1987). Hemimethylation prevents DNA replication in E. coli. Cell *50*, 1071–1079.

Russell, R.A., Wiegand, H.L., Moore, M.D., Schafer, A., McClure, M.O., and Cullen, B.R. (2005). Foamy Virus Bet Proteins Function as Novel Inhibitors of the APOBEC3 Family of Innate Antiretroviral Defense Factors. J. Virol. *79*, 8724–8731.

Sabag, O., Zamir, A., Keshet, I., Hecht, M., Ludwig, G., Tabib, A., Moss, J., and Cedar, H. (2014). Establishment of methylation patterns in ES cells. Nat. Struct. Mol. Biol. *21*, 110–112.

Sato, S., Yoshida, W., Soejima, H., Nakabayashi, K., and Hata, K. (2011). Methylation dynamics of IG-DMR and Gtl2-DMR during murine embryonic and placental development. Genomics *98*, 120–127.

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.

Schiesser, S., Hackner, B., Pfaffeneder, T., Müller, M., Hagemeier, C., Truss, M., and Carell, T. (2012). Mechanism and Stem-Cell Activity of 5-Carboxycytosine Decarboxylation Determined by Isotope Tracing. Angew. Chem. Int. Ed. *51*, 6516–6520.

Schnabel, L.V., Abratte, C.M., Schimenti, J.C., Southard, T.L., and Fortier, L.A. (2012). Genetic background affects induced pluripotent stem cell generation. Stem Cell Res. Ther. *3*, 30.

Schnaufer, A., Panigrahi, A.K., Panicucci, B., Igo, R.P., Wirtz, E., Salavati, R., and Stuart, K. (2001). An RNA ligase essential for RNA editing and survival of the bloodstream form of Trypanosoma brucei. Science *291*, 2159–2162.

Selker, E.U., Tountas, N.A., Cross, S.H., Margolin, B.S., Murphy, J.G., Bird, A.P., and Freitag, M. (2003). The methylated component of the Neurospora crassa genome. Nature *422*, 893–897.

Shaknovich, R., Cerchietti, L., Tsikitas, L., Kormaksson, M., De, S., Figueroa, M.E., Ballon, G., Yang, S.N., Weinhold, N., Reimers, M., et al. (2011). DNA methyltransferase 1 and DNA methylation patterning contribute to germinal center B-cell differentiation. Blood *118*, 3559–3569.

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, H.M., Peters, A., Baron, B., Zhu, X., and Storb, U. (1998). Mutation of BCL-6 gene in normal B cells by the process of somatic hypermutation of Ig genes. Science *280*, 1750–1752.

Sloof, P., Menke, H.H., Caspers, M.P., and Borst, P. (1983). Size fractionation of Trypanosoma brucei DNA: localization of the 177-bp repeat satellite DNA and a variant surface glycoprotein gene in a mini-chromosomal DNA fraction. Nucl. Acids Res. *11*, 3889–3901.

Song, J., Teplova, M., Ishibe-Murakami, S., and Patel, D.J. (2012). Structure-Based Mechanistic Insights into DNMT1-Mediated Maintenance DNA Methylation. Science *335*, 709–712.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.

Stenglein, M.D., Burns, M.B., Li, M., Lengyel, J., and Harris, R.S. (2010). APOBEC3 proteins mediate the clearance of foreign DNA from human cells. Nat. Struct. Mol. Biol. *17*, 222–229.

Stroud, H., Greenberg, M.V.C., Feng, S., Bernatavichute, Y.V., and Jacobsen, S.E. (2013). Comprehensive Analysis of Silencing Mutants Reveals Complex Regulation of the Arabidopsis Methylome. Cell *152*, 352–364.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., et al. (2009). Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. Science *324*, 930–935.

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

Teng, G., Hakimpour, P., Landgraf, P., Rice, A., Tuschl, T., Casellas, R., and Papavasiliou, F.N. (2008). MicroRNA-155 is a negative regulator of activation-induced cytidine deaminase. Immunity *28*, 621–629.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Trewick, S.C., Henshaw, T.F., Hausinger, R.P., Lindahl, T., and Sedgwick, B. (2002). Oxidative demethylation by Escherichia coli AlkB directly reverts DNA base damage. Nature *419*, 174–178.

Turelli, P., Mangeat, B., Jost, S., Vianin, S., and Trono, D. (2004). Inhibition of hepatitis B virus replication by APOBEC3G. Science *303*, 1829.

van Luenen, H.G.A.M., Farris, C., Jan, S., Genest, P.-A., Tripathi, P., Velds, A., Kerkhoven, R.M., Nieuwland, M., Haydock, A., Ramasamy, G., et al. (2012). Glucosylated Hydroxymethyluracil, DNA Base J, Prevents Transcriptional Readthrough in Leishmania. Cell *150*, 909–921.

Vesely, C., Tauber, S., Sedlazeck, F.J., Haeseler, von, A., and Jantsch, M.F. (2012). Adenosine deaminases that act on RNA induce reproducible changes in abundance and sequence of embryonic miRNAs. Genome Res. *22*, 1468–1476.

Viré, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., Morey, L., Van Eynde, A., Bernard, D., Vanderwinden, J.-M., et al. (2006). The Polycomb group protein EZH2 directly controls DNA methylation. Nature *439*, 871–874.

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. Dev. Biol. *350*, 13–23.

Wang, M., Yang, Z., Rada, C., and Neuberger, M.S. (2009). AID upmutants isolated using a high-throughput screen highlight the immunity/cancer balance limiting DNA deaminase activity. Nat. Struct. Mol. Biol. *16*, 769–776.

Wang, Y., Li, Y., Toth, J.I., Petroski, M.D., Zhang, Z., and Zhao, J.C. (2014). N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. Nat. Cell Biol. *16*, 191–198.

Wang, Y., Jorda, M., Jones, P.L., Maleszka, R., Ling, X., Robertson, H.M., Mizzen, C.A., Peinado, M.A., and Robinson, G.E. (2006). Functional CpG methylation system in a social insect. Science *314*, 645–647.

Weber, M., Davies, J.J., Wittig, D., Oakeley, E.J., Haase, M., Lam, W.L., and Schübeler, D. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat. Genet. *37*, 853–862.

Wijesinghe, P., and Bhagwat, A.S. (2012). Efficient deamination of 5-methylcytosines in DNA by human APOBEC3A, but not by AID or APOBEC3G. Nucl. Acids Res. *40*, 9206–9217.

Wu, D., Lamm, A.T., and Fire, A.Z. (2011). Competition between ADAR and RNAi pathways for an extensive class of RNA targets. Nat. Struct. Mol. Biol. *18*, 1094–1101.

Xu, Z., Zan, H., Pone, E.J., Mai, T., and Casali, P. (2012). Immunoglobulin class-switch DNA recombination: induction, targeting and beyond. Nat. Rev. Immunol. *12*, 517–531.

Yamaguchi, S., Shen, L., Liu, Y., Sendler, D., and Zhang, Y. (2013). Role of Tet1 in erasure of genomic imprinting. Nature *504*, 460–464.

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.

Yamane, A., Resch, W., Kuo, N., Kuchen, S., Li, Z., Sun, H.-W., Robbiani, D.F., McBride, K., Nussenzweig, M.C., and Casellas, R. (2010). Deep-sequencing identification of the genomic targets of the cytidine deaminase AID and its cofactor RPA in B lymphocytes. Nat. Immunol. *12*, 62–69.

Yang, W., Chendrimada, T.P., Wang, Q., Higuchi, M., Seeburg, P.H., Shiekhattar, R., and Nishikura, K. (2005). Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. Nat. Struct. Mol. Biol.

13, 13–21.

Yoder, J.A., Walsh, C.P., and Bestor, T.H. (1997). Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. *13*, 335–340.

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.

Zhu, B., Zheng, Y., Angliker, H., Schwarz, S., Thiry, S., Siegmann, M., and Jost, J.P. (2000). 5-Methylcytosine DNA glycosylase activity is also present in the human MBD4 (G/T mismatch glycosylase) and in a related avian sequence. Nucl. Acids Res. *28*, 4157–4165.

Zhu, J.-K. (2009). Active DNA Demethylation Mediated by DNA Glycosylases. Annu. Rev. Genet. *43*, 143–166.

Ziller, M.J., Gu, H., Müller, F., Donaghey, J., Tsai, L.T.Y., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E., et al. (2013). Charting a dynamic DNA methylation landscape of the human genome. Nature *500*, 477–481.

Ziller, M.J., Müller, F., Liao, J., Zhang, Y., Gu, H., Bock, C., Boyle, P., Epstein, C.B., Bernstein, B.E., Lengauer, T., et al. (2011). Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. PLoS Genet. *7*, e1002389.