CHARACTERIZATION OF INTRONIC POLYADENYLATION ISOFORMS IN NORMAL HUMAN CELLS AND B CELL MALIGNANCIES

A Dissertation

Presented to the Faculty of the Weill Cornell Graduate School

Of Medical Sciences

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Irtisha Singh

January 2017

CHARACTERIZATION OF INTRONIC POLYADENYLATION ISOFORMS IN NORMAL HUMAN CELLS AND B CELL MALIGNANCIES

Irtisha Singh, Ph.D.

Cornell University 2017

Alternative cleavage and polyadenylation (ApA) is most often viewed as the selection of alternative pA signals in the 3' UTR, generating 3' UTR isoforms that code for the same protein. However, ApA events can also occur in introns, generating either non-coding transcripts or truncated protein-coding isoforms due to the loss of C-terminal protein domains, leading to diversification of the proteome. Due to lack of a study that characterizes the intronic polyadenylation isoforms on a genome wide level, we decided to investigate the cell type specificity and potential functional consequences of isoforms generated by intronic ApA. We therefore carried out an analysis of 3'-seq and RNA-seq profiles from chronic lymphocytic leukemia (CLL) and multiple myeloma (MM) samples as compared to mature human B cells (naïve and CD5+) and plasma cells, respectively, together with our previous 3'-seq atlas generated from a wide variety of tissues and cell lines. We found that in more than 20 percent of human genes, intronic polyadenylation (IpA) sites are used to generate alternative 3' ends. This analysis shows that IpA is a normal and regulated process, most widely used in immune cells. IpA events are enriched near the start of the transcription unit, yielding non-coding transcripts or messages with minimal coding sequence (CDS). The expression of truncated mRNAs contributes to proteome diversity in normal cells. However, we found that cancer cells can take advantage of this mechanism to mimic genetic mutations. Many genes with truncating mutations in CLL express IpA isoforms. Cancer cells lacking genetic aberrations can disrupt IpA to generate truncated mRNAs. These IpA isoforms potentially have similar functional outcome as truncating mutations. In this study we unravel a new mechanism by which cancer cells can contribute to tumorigenesis.

BIOGRAPHICAL SKETCH

Irtisha Singh completed her undergraduate degree in the field of bioinformatics from Vellore Institute of Technology, India in 2008. After that she enrolled in Carnegie Mellon University (CMU) to pursue Master of Science in computational biology. The wealth of knowledge gained at CMU inspired her to delve deep into science. This motivated her to enroll in a doctorate program in United States. She worked as a research assistant at Language Technology Institute, CMU for a year. During the course of the PhD application process, she also worked as a bioinformatics analyst at the Biomedical Informatics Department of University of Pittsburgh. She was accepted for pursuing PhD at the Tri-Institutional Computational Biology and Medicine Program of Cornell University, Weill Cornell Medical College and MSKCC in 2011. She spent a year at Cornell University, Ithaca before joining Dr. Christina Leslie's lab as a graduate student in September 2012. She worked in close collaboration with Dr. Christine Mayr at MSKCC from 2013 to study the phenomenon of intronic cleavage and polyadenylation. She was also involved extensively in other collaborative projects during the course of her career as a graduate student in the Leslie lab. Post-graduation, Irtisha plans to find a post-doctorate position and continue her research in the field of computational biology.

This document is dedicated to all the existing and growing members of my family.

ACKNOWLEDGEMENTS

I have been incredibly fortunate to have parents that have struggled all their lives to provide me with the best environment to learn and develop a strong foundation for education. Fighting against all odds, they made every effort and inspired me to have a career that made me an independent entity of this society. I cannot thank them for leaving no stone unturned to make me a self-sustaining individual but can only acknowledge their efforts. I would like to thank my siblings, Isha and Anubhav for sharing this journey with me with the most cheering attitude. The next most important thankyou is for my advisor and mentor, Dr. Christina Leslie. I was extremely fortunate to pursue my doctorate as her student. Her wealth of knowledge and motivation for her students, always inspires to learn more and perform better. Her encouragement balanced with understanding provides the most adequate environment for growth of science. I am thankful to her for enhancing my desire to continue being a part of this exciting scientific world. I would also like to thank Dr. Christine Mayr for acting as my second advisor. My frequent interactions with her have honed my critical thinking abilities to navigate through hurdles. Her exceptional zeal for science and hard work serves as a source of encouragement.

My colleagues at lab played a major role in making this journey smooth. I would like to thank Steve - for giving the keys to the castle and helping in entering it, Julie – for her tangential conversations and all the support in times of need, Rafi – for all the fun arguments and discussions, Manu – for helping in settling in. I would also like to thank the other members of the lab Yuheng, Lauren, Alvaro and Mark for making it a fun place to work.

My next special thanks is for my friends, Aabhas and Rohit who have acted as my pillars of support in my moments of hopelessness. Their continuous encouragement has helped me to navigate through all tough times. I would also like to thank my mother-in-law and father-in-law for being so supportive in this journey.

Most importantly, I want to thank my husband, Akhilesh, for being my closest friend and doing everything possible to make this happen. He is my source of strength who never lets me give up. Lastly, I would like to thank my daughter, who is yet to come in this world. She is already making her presence felt in our lives and adding smile to our faces.

TABLE OF CONTENTS

Biographical Sketch iii
Dedication iv
Acknowledgements v
Table of Contents vii
List of Figures ix

1. INTRODUCTION 1

1.1 Background..........1
1.2 Overview of dissertation..........8
1.2.1 Accurate mapping and quantification of the 3' ends of transcripts..........10
1.2.2 A comprehensive atlas of intronic and 3' UTR cleavage events..........10
1.2.3 Characterization and regulation of IpA isoforms across normal tissue and cell types..........11
1.2.4 Deregulation of expression of IpA isoforms in B cell malignancies and its functional consequences 11
1.2.5 Functional roles of the wide-variety of mRNAs generated by intronic polyadenylation..........12

2. THE ACCURATE, QUANTITATIVE, AND HIGH-THROUGHPUT MAPPING OF MRNA CLEAVAGE EVENTS IN THE 3' UTR AND INTRONS OF TRANSCRIPTION UNITS 14

2.1 Introduction..........14
2.2 Description of tissue and cell types being analyzed 15
2.3 The 3'-seq protocol..........18
2.4 Processing the 3'-seq data 19
2.5 3'-seq sequencing protocol artifacts..........22

3. ACCURATELY CHARACTERIZING THE LANDSCAPE OF ALTERNATIVE CLEAVAGE AND POLYADENYLATION IN HUMANS 23

3.1 Atlas creation 23
3.2 External validation of IpA isoforms 29
3.3 Rescuing genes with convergent 3' UTRs 35
3.4 Testing for differential usage of IpA sites compared to pASs in 3' UTRs 41

4. CHARACTERIZATION AND REGULATION OF IpA ISOFORMS ACROSS NORMAL TISSUES AND IMMUNE CELL TYPES..........42

4.1 Introduction..42
4.1 IpA isoforms are expressed as a part of the regular expression program in normal cell types and display high evolutionary conservation................................43
4.2 High usage of IpA sites in immune cell types with recurrent sharing of the IpA sites between immune cell types..47
4.3 Predominant usage of promoter proximal IpA sites with increased recognition of promoter proximal IpA sites in cell types expressing a higher number of IpA isoforms...50
4.4 Intronic polyadenylation enriched in retained introns 55
4.5 Enriched pASs, depleted U1 snRNP, long introns and long transcription units define the landscape of the genes with IpA isoforms 60
4.6 Expression of IpA isoforms is differentially regulated between different cellular conditions...63

5. DEREGULATED IpA ISOFORMS EXPRESSION IN LEUKEMIAS 68

5.1 Truncated mRNAs generated by usage of IpA site can potentially mimic genetic mutations..69
5.2 Widespread usage of distal pAS in 3' UTRs...74
5.3 Loss of usage of IpA sites in multiple myeloma..75

6. INTRONIC POLYADENYLATION CREATES A POOL OF mRNAS WITH A WIDE VARIETY OF FUNCTIONS..78

6.1 5' IpA events create non-coding RNAs with a potential role in gene regulation78
6.2 Non-coding IpA isoforms enriched for binding sites of RNA binding proteins.83
6.3 IpA isoforms potentially diversify proteome but largely do not affect the transcriptional activity of the genes 85

7. SUMMARY AND CONCLUSION 89

7.1 Atlas with robustly expressed IpA isoforms across a wide variety of tissue and cell types 89
7.2 IpA isoforms are expressed as part of the normal expression program with variable expression patterns in different cell types..90
7.3 Diverse pattern of usage of IpA sites along transcription units in various tissue and cell types..92
7.4 Retained introns are associated with the occurrence of IpA events...................93
7.5 Genes with long transcription units, long introns and long 5' UTRs are enriched with pASs and devoid of U1 snRNP signals to facilitate intronic cleavage events..93
7.6 Intronic cleavage events creates a wide variety of mRNAs with potentially diverse functions 94
7.7 Truncated mRNAs generated by the usage of IpA sites can mimic genetic mutations...95

7.8 Conclusion 96

8. REFERENCES..99

LIST OF FIGURES

1.1: 3' UTR ApA isoforms……………………………………………………………….3

1.2: Types of IpA isoforms………………………………………………………………9

2.4 Distribution of 3'-seq unique reads mapping to different parts of the genes in entire genome…………………………………………………………………………………..21

3.1.1 Skewedness in distribution of reads because of antibody production…………25

3.1.2 Distribution of 3'-seq reads in the atlas……………………………………...29

3.2.1 Coverage of genomic reads in RNA-seq……………………………………31

3.2.2 Validation of IpA events using RNA-seq…………………………………....32

3.2.3 Validation of IpA events……………………………………………………….35

3.3.1 Feature matrix of sense and antisense peaks for learning from non-convergent genes…………………………………………………………………………………..39

3.3.2 Feature matrices of convergent genes before and after prediction……………..40

4.1.1 Expression of IpA isoforms across cell types…………………………………45

4.1.2 Robust expression of IpA isoforms……………………………………………..46

4.1.3 Higher conservation around the cleavage site of IpA isoforms………………..47

4.2.1 Immune cells express higher number of IpA isoforms………………………...49

4.2.2 Majority of IpA isoforms are shared between immune cell types……………..50

4.3.1 a) High usage of promoter proximal IpA sites…………………………………52

4.3.1 b) Tissues with higher fraction of genes with IpA isoforms have larger number of IpA isoforms ending close to the start of the transcription unit…………………52

4.3.2 Diversified usage of IpA site in different tissue and cell types………………..53

4.3.3 a) Tissues with shorter 3' UTRs have shorter IpA isoforms………………….54

4.3.3 b) Higher frequency of IpA isoforms associated with 3' UTR lengths……….54

4.4.1 Co-occurrence of IpA with intron retention……………………………………57

4.4.2 Association between intron retention and occurrence of IpA………………….59

4.4.3 a) Higher incidence of IpA events in introns with IR (intron retention)……….60

4.4.3 b) Significant enrichment of IpA in retained introns…………………………...60

4.4.4 IpA events with lower usage index co-occur with intron retention……………61

4.5.1 Genes with IpA isoforms enriched for pASs and depleted for U1 snRNP sites.63

4.5.2 Occurrence of IpA isoforms in genes with long transcription units, long introns and wider 5' UTRs……………………………………………………………………65

4.6.1 IpA isoform is expressed in Naïve B cells obtained from blood………………67

4.6.2 Significantly differently used IpA sites between naïve B cells from blood and tonsils……………………………………………………………………………68

5.1.1 a) High usage of IpA site in CLLs compared to naïve/CD5+ cells…………..72

5.1.1 b) Recurrently highly used IpA sites in CLL samples…………………………72

5.1.2 Truncating mutations enriched in genes with highly used IpA site in CLLs…..74

5.1.3 Robustly expressed IpA isoform of MGA……………………………………..75

5.2.1 Longer 3' UTRs in CLLs………………………………………………………77

5.3.1 a) Loss of usage of IpA sites in multiple myeloma cells……………………..78

6.1.1 Assembled transcript for the IpA isoform of CCT4…………………………...82

6.1.2 Most of the transcripts retaining less than 25% of coding sequencing are predicted to be non-coding……………………………………………………………83

6.2.1 Gained part of intron in non-coding IpA isoforms are enriched for RBP sites..86

6.3.1 Transcriptional activity of genes largely unaffected by expression of IpA isoforms…………………………………………………………………………………87

6.3.2 IpA can potentially diversify proteome……………………………………...88

CHAPTER 1

1. INTRODUCTION

1.1 Background

A key problem in molecular biology is the dissection of the different mechanisms that regulate co-transcriptional and post-transcriptional mRNA biogenesis. Decades of research has elucidated some of the molecular processes for the generation of a mature mRNA that can be efficiently transported to the cytosol from the nucleus where it can be translated into a protein. With the advent of next generation sequencing methodologies there has been increased focus on the characterization and regulation of these process, namely capping, splicing and cleavage and polyadenylation. The technological advances made in the recent years have established cleavage and polyadenylation as a crucial layer of gene regulation. Cleavage and polyadenylation precisely defines the 3' end of the transcripts in a two-step process: i) endonucleolytic cleavage of the transcribed transcript by the molecular machinery after the recognition of a functional poly (A) site, followed by ii) the addition of untemplated adenosines (polyA) at the 3' end of the transcript. This co-transcriptional processing is necessary for protein expression and is important for the nuclear export, stability, and translation of the transcript.

Discovery of poly (A) tails at the 3' end of mature mRNAs marked the beginning of discovery of the machinery that defines 3' ends of the transcripts along with its key players (Lim and Canellakis 1970). This observation of poly (A) tracts at the end of the transcripts was followed by studies that made use of the powerful Sanger sequencing technique to identify a common sequence AAUAAA within 20-30 nucleotides (nt) of the start of the poly (A) tail (Proudfoot and Brownlee 1976). Subsequent experiments focusing on mutation of this hexamer sequence signal, referred to as the poly(A) site (pAS), further illustrated that the presence of this signal is absolutely critical for an efficient 3' end polyadenylation (Fitzgerald and Shenk 1981). It has been shown that variants of AAUAAA also function as pAS. Apart from the hexamer signal, other cis elements are required to reconstitute a functional pAS. Presence of U- or GU- rich elements dowsnstream of the poly (A) signal enhances the 3' end formation (Gil and Proudfoot 1984). The next breakthrough in the understanding of 3' end processing came with the purification of two protein complexes: cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulatory factor (CstF) (Proudfoot 2011). These two key complexes have been shown to recognize the pAS followed by the endonucleolytic cleavage of the transcript. Further studies reconstituted the whole cleavage and polyadenylation machinery (C/P machinery) to be composed of 15-20 core polypeptides, four protein complexes and many single proteins (Tian and Manley 2013). CPSF-160 has been shown to interact with the pAS while while CPSF-73 is responsible for the endonucleatic cleavage of the pre-mRNA (Mandel et al. 2006). It was also established that the processing of 3' end of the transcripts is separate from the actual Pol II transcriptional termination. It was further shown that efficient 3' end processing is critical for transcription termination.

As the phenomenon of 3' end processing was being pieced together, further studies showed that there are eukaryotic genes that possess multiple functional pAS in their 3' UTRs and that the usage of these alternative pAS resulted in transcripts that differed in their 3' UTR lengths. This phenomenon of usage of alternative pAS is referred as alternative cleavage and polyadenylation (ApA). One of the first few examples that established the presence and usage of the alternative pASs was in the mouse dihydrofolate reducatase (DHFR) (Frayne et al. 1984). DHFR has been shown to express mRNAs with identical 5' ends but different 3' end through usage of alternative pAS in a growth dependent manner (Kaufman and Sharp 1983). Another example of a gene that exhibits ApA is eukaryotic initiation factor 2α (eIF-2α) where the usage of the alternative pAS is regulated across different tissues. By the end of the 20th century there were multiple independent studies establishing regulation of ApA in around 130 genes (Edwalds-Gilbert et al. 1997). Different studies showed the ApA isoforms of these genes to be highly regulated across different tissues as well as in different cellular conditions. These genes were shown to produce mRNAs with different 3' ends via methods like northern blots or nuclease protection assays. These methods focus on one gene at a time but provide robust evidence of the presence of different 3' ends of

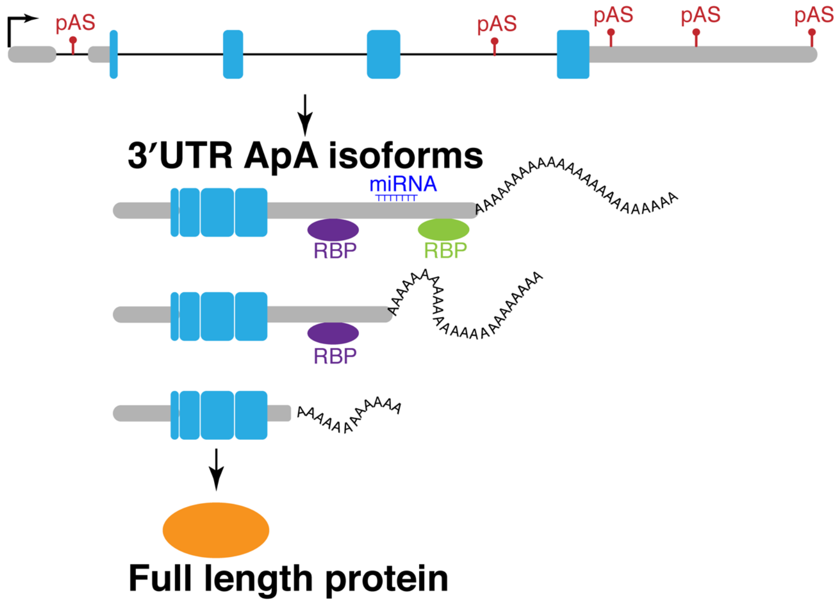

Figure 1.1: 3' UTR ApA isoforms
Recognition of the alternative pASs located in the 3' UTR generates mRNAs differing in their 3' UTR lengths. The mRNA isoforms created by ApA undergo different levels of regulation based on the regulatory sites present in their 3' UTRs. These 3' UTR ApA isoforms are translated into the same protein.

the mRNAs. Such studies established the occurrence of 3' UTR ApA where the recognition of the alternative pAS creates mRNAs that differ in their 3' UTR lengths but have the same protein coding regions (Figure 1.1).

Genome-wide mapping of ApA Expressed Sequence Tags (ESTs) were the first data resource that enhanced the analysis of ApA from a single gene study to multiple sequences at the same time. Studies made use of these ESTs to show that 54% of genes in humans while 34% genes in mouse undergo ApA (Tian et al. 2005). To overcome the limitation of ESTs, microarrays with multiple probes mapping the 3' UTRs were used. An early study used array analysis to demonstrate that activated immune cells express shorter UTRs, suggesting an association between proliferation an alternative ApA (Sandberg et al. 2008). This study also showed the regulation of mRNAs by ApA due the loss of microRNA binding sites from the extended 3' UTRs.

As microarrays suffered from the limitation of pre-designed probes, next generation sequencing technologies brought a significant boost in identifying novel regions of the genome. To focus on ApA, protocols were developed to map and quantify the usage of 3' ends of transcripts of the expressed mRNAs. Over the past few years researchers have progressively attempted to improve 3' end sequencing protocols so that they can not only be useful in identification of the 3' UTR ApA events but also provide quantitative usage of the cleavage sites. A number of sequencing methods for ApA analysis have been established so far, including 3' end RNA-seq (Yoon and Brem 2010), PAS-seq (Shepard et al. 2011), Poly(A)-seq (Derti et al. 2012), SAPAS (Fu et al. 2011), 3SEQ (Beck et al. 2010), A-seq (Martin et al. 2012), 3P-seq (Jan et al. 2011), 3' READS (Hoque et al. 2013), and 3'-seq (Lianoglou et al. 2013).

Regulation of ApA in normal physiological processes Genome-wide mapping of 3' UTR ApA events by the various high-throughput sequencing protocols have established tissue and condition specific alternative 3' UTR isoform expression patterns. These studies confirmed that approximately 50% of human genes are alternative cleaved and polyadenylated while the other 50% always end at the same position (Shepard et al. 2011; Lianoglou et al. 2013). Several independent studies established the expression of ApA isoforms to be highly regulated across tissues with evidence of tissue-biased expression of ApA isoforms (Zhang et al. 2005; Wang et al. 2008; Shepard et al. 2011; Derti et al. 2012; Li et al. 2012; Smibert et al. 2012; Ulitsky et al. 2012; Lianoglou et al. 2013). A comprehensive and systematic analysis of 3' UTR ApA events in ubiquitously expressed genes of diverse tissue types showed testes to express the shortest 3' UTR ApA isoforms with brain expressing the longest ApA isoforms (Lianoglou et al. 2013). This study proposed that ApA was a mechanism for tissue-specific regulation of ubiquitously expressed gene. These patterns have also been shown to exist in flies (Smibert et al. 2012). Further it has been shown that the pattern of usage of pAS in a particular tissue type across species is more correlated that the pattern of usage of pAS across different tissue types within the same species (Derti et al. 2012).

Multiple studies have also examined the regulation of 3' UTR ApA during proliferation and differentiation. Analysis of 3' UTR ApA events showed activation of T cells leads to creation of ApA isoforms with shortened 3' UTRs on a global level (Sandberg et al. 2008). There have also been studies suggesting ApA isoforms to have longer 3 'UTRs as the cell differentiates. Genes have been shown to have longer 3' UTRs during embryonic development in mouse (Ji et al. 2009). Another study with a focus on mouse embryonic stem (ES) cells and neural stem/progenitor (NSP) cells made similar observation of expression of ApA isoforms with longer 3' UTR during differentiation (Shepard et al. 2011). Such observations have laid the groundwork for a strong association between 3' UTR ApA and cellular differentiation, which potentially is a layer of post-transcriptional gene regulation to achieve cellular identity.

Deregulation of ApA in diseases and malignancies: Single nucleotide mutation in the pAS of the α2-globin gene and β-globin gene has been shown to give rise to two forms of rare thalassaemia leading to abnormal expression of the genes (Higgs et al. 1983; Orkin et al. 1985). With these two cases it was long ago affirmed that a precise 3' end of the mRNAs is essential for the efficient gene expression. Recently more pathological conditions associated with deregulation of 3' UTR ApA have been uncovered. Another such example is the FOXP3 gene where a mutation in the proximal pAS of FOXP3 creates mRNAs with longer 3' UTRs. This form of deregulation of 3'UTR ApA in FOXP3 causes IPEX, a disease with dysfunctional regulatory T cells followed by subsequent autoimmunity (Bennett et al. 2001).

Understanding of regulation of 3' UTR ApA became even more important when researchers found 3'UTR ApA to be deregulated in cancer cells. It has been found that apart from genetic alterations, deregulated ApA could also contribute to tumorigenesis. Specific oncogenes have been shown to have shorter 3'UTRs in transformed cells compared to normal cells, aiding in production of 10 times more protein on average from the same amount of mRNA (Mayr and Bartel 2009). This is achieved by loss of repressive elements, including microRNA binding sites and AU-rich elements, from the 3' UTR of these genes due to the widespread shortening of 3' UTRs. However, it is not true that deregulated ApA always increased usage of the shorter 3' UTR in cancer (Fu et al. 2011). While the breast cancer cell line MCF7 has shortening of 3' UTRs, another breast cancer line, MB231, shows lengthening of 3' UTRs for majority of the genes compared to mammary epithelial cell line, MCF10A. This study points towards a more complex regulation of 3' UTR ApA than a simple shortening of 3' UTRs in cancer. Interestingly a more recent pan-cancer study across seven different tumor types using RNA-seq of tumor samples from TCGA consortium reported broad shortening of the 3' UTRs (Xia et al. 2014). They concluded that lung (LUSC and LUAD), uterine (UCEC), breast (BRCA) and bladder (BLCA) undergo the most widespread changes in 3' UTR ApA. Overall these studies have just started focusing on the relevance of deregulation of 3' UTR ApA isoforms in cancer. More in-depth studies are required to discover the effects of these deregulated 3' UTR ApA isoform expression events in malignant cells as well as the factors responsible for producing these changes. As 3' UTR ApA deregulation appears to be mimicking genetic alterations in cancer, it opens avenues for discovering factors that could be targeted to make the cells have a normal 3' UTR ApA program.

Usage of polyA sites in introns is regulated in normal developmental processes

Approximately 20% of human genes have functional alternative pAS in the coding regions and introns of the transcript. Recognition of intronic polyadenylation (IpA) sites creates mRNA transcripts with reduced length and truncated protein product. Unlike the aberrant isoforms generated by the introduction of premature stop codons, these transcripts are not degraded by nonsense mediated decay (Lejeune and Maquat 2005) and are stably expressed. The usage of IpA sites can result in loss of C terminal exons, significantly altering the function of the truncated protein product compared to the full-length product. In B cell maturation, the usage of the IpA site of the immunoglobulin antibody heavy chain gene is regulated (Early et al. 1980; Rogers et al. 1980). In plasma cells, the IpA site is recognized, leading to the soluble protein product as opposed to its full-length product that contains a transmembrane domain in its C-terminus in mature B cells. This mechanism seems to be widespread as it is well known that intronic ApA in transmembrane proteins like receptor tyrosine kinases result in loss of the transmembrane anchoring domain of the protein, producing the soluble isoform of the protein (Vorlova et al. 2011) resulting in loss of signaling. It has also been shown that regulation of an IpA site can act as a developmental switch for the SREPF transcription factor to generate different protein isoforms in spermatogenesis (Wang et al. 2006). These examples show that the selective usage of the IpA site can potentially be a general phenomenon to regulate the protein isoforms with alternative C-terminal exons. Recognition of IpA sites upstream of a start codon can also result in loss of the open reading frame of the gene. This event can mimic genomic deletion as it generates a non-coding mRNA. Thus, alternative cleavage and polyadenylation in the introns can generate truncated mRNAs that have dramatically different functions.

Motivated by these single gene examples, we were interested in leveraging 3'-seq, a next-generation sequencing method for quantifying 3' ends, to characterize intronic polyadenylation isoforms in normal human cells and B cell malignancies. We wanted to investigate if usage of IpA sites is a more widespread phenomenon that potentially serves as a layer of gene regulation that has not yet been appreciated.

1.2 Overview of dissertation

To our knowledge there has been no study that has attempted to characterize alternative cleavage and polyadenylation in introns on a genome-wide scale. In this dissertation we characterized the landscape of intronic polyadenylation (IpA) isoforms across diverse normal human tissue types and B cell malignancies. Our aim was to study the regulation of the usage of IpA sites across diverse tissue types. We also investigated how the landscape of intronic cleavage events changes in malignant cells, which revealed a new mechanism of generating truncated mRNAs that could potentially mimic genetic mutations.

We hypothesized that in normal tissues, IpA isoforms generated by the usage of IpA sites towards the 3' end of the transcription unit (Type I) would be translated into proteins that lack C-terminal protein domains (Figure 1.2). These IpA isoforms would be translated into truncated proteins that could contribute towards the diversification

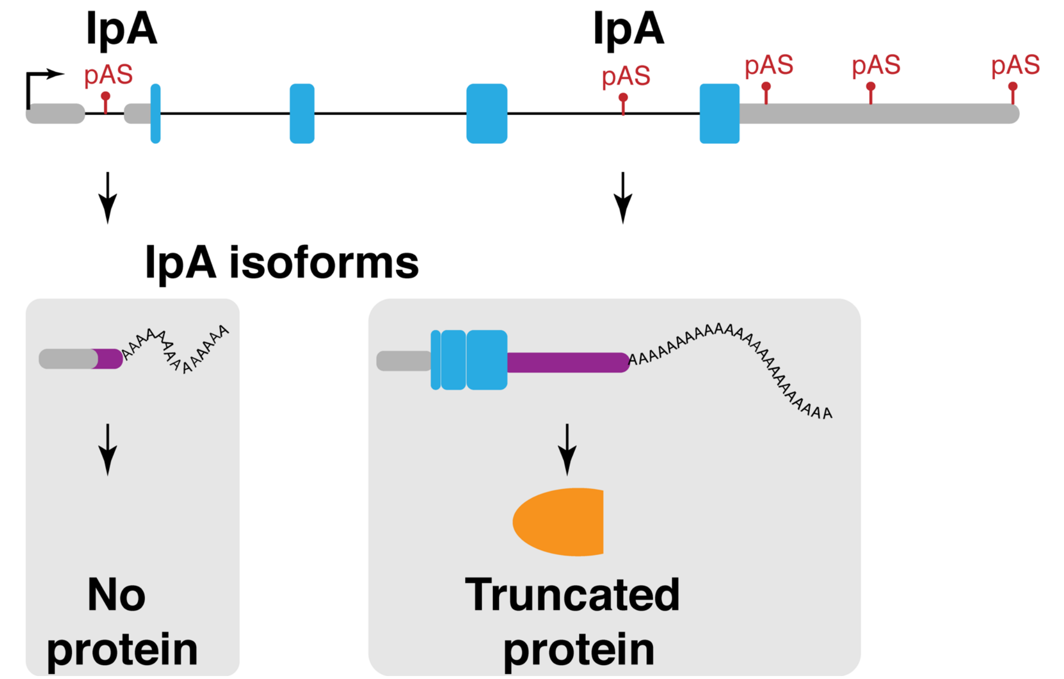

Figure 1.2: Types of IpA isoforms
Usage of pAS located in introns generates truncated mRNAs, referred to as IpA isoforms. An IpA isoform created by recognition of pAS in early introns is expected to produce an mRNA that lacks the ability to be translated into protein. IpA isoforms generated by later pASs would be translated into truncated proteins.

of the proteome or could create a dysfunctional protein due to loss of important protein domains. If the IpA isoforms are generated by the recognition of the pASs located close to the 5' end of the gene (Type II), then those isoforms would have marginal probability to be translated into a protein. Such IpA isoforms could potentially create non-coding RNAs. In cancer, the usage of IpA sites could generate truncated mRNAs that could have the potential to mimic genetic mutations. The type I IpA isoforms could mimic loss of function mutations, frame-shift mutations, or genomic translocations while type II could mimic deletions (Figure 1.2). To investigate our hypothesis, we first characterized the cleavage events in the introns and 3' UTRs and then assessed the potential functional relevance of the IpA isoforms.

1.2.1 Accurate mapping and quantification of the 3' ends of transcripts

In Chapter 2 of the dissertation, I describe the different tissue and cell types that we analyzed. This chapter gives an overview of the high throughput method called 3'-seq, which is used for the detection of the 3' ends of the transcripts. It also describes the steps for identification and quantification of the 3' ends of the transcripts. The known artifacts of the sequencing protocol are also discussed.

1.2.2 A comprehensive atlas of intronic and 3' UTR cleavage events

In Chapter 3, I outline the steps of preparing an atlas of robustly expressed IpA and 3' UTR ApA isoforms. We were cautious about all the sources of artifacts that could give false positive peaks or events that could not be assigned to a gene with high certainty. Any peaks that could have originated from artifacts were discarded. We investigated the possible sources that could skew the normalization for specific cell types and took them into account. We filtered for peaks that were robustly expressed using different criteria. We were able to create an atlas containing a comprehensive set of intronic and 3' UTR cleavage events in our cell types. To ensure that the intronic cleavage events are real, we gathered support for the existence of those events through other sources. To rescue genes with convergent 3' UTRs, we learned the shape of the peaks and rescued some genes that could not be analyzed previously. We also describe the statistical framework that we used for identifying the cleavage events that are regulated between conditions.

1.2.3 Characterization and regulation of IpA isoforms across normal tissue and cell types

In chapter 4 we establish that IpA isoforms are expressed as part of normal expression program. We investigated the frequency of occurrence of IpA isoforms across normal tissue and cell types and whether the usage of the pASs is shared across cell types. To understand IpA more comprehensively, we examined where in the gene locus IpA occurs and the properties of genes in which it occurs. We learned that IpA is most frequently observed in the beginning of the transcription units. In addition, we also investigated the regulation of expression of IpA isoforms between conditions using a rigorous statistical framework.

1.2.4 Deregulation of expression of IpA isoforms in B cell malignancies and its functional consequences

We were interested in investigating how the landscape of usage of IpA sites changes in malignant cells when compared to normal cells. In chapter 5, we identified the IpA isoforms that are differentially expressed in chronic lymphocytic leukemia (CLL) using 3'-seq in patient samples. We found that the genes that express IpA isoforms are also enriched for truncating mutations. This finding justifies our proposed hypothesis, as the truncated mRNAs created by the usage of IpA site could potentially mimic genetic mutations. We also examined the IpA isoforms of one of the frequently mutated genes in CLL, MGA, more rigorously to see if the mRNAs generated by truncating mutations and usage of IpA site share similarities.

1.2.5 Functional roles of the wide-variety of mRNAs generated by intronic polyadenylation

We hypothesized that usage of IpA site could generate a wide variety of mRNAs that could potentially have different functional roles. In chapter 6, we utilized RNA-seq data to establish that indeed the usage of IpA site located towards the 5' end of the transcription unit would create non-coding mRNAs. Further we showed that these non-coding RNAs are enriched for RNA-binding protein sites and thus potentially could play a role in gene regulation. We also showed that IpA isoforms generated by the usage of IpA sites towards the 3' end of the transcription unit would potentially contribute towards the diversity of the proteome. In addition, our results showed that only in one-third of cases does the expression of the full-length isoforms go down with higher expression of IpA isoforms.

By our rigorous analysis of the IpA isoforms we established that expression and regulation of these isoforms is part of the normal expression program. Expression of IpA isoforms is regulated between cellular conditions and tissue types. Their expression appears to be functionally relevant. We also unraveled a novel mechanism by which genetic mutations can be mimicked by the usage of an IpA site. We found expression of IpA isoforms to be deregulated in malignant cells compared to their normal cells. In chronic lymphocytic leukemia, these isoforms exhibit the potential to mimick genetic mutations.

CHAPTER 2

2. THE ACCURATE, QUANTITATIVE, AND HIGH-THROUGHPUT MAPPING OF MRNA CLEAVAGE EVENTS IN THE 3' UTR AND INTRONS OF TRANSCRIPTION UNITS

2.1 Introduction

Studies have shown that the cellular machinery uses alternative cleavage and polyadenylation in introns in important development events. In the B cell development, the full-length expression of the IgM gene in mature B cells switches to a truncated form by the usage of an IpA site in plasma cells (Early et al. 1980; Rogers et al. 1980). Another case exhibiting such regulation is the regulation of an IpA site that acts as a developmental switch for the SREPF transcription factor to generate different protein isoforms in spermatogenesis (Wang et al. 2006). These studies show that usage of IpA site can lead to alternative functional protein isoforms with important developmental roles. However, there has been no study that characterizes the global landscape, abundance and functional relevance of these isoforms across various tissue types. To understand the role of intronic polyadenylation (IpA) globally, we aimed to characterize IpA events along with 3' UTR ApA events on a genome-wide scale in a diverse human cell types, with a focus on the B lineage cells. Our goal was to examine the role of IpA in generating different protein isoforms for maintaining normal physiological states of the cell and how changes in the IpA landscape result in alternative protein isoforms in B cell malignancies. In this chapter, we describe the tissue and cell types that we analyzed and methods for mapping and quantifying intronic and 3' UTR cleavage events.

2.2 Description of tissue and cell types being analyzed

I) 3'-seq data-set

To accomplish our objective, we needed a high throughput method that precisely identifies the 3' end of the mRNA and accurately quantifies usage of cleavage sites in introns and 3' UTRs. Lianoglou et al. (Lianoglou et al. 2013) developed a next-generation sequencing method, 3'-seq, that identifies the 3' ends of the transcripts at single nucleotide resolution. This method not only accurately identifies the 3' UTR ApA and IpA isoforms of mRNA transcripts on a genome-wide level but also is quantitative at the level of ApA isoform expression. We decided to perform global mapping of IpA as well as 3' UTR ApA cleavage events by 3'-seq across a wide variety of immune cell types, primarily B cells, T cells and two B cell malignancies, chronic lymphocytic leukemia (CLL) and multiple myeloma (MM).

We performed 3'-seq on the B lineage cell types obtained from healthy donors, as described below.

B cell types: B cells were first isolated using a B cell kit. It was ensured that these B cells are CD3$^-$and CD19 positive. Subsequently, these B cells were sorted by FACS to obtain specific B cell populations. The B cells were selected for the following markers:

i. **Naïve B cells (2 donors)**: CD38$^-$ and CD27$^-$
ii. **Memory B cells (2 donors)**: CD38$^-$ and CD27$^+$
iii. **Germinal Center B cells (2 donors)**: CD38$^+$
iv. **CD5$^+$ B cells (4 donors)**: CD5$^+$

v. **Plasma cells (3 donors)**: CD138$^+$

All the B cells (except plasma B cells) described above were obtained from a lymphatic tissue (tonsils). The plasma B cells were obtained from bone marrow aspirates.

T cells: T cells were isolated from peripheral blood using CD3 beads. These beads captured CD3$^+$ T cells. It was ensured that these CD3$^+$ T cells were CD19$^-$ (a marker for B cells).

As we were interested in investigating the change in landscape of IpA in cancer, we performed 3'-seq on samples obtained from patients suffering from two different B cell malignancies, described below.

Chronic lymphocytic leukemia (13 CLL patients): These cells were obtained from 13 different patients suffering from chronic lymphocytic leukemia. The cells were isolated from peripheral blood using B-CLL Cell Isolation kit (Miltenyi).

Multiple Myeloma (15 MM patients): We obtained the multiple myeloma cells from 15 patients suffering from different stages of multiple myeloma. The cells were isolated from bone marrow aspirate.

Apart from these cell types we also included the 3'-seq samples from Lianoglou et al. (Lianoglou et al. 2013). This 3'-seq dataset comprised of: human ES cell line H9, naïve B cells (CD27$^-$/CD20$^+$) from peripheral blood samples of two different healthy donors, testis, ovary, brain, breast and skeletal muscle. It also contains the following cell lines: HEK293, HeLa, MCF10A, MCF7 and NTERA. We used all the above described 3'-seq samples for creating an atlas of IpA and 3' UTR ApA events.

II) RNA-seq dataset

We also performed RNA-seq for mRNA expression quantification on the above mentioned cell types which we utilized for different purposes in the analysis. For the majority of cases, RNA-seq was performed on the same samples for which we had the 3'-seq data with some exceptions that are noted below.

B cell types:

i) **$CD5^+$ B cells (3 donors)**: 2 samples were the same for which we had 3'-seq data-set. The remaining 1 sample was obtained from the blood of a healthy donor.

ii) **Naïve B cells (6 donors)**: In this case also we had 2 samples for which we also performed 3'-seq. Amongst the remaining 4 samples, 1 was obtained from tonsil of a healthy donor and the other 3 were obtained from the blood.

iii) **Memory B cells (5 donors)**: 1 sample was the same for which 3'-seq was performed. The other 3 were obtained from blood and 1 was obtained from the tonsil of a different donor.

iv) **Germinal Center B cells (4 donors)**: 2 samples were common to 3'-seq and RNA-seq. The other 2 samples were obtained from other donors.

v) **Plasma cells (13 donors):** All the samples in this case were obtained from different donors compared to the patients for we had the 3'-seq dataset. This RNA-seq was performed at a different facility than the remaining samples.

Leukemias:

i) **CLL (9 patients):** For 7 patients we had the RNA-seq and 3'-seq data. RNA-seq for the other 2 patients was performed at a different facility.

ii) **MM (12 patients):** These were same samples on which we performed 3'-seq. We did not perform RNA-seq on samples from 3 patients for which we had 3'-seq data. This RNA-seq was done at the same facility as the plasma cells.

Apart from the RNA-seq for our own samples, we also obtained RNA-seq for immune cells from published studies. The following samples were collected from different studies: 4 samples of naïve B cells from tonsil, 4 samples of germinal center B cells (tonsils), 2 samples of naïve B cells (blood) (ERR431624, ERR431586) (Ranzani et al. 2015), 1 sample of CD3$^+$ cells; GEO: GSM1576415 (Hoek et al. 2015).

2.3 The 3'-seq protocol

3'-seq is a high throughput tag-based sequencing protocol that aims to identify the 3' ends of the transcripts at single base pair resolution. The main advantage of 3'-seq is that apart from precise identification of cleavage events, it also provides accurate quantification of the IpA and 3' UTR ApA isoforms. 3'-seq uses an oligo(dT) bound to magnetic beads to pull down polyadenylated mRNAs and sequences the 50nt at the most 3' end of the mRNAs. It generates reads that align to the sense strand of the genes. Illumina HiSeq was used for sequencing the reads. This protocol was performed for all the samples as described in Lianoglou et al. to generate the 3'-seq dataset that forms the basis of all the analysis and results in this dissertation (Lianoglou et al. 2013).

2.4 Processing the 3'-seq data

Pre-processing 3'-seq libraries The raw reads generated by 3'-seq needed pre-processing before they could be aligned to the genome. The three main steps of pre-processing were:

i) Trimming low quality base pairs from the 3'-end

ii) Trimming the 3' sequence adapters

iii) Trimming of homopolymer As: When the oligo(dT) primed to the middle of the polyA tail instead of the 3' end of the transcript, then we ended up sequencing part of the polyA tail. These polyAs were trimmed from the reads.

After these steps of pre-processing, the reads that were 21 nt or longer were retained for aligning to the genome.

Alignment to the genome The pre-processed reads were aligned to the hg19 genome. As 3'-seq generates reads that come from the 3'-ends of the transcripts, there was little chance for these reads to be spliced. Thus, we used a short read aligner, Burrows-Wheeler Aligner (bwa) for aligning the 3'-seq reads against the hg19 genome (Li and Durbin 2009). The final outcome of this alignment was a BAM (Binary Alignment/Map) file, which was used for the subsequent steps of peak calling and isoform expression quantification. The pre-processing and alignment was performed for all the 3'-seq samples to create a BAM for each sample.

Peak calling of pooled samples To proceed with the subsequent analysis, rigorous identification of the peaks generated by read coverage of the genomic alignments was required. Thus for this purpose we made use of the peak caller developed by Lianoglou et al. that identifies peaks by detecting edges of the genomic alignments (Lianoglou et al. 2013). This study suggested the pooling of all the experiments provides very robust detection of peaks as i) peaks are more detectable at higher depth of coverage and ii) batch affect are averaged out across samples. As the majority of cleavage sites are the same across cell types, pooling of samples aids in identification of the location of cleavage sites. Encouraged by this rationale we combined all the 3'-seq samples and performed peak calling. Alignment of the reads against a genome provides two kinds of reads, i) reads that map uniquely to the genome and ii) reads that map to more than one location in the genome (multi-mapping). For quantification of expression levels, usage of uniquely mapping reads is widely accepted. However, Lianoglou et al. showed that peak calling provides the expected results if both unique as well as multi-mapping reads are used (Lianoglou et al. 2013). Thus, we used both types of reads for identification of peaks. Peak calling provided a comprehensive map of the genomic locations in the hg19 genome where a cleavage site was detected in at least one sample. These peaks were annotated with genomic context, defining whether a peak overlapped with either 3' UTR, extended 3' UTR (UTR3*, up to 5000 nt downstream of the annotated end of the 3' UTR), annotated 3' UTR of one isoform that is defined as intron in the other isoform (Intron.UTR3), introns, 5' UTR, extended 5' UTR (UTR5*, 5000 nt upstream of the annotated end of the 5' UTR), coding region (CDS) of a gene or non-coding genes (UTR). 3'-seq reads also map to intergenic regions of the genome. RefSeq annotation was used for annotation purposes. After we identified the location of cleavage sites, the peaks were quantified.

Quantification of peaks per sample Using the universe of genomic coordinates where a 3' end of a transcript was detected, quantification of the expression levels of

IpA and 3' UTR ApA isoforms was done. Quantification was performed by counting the number of unique reads per sample that mapped within the region defined as a peak during the peak-calling step. Figure 2.4 shows the typical distribution of the reads mapping to different regions of the genome for a sample. Reads that spanned multiple peaks were assigned to the peak with the maximum overlap. The total number of unique reads mapped to the genome constituted the library size for the sample. Library size provides as estimate of the depth of sequencing for the different sample libraries and can be used for normalizing the expression levels of the isoforms by the sequencing depth.

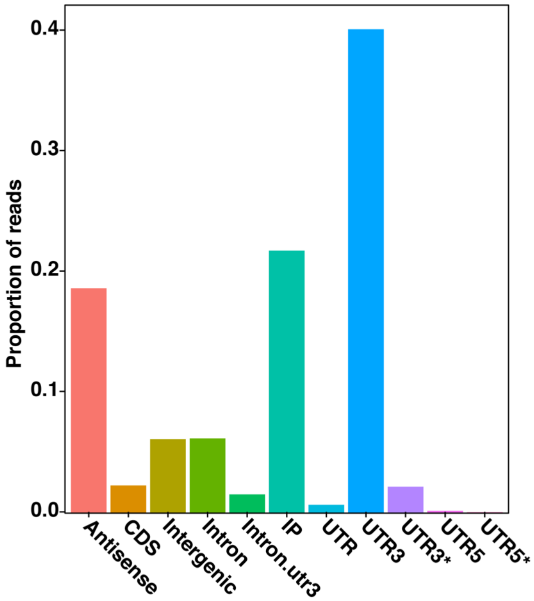

Figure 2.4: Distribution of 3'-seq unique reads mapping to different parts of the genes in entire genome
Majority of 3'-seq reads map to the 3' UTR as expected. However, a fraction of reads generated by sequencing protocol artifacts are the internally primed (IP) and antisense reads.

Interquartile range (IQR) of the start position of the reads 3'-seq identifies 3' ends at single nucleotide resolution, thus the end position of the reads mapping into the peak have highly similar end coordinates. However, the start position of reads always has much higher variability for real peaks. We noticed spurious peaks that had very similar start position for the reads, indicating that these reads were potentially PCR duplicates. Thus to filter such peaks we calculated the interquartile range of the start position of the reads that overlapped a peak as it would give an estimate of the variability of the start position of the reads. In the final atlas, these spurious peaks were eliminated by using this IQR value.

2.5 3'-seq sequencing protocol artifacts

Annotation of internally primed peaks Lianoglou et al. suggested internal priming (IP) to be a major source of artifactual 3' ends. They pointed out that genomic region with A-rich stretches have a high possibility to anneal to the oligo-dTs used in 3'-seq followed by reverse transcription of these regions. A scenario like this would create peaks that would mimic internal cleavage events. These false positive peaks should be removed before proceeding with the downstream analysis. Thus, we followed the same steps as described in Lianoglou et al. to annotate these peaks as internally primed. These would be eventually removed from the atlas in the later steps. Figure 2.4 shows an example of distribution of 3'-seq reads of a sample that has internally primed reads.

Antisense reads Another source of artifactual reads was the antisense reads. 3'-seq generates some reads that align to the opposite strand of the gene (Figure 2.4). These reads were annotated as antisense reads, and there is not very clear understanding of the reasons that leads to sequencing of these reads. The amount of antisense reads varied between the samples.

CHAPTER 3

3. ACCURATELY CHARACTERIZING THE LANDSCAPE OF ALTERNATIVE CLEAVAGE AND POLYADENYLATION IN HUMANS

The 3'-seq data pre-processing, alignment, peak calling and quantification gave a comprehensive set of 3' cleavage events identified from various tissue and cell types. We wanted to focus only on the robustly expressed isoforms free of possible artifacts. We followed a series of steps to create an atlas of robust 3' cleavage events that were detected in the cell types being analyzed.

3.1 Atlas creation

To create an atlas of robust 3' cleavage events, we started with all the possible peaks that were detected by peak calling of all the pooled samples and then followed the steps below to obtain a comprehensive atlas of robust cleavage events.

1. Next generation sequencing (e.g. ChIP-seq, MNase-seq, DNase-seq, FAIRE-seq) based functional genomics experiments often tend to produce artificial signal for certain regions in the genome. Certain such regions where high artificial signal (excessively high read mapping) has been observed repetitively across a large number of independent next-generation sequencing experiments have been annotated as “blacklisted” regions of the genome. These regions have unique mappability relative to the reference genome, and thus they are not removed by mappability filters. In our 3'-seq experiments we below observed very high read mapping for these regions in

certain samples. We obtained an exhaustive list of these blacklisted regions (https://sites.google.com/site/anshulkundaje/projects/blacklists) which was compiled using the next-generation sequencing experiments of the ENCODE consortium (Consortium 2012). The peaks in these regions were removed from the atlas (n = 3926; 0.14%). We also changed the library size of the samples by taking into account the number of reads of that were removed due to blacklisted peaks, as library size is important for sequencing depth normalization.

2. As described in Section 2.5, in a 3'-seq experiment the annealing of oligo d(T) to a stretch of As can result in false positive cleavage events. We followed criteria similar to Lianoglou et al. to flag a peak as internally primed (Lianoglou et al. 2013). All the peaks annotated as internally primed were removed, leaving the ones that had one of the thirteen functional pAS in the ~10 - ~45nt upstream window (Tian et al. 2005) or were annotated with a known 3' end. We rescued these peaks as they have an upstream functional pAS, which is one of the requirements for endonucleatic cleavage by the cleavage machinery. This step removed 37.79% (n = 1,025,231) of the peaks from the atlas. As internal priming is an artifact of the sequencing protocol we reduced the library size of the samples by the number of reads that were lost resulting from this artifactual signal removal.

3. Our dataset includes plasma cells, fully differentiated B cells that secrete antibodies. As these B cells produce massive amount of antibodies, a large fraction of 3'-seq reads come from parts of the genome (chromosome 2 and chromosome 14) that are transcribed to create the antibodies. It is important to account for this skewed expression of the specific genomic region for getting a reasonable expression of the other genes. Thus, we deleted the peaks (n = 11) overlapping with parts of the genome

coding for antibodies and reduced the library size of all the samples by the number of discarded reads. Figure 3.1.1 shows the distribution of reads before and after this correction. Even after this correction, two samples of plasma cells had a high number of intergenic reads (PC2 and PC3). Thus, we do not use these two samples for identification of robustly expressed isoforms but only to quantify them.

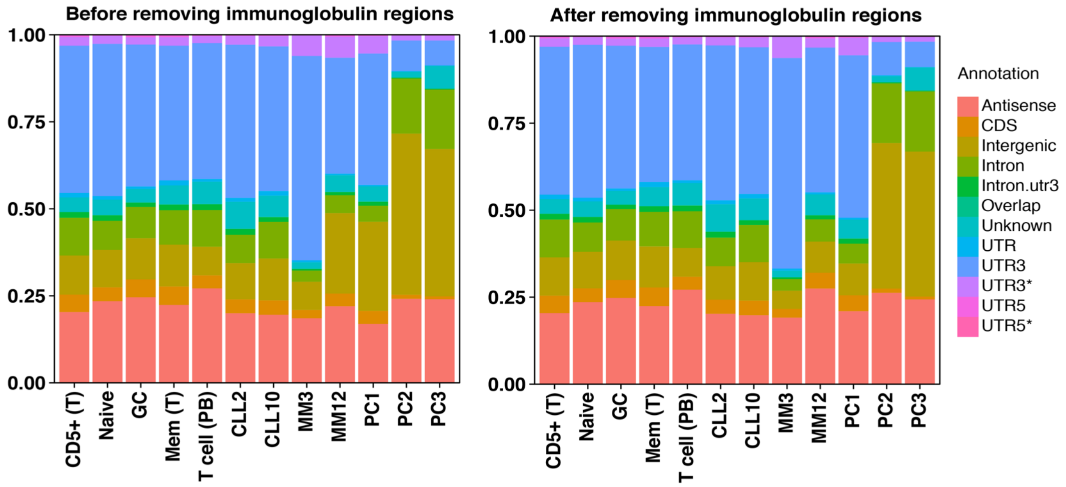

Figure 3.1.1: Skew in distribution of reads because of antibody production
This plot shows the distribution of 3'-seq reads mapping to different regions of the genome across different samples. A large fraction of reads map to intergenic regions for the plasma cells (PC) and multiple myeloma (MM) samples. These regions corresponded to parts of chromosome 2 and chromosome 14 that code for antibodies. To account for this skewed expression of specific regions, we discarded such peaks.

4. In the next step, we removed the peaks that fell in the intergenic regions of the genome. As we were interested only in the 3' cleavage events of the annotated genes of the genome, we removed peaks that were associated with any genomic region that was not annotated as a coding or non-coding gene.

5. To quantify the expression levels of the IpA and 3' UTR ApA isoforms, we determined the tags per million (TPM) of the regions that were called as peaks by our

peak-caller. The read count of the peak regions was normalized by the library size of the respective sample. This provides an assessment of the expression level of the isoforms taking into account the sequencing depth coverage of the 3'-seq library for the individual samples.

6. The genome has genes that fall on opposite strands but have convergent 3' UTRs. Although 3'-seq is a stranded protocol, it still suffers from artifacts of antisense reads. This poses problems with genes that fall on opposite strands but their 3' UTR ends are located within a span of 1000 nt from each other. This is especially a problem if the sense and antisense peaks overlap. To have a correct assessment of the sense and antisense peaks, we tried to assign the peaks to the respective genes based on the shape and expression of the peaks. For this we made use of a supervised learning method that can learn the shape of the sense peaks and distinguish the real sense and antisense peaks for these convergent genes. The entire learning process is described in detail in Section 3.3. We focused on genes that had decently expressed sense and antisense peaks. If the learning algorithm was able to classify the peaks as sense and antisense, then we included the sense peak in the atlas; otherwise the peak was discarded.

7. In the next step all the peaks that were annotated as antisense were removed (15.70%; n = 425,862). We reduced the library size for the samples as these antisense peaks were the result of sequencing protocol artifacts.

8. Some genes in genome overlap with each other. In such cases, it would be difficult to assign the 3'-seq reads to the genes accurately. Thus we decided to remove such genes from analysis (n = 338). This resulted in removal of 11,489 peaks from the atlas.

9. The genes that were on opposite strands but have a 3' UTR end in the intron (100 nt) of the convergent gene could create artifactual antisense peaks in the intron. Thus, peaks in introns that were close to the end of an opposite strand 3' UTR were also removed (n = 2,423). This corresponded to discarding peaks in the introns of 668 genes.

10. There are genes where the 3' UTR ends could continue in the intron of the gene that follows on the same strand. This would also create peaks in introns that would be contributed from the preceding gene. Peaks in the intron that were within 5000 nt of the 3' end of the 3' UTR of the previous gene were also discarded (n = 3036). Peaks from introns of 903 genes were removed.

11. A lot of microRNAs (miRNAs) and small nucleolar RNAs (snoRNAs) are located in the introns of other genes. As we were interested in investigating the IpA isoforms of the protein coding genes, peaks originating from the 3' supervisedends of the miRNAs and snoRNAs located in introns would not represent real IpA isoforms. Thus, we decided to remove peaks in introns that were within 500 nt of an annotated miRNA or snoRNA. We discarded 310 peaks that were potentially 3' ends of miRNAs or snoRNAs from 183 genes.

12. To avoid peaks that came from retrotransposons located in the introns, the peaks that overlapped with these retrotransposons were also removed (n = 5433, 2066 retrotransposons). This is based on the annotated retrotransposons in ucscRetroInfo5 track.

13. A gene could have many cleavage events with adequate expression levels. However, we wanted to examine cleavage events that represented one of the substantial isoforms with respect to all the other isoforms of the genes, thus we decided to filter these isoforms by a usage index (UI). A usage index is a statistic that gives an estimate of the relative expression of the isoform. As the 3' UTR ApA isoforms would create the same protein irrespective of the 3' UTR length, the usage of the isoforms was calculated with respect to the total expression of 3' UTR ApA isoforms. The IpA isoforms would result into distinct proteins, thus the usage of the IpA isoforms was calculated relative to other IpA isoforms and 3' UTR isoforms.

14. We were interested in analyzing functionally relevant isoforms, so we filtered the robustly expressed isoforms by using a TPM and a usage index filter. For the 3' UTR ApA isoforms, an isoform that was expressed with at least 3 TPM and with usage index of 0.1 or more was added to the atlas. To be able to focus on the more robust IpA isoforms we considered an IpA isoform to be robustly expressed only when it was expressed with 5 TPM or more and had at least 0.1 usage index in at least one sample. We also required the IQR of the start position of the reads to be 5 or more for the peaks in that particular sample to be defined as a real IpA isoform. These criteria helped to filter the lowly expressed isoforms as well as any possible known artifacts. Filtering for these expression criteria shrunk the atlas of 561,750 peaks to 48,635.

15. As we were interested in IpA and 3' UTR ApA isoforms that would have different functional consequences, the isoforms that have the ends very close would be highly alike. Thus driven by this rationale, we clustered the peaks that were within 200 nt to represent one 3' cleavage event. As we did not want to cluster noisy events, we discarded cleavage events that were not robustly expressed. Clustering reduced 48,635

peaks to 41,605. Figure 3.1.2 shows the final distribution of the reads in a sample after all cleaning steps.

After following the steps above, the atlas had 28,720 peaks from 15,855 genes for the cleavage events of the 3' UTRs. This atlas contained 3' ends of 6,465 IpA isoforms of 4,206 genes.

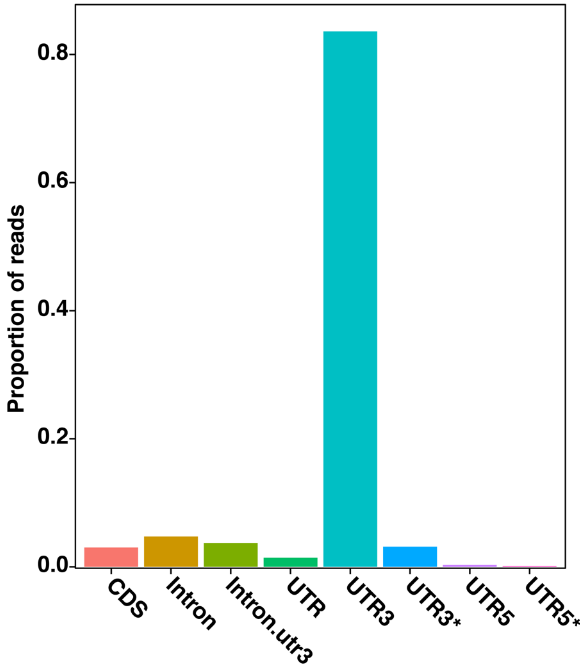

Figure 3.1.2: Distribution of 3'-seq reads in the atlas
The 3'-seq libraries were pre-processed and carefully filtered to remove all the known artifacts. Different criteria were used to identify robust cleavage events. After artifact removal and filtering of noisy events, majority of 3'-seq reads mapped to the 3' UTR as expected with a smaller fraction of reads mapping to other genomic loci.

3.2 External validation of IpA isoforms

The primary goal of this dissertation was to characterize the IpA isoforms and investigate the differential usage of IpA across the immune cell types and B cell leukemias. Thus, it was important to have extremely confident IpA events in the atlas. Introns tend to be low mapability regions, so in order to avoid any IpA isoforms that would be artifacts, we wanted to gather additional evidence that would support the existence of the IpA isoforms. We tried to corroborate the existing IpA events (n = 6465) with the sources of evidence described below:

1. External annotation There are be cases where the observed IpA isoform is not annotated in the RefSeq database but may be annotated as 3' end in other available genome annotations. If IpA isoforms could be verified with these external annotations, which have the transcript structures from other observed data or computational prediction, then it would strengthen the existence of the IpA isoforms as real 3' ends. Thus, last exons of the all the existing transcripts of hg19 annotation for UCSC and ENSEMBL were obtained. These last exons were resized to get a region 50 nt upstream and downstream of the annotated end. If the IpA isoform 3' end detected by our 3'-seq analysis overlapped with this span of 100 nt, then we annotated it to be substantiated by an external annotation. We found that 29.06% (n = 1,879) of all the IpA events had an annotated 3' end in the vicinity using an external annotation.

2. RNA-seq evidence RNA-seq read coverage is expected only over the exons and not over the introns, since the splicing machinery splices them out during co-transcriptional processing of the pre-mRNA. However, if there is an IpA isoform that ends in an intron, then there should be RNA-seq read coverage before the 3' end of the IpA isoform and no read coverage after the 3' end (Figure 3.2.1). We leveraged this information from RNA-seq read coverage and identified the IpA isoforms that were corroborated by RNA-seq in this manner.

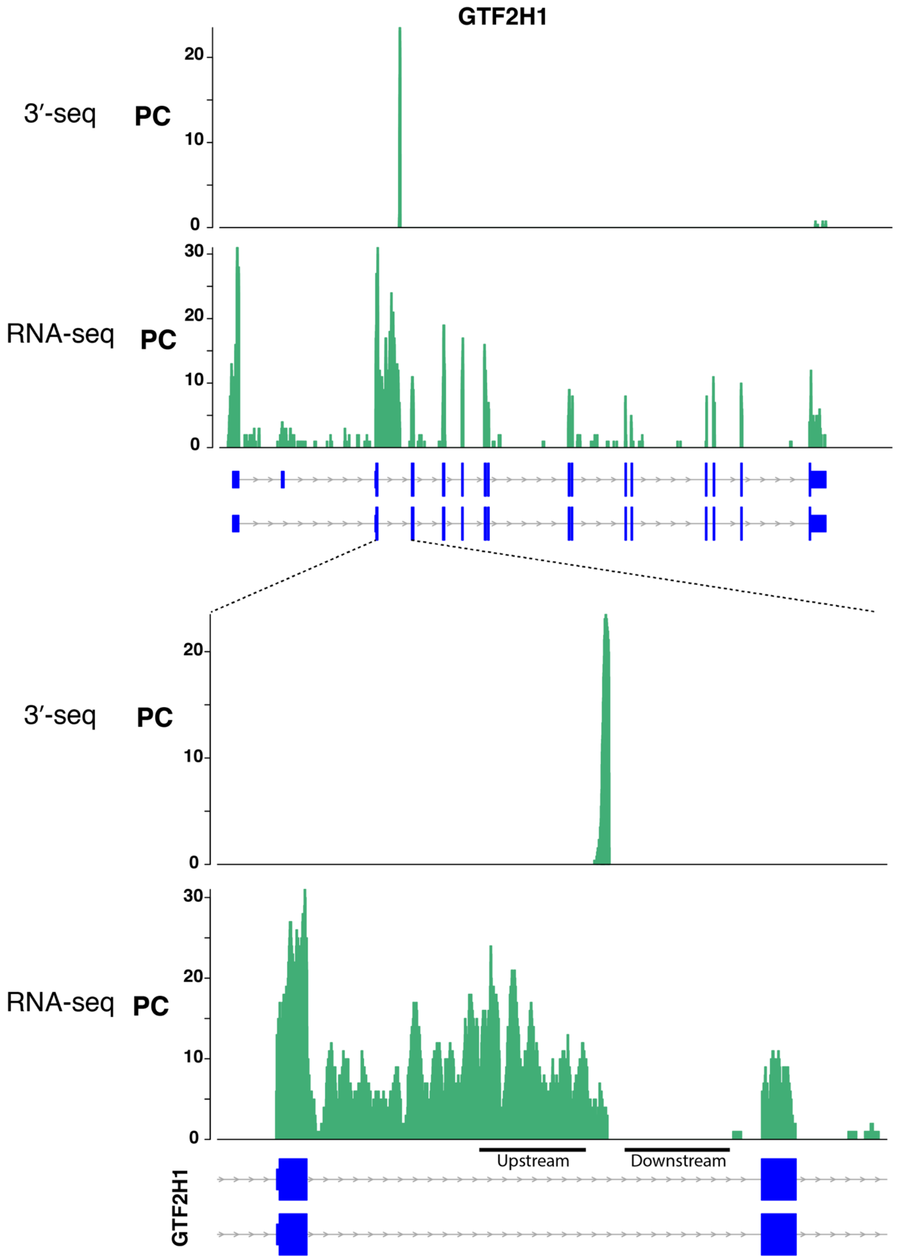

Figure 3.2.1: Coverage of intronic regions in RNA-seq

Higher coverage upstream of the cleavage event in the introns is expected in RNA-seq as evidemce of true IpA isoforms. Leveraging this knowledge, we can identify IpA events where we can observe higher genomic coverage upstream in RNA-seq profile compared to downstream. Using a GLM based model we identify the IpA events that have significantly higher coverage upstream than downstream of the 3' end in the introns of RNA-seq profiles. This example shows the case where a 3' end can be validated by RNA-seq.

- To test whether the upstream read coverage was more than the downstream read coverage we created two windows of 100 nt separated by 51 nt upstream and downstream of the IpA 3' end. These two windows served as replicates for the upstream and downstream coverage. As we did this within every RNA-seq sample, library size normalization was not required. Thus in this analysis the size factor that normalizes for the sequencing depth of the library was set as 1 for every comparison. Now we tested if there was significant differential expression upstream vs downstream using DESeq (Anders and Huber 2010). If the IpA isoform was validated by RNA-seq then the upstream read coverage should be significantly higher than the downstream coverage (adjusted $p \leq 0.1$) (Figure 3.2.2).

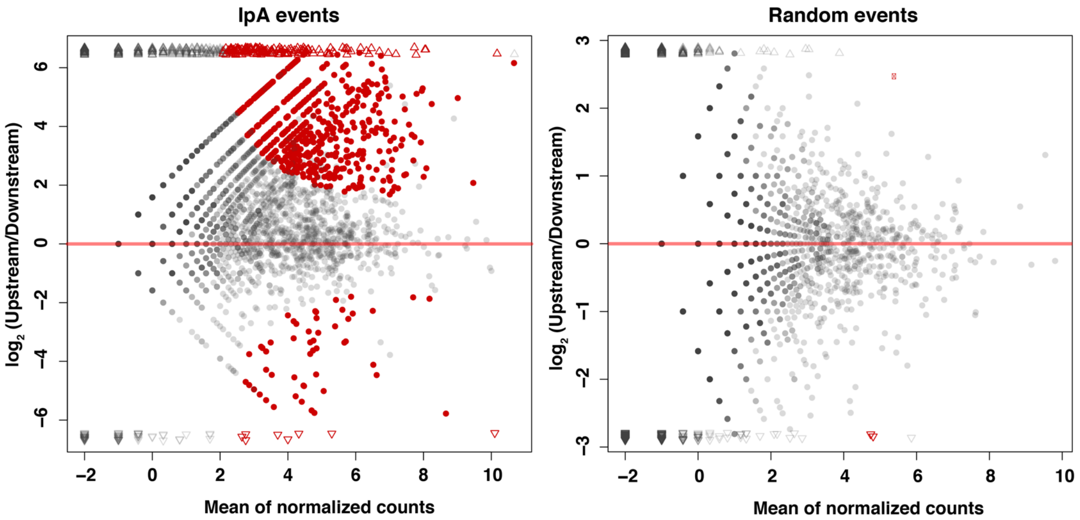

Figure 3.2.2: Validation of IpA events using RNA-seq

Higher coverage is expected upstream of the cleavage event in the introns when compared to downstream. Windows were defined upstream and downstream (Figure 3.2.1) of the cleavage event and were tested for differential coverage using a statistical framework. As expected we detected significantly higher coverage upstream than downstream of 20% cleavage events in the introns. We did not observe such differential coverage when we repeated a similar analysis for windows upstream and downstream of random points in introns

- Not all IpA isoforms could be validated by this approach. To avoid cases that could be affected by other confounding factors, a clean list of IpA events that could be validated by this method was prepared. IpA isoforms where the defined windows overlapped with an annotated exon were excluded from the further analysis. An additional filter was exclusion of cases where an intron had more than one IpA isoform. Having multiple IpA isoforms ending in the same intron would be confounding.

- As a control for this analysis, we chose random introns of expressed genes that did not contain 3' end peaks. From the center of the intron we created windows upstream and downstream windows as described above. For random introns we should see similar coverage upstream and downstream (Figure 3.2.2).

The RNA-seq validation was applied over all the RNA-seq samples. If an IpA event was validated in any sample, then we considered it to be supported by RNA-seq data. We were able to validate 20.12% (n = 1301) IpA events with this methodology.

3. Untemplated As from RNA-seq reads (polyA reads): In RNA-seq there are reads that come from the 3' end of the genes. Sometimes these reads have the untemplated As of the polyA tail and thus fail to map to the genome. We made use of the reads that did not map to the human genome to get additional support for the IpA 3'ends in our atlas. To make sure that these reads were the 3' end reads, the reads that had 4 of more As at the end were trimmed. Only the reads that were greater than 21 nt in length were kept for further processing. Unmapped reads of all the RNA-seq samples were trimmed and then all the reads with untemplated As were pooled. These reads were later aligned to the human genome. From the aligned BAM file, all the reads that were possible PCR duplicates were further filtered out. The uniquely mappable reads overlapping with the IpA peak were counted. These reads supported the presence of 3' end from an alternative sequencing protocol, RNA-seq that is used for measuring the expression levels of the transcriptome. If an IpA isoform had 4 or more reads that supported the presence of 3' end then the IpA isoform had another source of evidence. This approach supported the existence of 22.72% (n = 1469) of IpA isoforms.

4. Another 3' end sequencing protocol – A 3' end detected by our 3'-seq and observed also with another 3' end detection method would have additional confidence for the presence of the 3' end. This step of corroboration would make sure that the observed 3' end is not the result of the sequencing protocol biases as it is also being observed with another independent method. We found 60.97% (n = 3942) IpA isoforms expressed in our samples to be also expressed in atlas of 3' ends identified by an independent study (Derti et al. 2012).

5. Presence of a polyA site (pAS) – It has been established decades ago that a functional pAS is critical for an efficient 3' end processing of the mRNA. The signal is recognized by the cleavage factors followed by the endonucleatic cleavage of the transcript. This hexamer signal (AAUAAA) appears 10-40 nt upstream of the cleavage site. Presence of this hexamer or one its variant upstream of the IpA 3' ends would suggest the IpA events to be resulting from real 3' ends. 89.50% (n = 5786) IpA isoforms were verified by a pAS signal.

By the above five approaches we were able to validate 6151 IpA events of the atlas.

Some of these intronic 3' ends are already annotated in RefSeq. We compared the proportion of annotated and unannotated intronic 3' ends that could be validated by the above five approaches (Figure 3.2.3). Validation by RNA-seq, polyA reads and polyA signal were fairly similar for both the categories, which shows that the unannoated IpA isoforms are not resulting from artifacts. We validated fewer unannotated isoforms by external annotation as we were aware that most of these isoforms have not been annotated in any database. Even another independent protocol supported the presence of fewer unannotated isoforms as this data set does not include all the cell types present in our data set. The remaining 314 IpA events were discarded. In the subsequent analysis we investigated these IpA events and discarded the remaining IpA

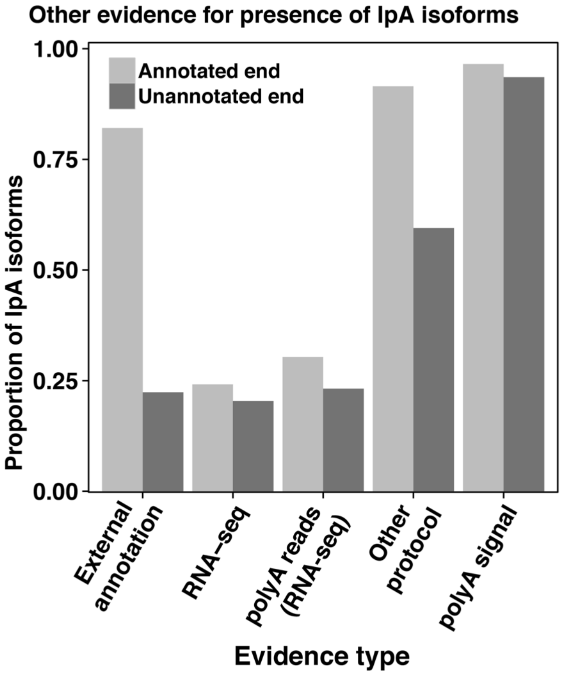

Figure 3.2.3: Validation of IpA events

IpA events were validated using other resources. Some of the IpA events are already annotated as 3' ends in the introns in RefSeq. We used these annotated 3' ends as baseline to assess how well the unannotated 3' ends could be validated by using these other resources. Similar proportions of annotated and unannotated cleavage events were supported by RNA-seq, polyA reads of RNA-seq and polyA signal. Higher numbers of annotated ends were supported by external annotation and other protocol. This was expected, as we knew that majority of IpA events have not been annotated in any database. The lack of support from other protocol was potentially because we had many more cell types in our dataset.

events that were not corroborated any of the above methods. This gave us the final version of the atlas that would comprise the universe of the 3' UTR ApA and IpA events that would be investigated in the later analysis.

3.3 Rescuing genes with convergent 3' UTRs

The genome has certain genes falling on the opposite strands but having convergent 3' UTRs. In these cases, it is difficult to conclude whether the 3'-seq genomic read alignment is the outcome of expression of the gene on the sense strand or of antisense reads of the gene from the opposite strand. However, from examination of 3'-seq genomic read alignment we perceived that the shape of an authentic sense peak and could be learned using a supervised learning algorithm. This learning procedure could be used to distinguish between a true sense and antisense peak. A successful discrimination between an actual sense and antisense peak would help to include these genes with convergent 3' UTRs in our analysis. By closely examining the shape of the sense peaks, we realized that the shape of the peaks could be learned by using the start and end position of the reads overlapping with the genomic peak region. Thus we decided to use the start and position of the reads as features for the learning algorithm. We reasoned that the positive and negative cases for learning the model that could discriminate between the sense and antisense peaks will be the non-convergent genes that have a well expressed sense and a robustly expressed antisense peak. Sense and antisense peaks of such non-convergent genes could be used to learn the parameters of the model. These learned parameters could then be applied to convergent genes to identify real sense and antisense peaks. As retaining the start and end positions of all the reads mapping to the genome would have been extremely memory intensive, so we decided to do the learning with only few samples from the complete 3'-seq dataset. We chose a diverse range of immune cell types for this purpose: GC B cells, T cells, naïve B cells, CD5^{+} B cells and two CLL samples. As described earlier, firstly the peak calling was performed, and then the quantification of these peaks was performed. In this process, apart from the number of reads that map to the peak region, we also saved the start and end position of all the reads mapping into a peak.

Convergent and Non-Convergent Genes To proceed with the learning process we first needed the set of convergent and non-convergent genes. For this the raw atlas of peaks without any filtering was used as the starting point. The peaks possibly arising from artifacts were first removed by eliminating events that fell in the blacklisted region or were annotated as internally primed. Thereafter the peaks that mapped into intergenic regions of the genome or overlapping genes were removed. To get the expression level of the sense and antisense peaks, the TPM of the peaks was calculated. The genes that had convergent 3' UTRs within a distance of 1000 nt were called the convergent genes (n = 2,395) and the remaining genes were the non-convergent genes. To progress with the model learning, only the genes that had a robustly expressed ($\geq$ 5 TPM) sense and antisense peak in any sample were used.

Constructing the features for learning A learning algorithm can be used to distinguish between the shape of sense and antisense peaks. To do this we captured information about where the reads in a peak started and ended with respect to the annotated start and end of the peak. The following steps were used to get this information.

i) The distance of read starts from the annotated start of the peak was calculated. We focused only on reads that ended within 20 nucleotides upstream or downstream of the annotated start of the peak. This information was converted into fraction of reads starting at different distances from the annotated start. This created a 41 length feature vector: -20…0…20 positions upstream and downstream of the start, where 0 represents the start position of the annotated peak.

ii) Similarly, distance of the read ends from the annotated end of the peak was calculated. The reads that ended within 20 nucleotides upstream or downstream of the annotated end of the peak were used. Again this was turned into fraction of reads that end at a certain distance from the annotated end. We observed that, for a true sense peak, the majority of reads end at the same position, which could be slightly different than annotated peak end. Thus to leverage this knowledge, the cleavage distance was centered at the position where the maximum number of reads ended. This also created a feature vector: -20…0…20, where 0 represented the position where the maximum number of reads ended. In this case -20 and 20 represented the positions with the fraction of reads ending 20 nt upstream or downstream of the centered end.

We learned the model using genes having a highly expressed sense peak as well a highly expressed antisense peak (≥ 5 TPM). In order to have a clean positive set for learning, sense peaks with at least 50% of reads ending at the same position in any sample (4720 peak pairs) were used. These sense peaks also had a highly expressed (≥ 5 TPM) corresponding antisense peak, which we were aware of being an artifact. For these set of sense and antisense peaks we constructed the feature matrix using i) the fraction of reads that started within 20 nt to 20 nt downstream of the annotated start of the peak and ii) the fraction of reads that ended within 20 nt to 20 nt downstream of the annotated end of the peak, followed by centering at the position where the maximum number of reads ended. The two vectors were appended to create the final feature matrix. This gave us a feature matrix 4720 × 82 (Figure 3.3.1).

We made use of L1-regularized L2-loss support vector classification (Fan et al. 2008) to learn the model that could classify the sense peaks from antisense peaks for the non-convergent genes. The feature matrix was log transformed for the learning purpose.

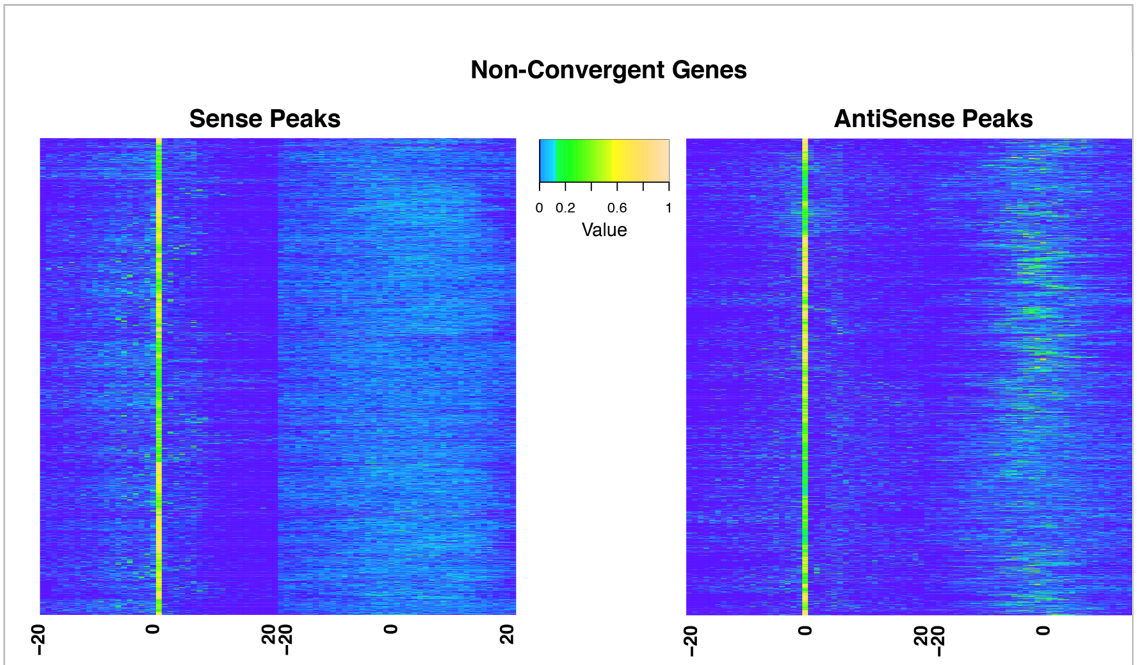

Figure 3.3.1: Feature matrix of sense and antisense peaks for learning from non-convergent genes

Feature matrix for 4720 non convergent genes (rows) with 82 columns. The first -20..0..20 represents the fraction of reads that end within 20 nts of the annotated end of the peak. The 0 in this case was centered at the position where the maximum reads ended. The next -20..0..20 shows the fraction reads that start within 20 nts of the annotated start of the peak, where 0 marks the annotated start of the peak. The differences between the two matrices for sense and antisense peaks can be seen in this figure.

The classification task was performed with an average accuracy of 88.72% by 10 fold cross validation. We also assessed the performance by learning the model on sense and antisense of peaks of 5 samples and predicting it on the 6th sample. However, our aim was to classify the sense and antisense peaks for convergent genes. So for this purpose we learned the classification model on the entire set of sense and antisense peaks of the non-convergent genes. We made use of the weights learnt on this

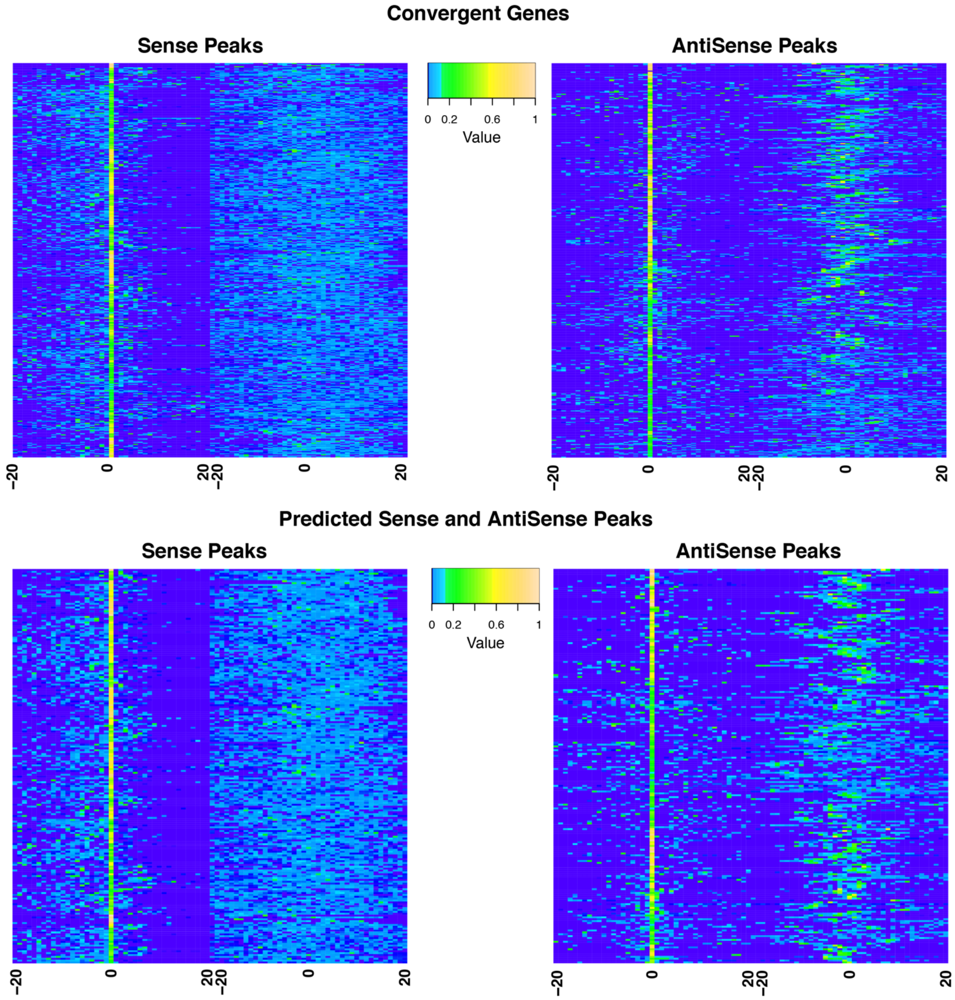

Figure 3.3.2 Feature matrices of convergent genes before and after prediction
The above two matrices (2395 × 82) show the feature matrices for the sense and antisense peaks for all the convergent genes. The weights for these features were learnt using a L1-regularized L2-loss support vector classification scheme. These weights were uses to classify real sense and antisense peaks. This helped to rescue 275 genes (shown in lower half).

complete dataset to classify the peaks of the convergent genes. By this procedure, we classified the peaks that were sense and antisense for the convergent genes. The sense peaks that had a positive score and their corresponding antisense peaks had a negative score, were the correctly classified peaks. Figure 3.3.2 shows the feature matrices of the sense and antisense peaks for convergent genes as well as the sense and antisense peaks that were predicted accurately. We could use such genes in our further analysis. As our expectation was that sense peaks should be highly expressed compared to the antisense peaks, we included peaks that were misclassified (sense as well as the antisense had a positive score). Peaks that were misclassified with the above criteria, but had log2(count) sense peak/log2 (count) antisense peak > 0.2 were added to our correct classification category. This was repeated for every single sample. The peaks that were correctly classified in atleast 66% of samples made the final set of convergent genes that could be further used for the analysis. Using the learning strategy 275 genes with convergent 3' UTRs were added to the atlas of 3'UTR ApA events.

3.4 Testing for differential usage of IpA sites compared to pASs in 3' UTRs

We were interested in identifying statistically significant changes in the relative usage of the intronic pAS and pASs in the 3' UTR of the genes independent of the gene expression changes. All the full-length 3' UTR ApA isoforms would be translated into the same protein so we summed up the expression of all the ApA isoforms to represent the full length mRNAs. The IpA isoforms would be translated into a different protein compared to the 3' UTR ApA isoforms. If a gene has multiple IpA isoforms then we tested the relative expression of each IpA isoform and full-length mRNA independently as each IpA isoform in translated into a different protein. To identify the statistically significant changes in the usage of pASs we made use of generalized linear model (GLM), where we model read counts of both the isoforms as negative binomial distribution. The model is used for testing the significance of the interaction between the expression of the two isoforms and the condition in which it is expressed. This form of modeling approach was borrowed from DEXSeq, which is formulated for testing the differential usage of exons (Anders et al. 2012). Lianoglou et. al. used this approach to test for differential usage of pASs in the 3' UTR of the genes. We used a similar approach, the only difference being that we tested for differential usage of IpA site relative to all the 3' UTR isoforms being treated at one isoform between different conditions.

CHAPTER 4

4. CHARACTERIZATION AND REGULATION OF IpA ISOFORMS ACROSS NORMAL TISSUES AND IMMUNE CELL TYPES

4.1 Introduction

Previous studies have shown that the cellular machinery uses alternative cleavage and polyadenylation in introns in important development events. In the B cell development, the full-length expression of the IgM gene in mature B cells switches to a truncated form by usage of an intronic pAS (IpA site) in fully differentiated plasma cells (Early et al. 1980; Rogers et al. 1980). Another extensively studied case is the calcitonin/calcitonin gene-related peptide gene (CALCA), where the usage of an IpA site is regulated by the splicing factor SRp20 in a tissue-specific manner to make mRNA variants. Further, it has been shown that regulation of an IpA site can act as a developmental switch for the SREPF transcription factor to generate different protein isoforms in spermatogenesis (Wang et al. 2006). It was also demonstrated that during expression of FLT1, recognition of intronic polyadenylation site creates soluble variants of FLT1, which are highly expressed in the placenta (Thomas et al. 2007). These studies show that intronic polyadenylation can lead to alternative functional protein isoforms with important developmental roles. Usage of IpA sites was shown to be a widespread phenomenon using EST/genome sequence data from all the known human genes (Tian et al. 2007). To our knowledge no previous study that has attempted to characterize alternative cleavage and polyadenylation in introns on a genome-wide scale across a wide variety of cell types by leveraging 3'-end sequencing methods. As mentioned earlier, we aimed to study the regulation of the usage of IpA sites across diverse tissue types that could generate protein isoforms with altered functions. Thus, to understand the role and extent of intronic polyadenylation globally, we wanted to characterize intronic alternative cleavage and polyadenylation (IpA) events on a genome-wide scale in a number of diverse cell types. To accomplish these goals we created an atlas of robust IpA events. Using this atlas of IpA events we tried to answer a range of questions, such as frequency of expression of IpA isoforms in various cell types and their preferred location within transcription units. In this chapter we also investigated the differential regulation of the IpA isoforms from a single a cell type (naïve B cells) obtained from two different environments (blood and tonsil). This analysis helped us to establish that IpA isoforms are tightly regulated between cell types and different environments.

4.1 IpA isoforms are expressed as a part of the regular expression program in normal cell types and display high evolutionary conservation

Normal expression program In order to proceed with further analysis we decided to focus on cleavage events of coding genes. We also focused only on IpA isoforms for which the full-length isoform was detected in at least one of the samples. Our atlas constructed using a diverse range of tissue types and immune cell types identified 5434 IpA events. We found that out of 15525 of all the expressed genes in the atlas, 3615 genes expressed at least one IpA isoform. The atlas contained widespread usage of the IpA site in normal cell types. There were 1201 genes that used more than one intronic pAS across the transcription unit. Observing this recurrent usage of the intronic pAS, we concluded that recognition of these pASs by the cleavage machinery was a part of the regular expression program leading to diversification of the transcriptome/proteome. Extensive studies have shown U1snRNP to be critical factor preventing the usage of the pASs in the introns causing premature cleavage and polyadenylation (Kaida et al. 2010; Berg et al. 2012). U1 snRNP has been shown to play an essential role in defining the length of mRNAs and isoform expression. These studies suggest that deficient levels of U1 snRNP induce premature cleavage and polyadenylation (PCPA) by the recognition of cryptic pASs located in the introns of the transcription unit (Berg et al. 2012). However, from our observation of frequent occurrence of the IpA isoforms we deduced that these isoforms should not be regarded to be originating from cryptic pASs. Figure 4.1.1 shows the expression of IpA isoforms of two genes (GTF2H1 and RAB10) in different cell types. We observe a wide range of expression across these cell types. Our subsequent analysis showed usage of IpA sites to be highly regulated across multiple cell types to generate mRNA

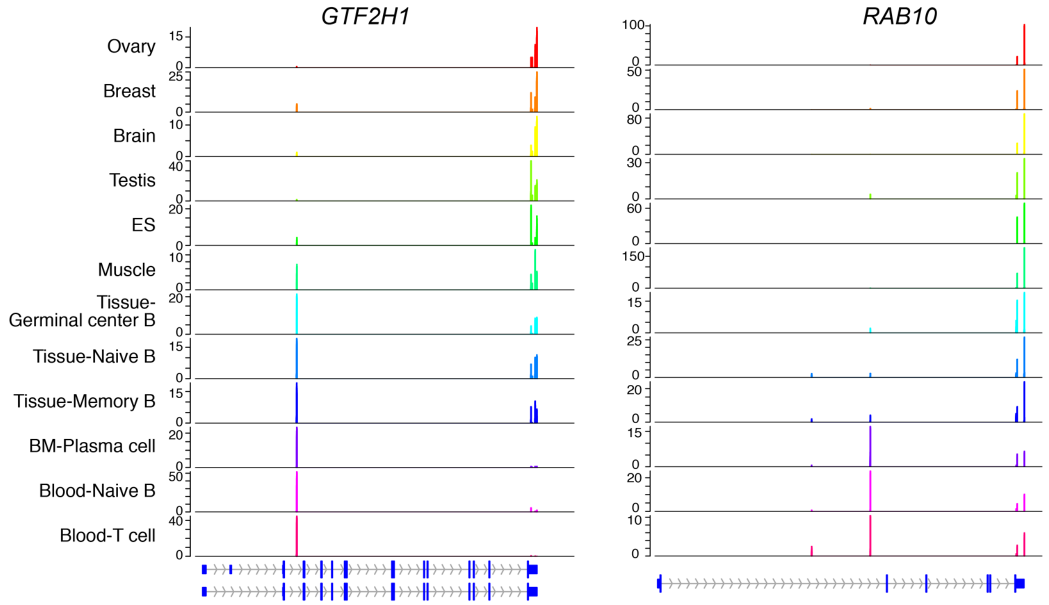

Figure 4.1.1: Expression of IpA isoforms across cell types
The expression of IpA isoforms varies widely across cell types. Usage of IpA sites appear to be regulated across different cell types. IpA isoforms are expressed as part of normal expression program.

isoforms in normal physiological states of the cell.

Robust expression The atlas of IpA isoforms was created to filter out transcriptional noise, eliminating events unlikely to produce highly expressed alternative protein isoforms. However, it was important to assess the robustness of expression of these isoforms relative to the full-length expression of the genes. The IpA isoforms were not as highly expressed as the full-length isoforms but their expression was high enough to be able to conclude that they would potentially contribute substantially to the diversification of the transcriptome. Figure 4.1.2 shows the expression level of the full-length isoforms and IpA isoforms for three cell types. The median expression level ($\log_2$TPM) for full-length isoforms for the PC, naïve B cells (PB) and T cells is 4.54, 5.12 and 5.10 respectively while for the IpA isoforms it is 3.65, 3.80 and 3.65.

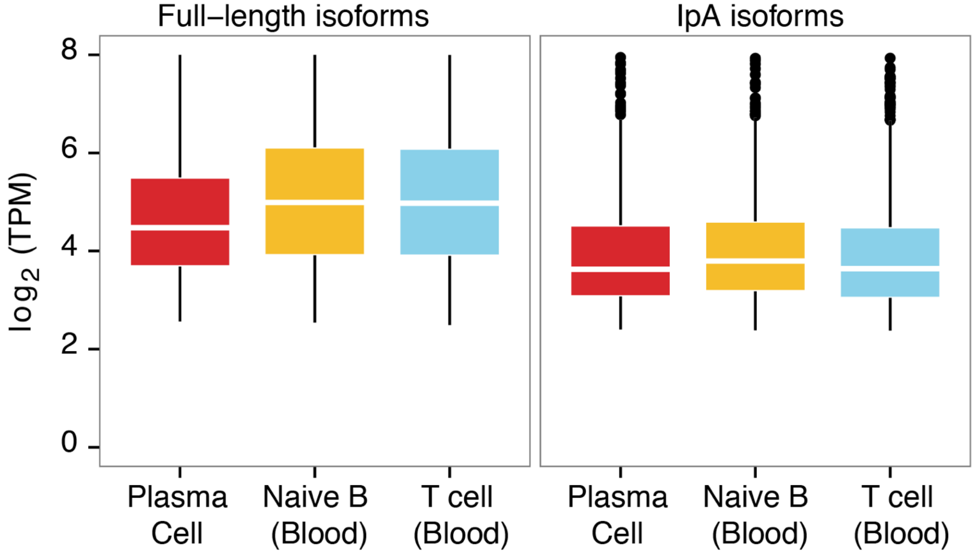

Figure 4.1.2 Robust expression of IpA isoforms
The IpA isoforms are robustly expressed in comparison to the full-length isoforms in the various cell types. The median expression level ($\log_2$TPM) for full-length isoforms for the PC, naïve B cells (PB) and T cells is 4.54, 5.12 and 5.10 respectively while for the IpA isoforms it is 3.65, 3.80 and 3.65.

IpA isoforms are highly conserved compared to introns We wanted to determine if the 3' ends of the IpA isoforms have a different level of conservation compared to the other random introns. The IpA isoforms that did not have an exon upstream and downstream within 200 nts of the 3' end were used for further analysis (n = 4611). We obtained the phastCons 46-way conservation score for 200 nts up and down of the 3' end for these IpA isoforms (Siepel et al. 2005). We wanted to compare the mean conservation score around the cleavage sites of the IpA isoforms against the introns where IpA did not occur. For this purpose, we randomly selected introns (n = 5,000) of the genes that expressed IpA isoforms but did not have peak in that intron. Out of these random introns, we filtered for the ones that had at least one pAS (AAUAAA) in that intron. This would ensure that there was a chance for IpA to happen in that intron. After this step, one of the pAS was randomly selected and we obtained the phastCons

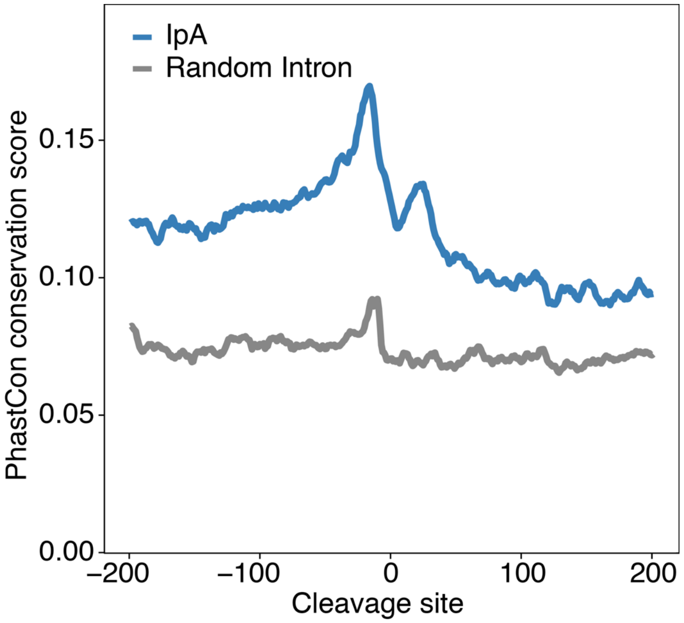

Figure 4.1.3: Higher conservation around the cleavage site of IpA isoforms
The mean sequence conservation upstream and downstream of the cleavage site of the IpA isoforms (n = 4611) is significantly higher compared to the conservation of the sequence upstream and downstream of randomly selected pAS (AAUAAA) from an intron that did not have the 3' end of any IpA isoform (One sided KS test: 1.45 x 10^{-170}).

46-way conservation score for 200 nts upstream and downstream of the pAS. We found that the ends of IpA isoforms are significantly highly conserved (One sided KS test: 1.45 x 10^{-170}) compared to these random introns at pASs (Figure 4.1.3). This observation suggests that IpA isoforms have been under evolutionary selective pressure to be expressed in different cell types and are potentially play important functional roles.

4.2 High usage of IpA sites in immune cell types with recurrent sharing of the IpA sites between immune cell types

Highly used IpA sites in immune cell types We evaluated a wide variety of tissue and immune cell types for the number of genes that express IpA isoforms. We were interested in determining how often IpA isoforms were expressed in different tissue and cell types in our atlas and if there was a biased landscape of usage of IpA sites in specifc cell types. We calculated the fraction of genes expressing at least one IpA isoform compared to all the expressed genes in each cell type. The fraction of expressed genes with IpA isoforms would provide an assessment of the frequency of the IpA isoforms in every cell type. Figure 4.2.1 shows that IpA isoforms are most abundantly expressed in the immune cell types with limited occurrence in the complex tissues. Naïve B cells (blood) had the highest fraction (0.15) of genes with IpA isoforms while brain had the lowest number (0.03) of IpA isoforms. The expression of IpA isoforms also appeared to vary between the same cell types obtained from different environments, as naïve B cells obtained from tonsil had lower number of genes (0.06) with IpA isoforms compared to naïve B cells from blood (0.13). Amongst the immune cell types, the fraction of genes with IpA isoforms varied from 0.05 to 0.15, while the complex tissues ranged from 0.03 to 0.07. With this observation we concluded that IpA isoforms appear to be most abundant in immune cells. This observation reinforces the previous claim of IpA isoforms being expressed as a part of normal expression program.

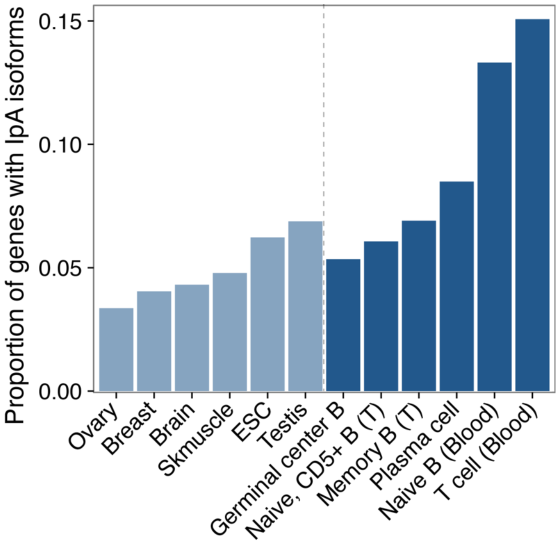

Figure 4.2.1: Immune cells express higher number of IpA isoforms
Higher proportion of expressed genes express IpA isoforms in immune cell types (0.05 to 0.15) in comparison to the solid tissues. Only 0.03 to 0.07 genes express IpA isoforms in the solid tissue.

IpA sites shared across immune cell types We wanted to find out if the cleavage machinery recognizes the same IpA sites across different tissue types and to determine if there were IpA sites that were specifically being used in a particular tissue or cell type. To get a global view of the tissue specificity of IpA sites we identified the IpA isoforms that were expressed in any tissue or cell type. Amongst these IpA isoforms we determined for each tissue/cell type whether the IpA isoform was expressed, the IpA isoform was not expressed when the gene was expressed, or if the gene itself was not expressed. Visualization (Figure 4.2.2) of the IpA isoforms (n = 3197) expression pattern shows that majority of IpA are expressed in the immune cell types and that these IpA sites are used in at least two immune cell types. The IpA isoforms that are expressed in at least 75% of the samples of a given cell type/condition were used. A gene is considered to be expressed if either the IpA isoform (≥ 5 TPM) or the full-length transcript (≥ 5.5 TPM) were expressed in ¾ of the samples of the particular cell type. The mean expression level of the IpA isoform across all the samples had to be more than 5 TPM to be flagged as an expressed IpA isoform. Non-immune tissues like testis and ES cells express tissue-specific IpA isoforms, but the majority of these isoforms are expressed in tissue-specific genes, eliminating the scope for the isoform to be expressed in other tissue types. Presence or absence of IpA isoforms of genes expressed consistently across a variety of tissue types suggests the highly regulated usage of the IpA sites between different cell types. This pattern of expression of IpA isoforms also implies that these IpA isoforms might have consequential functions in the cell types in which they are expressed.

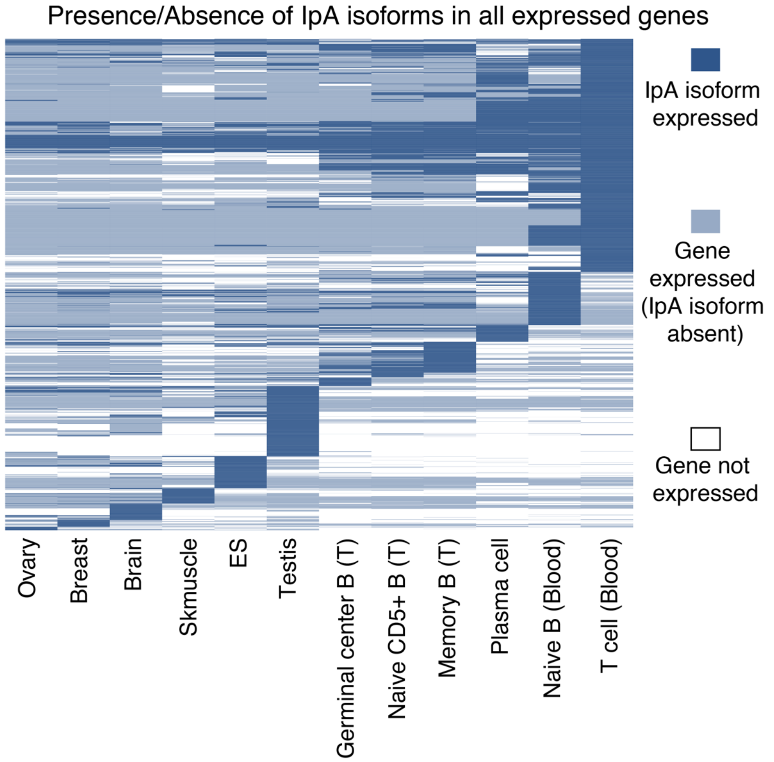

Figure 4.2.2: Majority of IpA isoforms are shared between immune cell types
Most of the IpA isoforms are expressed in two or more immune cell types. The tissue specific IpA isoforms are mostly expressed in tissue specific genes.

4.3 Predominant usage of promoter proximal IpA sites with increased recognition of promoter proximal IpA sites in cell types expressing a higher number of IpA isoforms

The majority of IpA isoforms occur at the beginning of transcription unit The position of the 3' ends of the IpA isoforms in the transcription unit is extremely important in determining the function of the protein translated from the IpA isoform. If the mRNA isoforms lose fewer C-terminal exons by the usage of an IpA site, then the protein product would still retain most of the protein domains responsible for its functional activity. In such cases, the activity and functional capability of the protein would largely depend on the protein domains that are lost vs. retained. For example, loss of an active site in enzymes would make a non-functional enzyme but loss of a transmembrane domain from a membrane receptor would still create a soluble protein with an altered function. This would completely depend on the gene in question and its corresponding protein. By contrast, if the IpA isoforms result from recognition of IpA site present early in the transcription unit then these mRNA isoforms would lose most of the sequence that is translated into protein domains. In such cases, it would be highly unlikely for these IpA isoforms to create proteins that retain the original function. As recognition of early IpA sites would result into mRNA isoforms of shorter length, the probability that these mRNAs would be translated into protein would also go down significantly. Thus, with this insight we wanted to investigate the position of the 3' ends of the IpA isoforms present in our atlas. As we were interested in the variability of the protein translated from the IpA isoforms, we determined the fraction of protein coding sequence that would be retained by all the IpA isoforms of the atlas. We hypothesized that IpA events should occur more towards the end of the transcription unit, leading to loss of only few C-terminal exons in order to contribute towards proteome diversity. However, we made the opposite observation (Figure 4.3.1a) and in fact found that more than one-third (38.64%; n = 2,100) of IpA isoforms in our atlas are created by the usage of pASs located in the early introns (with retained CDS ≤ 0.25) of the transcription unit. Interestingly, plenty (n = 564) of IpA isoforms use a pAS in the intron located upstream of the start codon, probably having no potential to be translated into a protein. For the remainder of the study we refer these early occurring/promoter proximal IpA events (with retained CDS < 0.25) as 5' IpA events as they occur more towards the 5' end of the transcription unit. The IpA events occurring towards the 3' end of the transcription unit will be referred as 3' IpA events (with retained CDS ≥ 0.50). We describe the functional consequences of these 5' IpA isoforms in Section 6.1.

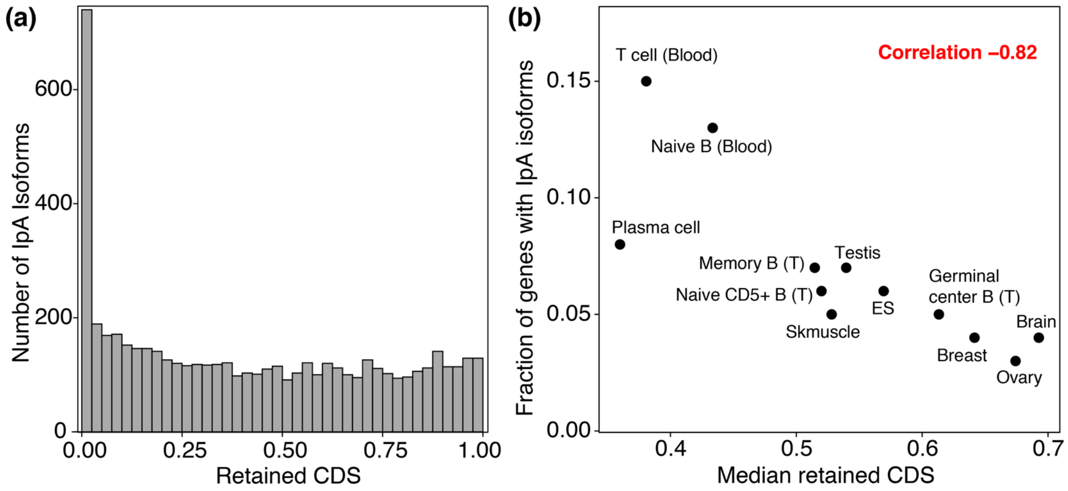

Figure 4.3.1 a) High usage of promoter proximal IpA sites
The length of retained coding sequence (CDS) is the fraction of coding sequence of the original transcription unit that the IpA isoform has after an early termination event in the intron. A large fraction of IpA isoforms use the IpA site located close to the start of the transcription unit retaining a small fraction of coding sequence (n= 2,100 with retained CDS < 0.25);
b) Tissues with higher fraction of genes with IpA isoforms have larger number of IpA isoforms ending close to the start of the transcription unit.
The fraction of genes with IpA isoforms is negatively correlated (Pearson correlation: -0.82) with the median coding sequence that was retained by the expressed IpA isoforms in each tissue.

Higher frequency of IpA events correlated with higher incidence 5' IpA events

We noticed that cell types with higher frequency of IpA isoforms had higher incidence of 5' IpA events. We wanted to see if there was a relationship between the frequency of IpA isoforms and location of 3' ends of the IpA isoforms across the different cell types. To confirm this, we calculated the median retained coding sequence (CDS) in nucleotides for each tissue. The fraction of genes with IpA isoforms was found to be strongly negatively correlated (Figure 4.3.1b; Pearson correlation: -0.82) with the median retained coding sequence in nucleotides across the different cell types. This means that if a cell type has a higher occurrence of IpA isoforms then it is likely that these IpA isoforms are created by the recognition of the IpA sites located early in the transcription units.

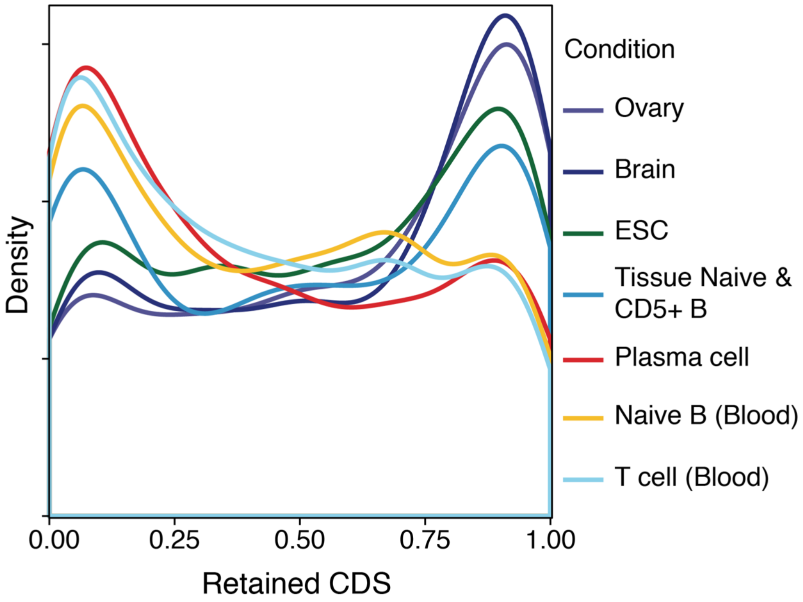

Figure 4.3.2: Diversified usage of IpA site in different tissue and cell types
The pattern of usage of IpA site varies widely across the tissue and cell types. Ovary and brain have majority of IpA events towards the 3' end of the transcription unit, naïve B cells (blood), T cell (blood) and plasma cell have most of the events near start of the transcription unit while ESCs, naïve and CD5+ B cells have IpA cleavage events at both start and end of transcription units.

Diversified usage of IpA sites in wide variety of tissues and immune cell types

Global IpA events exhibited a predominantly high usage of the IpA sites located at the start of the transcription unit. However, it is important to examine if this pattern of usage of the IpA site is consistently true across the different tissues and immune cell types. Thus, we examined the location of 3' ends of the IpA isoforms across all the tissues and cell types. Strikingly, we observed very different patterns of usage of IpA sites in different cell types (Figure 4.3.2). The immune cell types like the naïve B cells (blood) and T cells have very high occurrence of IpA events at the 5' end resulting in loss of the majority of protein domains while tissues like brain and breast have most of their IpA isoforms at the 3' end, losing very few C-terminal exons. Unlike, naïve B cells (blood) or brain, ES cells express IpA isoforms using both the 5' and 3' IpA sites. This diversification of usage of IpA sites in different tissues and immune cell types again points to a potential cell type specific functional role.

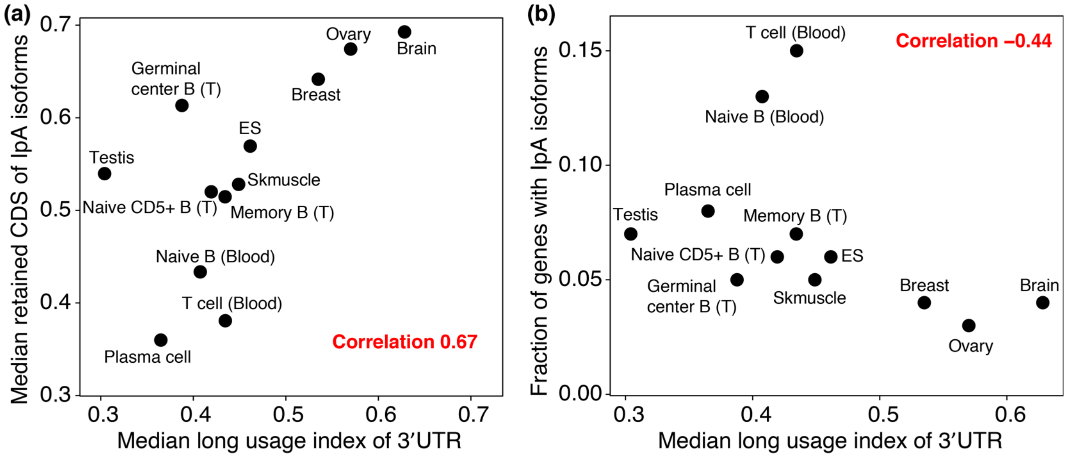

Figure 4.3.3a) Tissues with shorter 3' UTRs have shorter IpA isoforms

The tissues that have shorter 3' UTRs also express IpA isoforms shorter in length (Pearson correlation: 0.67). This suggests that the factors that are involved in usage of proximal pASs in the 3' UTRs might also be involved usage of IpA sites

b) Higher frequency of IpA isoforms associated with 3' UTR lengths

Tissues with longer 3' UTRs have lower incidence of IpA events (Pearson correlation: -0.44). Tissues with longer 3' UTRs tend to have lower usage of the IpA sites.

Length of IpA isoforms moderately correlated with 3' UTR lengths Global trends of the length of 3' UTR ApA isoforms for the multi-UTR genes have been reported previously (Lianoglou et al. 2013). Studies have shown that brain expresses 3' UTR ApA isoforms with the longest 3' UTRs while testis expresses 3' UTR ApA isoforms with the shortest 3' UTRs, while other tissue types span the middle range (Zhang et al. 2005; Ramskold et al. 2009; Shepard et al. 2011; Li et al. 2012; Smibert et al. 2012; Ulitsky et al. 2012; Lianoglou et al. 2013; Miura et al. 2013). With these established patterns of 3' UTR lengths for the various tissues, we wanted to address if these patterns held up for the IpA isoforms. More precisely, we were interested in determining whether the tissues that had shorter 3' UTR lengths also had IpA isoforms retaining less of the coding sequence. To query the relationship between the 3' UTR length of ApA isoforms and the position of 3' ends of IpA isoforms for the different cell types, we examined the correlation between these two variables. We made use of long usage index (LUI) for this purpose, a statistic that reflects relative abundance of the distal 3' UTR isoform compared to other 3' UTR isoforms. The median long usage index (LUI) was thought to be a fair representation for the overall 3' UTR lengthening/shortening of the multi-UTR genes expressed in each tissue. Thus, we examined the correlation between the median LUI and median retained CDS. Figure 4.3.3a shows that there is a moderate positive correlation (Pearson correlation: 0.67) between shorter 3' UTRs and shorter IpA isoforms. This correlation perhaps suggests that protein factors playing a role in the defining the length of 3' UTRs for the genes on a global level might also be involved in defining the length of IpA isoforms. We also observed a milder negative correlation (Pearson correlation: -0.44) between fraction of genes with IpA isoforms and median LUI, indicating that the tissues that have longer 3' UTRs also have lower incidence of IpA isoforms (Figure 4.3.3b). This relationship could be another level of evidence hinting at the role of same machinery in defining both features.

4.4 Intronic polyadenylation enriched in retained introns

Visualization of RNA-seq tracks along with 3'-seq tracks showed a frequent incidence of introns containing a 3'-seq peak to be retained in RNA-seq (Figure 4.4.1). Intron retention is described as form of alternative splicing where the transcribed intron is not spliced out during pre-mRNA processing (Black 2003). Generally, these transcripts with a retained intron have a premature stop codon causing them to be targeted by nonsense-mediated decay (NMD) pathway. The co-occurrence of IpA isoforms with intron retention might implicate a potential association between the two processes. As we had RNA-seq data for many of the same samples as our 3'-seq data, we decided to investigate if this pattern existed globally. We first asked if usage of IpA sites is enriched in introns that are retained. To examine this thoroughly we needed to determine the introns that are retained in the various cell types. For some of the tissues we did not have our own RNA-seq data, so we gathered RNA-seq data for those tissues from other studies (Section 2.2). For this analysis we also collected RNA-seq for other tissues from other studies. We decided to identify the retained introns using a modified version of the IRFinder algorithm, which is also meant to detect the introns that are retained (Wong et al. 2013). If an intron is retained, then we should observe some read coverage over the intron in the RNA-seq data, and this read coverage should be relatively high compared to flanking exons. Thus, we took advantage of this knowledge and designed steps for filtering retained introns from RNA-seq data. To avoid genes with a complex genomic architecture we removed genes that overlap with other genes in either the sense of antisense strand. An intron was categorized as retained if it satisfied the criteria described below:

1) There should be at least 3 reads spanning both a) the upstream exon (E1) and intron junction and b) the downstream exon (E2) and intron junction. This criterion ensures that there were reads supporting the retention of the intron by making use of reads spanning the intron and the flanking exons.
2) At least 50% of the intron length should be covered by 3 or more unique reads. Mappability of introns could be a limitation in this case, thus we focused only on introns that had at least 50% uniquely mappable sequence relative to its complete length.
3) To ensure adequate expression of the flanking exons, the median coverage over the flanking exons was required to be 10 reads or more.
4) Since the introns should have more coverage than the background noise, we considered introns to be retained if the ratio of median coverage over the intron

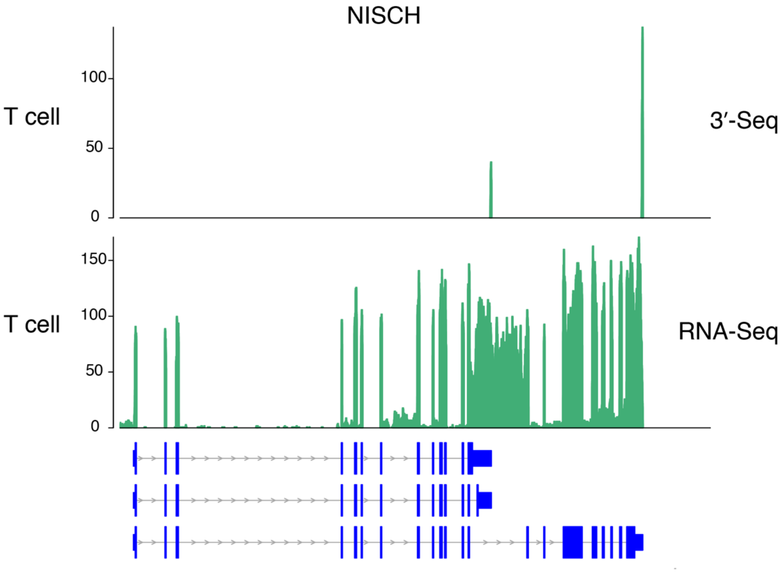

Figure 4.4.1: Co-occurrence of IpA with intron retention
Visualization of 3'-seq and RNA-seq profiles showed the pattern of occurrence of 3' end of the IpA isoform in retained introns.

to median coverage of the upstream exon was at least 10%. A similar criterion was established for intron coverage relative to the downstream exon.

In the next step we determined all the introns that are retained in every cell type. The analysis was restricted to introns of expressed genes in every cell type. An intron was annotated as retained if it fulfilled the above-mentioned criterias in at least 66% of the RNA-seq samples of the particular cell type. Introns retained in 33% or fewer samples were flagged as not retained while the introns that were retained in more than 33% samples but less than 66% of RNA-seq samples were removed from the analysis as nothing could be concluded about these introns. For a 3' end of an IpA isoform to occur in a particular intron, the intron should have a pAS that can be recognized by the cleavage machinery. Thus, we concentrated the remaining analysis only on the introns that were retained and had one of the known pASs (Tian et al. 2005).

Artifacts: Our data showed that some genes had very high coverage over almost all the introns of the gene. This type of high coverage over all the introns appeared to be the outcome of sequencing artifacts. Thus to eliminate such noisy genes from the analysis, we removed these genes using another statistic. We determined the median (median coverage over all the introns)/ median (median coverage over all the exons), if this ratio was ≥ 0.2 then these genes were flagged for removal. A high value of this statistic would imply that the coverage over all the introns of the transcription unit is similar to or approaching the coverage level over all the exons of the transcription unit.

If the association between intron retention and IpA existed, then the tissues with high IpA events should also have high intron retention cases. We observed a moderately high correlation (Pearson correlation: 0.58) between the fraction of genes that had IpA events and the fraction of genes with intron retention (Figure 4.4.2). This observation motivated more analysis to demonstrate the association. As we were interested in determining if 3' ends of IpA were enriched in retained introns, we first determined the number of introns that are retained and also have a 3' end of an IpA isoform in the particular intron just by random chance. Then we determined the actual number of introns that had the 3' end of an IpA isoform and are retained. We found the co-occurrence of IpA with intron retention was more frequent than just by random expectation (Figure 4.4.3a). The association between the two processes was authenticated by statistical enrichment (Fisher's exact test), showing IpA to be enriched in introns that are retained (Figure 4.4.3b).

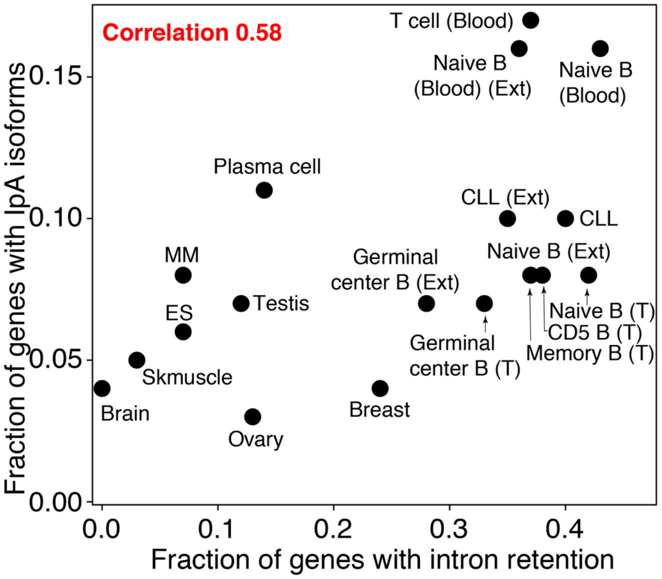

Figure 4.4.2: Association between intron retention and occurrence of IpA
Tissues with higher number of genes with intron retention also have more genes that express IpA isoforms (Pearson correlation: 0.58). This observation suggests of a possibility an association between the two phenomenon.

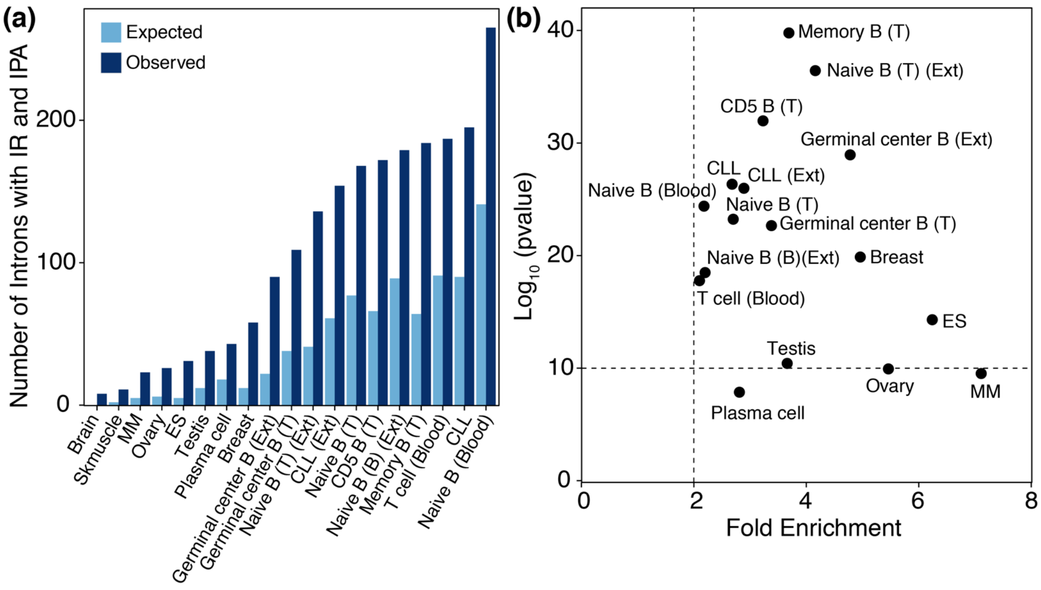

Figure 4.4.3 a) Higher incidence of IpA events in introns with IR (intron retention)

The number of introns where IpA and IR is observed simultaneously is much higher than expected by chance;

b) Significant enrichment of IpA in retained introns

Motivated by the observation made in 4.4.3a we found that IpA events are significantly enriched in introns that are retained in contrast to introns that are not retained. This establishes an association between IR and IpA. However, it is difficult to establish the causality of the events. We hypothesize that cleavage machinery recognizes the IpA sites in the retained introns causing the formation of 3' end.

Co-occurrence of intron retention and IpA isoforms in the same introns suggests an association between the two phenomena. We hypothesized that retention of intron (inclusion of the intron in the transcript) could be a requirement for IpA to occur in these introns; however it is difficult to establish the direction of causality with the available data and would need experimental work to make strong conclusions. We found that the median usage of the IpA isoforms that co-occurred with intron retention is lower compared to IpA isoforms in introns that are not retained (Figure 4.4.4). This observation suggests perhaps we could not detect intron retention in these highly used IpA isoforms because 3' end cleavage was highly efficient, leaving no reads after the cleavage event.

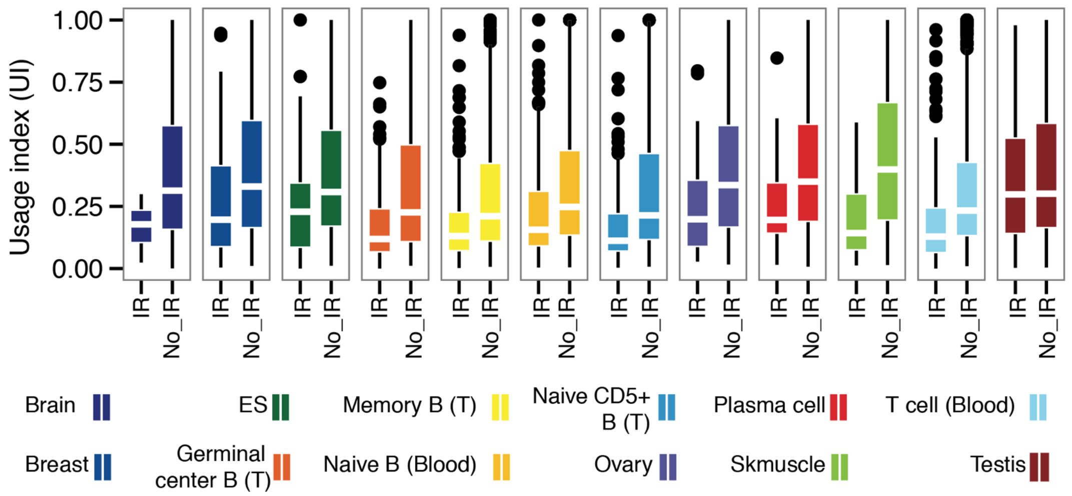

Figure 4.4.4: IpA events with lower usage index co-occur with intron retention The median usage of the IpA isoforms that co-occurred with intron retention is lower compared to the introns that are not retained. This observation suggests that it could be possible that we could not detect intron retention in these highly used IpA isoforms as 3' end cleavage was highly efficient, leaving no reads after the cleavage event.

4.5 Enriched pASs, depleted U1 snRNP, long introns and long transcription units define the landscape of the genes with IpA isoforms

Having observed a widespread usage of IpA sites in the expressed genes, we wanted to determine if the genes exhibiting these isoforms had any special features that make them different from the genes that always express full-length isoforms. We investigated the differences in the sequence signals and the genomic architecture between the genes that express IpA isoforms and the genes that only express full-length 3' UTRs.

IpA isoforms enriched in multi-UTR genes It has been shown that approximately

50% of the genes always have single UTRs while the remaining 50% undergo 3' UTR ApA and thus are referred as multi-UTR genes (Shepard et al. 2011; Lianoglou et al. 2013). We were interested in determining if the genes that expressed IpA isoforms were enriched in either of these two categories. We found that 30% of the multi-UTR genes express IpA isoforms while only 17% of the single UTR genes express IpA isoforms. Our results showed that genes with IpA isoforms are significantly enriched (Fisher's exact test: $p < 2.69 \times 10^{-82}$) amongst the multi-UTR genes compared to single UTR genes. This enrichment elucidates a tighter regulation of expression of the single UTR genes with multi-UTR genes having the genomic landscape to provide additional layers of regulation via IpA and 3' UTR ApA isoforms.

IpA isoforms occur in transcription units enriched for pASs and depleted for U1 snRNP signals Presence of a functional pAS has been shown to be essential for the efficient 3' end processing and cleavage of the transcript (Proudfoot 2011). As formation of the 3' ends of the IpA isoforms would also require pASs, it is important to investigate the sequence context of the genes with IpA isoforms to unravel what makes them different from the genes that only express full-length 3' UTRs. Another factor that potentially plays a crucial role in the creation of the IpA isoforms is a small nuclear ribonucleoprotein, U1 snRNP. U1 snRNPs have been shown to play an essential role in preventing the premature cleavage and polyadenylation (PCPA) during pre-mRNA processing (Kaida et al. 2010). Deficiency of U1 snRNP leads to PCPA (premature cleavage and polyadenylation) and creation of mRNA isoforms of varied lengths (Berg et al. 2012). U1 snRNP binds to hexamer sequence signals (GGUAAG, GGUGAG, GUGAGU) and prevents the recognition of the pASs in the vicinity (Mount et al. 1983). Further it has been shown that the appropriate direction of transcription is maintained by the enrichment of U1 snRNP signals and depletion of pASs in the promoter proximal region of the transcription unit (Almada et al. 2013). Thus, we wanted to assess the frequency of the pAS and U1 snRNP in transcription units with IpA isoforms. We only looked for the top two most used pASs (AAUAAA, AUUAAA) (Tian et al. 2005) to avoid accounting for very pervasive pASs that are not very functionally efficient. Indeed, we found the occurrence of pASs to be higher (one sided KS test, $p < 5.49 \times 10^{-53}$) in the gene bodies of genes that express IpA isoforms (n = 3671) compared to genes that only express full-length 3' UTRs (n = 11775) (Figure 4.5.1a). Additionally, genes with IpA isoforms are also depleted (one-sided KS test: $p < 3.97 \times 10^{-40}$) for the U1 snRNP sequence signals in comparison to genes that always express full-length 3' UTRs (Figure 4.5.1b). This enrichment of pASs and depletion of U1 snRNP binding sequences may provide an adequate landscape for the creation of the IpA isoforms in these transcription units.

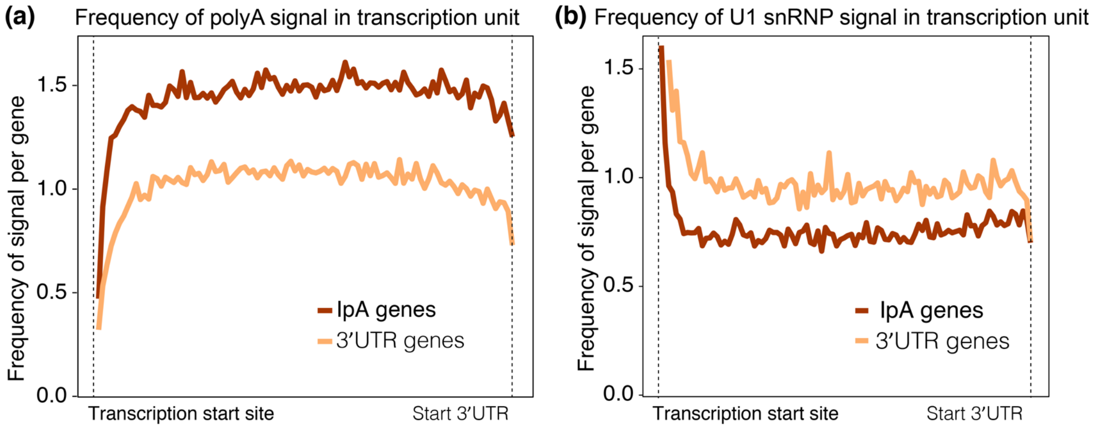

Figure 4.5.1: Genes with IpA isoforms enriched for pASs and depleted for U1 snRNP sites

Expression of IpA isoforms is facilitated by the sequence composition of the genes. The genes with IpA cleavage events are enriched for pASs (one sided KS test, $p < 5.49 \times 10^{-53}$), that is recognized by the cleavage machinery for 3' end processing. These genes are also depleted (one KS test, $p < 3.97 \times 10^{-40}$) for U1 snRNP signals where U1 snRNPs bind to prevent premature cleavage events. The presence of pASs and absence of U1 snRNP signals makes these genes suitable candidates for expression IpA isoforms. Sequence composition of the IpA genes (n = 3671) was compared to the genes that always express full-length isoforms (n = 11775).

Longer introns and longer transcription units aid the creation of IpA isoforms

Apart from the sequence context of the genes with IpA isoforms we were also interested in the structural differences between transcription units that do or do not express IpA isoforms. These features could help us to determine the factors that aid the creation of IpA isoforms. As our previous results showed that IpA events are enriched at the start of transcription units (5' IpA events), we also wanted to examine the differences in the genomic architecture of the 5' and 3' IpA events. To have a clean signal we used the following definition for the 5' and 3' IpA events: 5' (fraction of retained CDS in nt < 0.25) and three-prime (0.50 < fraction of retained CDS in nt < 1.0). We compared the width of the introns, width of the transcription units and width of the 5' UTRs. IpA isoforms occur in significantly longer transcription units (one sided KS test: $p \sim 0$) with wider 5' UTRs (one-sided KS test: $p < 4.93 \times 10^{-29}$) when compared to genes that only express full-length 3'-UTRs (Figure 4.5.2), which is consistent with a previously reported observation (Tian et al. 2007). When a similar comparison was performed between the 5' and 3' IpA events, we found 5' IpA events tend to occur in longer introns, longer transcription units and longer 5' UTRs (one-sided KS test, $p < 9.59 \times 10^{-66}$; $p < 9.68 \times 10^{-08}$; $p < 1.39 \times 10^{-09}$, respectively; Figure 4.5.2). This suggests that large intron size and long transcription units could be a determining factor for the usage of the IpA site leading to the formation of the IpA isoforms. The 5' IpA events also tend to occur in significantly longer 5' UTRs, with many of them having an intron in the 5' UTR, providing an opportunity for formation of the IpA isoform.

4.6 Expression of IpA isoforms is differentially regulated between different cellular conditions

So far we focused our analysis on the characterization of the IpA isoforms across the wide variety of tissue and immune cell types. We also tried to distinguish the genomic features that aid in the formation of the IpA isoforms. Having answered these questions, we wanted to determine how significantly different the IpA isoform expression levels were for naïve B cells obtained from two different environments of the healthy donors – blood and tonsils. We used a generalized linear model (GLM) as described earlier (Section 3.3) to determine if there was a differential usage of the IpA sites when compared to full-length 3' UTR isoform (Anders et al. 2012; Lianoglou et al. 2013). This form of GLM adjusts for the differences in the sequencing depth of the samples (library sizes) and the biological variation between the replicates of the same

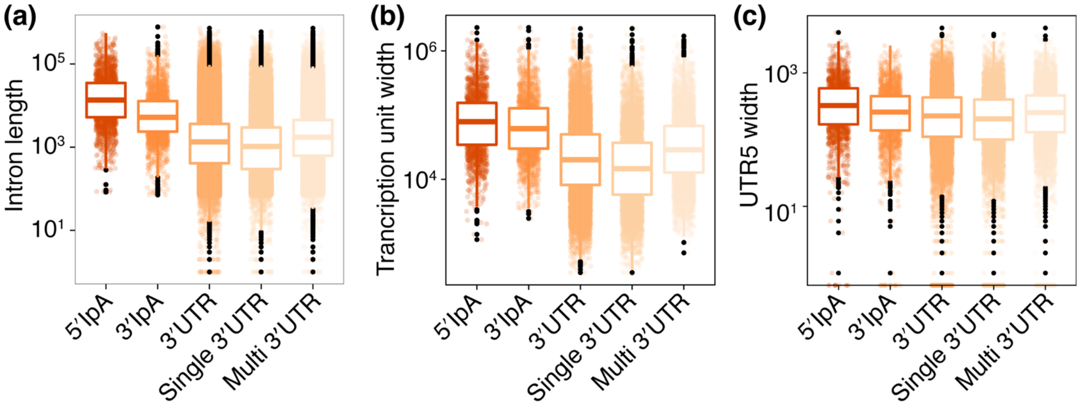

Figure 4.5.2: Occurrence of IpA isoforms in genes with long transcription units, long introns and wider 5' UTRs

5' IpA events tend to occur in long introns. Genes that have IpA isoforms have longer transcription units and wider 5' UTRs compared to genes that only express full-length 3' UTRs. The genes expressing IpA isoforms have these special characteristics that are different from the genes that only express full-length isoforms. This suggests that large intron size and long transcription units could be a determining factor for the usage of the IpA site leading to the formation of the IpA isoforms.

condition. It allows for testing for significant differences for relative expression of IpA isoform and full-length isoform after accounting for the differences in the gene expression in the different conditions. Using this GLM model we compared two independent samples of naïve B cells (blood) against the naïve B cells (tonsils). An IpA isoform goes into the analysis only if either the full-length is expressed (> 5.5 TPM) or IpA isoform is expressed (> 5 TPM) in 75% of normal B cells or any of the cancer samples (IpA = 1,164). The IpA sites with a usage difference between the two cell types was more than 10% with adjusted $p < 0.1$ were considered to be significantly different between the conditions (n = 364, high usage = 331, loss of usage = 33 in naïve B cells (blood), Figure 4.6.2). This comparison revealed that in majority of the cases there is an increase in the usage of IpA sites of naïve B cells (blood) with fewer cases where there is an increased usage of IpA sites of naïve B cells (tonsils). We found that this switch in isoform expression potentially contributes towards cellular phenotype. An example that supports this idea is the spectrin (SPTBN1) gene which has a significantly higher usage of the IpA site in blood B cells (Figure 4.6.1). Spectrin plays an important role in defining the cell shape along with the arrangement of transmembrane proteins and organization of organelles. This actin crosslinking and molecular scaffold protein that links the plasma membrane to the actin cytoskeleton has a Pleckstrin homology domain (PH domain) (Das et al. 2008). It has been shown that the PH domain is critical for the localization of the protein on the plasma membrane. The naive B cells in the tissue environment (tonsils) would need spectrin on its plasma membrane for the maintenance of cell shape and tissue structure. However, the naïve Bcells in the blood environment do not need to maintain the tissue structure and thus do not need the spectrin protein on its plasma membrane. This explains the high expression of the IpA isoform that lacks the PH domain in naïve B cells obtained from blood. More detailed screening of the list of the genes that significantly change their isoform usage might reveal more candidate genes are directly involved in cellular functioning. These results demonstrate tight regulation of IpA isoforms between cell types and environmental conditions. They also suggest that these significantly regulated IpA isoforms are potentially functionally relevant. These observations clearly indicate that expression of IpA isoforms should not be viewed as usage of cryptic IpA sites or as premature cleavage polyadenylation induced by the malfunctioning of the cellular machinery.

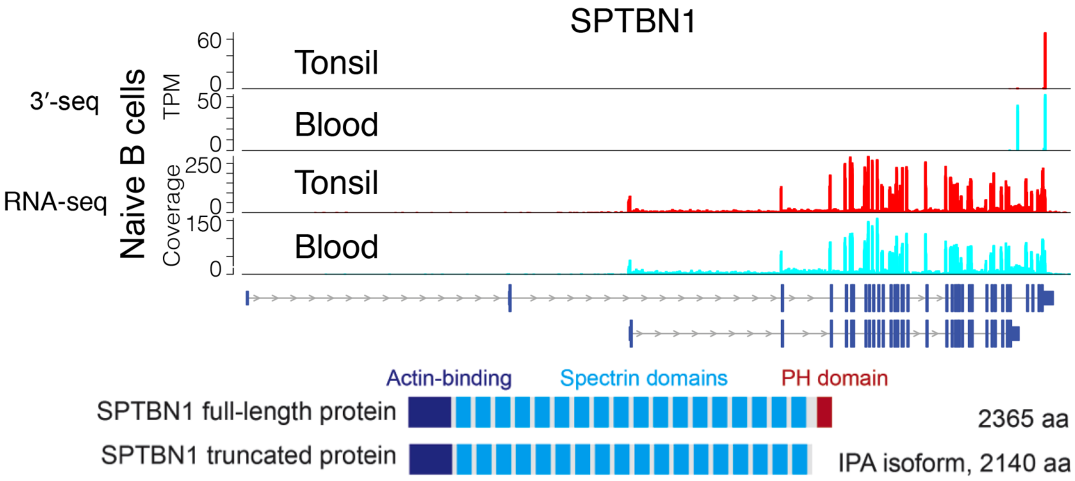

Figure 4.6.1: IpA isoform is expressed in Naïve B cells obtained from blood
Spectrin expresses an IpA isoform in naïve B cells obtained from blood. This isoform is absent in the naïve B cells that are obtained from tonsils. The expression of this IpA isoform is regulated between the different cellular environments suggesting of a functional consequence between the conditions. The IpA isoform lost the PH domain that is required for the transport of SPTBN1 to the plasma membrane to maintain the cell-shape

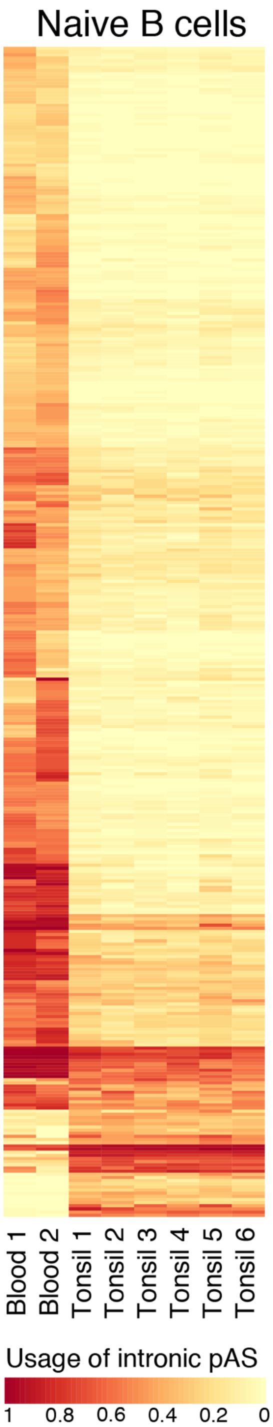

Figure 4.6.2: Significantly differently used IpA sites between naïve B cells from blood and tonsils (N = 364)

The usage of IpA sites is regulated between cellular conditions. Naïve B cells from blood have significantly higher usage of IpA sites compared to naïve B cells obtained from tonsils (n = 331). Only very few genes show loss of recognition of IpA site in naïve B cells obtained from blood (n = 33).

CHAPTER 5

5. DEREGULATED IpA ISOFORMS EXPRESSION IN LEUKEMIAS

Cancer is largely associated with genetic abnormalities that include mutations, chromosomal translocations, deletions and amplifications. These genetic abnormalities in proto-oncogenes are considered to be drivers that lead to malignant cells with uncontrollable growth. In many cases, overexpression of oncogenes is also observed without any genetic alteration. Shortening of the 3' UTR by ApA has been shown to contribute to tumorigenesis by increased production of protein from the same amount of mRNA (Mayr and Bartel 2009). Studies have observed deregulated 3' UTRs in cancers, with a shift towards shortening of 3' UTRs across a wide variety of cancers with some exceptions (Fu et al. 2011; Morris et al. 2012; Xia et al. 2014). All the studies so far have focused on the shift in patterns of length of 3' UTRs. There has been no study that has characterized the changes in the landscape of usage of IpA sites in cancer. Thus we wanted to investigate the changes in relative expression of IpA isoforms vs. full-length isoforms for B cell derived malignancies compared to normal B cells. As discussed earlier, we hypothesize that a relatively higher usage of IpA sites would create truncated mRNAs that could have the potential to mimic genetic alterations, like mutations, deletion, and chromosomal translocations. Generating an IpA isoform could be an alternate way for the cancer cell to have similar effects as the genetic alterations. To investigate this hypothesis, we performed the following analysis.

We were interested in examining the extent of differential expression of IpA isoforms compared to full-length isoform in chronic lymphocytic leukemia (CLL) and multiple myeloma (MM). CLL is the leukemia of naïve/CD5$^+$ B cells that normally occurs in adults. MM is a cancer formed by malignant plasma cells, which is very aggressive with very low survival rate. We compared each of these leukemias to their respective normal cell types and identified the significant changes in the relative usage of the IpA site and full-length 3' UTR pAS.

5.1 Truncated mRNAs generated by usage of IpA site can potentially mimic genetic mutations

Higher usage of IpA sites in CLLs: We were interested in identifying statistically significant changes in the relative usage of the IpA sites and pASs in the 3' UTR of genes independent of gene expression changes between CLL samples and naïve/CD5$^+$ B cells. We had CLL samples from 13 patients and 6 naïve/CD5+ B cells from healthy donors. Before testing for statistically significant changes, we wanted to determine whether all the CLL patients could be treated as one single condition or whether they should be classified into separate groups. Thus we performed hierarchical clustering of the CLL samples using their IpA expression levels (TPM). This helped us to define three separate groups of CLL patients (group 1, 3 patients; group 2, 5 patients; and group 3, 5 patients). Each group of CLL patients was compared against the normal B cells (which we call "group-wise" comparisons). As described earlier, an IpA isoform goes into the analysis only if either the full-length isoform is expressed (> 5.5 TPM) or the IpA isoform is expressed (> 5 TPM) in 75% of normal B cells or any of the cancer samples (IpA isoforms = 1,840; genes = 1,406). These genes were then tested for differences in relative usage of the IpA sites or pASs in the 3' UTR. We also repeated this analysis in a slightly different manner, where we compared every single sample against all normal B cells (we call these "sample-wise" comparisons). In the group-wise comparison, we found that two of the groups (group 2 and group 3) of patients had fewer changes compared to normal B cells in the usage of the IpA sites compared to full-length while one group (group 1) showed significantly higher usage in CLL patients (adjusted $p < 0.1$; usage difference > 0.1). 345 IpA isoforms showed higher usage of the IpA site in at least one the defined CLL groups while there were only 75 IpA isoforms with lower usage of the IpA site. To select more robustly changing IpA events, an additional criterion of usage fold change between case and control (Usage index FC > 3) was used (Figure 5.1.1a). 171 IpA isoforms have increased usage of the IpA site while 19 IpA isoforms show loss of recognition of IpA site with the added criterion. We identified the IpA isoforms that are most recurrently highly expressed across at least one quarter of the CLL patients (n >= 4) using the sample-wise comparisons. Figure 5.1.1b shows the usage index of these IpA isoforms across the different CLL and naïve/CD5+ B cell samples.

Enrichment for truncating mutations: We had hypothesized that usage of IpA sites could have the potential to mimic genetic alterations. The truncated mRNAs that lose their ability to be translated into proteins would be phenocopy genomic deletions or nonsense mutations. IpA isoforms loosing C-terminal exons could potentially mimic loss-of-function mutations, frame-shift mutations or chromosomal translocations. To investigate this hypothesis, we collected information about genes that are mutated in CLL patients from the available studies (Puente et al. 2011; Quesada et al. 2012; Landau et al. 2015). Since IpA isoforms would certainly be translated into different protein sequences than those obtained from full-length isoforms, we focused on truncating mutations (nonsense mutations, frame-shift mutations and splice-site mutations) in expressed genes that would likely have a similar outcome as IpA isoforms. We found 637 expressed genes to harbor one these mutations in CLL patients. Amongst the 330 genes that have significantly higher usage of an IpA site, 34 genes were also found to have one the truncating mutations. In support of our hypothesis we found that genes with high usage of IpA sites are significantly enriched

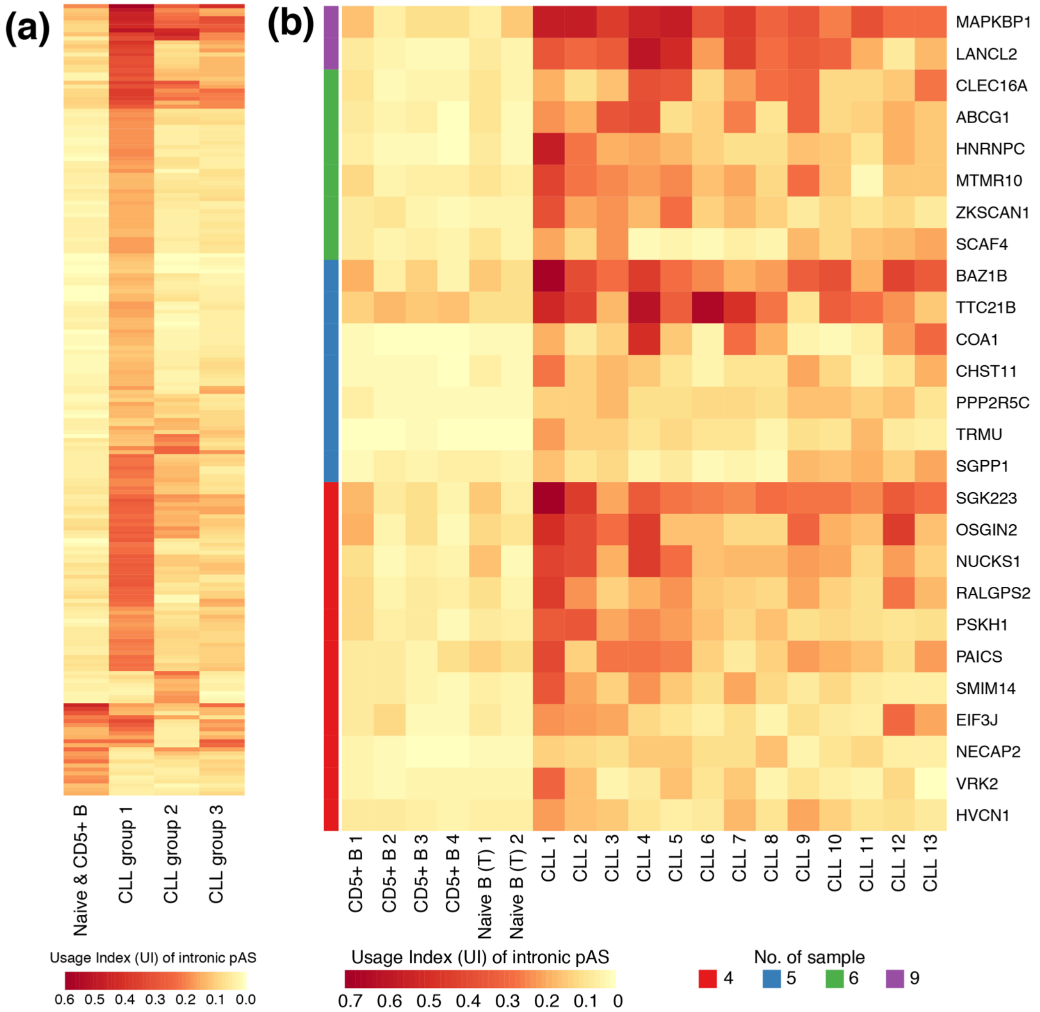

Figure 5.1.1a) High usage of IpA site in CLLs compared to naïve/CD5+ cells: 171 IpA isoforms are more highly expressed in CLL while for only 19 IpA isoforms there is decreased usage of IpA site (adjusted $p < 0.1$, usage index difference < 0.1 and usage fold change > 3);

b) Recurrently highly used IpA sites in CLL samples
26 IpA site are significantly highly used in at least one quarter of CLL samples in sample-wise comparison relative to pAS in 3' UTR.

for truncating mutations (Fisher's exact test, $p < 0.04$). It was important to assess if the mutation occurs before or after the 3' end of the IpA isoform. If the mutation occurs after the 3' end of the IpA isoform, then there would be higher potential for the IpA isoform to have comparable functional consequences as the mutated isoform. Figure 5.1.2 shows the sequence that is retained by the IpA isoform and the position of the mutations along the gene length. In most of the cases we observe that the mutation occurs after 3' end of the IpA isoform, suggesting that IpA isoforms might have comparable functional effects as the genetic mutations. This observation also points towards the possibility for the malignant cell to use IpA in order to achieve similar functional consequences contributing towards tumorigenesis without presence of any genetic alteration.

We also wanted to compare the proportion of CLL patients with increased usage of IpA site with the proportion of patients that harbor truncating mutations in these 34 genes. We also included available copy number variation data for this purpose (Pfeifer et al. 2007). Strikingly, we found that higher usage of the IpA site occurs in a much higher proportion of patients compared to patients with mutations (Figure 5.1.2). This observation further supports the hypothesis that disruption of IpA could be an alternative way for the cancer cell to phenocopy somatic mutations.

MGA IpA isoform is similar to mutated isoforms Amongst the 34 IpA genes that are mutated in CLL patients, we found MGA to be most recurrently mutated in CLL patients. The MGA IpA isoform is robustly expressed in CLL patients (Figure 5.1.3). Mga is a transcription factor that is famously known as a dimerization partner for Max. Mga has been shown repress Myc's oncogenic transcriptional activity by occupying the promoters of Myc's target genes along with Max. Studies have also shown that the presence of Mga reduces Myc-induced oncogenic transformation (Hurlin et al. 1999). This finding suggests that loss of functional MGA via a truncating mutation would enhance Myc-induced oncogenic transformation in cancer cells.

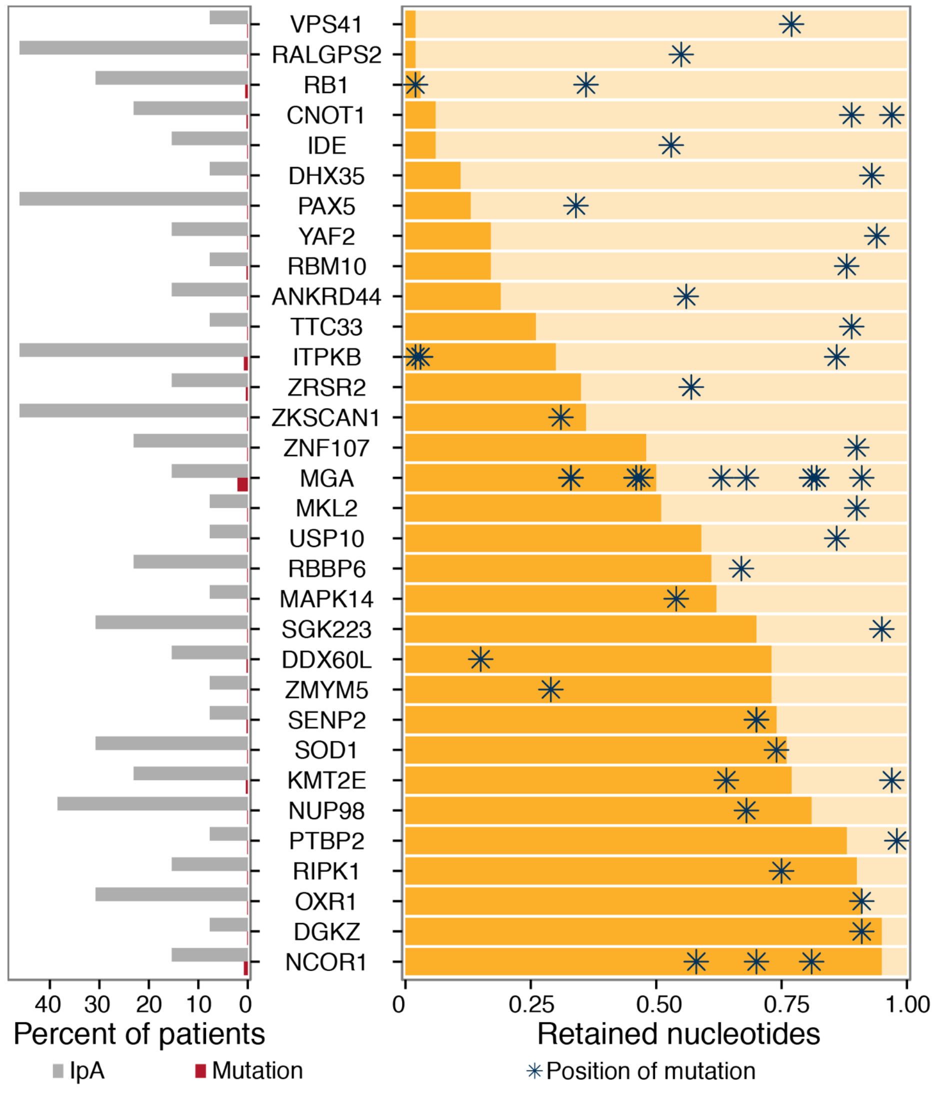

Figure 5.1.2: Truncating mutations enriched in genes with highly used IpA site in CLLs

The genes with high usage of IpA sites are enriched for truncating mutations (Fisher's exact test; *p* < 0.04). In the majority of cases we observe truncating mutations to occur after the 3' end of the IpA isoform, suggesting that high usage of IpA site leads to loss of the same or more exons as the mutation. Higher usage of the IpA site occurs more frequently across CLL patients compared to the occurrence of corresponding mutation in CLL patients. This suggests that usage of IpA sites may be an alternative way adopted by cancer cells to phenocopy the effect of genetic mutations.

Figure 5.1.2 shows that the majority of the truncating mutations of MGA occur after the 3' end of the IpA isoform, indicating that the IpA isoform would lose same or a greater number of C-terminal exons as the mutated isoform. Only 2% of patients had a truncating mutation in MGA while 15% of patients exhibited high usage of the IpA site in MGA. This observation again suggests that the malignant cells potentially have adopted an alternative way of disrupting Mga expression to enhance Myc-incduced transformation without any genomic mutation.

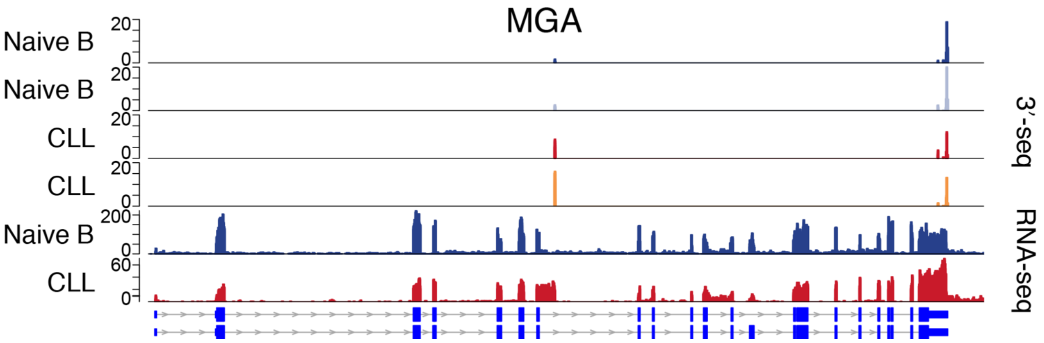

Figure 5.1.3: Robustly expressed IpA isoform of MGA
MGA is recurrently mutated in CLL patients. The usage of an IpA site in MGA increases significantly in CLL patients. The IpA isoform appears to result in a truncated protein similar to the mutated isoform. This suggests that expression of IpA isoform could be an alternative way CLL cells to enhance Myc-induced transformation in the absence of an MGA genetic mutation.

5.2 Widespread usage of distal pAS in 3' UTRs

As we observed increased usage of IpA site in CLLs, we wanted to investigate if there was a global pattern of shift towards generating shorter mRNAs in CLL. Global shortening of 3' UTRs has been reported in other cancer studies (Mayr and Bartel

2009; Morris et al. 2012; Xia et al. 2014). For this analysis we focused on determining if there was a significant difference in the relative usage of the alternative pASs in the 3' UTR between the CLL groups and naïve/CD5+ B cells. We used the GLM based approach as described in Lianoglou et al. Only the genes that were expressed (> 5.5 TPM) in every sample went into the analysis (n = 3449). As B cells express multiple ApA isoforms, we decided to test the two most variable alternative pASs between the samples. In contrast to our expectations, we found that there is a global shift towards longer 3' UTRs in CLL (adjusted $p < 0.1$, usage difference > 0.1, longer 3' UTRs = 878, shorter 3' UTRs = 347). Figure 5.2.1 shows a heatmap with global changes in the usage of the alternative pASs in the 3' UTR. A majority of ApA studies have reported a shift towards usage of shorter 3' UTRs in cancer. However, in CLL cells, ApA appears to be disrupted to create longer 3' UTRs. Theresome, in CLL, the cleavage and polyadenylation process is deregulated in a manner such that, on one hand truncated IpA isoforms are created while on the other hand genes express mRNAs with longer 3' UTRs. CLL cells potentially could be using both forms of deregulation for their survival and growth.

5.3 Loss of usage of IpA sites in multiple myeloma

We also examined the change in landscape of usage of IpA sites in multiple myeloma cells compared to plasma cells. We performed to 3'-seq on 15 samples obtained from multiple myeloma patients and 2 samples of plasma cells from healthy donors. As the samples were obtained from patients at different stages of multiple myeloma, we wanted to define groups of patients that were most similar to each other. Using hierarchical clustering based on the expression of the 50% most variable IpA isoforms, patients were divided into three groups (group 1, group 2, and group 3). We carried

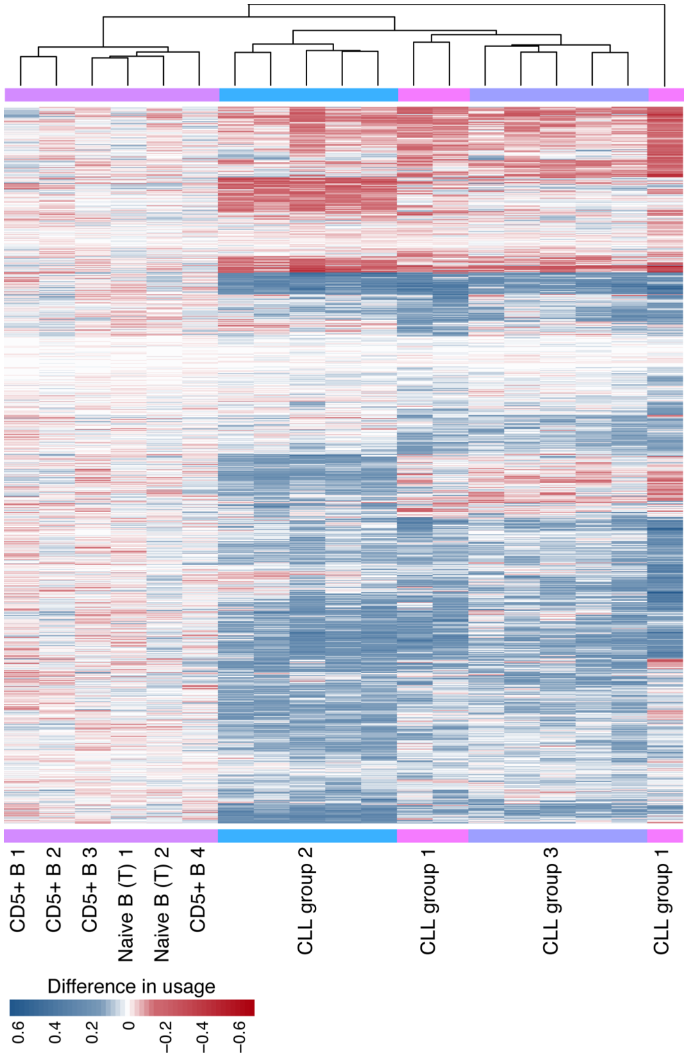

Figure 5.2.1: Longer 3' UTRs in CLLs
CLLs show higher usage of the distal pAS in the 3' UTR compared to the normal naïve/CD5+ B cells (longer 3' UTR, 878 genes; shorter 3' UTRs, 347 genes). The figure shows the difference in the long usage index of CLLs and mean long usage index of normal B cells. This is in contrast to the widespread shortening observed in cancer cells in many studies.

out the GLM modeling as described above to determine the differential usage of IpA sites compared to pASs generating full-length isoforms for these three MM groups compared to normal plasma cells. In contrast to CLLs, we found that two groups of MM patients showed loss of usage of IpA sites compared to plasma cells. Out of the 1118 IpA isoforms that were analyzed (adjusted $p < 0.1$, difference in usage index <

0.1), 356 IpA sites have loss of usage of an IpA site while 10 IpA isoforms show increased usage of an IpA site in MM. Figure 5.3.1a shows the mean usage index of IpA sites in the three MM groups and plasma cells. This loss of usage of IpA sites also affects important genes in MM biology, like the transcription factor IKZF1 (Figure 5.3.1b). These results show that deregulated cleavage and polyadenylation process does not result in similar global shifts in different malignancies.

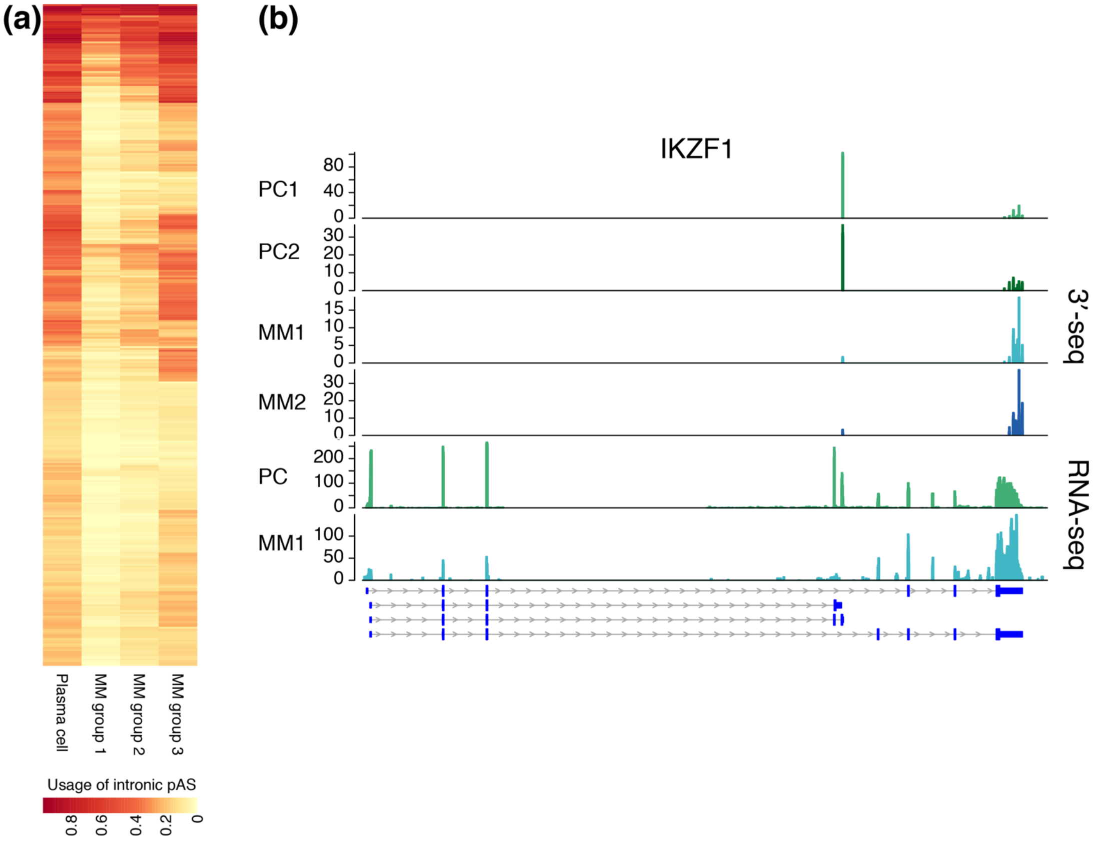

Figure 5.3.1a) Loss of usage of IpA sites in multiple myeloma cells
In contrast to CLL we observe loss of usage of IpA sites in multiple myeloma compared to plasma cells (adjusted $p < 0.1$, difference in usage index < 0.1, loss of usage of IpA sites: 356, high usage of IpA sites: 10); b) IKZF1, an important transcription factor in MM biology shows loss of usage of IpA sites in MM.

CHAPTER 6

6. INTRONIC POLYADENYLATION CREATES A POOL OF mRNAS WITH A WIDE VARIETY OF FUNCTIONS

IpA isoforms generated by the usage of IpA sites could have the potential to play a variety of functional roles. Their robust expression and cell type specific regulation suggest that alternative IpA isoforms may have different biological functions. Demonstrating the role of every IpA isoform would require a comprehensive set of experiments establishing the precise function of the isoform. We attempted to get an overall view of the potential functions of the mRNAs generated by the usage of IpA sites through computational analysis.

6.1 5' IpA events create non-coding RNAs with a potential role in gene regulation

5' IpA isoforms are likely non-coding We found (Figure 4.3.1a) that a large fraction of IpA isoforms have 3' ends near the start of the transcription units. Positional analysis showed enrichment for the usage of the IpA site located close to the 5' end of the transcription unit (5' IpA isoforms). All these 5' IpA events create IpA isoforms that retain 0 or < 0.25 of protein coding sequence. We hypothesized that such 5' IpA isoforms would have limited potential to be translated into a protein and thus would probably be non-coding RNAs. Previous studies have created RNA maps where they observe a similar class of short RNAs resulting from pervasive transcription (Kapranov et al. 2007). Recently there have been a number of studies showing the role of promoter proximal non-coding RNAs in cis-regulation of the genes. For example, a non-coding RNA generated from the minor promoter of dihydrofolate reductase (DHFR) has been shown to repress the transcription of the gene (Martianov et al. 2007). The non-coding RNA interacts with the major promoter causing promoter-specific transcriptional repression of DHFR. Another study demonstrated the role of a novel RNA in regulation of CEBPA, a transcription factor involved in the differentiation of immune cells (Di Ruscio et al. 2013). This RNA is transcribed from the CEBPA gene locus and plays an essential role in regulating the local DNA methylation profile. It was shown that this RNA binds to DNMT1, a DNA methyltransferase preventing CEBPA gene locus methylation. In another example, promoter associated RNAs have been shown to mediate transcriptional repression by some form of interaction with the promoter region of the gene (Han et al. 2007; Schmitz et al. 2010). Motivated by these examples we wanted to determine if the 5' IpA isoforms would also form a set of non-coding RNAs that might have the potential to mediate gene regulation. To investigate this hypothesis the complete transcript structure with the 3' end of the IpA of the isoform was required. Having the complete transcript structure would enable prediction about the coding potential of the transcript. Thus for this purpose we utilized RNA-seq data and performed *de novo* transcript assembly. 3'-seq data was used for precisely defining the 3' end of the transcripts.

De-novo **transcript assembly** As the majority of IpA isoforms in our atlas are present in immune cells, we decided to use a compendium of immune cell RNA-seq data, generated by us as well as gathered through other resources, for our analysis. During our intron retention analysis we observed that certain RNA-seq libraries had very high coverage in introns relatively to flanking exons. These samples were excluded in order to avoid assembly of transcripts with spurious structures. The complete transcript structure was obtained by following steps:

1. We used STRINGTIE, a novel and improved method for more accurate *de novo* assembly of the transcripts (Pertea et al. 2015). The *de novo* assembly was performed on every RNA-seq sample with default settings using hg19 RefSeq annotation (obtained from UCSC).
2. These transcripts from the multiple assemblies were subsequently unified using Cuffcompare, which removes all the redundant transcripts (Trapnell et al. 2012). It provides a set of unique transcript structures after combining all the assemblies.
3. For every single gene, we obtained the transcripts that overlapped the gene coordinates. We preferred multi-exonic transcripts to single exon transcripts. For the single exon transcripts we allowed the start/end to be within 100 nt of the TSS (Transcription Start Site). We gave this advantage to the single exonic transcripts because the direction of the transcription for these transcripts is not certain.
4. Now using the 3' ends of IpA isoforms (from our 3'-seq data), we assigned transcripts with nearest ends to these IpA isoforms. Firstly, we identified transcripts that ended within 50 nucleotides of the 3' ends. If there were several assembled transcripts that ended within 50 nucleotides of the 3' end then we chose the transcript that had the maximum number of exons. If there was a tie in the number of exons then we chose transcripts that started closest to annotated TSS. For the remaining 3' ends, we assigned the nearest ending transcript.

Finally, using the above defined criteria for selecting the transcript structures, we

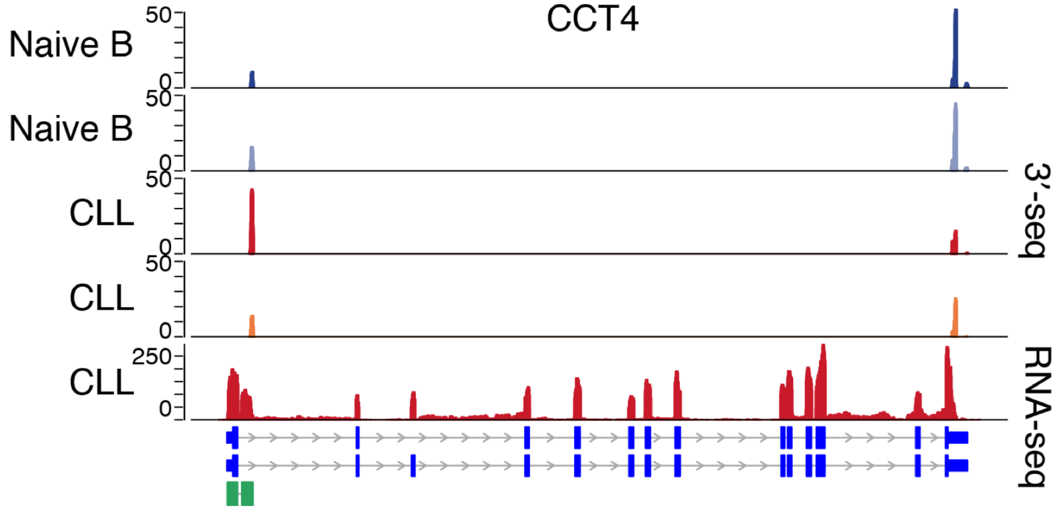

Figure 6.1.1: Assembled transcript for the IpA isoform of CCT4
The green transcript was assembled through de novo assembly using RNA-seq data.

determined which IpA isoforms corresponded to these assembled transcripts. If the 3' end of the IpA isoform was within 500 nt of the defined transcript end, then we assumed that this particular transcript represents the full structure of the transcript that ended there. For some IpA isoforms we observed usage of alternative pASs within the introns. Thus to account for such cases, for the 3' ends of the IpA isoforms that did not have a transcript end within 500 nt, we determined if it overlapped a transcript that ended within 5000 nt. If this was the case, then we assigned this transcript to that 3' end. We were able to define the transcript architecture for n = 3202 3' ends of the IpA isoforms (annotated and unannotated 3' ends). If the transcripts ends differed from the 3' ends of the IpA isoforms, then we defined the 3' end determined from the 3'-seq to be the real end. This was done as 3'-seq identifies 3' ends of polyadenylated mRNAs at a single nucleotide resolution and thus position of these ends would be much more accurate than the ends of transcripts obtained from transcript assembly. Figure 6.1.1 shows an assembled transcript for CCT4 IpA isoform.

Coding potential prediction Our goal was to determine the probability that the 5' IpA events were non-coding transcripts. To accomplish this aim, we made use of CPAT, a coding potential assessment tool that predicts the coding potential of the transcript based on four sequence features: open reading frame size, open reading frame coverage, Fickett TESTCODE statistic and hexamer usage bias (Wang et al. 2013). Once the coding potential of the transcripts was predicted, we found that early ending transcripts had a lower probability to be coding, as we hypothesized (Figure 6.1.2). This implied that the 5' IpA events have a higher probability to be non-coding RNAs. For the remaining analysis we considered non-coding IpA isoforms to be the ones that had coding potential probability less than 0.3 and had retained coding sequence less than 25% (n = 711). These putative non-coding 5' IpA isoforms could be involved in various functional roles. A number of studies have shown that non-coding RNAs generated from the gene locus critically regulate the same gene (Han et al. 2007;

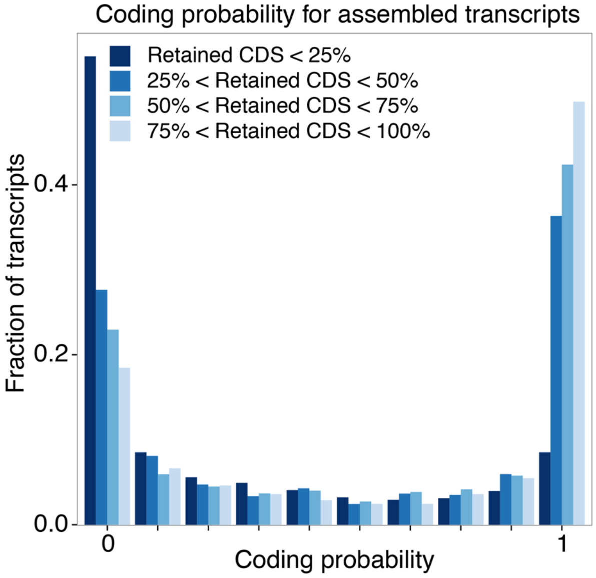

Figure 6.1.2: Most of the transcripts retaining less than 25% of the coding sequence are predicted to be non-coding
The transcripts that retain less than 25% of the coding sequence of the complete transcript have very low probability to be translated into protein.

Martianov et al. 2007; Schmitz et al. 2010; Di Ruscio et al. 2013). From our results, the presence of non-coding RNAs across a large pool of genes suggests that this phenomenon of cis-regulation could be more widespread than previously appreciated.

6.2 Non-coding IpA isoforms enriched for binding sites of RNA binding proteins

Over the years a number of independent studies have established the role of different long non-coding RNAs (lncRNAs) involved in regulating development and differentiation (Fatica and Bozzoni 2014). A highly studied case is role of lncRNAs in X chromosome inactivation and genomic imprinting. Three different lncRNAs, X-inactive specific transcript (*Xist*), *Kcnq1* overlapping transcript 1 (*Kcnq1ot1*) and *Airn* (antisense *Igf2r* (insulin-like growth factor 2 receptor) RNA) co-ordinate to establish repressive chromatin (Lee and Bartolomei 2013). They do so by the recruitment of DNA methyltransferase 3 (DNMT3), which induces DNA methylation; Polycomb repressive complex 2 (PRC2), which produces histone H3 lysine 27 trimethylation (H3K27me3); and histone lysine *N*-methyltransferase EHMT2, which is responsible for producing H3K9me2 and H3K9me3. Another case of cis regulation by a lncRNA is the HOXA distal transcript antisense RNA (*HOTTIP*), which is generated from the 5' tip of the HOXA locus (Wang et al. 2011). HOTTIP recruits the MLL1 complex leading to deposition of the active H3K4me3 mark. LncRNAs have also been shown to play a crucial role in trans-regulation. A well-studied case for trans-regulation is the recruitment of two repressive complexes, PRC2 and the H3K4 demethylating complex over the HOXD genes by HOXA transcript antisense RNA (*HOTAIR*) (Rinn et al. 2007). Encouraged by these examples of lncRNA-mediated gene regulation, we wanted to investigate if the IpA isoforms that we predict to be non-coding RNAs could have the ability to recruit RNA binding proteins and play a role in regulation. We were particularity interested to find out if the IpA isoform gained intronic sequence, that was enriched with RNA binding protein sites.

To investigate our hypothesis, we used available RBP CLIP (cross-linking immunoprecipitation) data from different studies. The different CLIP protocols are next-generation sequencing based assays that provide a transcriptome-wide map of RNA binding proteins sites. A majority of CLIP studies have been done in a cell line derived from human embryonic kidney cells (HEK293). We obtained the binding site information from a resource database, doRiNA that provides processed CLIP experimental results from many different studies (Blin et al. 2015). As the binding site information is only available for HEK293 cells, we focused on the non-coding IpA isoforms expressed in HEK293. Only IpA isoforms that had the gained part of an intron longer than 50nt were used for the analysis (IpA isoforms = 79, genes = 72). We wanted to determine if this gained part of the intron was enriched for RNA binding protein sites compared to the other introns of the same genes that were not included in the non-coding isoforms. The entire length of the gained part of the non-coding isoform and the other introns was converted into windows of lengths 50 nts. For every CLIP experiment, we determined the probability of having a binding site in the other introns (background) to get the expected number of binding sites in the gained part of the non-coding IpA isoforms (expected value). Using the expected and observed number of sites we calculated the binomial Z-score of each CLIP experiment. We repeated the same procedure to get the binomial z-score with coding exons of the same genes as background. We found that these gained parts of the introns are enriched for various RBP binding sites. The RBPs for which the Z-score ≥ 10 with other introns as the background and Z-score ≥ 2 with coding sequence as the background were the RBPs that could potentially be recruited by these non-coding IpA isoforms (PUM2, FUS, AGO2, LIN28B, HUR, ZC3H7B, TIAL1, TAF15 and TIA1). Figure 6.2.1 shows IpA isoform of CUL1 with binding sites for RBPs. This observation suggests that the non-coding IpA isoforms could potentially recruit RBPs to the same gene locus, to a target locus in *trans*, or to some other subcellular compartment.

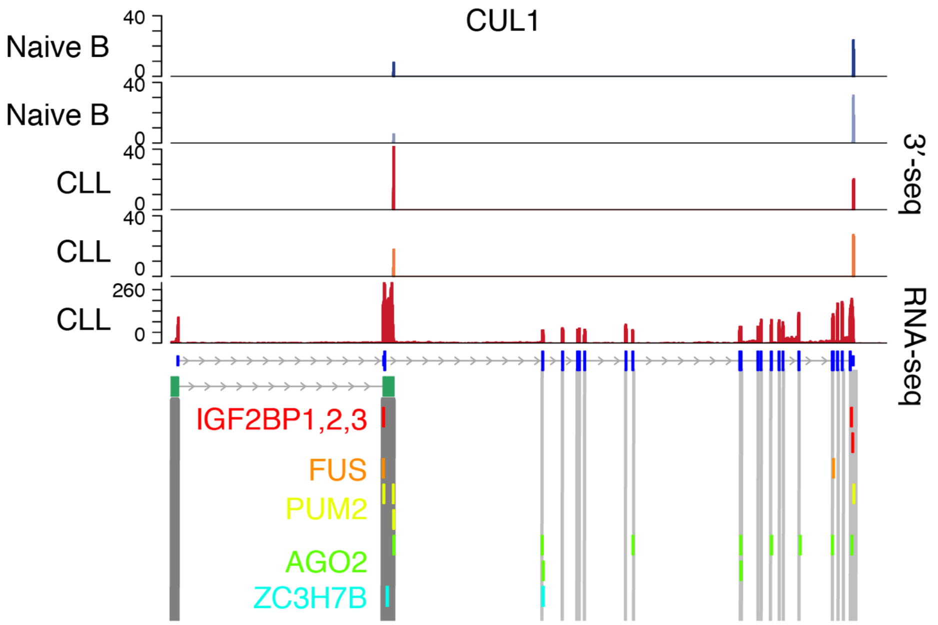

Figure 6.2.1: Gained part of intron in non-coding IpA isoforms is enriched for RBP sites

The gained part of intron in the non-coding IpA isoforms is enriched for RNA binding protein sites compared to the other introns of the same genes. This shows an example of a non-coding IpA isoform of CUL1 that has gained part of the second intron. The gained part has binding sites for PUM2, AGO2 and ZC3H7B.

6.3 IpA isoforms potentially diversify proteome but largely do not affect the transcriptional activity of the genes

Transcriptional activity of genes is largely unaffected by expression of IpA

isoforms One of the most intuitive functions of the IpA isoforms could be to change the full-length expression of the genes. Previous studies have suggested that higher expression of IpA isoforms would cause switch-like reduction of expression of the full-length transcript (Tian et al. 2007; Andersen et al. 2012). We also hypothesized that expression of IpA isoforms would lead to repression of the full-length expression of the genes and vice-versa. To test this hypothesis, we compared expression of the full-length transcript between two conditions (naïve B cells (blood) vs naïve/CD5+ B cells (tonsil), CLL group 1 vs N5, MM group 1 vs PC) for all the expressed genes with IpA isoforms. Contrary to our hypothesis we found that in all three comparisons less than half of IpA isoforms (38% in B cells, 20% in CLL group 1 and 49% in MM

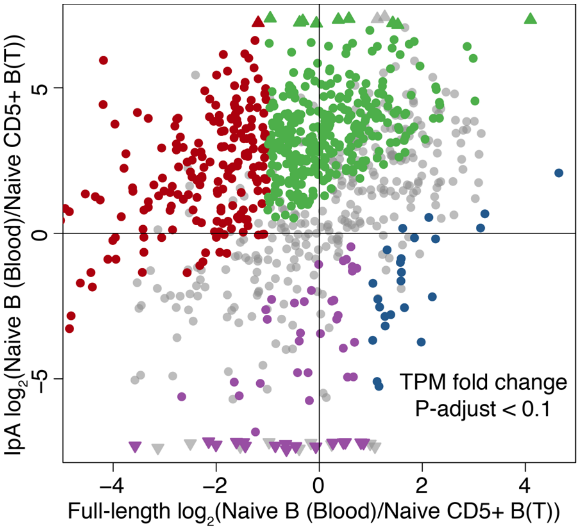

Figure 6.3.1: Transcriptional activity of genes largely unaffected by expression of IpA isoforms

Less than 50% of genes show increase/decrease in the full-length transcriptional activity by the differential usage of the IpA site. This figure shows the fold change of the IpA isoform and the full-length isoform in naïve B cells (blood) vs naïve/CD5 + B cells (tonsils). The genes that have IpA isoforms with significantly different usage (adjusted $p < 0.1$ and difference in usage < 0.1) of the IpA site are highlighted in red, green, purple or blue. Red genes: Full-length express was downregulated with high expression of IpA isoforms; green genes: IpA isoform was expressed without any effect on the full-length isoform expression; blue genes: Loss of expression of IpA isoform caused upregulation of the full-length isoform and purple genes: Loss of usage of IpA site had no effect on the expression of full-length isoform.

group 1) appeared to have affected the full-length expression (Figure 6.3.1). This observation indicates that the expression of majority of IpA isoforms is independent from the expression of the full-length transcript, contrary to previous reports (Tian et al. 2007; Andersen et al. 2012).

IpA isoforms contribute towards diversification of proteome Our previous results in Section 6.1 suggest that IpA isoforms retaining more than 25% of coding sequence (n = 3240) likely have a higher potential to be translated into protein. IpA isoforms that end in the same intron were treated as one case. These IpA isoforms would still create proteins that are different from the proteins generated from the full-length mRNAs. Such IpA isoforms would have lost certain protein domains due to early cleavage, like the IgM gene or the receptor tyrosine kinases (Early et al. 1980; Rogers

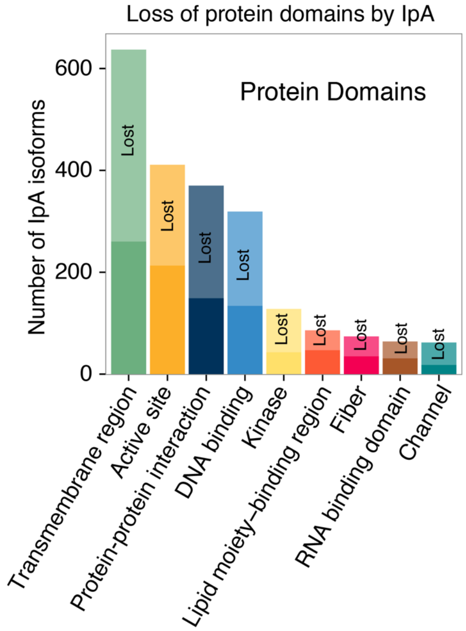

Figure 6.3.2: IpA can potentially diversify proteome
Usage of IpA sites creates mRNA that would be translated into proteins different than the full-length isoforms. This figure shows the different domains that are present in the genes that express IpA isoforms. It shows the number of IpA isoforms that loose one of these domains and would potentially contribute towards diversification of proteome.

et al. 1980; Vorlova et al. 2011). If a transcription factor mRNA isoform loses a DNA-binding domain or an enzyme mRNA isoform loses the active site, then such isoforms would create proteins that differ significantly in their function. We believe that IpA could be a way to diversify the proteome. To get an assessment of the protein domains that are lost from the IpA isoforms, we determined the distribution of IpA isoforms that loose one of the important domains. The IpA isoforms that retained more than 25% of their coding sequence were used for the analysis. We found that 50% of such isoforms (n = 1657) have at least one of the following domains: transmembrane domain, an active site, a protein-protein interaction domain, DNA binding domain, kinase domain, lipid binding region, fiber, RNA-binding domain or channel. Amongst these more than half of IpA isoforms (61.19%) loose at least one of the above domains (n = 1014). Figure 6.3.2 shows the distribution of loss of these domains by the IpA cleavage event.

CHAPTER 7

7. SUMMARY AND CONCLUSION

To my knowledge our study is the first to investigates the usage of IpA site on a genome-wide level across diverse tissue and cell types. In this dissertation we characterized IpA isoforms across a wide variety of human cell types and B cell malignancies. We made use of high throughput 3' end sequencing to gain a finer understanding of the usage of IpA sites, its contribution to generate non-coding RNAs and alternative protein isoforms, and the functional importance of these IpA isoforms. We carefully created an atlas of IpA isoforms that are robustly expressed in the various cell types. We also investigated various properties of these IpA isoforms and their potential functional consequences. Using the atlas, we learned how the landscape of IpA isoform expression changes across different normal cellular conditions and in B cell malignancies compared to the normal cell type of origin. We also tried to understand how the differentially expressed IpA isoforms could potentially contribute to tumorigenesis by mimicking somatic alterations. A summary of our findings from this study is presented below:

7.1 Atlas with robustly expressed IpA isoforms across a wide variety of tissue and cell types

We performed a high throughput sequencing protocol called 3'-seq that precisely identifies the 3' ends of the transcripts on a global level over a large number of cell types and tissue. 3'-seq is a highly quantitative, tag-based sequencing approach that provides an accurate quantification of the relative expression of 3' end isoforms. Using

3'-seq, we created an atlas of robustly expressed IpA isoforms across diverse tissues, normal immune cells, and B cell malignancies. The pipeline for creating the atlas accounted for all known potential artifacts and used various steps to filter out spurious IpA isoforms. Cases where the genomic context was complicated and it was difficult to assign isoforms to a single gene with high confidence were also removed. At the end, IpA isoforms that were potentially produced by "transcriptional noise" and unlikely to produce highly expressed non-coding RNAs or alternative coding isoforms were filtered out, leaving only robust cleavage events in the atlas. This atlas became the basis of all the future analysis. Similar to an earlier study that mapped pASs using cDNA/ESTs on the human genome (Tian et al. 2007), we found usage of IpA sites in approximately 23% of expressed genes in wide variety of tissue and cell types.

7.2 IpA isoforms are expressed as part of the normal expression program with variable expression patterns in different cell types

We found out that the fraction of genes expressing IpA isoforms varies between different cell types. Immune cells have a higher proportion of expressed genes with intronic cleavage events compared to solid tissues. Variability in the usage of IpA sites in different cell lines has also been reported previously (Tian et al. 2007). We observed almost no IpA isoforms that were uniquely expressed in a single tissue. Rather, the majority of IpA isoforms are expressed in at least two cell types where the gene is expressed. The tissue-specific isoforms are expressed in tissue-specific genes of the particular tissue type. Since IpA isoforms are robustly expressed and regulated across normal cells, usage of IpA sites should be viewed as part of the normal expression program. Regulated usage of IpA site in the heavy chain IgM gene in the B cell maturation pathway is one of the classic case of intronic alternative cleavage and polyadenylation (Early et al. 1980; Rogers et al. 1980). However, since then very few studies have appreciated the potential of intronic cleavage and polyadenylation to serve as a layer of gene regulation. Amongst these studies, the few well known cases where the regulation of IpA site plays a role in generating different protein isoforms are calcitonin/calcitonin gene-related peptide genes (CALCA), transcriptional factor SREPF and FLT1 (Wang et al. 2006; Thomas et al. 2007). There have been some studies describing the usage of IpA sites as recognition of cryptic polyadenylation sites or a premature cleavage and polyadenylation event that should not occur in normal circumstances (Kaida et al. 2010; Berg et al. 2012). Thus, our study is critical in establishing that IpA isoforms are not generated by noisy or undesired usage of cryptic pAS in the introns, as previously believed. A recent study has described generation of new alternative 3' ends in introns by exonisation of *Alu* elements (Tajnik et al. 2015). Similar to our observation, this study describes the usage of these IpA sites to be tissue-specific. We found the expression of a large number of IpA isoforms to be tightly regulated between the same cells obtained from different cellular environments (naive B cells from tonsils and blood) obtained from different cellular environments. We found an example of a differentially regulated IpA isoform of SPTBN1, which is translated into two different protein isoforms that might have a direct role in defining cellular shape in these different cellular environments. The strong regulation of expression of IpA isoforms between normal cell types provides evidence of their functional role. Levels of hnRNP C and U2AF65 have been shown to play a role in defining the usage of IpA sites originating by *Alu* exonisation (Tajnik et al. 2015). We also believe that the level of the factors involved in splicing and cleavage would play a pivotal role in determining the usage of IpA sites in wide variety of tissue and cell types. Multiple studies have suggested an interplay between splicing and alternative cleavage, specifically pASs located in the introns of the genes

(Tikhonov et al. 2013; Movassat et al. 2016). A systematic screening of the splicing machinery and cleavage factors is required to identify the key players involved in the usage of the IpA sites.

7.3 Diverse pattern of usage of IpA sites along transcription units in various tissue and cell types

The positions where the intronic cleavage events take place are critical, as they define the protein that would be translated from the IpA isoforms. We observed that there is a higher tendency for the IpA isoforms to have their cleavage sites close to the 5' UTR of the transcription unit, thus retaining at most a small portion of the coding sequence. However, this positional preference of IpA sites varies between different cell types. The majority of IpA isoforms expressed in tissues like ovary and brain use IpA sites located close to the 3' end of the transcription unit (3' IpA events). Such isoforms would be translated into alterative protein isoforms due to loss of 3' terminal exons. Cell types like T cells and plasma cells have most of the 3' ends of the IpA isoforms in early introns close to the start of the transcription unit (5' IpA events). We also observed that cell types with a higher fraction of genes expressing IpA isoforms also have a higher number of 5' IpA events. This indicates that cell types that express a higher number of IpA isoforms tend to increase the usage of IpA sites near the start of the transcription unit. The controlled usage of IpA sites in different tissue types suggests that these IpA isoforms potentially create wide variety of mRNAs that might be involved in functional roles. Additionally, we observed a correlation between the length of the 3' UTRs and the proportion of IpA isoforms that are expressed across cell types. We found that the tissues with shorter 3' UTRs have higher proportion of genes with IpA isoforms. This observation suggests that the factors involved in defining the length of 3' UTRs potentially also play a role in regulating the usage of IpA sites.

7.4 Retained introns are associated with the occurrence of IpA events

Utilizing the RNA-seq data we found that IpA events are enriched in introns that are retained. We also observed that the tissues and cell types with a higher number of retained introns had a higher fraction of genes with IpA isoforms. This observation suggests of an association between the two processes: intron retention and IpA. We hypothesize that retention of the intron could facilitate the recognition of an IpA site by the cleavage machinery leading to the formation of the 3' end. Splicing machinery factors have been shown to be involved in retention of intron (Wong et al. 2013). In particular, expression levels of factors that define the exon junctions, U5, U4/U6, U1, U2, U2AF along with SR proteins affect the intron retention levels. We believe that the expression levels of these factors in conjunction with cleavage and polyadenylation machinery factors might be important in defining the usage of IpA site.

7.5 Genes with long transcription units, long introns and long 5' UTRs are enriched with pASs and devoid of U1 snRNP signals to facilitate intronic cleavage events

We observed that genes with IpA events have a special genomic architecture along with a different sequence composition that together facilitate the formation of 3' ends in introns. We found that IpA events are most prevalent in longer transcription units that have long introns. Occurrence of higher usage of IpA sites in long introns in humans has also been reported previously (Tian et al. 2007). Similar to our results, usage of IpA sites in long introns of genes in *Arabidopsis Thaliana* has been observed (Guo et al. 2016). These genes also have longer 5' UTRs in comparison to the genes that do not express IpA isoforms. This observation of higher incidence of recognition of IpA sites in long introns asserts that the long introns provide an increased opportunity for the cleavage machinery to define a cleavage event. Apart from these architectural features, the genes that express IpA isoforms also differ in their sequence composition. There is a higher frequency of pASs in these genes while these genes are depleted for U1 snRNP signals. The combination of the presence and absence of these two kinds of signals makes these genes ideal candidates to express IpA isoforms. U1 snRNP has been shown to play a crucial role in preventing premature cleavage and polyadenylation while recognition of pASs leads to a 3' cleavage event (Kaida et al. 2010; Berg et al. 2012). Thus the genes that express IpA isoforms have special architectural and compositional features in comparison to the genes that always make full-length isoforms. However, it is important to realize that the usage of these IpA sites is regulated strongly between cell types, resulting in varied expression of IpA isoforms across different tissue and cell types. To establish this more firmly we would need to measure the levels of U1 snRNP in our samples to affirm the role of U1 snRNP is recognition of IpA sites.

7.6 Intronic cleavage events creates a wide variety of mRNAs with potentially diverse functions

In our atlas, the majority of IpA isoforms use an IpA site close to the start of the transcription unit, retaining at most a small fraction of the coding sequence. We established that there is a higher probability for these early ending IpA isoforms to be non-coding mRNAs. Other studies have demonstrated the role non-coding RNAs generated from the gene locus in regulation of genes (Martianov et al. 2007; Di Ruscio et al. 2013). We also investigated the potential role of these non-coding mRNAs in cis/trans gene regulation. We found that intronic sequences that become part of the non-coding transcript are enriched for binding sites of RNA binding proteins compared to the other introns of the genes that were not incorporated in the IpA isoform. This enrichment of binding sites for RNA binding proteins may suggest a potential role of these non-coding IpA isoforms in gene regulation. The IpA isoforms with cleavage sites near the 3' end of the transcription unit lose C-terminal exons. Such mRNAs would create proteins with altered functions and would potentially contribute towards diversifying the proteome. Loss of important protein domains like the active site of an enzyme or a DNA-binding domain of a transcription factor could also make the protein dysfunctional. We hypothesized in the beginning of the study that the likefest function of non-coding IpA isoforms would be to regulate the expression of full-length isoforms. However, by comparing various conditions, we found that the full-length transcript expression level is only affected by an increase/loss in usage of the IpA sites in less than 50% of genes. This observation suggests that in a majority of cases the expression of IpA isoforms is independent of the expression of the full-length isoform.

7.7 Truncated mRNAs generated by the usage of IpA sites can mimic genetic mutations

We examined the change in landscape of usage of IpA sites in chronic lymphocytic leukemia compared to naïve and CD5+ B cells obtained from tonsils. We found a group of CLL samples to have significantly higher usage of IpA sites relative to full-length isoforms. We hypothesized that in cancer, IpA isoforms could have the potential to contribute to tumorigenesis by mimicking genetic alterations like loss-of-function mutations, chromosomal translocations or deletions. The IpA isoforms creating alternative protein isoforms by loss of C-terminal exons could potentially mimic loss-of-function mutation or chromosomal translocation, while the IpA isoforms resulting in non-coding RNAs could mimic the deletion of a gene. We investigated this hypothesis in CLL and indeed found that a subset of genes with truncating mutations in CLL also have increased usage of IpA sites relative to the full-length isoform. Strikingly, we found that a much higher proportion of CLL patients express IpA isoforms compared to the proprtion that harbor truncating mutations. This finding suggests that the cancer cells might use IpA as an alternative mechanism to achieve proliferation and survival in absence of a somatic mutation. We confirmed this finding by showing that the truncated protein created by IpA isoforms is similar to the truncated protein created by identified mutations in the MGA gene. Another study has demonstrated the recognition of IpA site to generate MAGI3 truncated protein in human breast cancer (Ni and Kuperwasser 2016). This truncated protein acts a dominant-negative oncogene promoting the malignant transformation human mammary epithelial cells. Our results also suggest that usage of IpA sites could lead to creation of protein isoforms contributing towards tumorigenesis in CLL. We also examined the change in usage of IpA sites in multiple myeloma. In contrast to CLL, we found that multiple myeloma has decreased usage of IpA sites compared to normal plasma cells.

7.8 Conclusion

In this dissertation we conducted an in-depth analysis of 3' end sequencing libraries to characterize the IpA isoforms across a wide variety of tissue and cell types. With this analysis we were able to establish that IpA isoforms are expressed and regulated as a part of the normal expression program in human cells. These IpA isoforms represent a wide variety of mRNAs that could potentially have diverse functions. Truncated transcripts generated by IpA could create alternative protein isoforms with altered functions or create non-coding RNAs that potentially serve as a layer of gene regulation. This study provided evidence that cancer cells can regulate the usage of IpA sites to mimic genetic mutations, providing an alternative mechanism for activated oncogenic expression programs. Overall this study contributes significantly towards a better understanding of potential functional role of IpA sites in normal and cancer cells.

8. REFERENCES

Almada AE, Wu X, Kriz AJ, Burge CB, Sharp PA. 2013. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**: 360-363.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**: R106.

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

Andersen PK, Lykke-Andersen S, Jensen TH. 2012. Promoter-proximal polyadenylation sites reduce transcription activity. *Genes Dev* **26**: 2169-2179.

Beck AH, Weng Z, Witten DM, Zhu S, Foley JW, Lacroute P, Smith CL, Tibshirani R, van de Rijn M, Sidow A et al. 2010. 3'-end sequencing for expression quantification (3SEQ) from archival tumor samples. *PLoS One* **5**: e8768.

Bennett CL, Brunkow ME, Ramsdell F, O'Briant KC, Zhu Q, Fuleihan RL, Shigeoka AO, Ochs HD, Chance PF. 2001. A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA-->AAUGAA) leads to the IPEX syndrome. *Immunogenetics* **53**: 435-439.

Berg MG, Singh LN, Younis I, Liu Q, Pinto AM, Kaida D, Zhang Z, Cho S, Sherrill-Mix S, Wan L et al. 2012. U1 snRNP determines mRNA length and regulates isoform expression. *Cell* **150**: 53-64.

Black DL. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72**: 291-336.

Blin K, Dieterich C, Wurmus R, Rajewsky N, Landthaler M, Akalin A. 2015. DoRiNA 2.0--upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res* **43**: D160-167.

Consortium EP. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

Das A, Base C, Manna D, Cho W, Dubreuil RR. 2008. Unexpected complexity in the mechanisms that target assembly of the spectrin cytoskeleton. *J Biol Chem* **283**: 12643-12653.

Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22**: 1173-1183.

Di Ruscio A, Ebralidze AK, Benoukraf T, Amabile G, Goff LA, Terragni J, Figueroa ME, De Figueiredo Pontes LL, Alberich-Jorda M, Zhang P et al. 2013. DNMT1-interacting RNAs block gene-specific DNA methylation. *Nature* **503**: 371-376.

Early P, Rogers J, Davis M, Calame K, Bond M, Wall R, Hood L. 1980. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell* **20**: 313-319.

Edwalds-Gilbert G, Veraldi KL, Milcarek C. 1997. Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Res* **25**: 2547-2561.

Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. 2008. LIBLINEAR: A Library for Large Linear Classification. *J Mach Learn Res* **9**: 1871-1874.

Fatica A, Bozzoni I. 2014. Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet* **15**: 7-21.

Fitzgerald M, Shenk T. 1981. The sequence 5'-AAUAAA-3'forms parts of the recognition site for polyadenylation of late SV40 mRNAs. *Cell* **24**: 251-260.

Frayne EG, Leys EJ, Crouse GF, Hook AG, Kellems RE. 1984. Transcription of the mouse dihydrofolate reductase gene proceeds unabated through seven polyadenylation sites and terminates near a region of repeated DNA. *Mol Cell Biol* **4**: 2921-2924.

Fu Y, Sun Y, Li Y, Li J, Rao X, Chen C, Xu A. 2011. Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res* **21**: 741-747.

Gil A, Proudfoot NJ. 1984. A sequence downstream of AAUAAA is required for rabbit beta-globin mRNA 3'-end formation. *Nature* **312**: 473-474.

Guo C, Spinelli M, Liu M, Li QQ, Liang C. 2016. A Genome-wide Study of "Non-3UTR" Polyadenylation Sites in Arabidopsis thaliana. *Sci Rep* **6**: 28060.

Han J, Kim D, Morris KV. 2007. Promoter-associated RNA is required for RNA-directed transcriptional gene silencing in human cells. *Proc Natl Acad Sci U S A* **104**: 12422-12427.

Higgs DR, Goodbourn SE, Lamb J, Clegg JB, Weatherall DJ, Proudfoot NJ. 1983. Alpha-thalassaemia caused by a polyadenylation signal mutation. *Nature* **306**: 398-400.

Hoek KL, Samir P, Howard LM, Niu X, Prasad N, Galassie A, Liu Q, Allos TM, Floyd KA, Guo Y et al. 2015. A cell-based systems biology assessment of human blood to monitor immune responses after influenza vaccination. *PLoS One* **10**: e0118528.

Hoque M, Ji Z, Zheng D, Luo W, Li W, You B, Park JY, Yehia G, Tian B. 2013. Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat Methods* **10**: 133-139.

Hurlin PJ, Steingrimsson E, Copeland NG, Jenkins NA, Eisenman RN. 1999. Mga, a dual-specificity transcription factor that interacts with Max and contains a T-domain DNA-binding motif. *EMBO J* **18**: 7019-7028.

Jan CH, Friedman RC, Ruby JG, Bartel DP. 2011. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature* **469**: 97-101.

Ji Z, Lee JY, Pan Z, Jiang B, Tian B. 2009. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc Natl Acad Sci U S A* **106**: 7028-7033.

Kaida D, Berg MG, Younis I, Kasim M, Singh LN, Wan L, Dreyfuss G. 2010. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**: 664-668.

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**: 1484-1488.

Kaufman RJ, Sharp PA. 1983. Growth-dependent expression of dihydrofolate reductase mRNA from modular cDNA genes. *Mol Cell Biol* **3**: 1598-1608.

Landau DA, Tausch E, Taylor-Weiner AN, Stewart C, Reiter JG, Bahlo J, Kluth S, Bozic I, Lawrence M, Bottcher S et al. 2015. Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**: 525-530.

Lee JT, Bartolomei MS. 2013. X-inactivation, imprinting, and long noncoding RNAs in health and disease. *Cell* **152**: 1308-1323.

Lejeune F, Maquat LE. 2005. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. *Curr Opin Cell Biol* **17**: 309-315.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Li Y, Sun Y, Fu Y, Li M, Huang G, Zhang C, Liang J, Huang S, Shen G, Yuan S et al. 2012. Dynamic landscape of tandem 3' UTRs during zebrafish development. *Genome Res* **22**: 1899-1906.

Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. 2013. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev* **27**: 2380-2396.

Lim L, Canellakis ES. 1970. Adenine-rich polymer associated with rabbit reticulocyte messenger RNA. *Nature* **227**: 710-712.

Mandel CR, Kaneko S, Zhang H, Gebauer D, Vethantham V, Manley JL, Tong L. 2006. Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature* **444**: 953-956.

Martianov I, Ramadass A, Serra Barros A, Chow N, Akoulitchev A. 2007. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* **445**: 666-670.

Martin G, Gruber AR, Keller W, Zavolan M. 2012. Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell Rep* **1**: 753-763.

Mayr C, Bartel DP. 2009. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**: 673-684.

Miura P, Shenker S, Andreu-Agullo C, Westholm JO, Lai EC. 2013. Widespread and extensive lengthening of 3' UTRs in the mammalian brain. *Genome Res* **23**: 812-825.

Morris AR, Bos A, Diosdado B, Rooijers K, Elkon R, Bolijn AS, Carvalho B, Meijer GA, Agami R. 2012. Alternative cleavage and polyadenylation during colorectal cancer development. *Clin Cancer Res* **18**: 5256-5266.

Mount SM, Pettersson I, Hinterberger M, Karmas A, Steitz JA. 1983. The U1 small nuclear RNA-protein complex selectively binds a 5' splice site in vitro. *Cell* **33**: 509-518.

Movassat M, Crabb TL, Busch A, Yao C, Reynolds DJ, Shi Y, Hertel KJ. 2016. Coupling between alternative polyadenylation and alternative splicing is limited to terminal introns. *RNA Biol* **13**: 646-655.

Ni TK, Kuperwasser C. 2016. Premature polyadenylation of MAGI3 produces a dominantly-acting oncogene in human breast cancer. *Elife* **5**.

Orkin SH, Cheng TC, Antonarakis SE, Kazazian HH, Jr. 1985. Thalassemia due to a mutation in the cleavage-polyadenylation signal of the human beta-globin gene. *EMBO J* **4**: 453-456.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**: 290-295.

Pfeifer D, Pantic M, Skatulla I, Rawluk J, Kreutz C, Martens UM, Fisch P, Timmer J, Veelken H. 2007. Genome-wide analysis of DNA copy number changes and LOH in CLL using high-density SNP arrays. *Blood* **109**: 1202-1210.

Proudfoot NJ. 2011. Ending the message: poly(A) signals then and now. *Genes Dev* **25**: 1770-1782.

Proudfoot NJ, Brownlee GG. 1976. 3' non-coding region sequences in eukaryotic messenger RNA. *Nature* **263**: 211-214.

Puente XS, Pinyol M, Quesada V, Conde L, Ordonez GR, Villamor N, Escaramis G, Jares P, Bea S, Gonzalez-Diaz M et al. 2011. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475**: 101-105.

Quesada V, Conde L, Villamor N, Ordonez GR, Jares P, Bassaganyas L, Ramsay AJ, Bea S, Pinyol M, Martinez-Trillos A et al. 2012. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet* **44**: 47-52.

Ramskold D, Wang ET, Burge CB, Sandberg R. 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* **5**: e1000598.

Ranzani V, Rossetti G, Panzeri I, Arrigoni A, Bonnal RJ, Curti S, Gruarin P, Provasi E, Sugliano E, Marconi M et al. 2015. The long intergenic noncoding RNA landscape of human lymphocytes highlights the regulation of T cell differentiation by linc-MAF-4. *Nat Immunol* **16**: 318-325.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E et al. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**: 1311-1323.

Rogers J, Early P, Carter C, Calame K, Bond M, Hood L, Wall R. 1980. Two mRNAs with different 3' ends encode membrane-bound and secreted forms of immunoglobulin mu chain. *Cell* **20**: 303-312.

Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. 2008. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science* **320**: 1643-1647.

Schmitz KM, Mayer C, Postepska A, Grummt I. 2010. Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev* **24**: 2264-2269.

Shepard PJ, Choi EA, Lu J, Flanagan LA, Hertel KJ, Shi Y. 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**: 761-772.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034-1050.

Smibert P, Miura P, Westholm JO, Shenker S, May G, Duff MO, Zhang D, Eads BD, Carlson J, Brown JB et al. 2012. Global patterns of tissue-specific alternative polyadenylation in Drosophila. *Cell Rep* **1**: 277-289.

Tajnik M, Vigilante A, Braun S, Hanel H, Luscombe NM, Ule J, Zarnack K, Konig J. 2015. Intergenic Alu exonisation facilitates the evolution of tissue-specific transcript ends. *Nucleic Acids Res* **43**: 10492-10505.

Thomas CP, Andrews JI, Liu KZ. 2007. Intronic polyadenylation signal sequences and alternate splicing generate human soluble Flt1 variants and regulate the abundance of soluble Flt1 in the placenta. *FASEB J* **21**: 3885-3895.

Tian B, Hu J, Zhang H, Lutz CS. 2005. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* **33**: 201-212.

Tian B, Manley JL. 2013. Alternative cleavage and polyadenylation: the long and short of it. *Trends Biochem Sci* **38**: 312-320.

Tian B, Pan Z, Lee JY. 2007. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res* **17**: 156-165.

Tikhonov M, Georgiev P, Maksimenko O. 2013. Competition within Introns: Splicing Wins over Polyadenylation via a General Mechanism. *Acta Naturae* **5**: 52-61.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**: 562-578.

Ulitsky I, Shkumatava A, Jan CH, Subtelny AO, Koppstein D, Bell GW, Sive H, Bartel DP. 2012. Extensive alternative polyadenylation during zebrafish development. *Genome Res* **22**: 2054-2066.

Vorlova S, Rocco G, Lefave CV, Jodelka FM, Hess K, Hastings ML, Henke E, Cartegni L. 2011. Induction of antagonistic soluble decoy receptor tyrosine kinases by intronic polyA activation. *Mol Cell* **43**: 927-939.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470-476.

Wang H, Sartini BL, Millette CF, Kilpatrick DL. 2006. A developmental switch in transcription factor isoforms during spermatogenesis controlled by alternative messenger RNA 3'-end formation. *Biol Reprod* **75**: 318-323.

Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA et al. 2011. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**: 120-124.

Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* **41**: e74.

Wong JJ, Ritchie W, Ebner OA, Selbach M, Wong JW, Huang Y, Gao D, Pinello N, Gonzalez M, Baidya K et al. 2013. Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* **154**: 583-595.

Xia Z, Donehower LA, Cooper TA, Neilson JR, Wheeler DA, Wagner EJ, Li W. 2014. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat Commun* **5**: 5274.

Yoon OK, Brem RB. 2010. Noncanonical transcript forms in yeast and their regulation during environmental stress. *RNA* **16**: 1256-1267.

Zhang H, Lee JY, Tian B. 2005. Biased alternative polyadenylation in human tissues. *Genome Biol* **6**: R100.