Title

The Identification and Characterization of Alterations to DNA and RNA in Cancer Using Next-Generation Sequencing Data

Permalink

https://escholarship.org/uc/item/0dt1w1fx

Author

Radenbaugh, Amie

Publication Date

2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

THE IDENTIFICATION AND CHARACTERIZATION OF ALTERATIONS TO DNA AND RNA IN CANCER USING NEXT-GENERATION SEQUENCING DATA

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Amie J. Radenbaugh

June 2015

The dissertation of Amie J. Radenbaugh is approved:

Professor David Haussler, Chair

Professor Joshua Stuart

Assistant Professor Eric Collisson, M.D.

Professor Todd Lowe

Dean Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

The Identification and Characterization of Alterations to DNA and RNA in Cancer Using Next-Generation Sequencing Data

by

Amie J. Radenbaugh

Much of our current understanding of cancer has come from investigating how normal cells are transformed into malignant cancers through the stepwise acquisition of somatic genomic abnormalities. These abnormalities include single nucleotide variants (SNVs), insertions and deletions (INDELs), chromosomal rearrangements, and copy number aberrations. The detection of SNVs is a crucial component to the characterization of the cancer genome. They assist in identifying key genes as possible drug targets, diagnostic markers for early detection, and prognostic markers for monitoring a patient's response to therapy. Variant calling algorithms thus far have focused on comparing the normal and tumor genomes from the same individual. In recent years, it has become routine for projects like The Cancer Genome Atlas (TCGA) to also sequence the tumor RNA. A novel computational method called RADIA (RNA and DNA Integrated Analysis) that combines the patient-matched normal and tumor DNA with the tumor RNA to detect SNVs is presented here. RADIA has detected somatic mutations for nearly 4,500 patients across 22 different cancers, and including the RNA provided a 2-7% increase in sensitivity.

RNA editing is an additional epigenetic mechanism involved in cancer development and progression. RNA editing of the *AZIN1* gene has been identified as a driver in the pathogenesis of hepatocellular carcinoma and may be a potential driver for other human cancers as well. An investigation of *AZIN1* RNA editing in data collected from nearly 5,000 patients across 12 cancers has been performed. Higher editing frequencies significantly correlated with clinical data such as larger tumor sizes, greater lymph node involvement, the presence of metastases, and higher tumor grades. They were also associated with subtypes that often have the worst prognosis. Over-editing in many cancers is correlated with poor overall and recurrence free survival.

With projects like TCGA providing sequencing data for both DNA and RNA from the same patients across multiple cancers, it is now possible to characterize germline variants, somatic mutations, and RNA editing events on a genome-wide scale. The identification of SNVs that occur in specific genes across multiple cancers provides a powerful way to discover genes that are important to these diseases.

To my wife,

Kathryn Tito

for her endless encouragement, support, and patience,

and my parents,

Dale and Judith Radenbaugh

who always encouraged me to follow my dreams.

Acknowledgments

I would like to thank my advisor, David Haussler, for his invaluable advice and encouragement on this thesis. Much appreciation also goes to my thesis committee members: Joshua Stuart, Eric Collisson and Todd Lowe, for all of the helpful discussions and suggestions. I would like to thank the RADIA team consisting of Jing Zhu, Singer Ma, Adam Ewing, Joshua Stuart, Eric Collisson, and David Haussler and the RNA editing team consisting of Sofie Salama, Divyashree Kushnoor, Wendy Lee, Kiley Graim, Joshua Stuart, Eric Collisson, and David Haussler. In addition, I thank my fellow UCSC cancer group members, especially Sofie Salama, Jing Zhu, Singer Ma, Chris Wilks, Zack Sanborn, Steve Benz, Charlie Vaske, Chris Szeto, Dent Earl, Ngan Nguyen, Tracy Ballinger, Sam Ng, Kyle Ellrott, Olena Morozova, Arjun Rao, Mia Grifford, Kiley Graim, Yulia Newton, Mark Diekhans, Melissa Cline, and Evan Paull, and a special thanks to our extremely dedicated and helpful IT team: Erich Weiler, Jorge Garcia, and Gary Moro. I am very grateful to Sami Khuri for inspiring me to pursue a PhD and Wendy Lee for all of her support.

Chapter 1

Introduction

Cancer is a group of diseases characterized by uncontrolled cell growth with the potential to spread to other tissues in the body. According to the American Cancer Society (ACS), nearly 600,000 Americans, or more than 1,600 people a day, are expected to die of cancer in 2015, and an estimated 1.7 million new cancer cases are expected to be diagnosed [5]. In the US, approximately one in two men and one in three women will develop cancer in their lifetime, with the leading tissues of origin being prostate, lung, and colon for men and breast, lung, and colon for women [5]. The chance of an individual developing cancer is based on both controllable and uncontrollable risk factors [29]. Controllable factors include tobacco use, heavy alcohol consumption, unhealthy diets, physical inactivity, infectious organisms, and exposure to sunlight or harmful environmental pollutants [5, 29]. Uncontrollable factors include germline mutations inherited from one's parents, hormones, and immune conditions [5].

Since the 1990s, the death rates for most cancers have experienced a gradual decline. This decline is primarily due to behavioral changes, early detection, and advances in cancer treatment. A reduction in smoking rates has lowered the lung cancer death rate by approximately 30% [5]. Cancers caused by infectious agents such as human papillomavirus (HPV), hepatitis B virus (HBV), hepatitis C virus (HCV), human immunodeficiency virus (HIV), and *Helicobacter pylori (H. pylori)* have been reduced due to vaccinations and the treatment of infections [5]. Early detection of cervical cancer through the Papanicolaou test (Pap test), prostate cancer from the prostate-specific antigen (PSA) test, and colon cancer by a colonoscopy, has helped prevent and reduce death rates for these cancers. Lastly, advancements in treatment, such as the discovery of *Gleevac* to treat chronic myeloid leukemia (CML) and *Herceptin* to treat Human Epidermal growth factor Receptor 2-positive (HER2+) breast cancer, have drastically improved the survival rates of patients with these specific types of cancer. Despite all of these advancements, there is a great need for a better understanding of the biological mechanisms that play a role in cancer.

There are ten "hallmarks of cancer" that are generally accepted to be the underlying rules for the transformation of normal cells into malignant cancers [46, 47]. They are 1) sustaining proliferative signaling, 2) evading growth suppressors, 3) avoiding immune destruction, 4) enabling replicative immortality, 5) tumor-promoting inflammation, 6) activating invasion and metastasis, 7) inducing angiogenesis, 8) genome instability and mutation, 9) resisting cell death, and 10) deregulating cellular energetics (Figure 1.1) [46, 47].

Transforming a normal human cell into a malignant, immortal cancer cell line

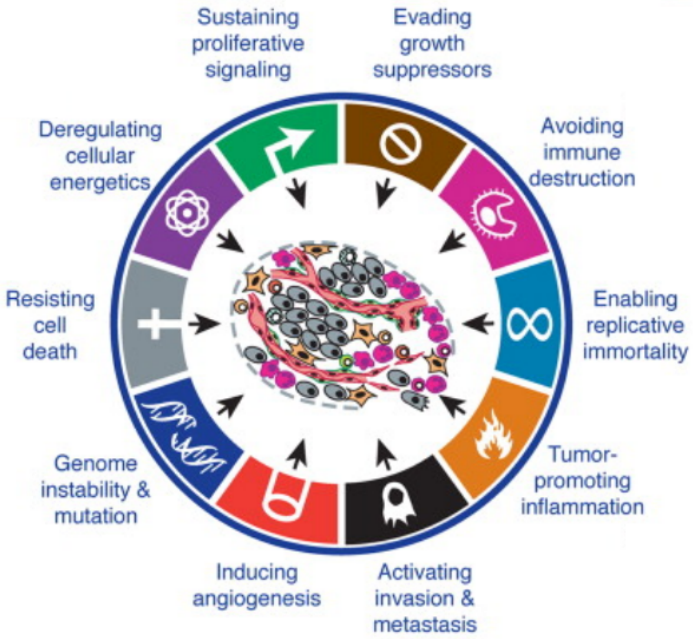

Sustaining proliferative signaling
Evading growth suppressors
Avoiding immune destruction
Enabling replicative immortality
Tumor-promoting inflammation
Activating invasion & metastasis
Inducing angiogenesis
Genome instability & mutation
Resisting cell death
Deregulating cellular energetics

Figure 1.1: The ten hallmarks of cancer that are thought to be the underlying rules for the transformation of normal cells into malignant cancers. Figure from [47].

requires an estimated five to seven genetic alterations in key genes and pathways [46]. Normal cell growth is strictly regulated, and many normal cells need signals such as growth factors to initiate cell division. Normal cells also communicate with nearby cells to determine if it is acceptable for them to divide. Cancer cells, on the other hand, grow uncontrollably. They no longer need special signals from growth factors to divide, and along the way they gain additional abilities to survive and grow in an environment that would normally not tolerate such growth. Not surprisingly, much research has been devoted to determining how cancer cells are able to acquire their abilities through the accumulation of somatic mutations.

In order to identify somatic alterations that are essential for cancer detection, development, and treatment, researchers have focused on determining the genomic differences between tumor cells and normal cells in the same individuals. The Cancer Genome Atlas (TCGA) project has produced exome-wide data from thousands of tumors and patient-matched normal tissues. With the development of RNA Sequencing (RNA-Seq) [139], TCGA began providing an additional high-throughput tumor sequence dataset. These three datasets consisting of tumor and patient-matched normal DNA and tumor RNA have become a new standard in cancer genomics. RNA-Seq enables one to investigate the consequences of genomic changes in the RNA transcripts they encode to better characterize 1) germline variants, 2) somatic mutations, and 3) variants in the RNA that are not found in the DNA that could be the result of RNA editing [43].

With the cost of sequencing steadily decreasing, many more whole-genome and whole-exome DNA and RNA-Seq BAM (the binary version of (SAM) Sequence Alignment/Map [84]) files will become available. TCGA has collected over 10,000 tissue samples from more than 20 types of cancer. There is a clear need for an efficient method for the combined analysis of patient-matched tumor DNA, normal DNA, and tumor RNA. Here, a novel method called RADIA (RNA and DNA Integrated Analysis) has been developed to identify and characterize Single Nucleotide Variants (SNVs) in cancer using DNA and RNA obtained by high-throughput sequencing data. Traditional somatic mutation calling algorithms use the patient-matched pairs of tumor and normal DNA. The inclusion of the RNA in RADIA increases the power to detect somatic mutations that are typically missed by traditional algorithms. RADIA identifies mutations in the most significantly mutated genes that are missed by other algorithms. This increase in sensitivity is essential for the complete characterization of a patient's tumor and assists in grouping patients into subtypes that are partially defined by the presence or absence of mutations in key genes.

Finally, RADIA has been used to identify RNA editing events that have recently been linked to cancer development and progression. An RNA editing event in the *AZIN1* gene has been identified as a possible driver of hepatocellular carcinoma and may be involved in the pathogenesis of other human cancers as well. An in-depth study of *AZIN1* editing was performed on 12 different cancers, and 10 of the 12 cancers exhibted significant over-editing in the *AZIN1* gene. Over-editing of the *AZIN1* gene leads to the overexpression of cyclin D1 protein and an increase in cell proliferation. A particular focus on the luminal subtypes of breast cancer, known for overexpression of cyclin D1 will be given. Higher *AZIN1* editing frequencies significantly correlated with many of the clinical variables that represent more aggressive and advanced characteristics of tumorigenesis, and patients with over-editing of *AZIN1* often have a poor overall or recurrence free survival.

Chapter 2

Background

This chapter provides background information about Single Nucleotide Variants (SNVs) specifically germline variants, somatic mutations, and RNA editing events. SNVs describe variations at a single nucleotide position of a DNA or RNA sequence. Germline SNVs are present in the DNA of a parent and inherited by their offspring, somatic SNVs accumulate in the DNA of an individual during their lifetime, and RNA editing SNVs are modifications to the RNA sequence that are not present in the genomic DNA.

SNVs can have both benign or severe effects on the resulting protein that in return can have neutral or drastic effects on an individual. SNVs that occur in the coding regions can cause: 1) silent mutations, also known as synonymous mutations, resulting in a different codon for the same amino acid, 2) missense mutations, or nonsynonymous mutations, resulting in a different amino acid and/or leading to the readthrough of a stop codon resulting in an elongated protein, 3) nonsense mutations that code for a premature stop codon resulting in a truncated protein. SNVs in splice sites can affect alternative splicing by creating new splice sites or causing a read-through of existing splice sites.

Germline variants can lead to an increase in risk of developing cancer earlier in life. There are many instances of somatic SNVs that can activate oncogenes or inactivate tumor suppressors. A deregulation of RNA editing can promote oncogenic activity or knock out tumor suppressors.

2.1 Germline Variants

There are several factors that influence whether an individual will develop cancer or not, including the environment, life-style, and set of genomic sequence variants that the person is born with. These so called "germline" variants can influence one's susceptibility to cancer in a number of ways, including altering the growth of the tumor, the mutation rate in somatic cells, or the metabolism of carcinogens [130].

Recurring mutations are observed in the germline of multiple cancer patients, indicating that these mutations may be contributing factors to an individual's susceptibility to cancer. For example, women who have germline mutations in the *BRCA1* and *BRCA2* genes are more likely to develop breast or ovarian cancer by age 70 [131]. These germline mutations are thought to affect the cell's ability to repair double-stranded breaks and ultimately lead to the inactivation of the tumor suppressor, *PTEN* [122].

2.2 Somatic Mutations

Somatic mutations occur in the genomes of normal somatic (non-sex) cells during cell division. The rate of accumulation and the types of mutations that occur varies. Although somatic mutations are thought to arise randomly, some mutations called “driver mutations” occur by chance in key genes called “cancer genes”. Driver mutations permit the cell to avoid the normal control of cell proliferation, differentiation, and death [130]. After a driver mutation is acquired, passenger mutations accumulate. Passenger mutations are neither particularly helpful nor harmful to cancer cells. They are neutral mutations that accumulate during the clonal expansion of a cell [130].

A number of driver mutations and cancer genes have already been identified. An assumption is made that passenger mutations are randomly distributed throughout the genome while driver mutations are concentrated in cancer genes. By searching through a large number of samples for a specific type of cancer, genes that have a higher mutation rate than expected by chance, contain possible driver mutations [130]. Roughly 400 somatically mutated cancer genes for multiple types of cancer have been identified [130]. Cancer genes are classified as either dominant or recessive. A common analogy to describe oncogenes (dominant) and tumor suppressors (recessive) is to use the brake and gas pedals in a car. Mutations that transform a gene into an oncogene are like cars with a jammed gas pedal; they increase cell division. Mutations that inactivate tumor suppressors are like cars with defective breaks; the cells divide uncontrollably.

2.2.1 Oncogenes

Oncogenes only require one of the parental alleles to be mutated, and the mutation typically leads to activation of the encoded protein, also known as gain-of-function mutations. Several dominant cancer genes such as *BRAF*, *EGFR*, *ERBB2*, *PIK3CA*, *IDH1*, *IDH2*, *KRAS*, and *JAK2* have been identified [130].

2.2.2 Tumor Suppressors

Tumor suppressors, on the other hand, require both of the parental alleles to be mutated, and the mutation normally leads to inactivation of the encoded protein providing a loss-of-function. One of the most well known tumor suppressors, *TP53*, is mutated in nearly all cancers, demonstrating its general importance to cancer. Other common tumor suppressors are *PTEN*, *STK11*, *AZI*, *SETD2*, *KDM6A*, *KDM5C*, *PBRM1*, *BAP1*, *ARID1A*, *DNMT3A*, *GATA3*, and *MLL2*

Cancer cells find many ways to acquire the abilities that they need to grow and survive in an otherwise well-regulated environment. Some mutations resulting in the activation of *BCL2*, inactivation of *APAF1* or *TP53*, and alterations to the PI3K-AKT signaling pathway, allow cells to avoid cell death (apoptosis) [105]. Other mutations can lead to over-amplification of cancer genes that promote tumor cell growth such as *EGFR* or disrupt the regulation of angiogenesis via *VEGF* [142].

2.3 RNA Editing

RNA editing can be broadly defined as any alteration to a specific nucleotide in an RNA sequence that is unexpected given the corresponding DNA template. The central dogma of biology as defined by Francis Crick in 1958 [32] states that genetic information is transferred from DNA to RNA through a process called transcription and from RNA to three-dimensional proteins through a process called translation. Since then, many additional post-transcriptional processes such as splicing, 5' capping, and 3' polyadenylation have been identified as necessary for the conversion of precursor RNA to mature RNA.

RNA editing is an epigenetic post-transcriptional event that results in the conversion (or substitution) of one type of nucleotide into another. These alterations to the RNA can lead to an increase in transcriptome diversity. The consequences of RNA editing are similar to those of other single nucleotide variants. They main functional effects include: 1) synonymous mutations resulting in a different codon for the same amino acid, 2) nonsynonymous mutations that lead to amino acid changes in the final protein and/or cause a read through of existing stop codons resulting in elongated proteins, 3) nonsense mutations that lead to premature stop codons and truncated proteins. In addition, RNA editing can affect other post-transcriptional processes such as splicing, translation, RNA degradation, RNA interference, and protection from transposons while RNA editing in the 5' and 3' Untranslated Regions (UTRs) and microRNAs (miRNAs) themselves can affect miRNA targeting and the overall stability of the mRNA [106].

RNA editing has been observed in plants, animals, and both vertebrate and invertebrate species and can occur in a variety of types of RNAs including messengerRNAs, transferRNAs, ribosomalRNAs, microRNAs and non-coding RNAs [43]. Although all 12 types of base substitutions are theoretically possible and have been observed in various studies, the underlying biological mechanisms for most of them are largely unknown. The most well-studied types of RNA editing, resulting in base substitution, are Adenosine-to-Inosine (A-to-I), Cytidine-to-Uridine (C-to-U), and Uridine-to-Cytidine (U-to-C). The mechanisms behind these types of RNA editing are believed to be deamination for A-to-I and C-to-U and amination for U-to-C [22, 43, 99, 121, 126].

2.3.1 A-to-I Editing Mechanism

The most common type of RNA editing is called A-to-I editing where an adenosine is converted into inosine by hydrolytic deamination at the C6 position of the purine ring (Figure 2.1) [43, 93, 111, 119, 141]. A-to-I substitutions are also known as A-to-G substitutions, because inosine preferentially base pairs with cytidine and is interpreted as guanine by the splicing and translational machineries. A-to-G editing is typically identified by comparing genomic DNA with cDNA. Since inosine can base pair with cytosine, it is replaced by guanosine during reverse transcription and Polymerase Chain Reaction (PCR) amplification during cDNA synthesis.

Previous studies attempting to characterize A-to-I editing discovered that instead of targeting specific nucleotide sequences, adenosine deaminases were largely directed toward double-stranded RNA (dsRNA) [34, 42, 43, 50]. Several studies showed

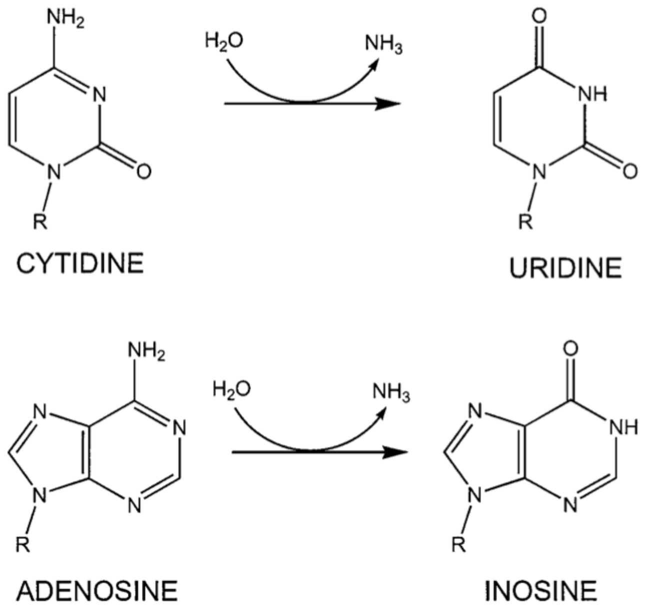


Figure 2.1: The mechanism of RNA editing for cytidine to uridine (C-to-U) and adenosine to inosine (A-to-I) is hydrolytic deamination. Figure from [43].

that either a perfect or imperfect inverted repeat in the RNA was necessary for most A-to-I editing as shown in Figure 2.2 [18, 34, 43, 50, 95, 120]. The repeat element is thought to form a duplex RNA structure that has single-stranded bulges and loops that are necessary for RNA editing site specificity. The dsRNA can be created in transcripts by base-pairing between exons and introns, especially with UTRs where there are many repetitive regions. Since A-to-I editing occurs in the introns of pre-mRNAs, researchers believe that A-to-I editing happens before splicing [43, 50, 119, 120] and can affect splicing.

An enzyme that had the necessary characteristics for A-to-I editing was biochemically purified, and the cloned cDNA showed that the enzyme (later named adenosine deaminase that acts on RNA 1 (ADAR1)) contained a nuclear localization signal (NLS),

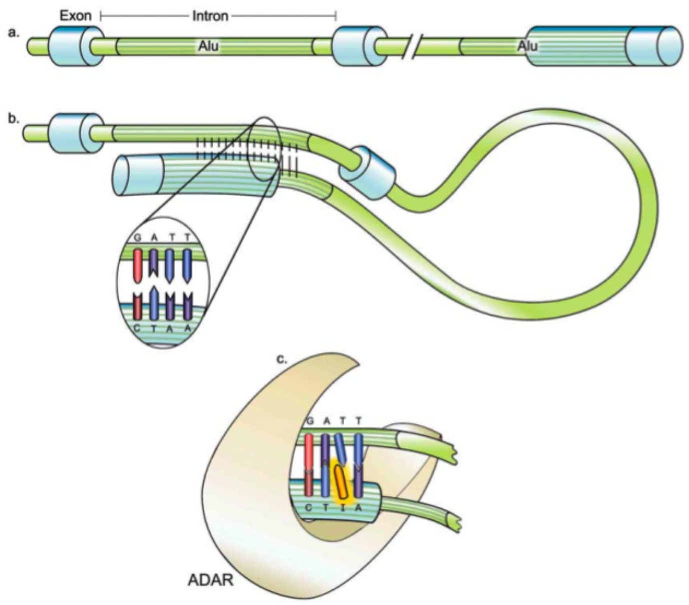


Figure 2.2: Schematic of A-to-I editing with adenosine deaminases acting on RNA (ADARs). a) A pre-mRNA containing two *Alu* repeats on opposite strands, one that overlaps with an exon. b) The two *Alu* repeats on opposite strands base pair with each other to form a dsRNA duplex. c) An ADAR enzyme deaminates some of the adenosines in the dsRNA duplex into inosines. Figure from [80].

dsRNA-binding motifs, and regions that were homologous to the catalytic domain of other known deaminases [43, 52, 73, 103, 104]. It was shown that ADAR1 was able to edit many specific sites but not all of the possible editing sites could be modified by this enzyme alone. Homology studies quickly identified two other enzymes with dsRNA-binding domains and a catalytic deaminase domain called ADAR2 and ADAR3. [43, 91, 92]. This family of enzymes called adenosine deaminases that act on RNA (ADARs) along with their main functional domains are shown in Figure 2.3.

The exact characteristics of the RNA that are needed for an adenosine to be selectively modified are still unknown. ADAR1 prefers to deaminate adenosines where

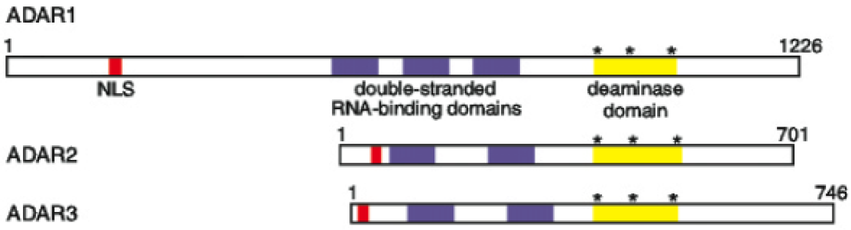


Figure 2.3: The size and main functional domains of the ADAR family members. The red boxes indicate the location of the nuclear localization signal (NLS), the purple boxes represent the double-stranded binding domains, the yellow boxes indicate the location of the deaminase domain. The asterisks represent conserved amino acids in all three domains that are thought to be critical for active-site zinc coordination. Figure from [43].

the 5' neighboring base is either A, U, or C. It also disfavors adenosines near the 3' end of the duplex [43, 109]. ADAR1 and ADAR2 have both overlapping and unique specificities [43, 92]. Using ADAR1 and ADAR2 knockout mice, researchers [115] were able to characterize sites that are targeted by only ADAR1 or ADAR2 and sites that are targeted to varying degrees by both. It was hypothesized that ADAR1 and ADAR2 may also have some regulatory affects on each other. Not much is known about specific sites that ADAR3 may target. ADAR3 is able to bind dsRNA, but it is distinct from ADAR1 and ADAR2 in that it is also able to bind single-stranded DNA [23]. It was also shown *in vitro* that ADAR3 was able to inhibit RNA editing by ADAR1 and ADAR2 suggesting that ADAR3 may also play a regulatory role in RNA editing [23].

2.3.2 A-to-I Editing in *Alu* sequences

Much of the early research on A-to-I editing focused on non-coding regions, especially within *Alu* elements. *Alu* elements are repetitive sequences roughly 300 bp in length and consist of two monomers that are joined by an A-rich region. They are found in various regions throughout the human genome, and 75% of known genes have them within their introns and UTRs [71]. Multiple A-to-I editing events were discovered in the repeat regions of introns and 3' UTRs in human brain RNAs [94] which prompted researchers to investigate RNA editing in repeat regions on a global scale [71]. An abundance of A-to-I editing within the embedded *Alu* sequences in transcripts of >30 tissue sources were found [71]. Experimental evidence for RNA editing in liver, lung, kidney, prostrate, and colon was shown for the first time. For most genes, the same editing was found in all of the tissues at a varying rate, and most often the unedited signal dominated the edited signal [80].

In 2009, a study identified A-to-I RNA editing sites that were not in repetitive regions on a genome-wide scale [85]. Until then, there were only 13 known genes that did not have obvious repetitive regions but still had evidence of RNA editing [85]. The study identified over 700 edited sites, with nearly 250 occurring in coding regions. Although they were not in repetitive regions, many of the gene sequences could form dsRNA structures suitable for A-to-I editing by the ADAR family.

The amount of A-to-I editing that occurs per transcript can vary greatly. Some transcripts may only have a single A-to-I event, while others can have more than 50% of the adenosines edited [43, 121]. It is hypothesized that there are at least two different underlying mechanisms behind A-to-I editing. Much of the focus has been on A-to-I hyperediting in areas where there are either perfect or imperfect repeats that can form double-stranded RNA (dsRNA). On the other hand, the exact mechanism for A-to-I editing in areas without such obvious repeats is yet to be discovered.

2.3.3 A-to-I Editing in Diseases

Proteins that are able to edit nucleotides must be very well regulated in order to avoid potentially dangerous scenarios. RNA editing is essential for normal brain function and normal central nervous system functions. Mice that lack the enzymes that are responsible for A-to-I editing are embryonic lethal, indicating that A-to-I editing is required for normal development. In mammals, most of the edited transcripts are expressed in the central nervous system (CNS) [39]. It is not known if the variety of proteins generated by editing are required by the CNS or if they are simply more tolerated since it is an immune-privileged system. Diseases of the CNS such as depression, epilepsy, schizophrenia and amyotrophic lateral sclerosis (ALS) have been reported to have a deregulation of RNA editing [39, 65, 89].

Although the editing process of ADARs is very well regulated *in vivo* [19, 21, 39, 78], it is unclear what the regulation factors are. It has been shown that ADARs can form homo- and hetero-dimers [39, 40, 112, 134] which has a significant impact on the editing activity [39, 40].

2.3.4 A-to-I Editing in Cancer

The relationship between RNA editing and cancer is just beginning to unravel. The most significant differences in editing between normal and tumor tissues until now have been found in brain and liver samples, but alterations in editing levels have also been shown for other types of cancer as well. Researchers are beginning to provide evidence that alterations to RNA editing in cancer-related genes may be relevant to the initiation and progression of cancer. An increase in the expression of oncogenes or a decrease in the expression of tumor suppressors due to RNA editing is emerging. An analysis of RNA editing for multiple cancers will help reveal the relationship between RNA editing and cancer. It may unveil possible drug targets or identify new diagnostic and prognostic markers that may lead to early detection, assistance in monitoring a patient's response to therapy, or aid in the detection of residual parts of the disease after treatment.

2.3.4.1 A-to-I Hypoediting in Brain Cancers

A global A-to-I hypoediting of *Alu* elements was discovered primarily in brain tumors but also in other tumor tissues including prostate, kidney, lung, and testis [106]. A correlation between a decrease in ADAR expression and the grade of tumor malignancy, where the expression levels were lowest in higher grade brain tumors was shown. This was the first strong evidence for a connection between the deregulation of ADARs and cancer progression. A reduction in expression levels for all three of the ADAR enzymes in brain tumors was found (Figure 2.4). This included a 99% decrease in ADAR2 expression, which is only expressed in the brain, for glioblastoma multiforme (GBM), the most aggressive brain tumor [106]. This finding suggests that the overall decrease in editing activity could be due to a decrease in ADAR expression [106]. In addition, a decrease in the proliferation of a GBM cell line due to the over-expression of the ADAR enzymes suggests that A-to-I editing in brain tumors is involved in the progression of cancer [106].

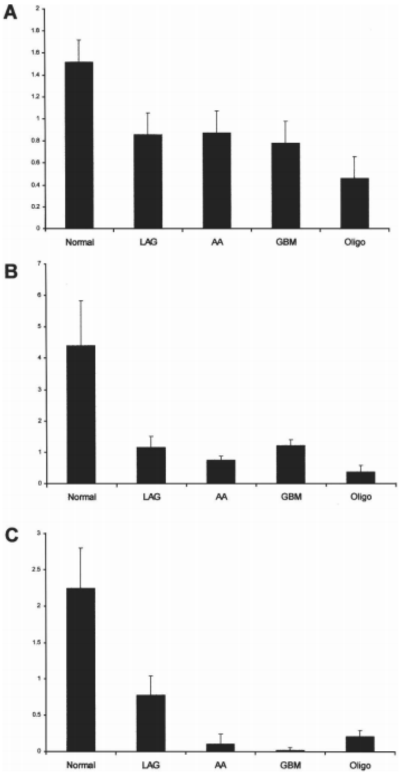


Figure 2.4: Reduced levels of mRNA expression of the ADAR family of enzymes in various brain cancers. (A) Reduction of ADAR1 expression, (B) Reduction of ADAR2 expression, (C) Reduction of ADAR3 expression. LAG=Low-Grade Astrocytomas, AA=anaplastic astrocytomas, GBM, glioblastomas multiforme, Oligo, oligodendrogliomas. Figure from [106].

Defects in ADAR2 activity have been directly linked to GBM [39, 49, 90]. In GBM patients, there is a decrease in RNA editing at the Q/R site of the *GluR-B* gene which is exclusively edited by ADAR2. Due to the fact that tumors show significant alterations in many biological pathways, it is still unknown if the decrease in editing is a consequence or cause of tumor progression. A previous study showed that the Q/R site of *GluR-B* is essential for suppressing migration *in vivo* and alterations may contribute to the aggressive growth of GBMs through the ATK pathway activation [39, 56, 57].

Although there is not much information about RNA editing in children, a correlation between a decrease in ADAR2 editing activity and the grade of tumor malignancy in pediatric astrocytomas has been shown [21, 39]. There was little to no ADAR2 editing activity detected in both astrocytoma tissues and cell lines, but the cell malignant behavior can be significantly reduced by restoring the normal ADAR2 activity [21, 39]. The elevation of ADAR1 expression levels appears to interfere with ADAR2 activity, suggesting that the balancing of ADAR activity is essential and could be at the origin of cell transformation [21, 39].

2.3.4.2 A-to-I Hypoediting and Hyperediting of Cancer Specific Genes

A more complex picture emerges when evaluating specific cancer-related genes [106]. Some genes (*BRCA1* and *BLCAP*) were found to be edited more in tumors vs. normal tissues while others (*MED13*, *FLNA* and *CYFIP2*) were edited less [106]. For *BRCA1*, a breast cancer tumor suppressor gene, editing in tumors was significantly higher than in normal samples [106]. For the *BLCAP* (bladder cancer-associated protein) gene, the trend was the same but less significant, with an average editing level in normal tissues of 16% while tumors had an average of 21% [106]. The *MED13* (thyroid hormone receptor associated protein 1) gene is known to be in a genomic location that is amplified in breast cancer, and the normal samples had higher average editing levels than the tumor tissues [106]. The *FLNA* (filamin A, alpha (actin binding protein 280)) gene down-regulates activity of the androgen receptor that is associated with prostate cancer, and normal brain samples had an average of 21.5% editing, while the tumor samples had an average of 8.5% editing [106]. The *CYFIP2* (cytoplasmic FMR1 interacting protein 2) gene is down-regulated by the well-known tumor suppressor *TP53*, and has an average editing level of 52% in normal brain tissues vs. 22% in brain tumors [106].

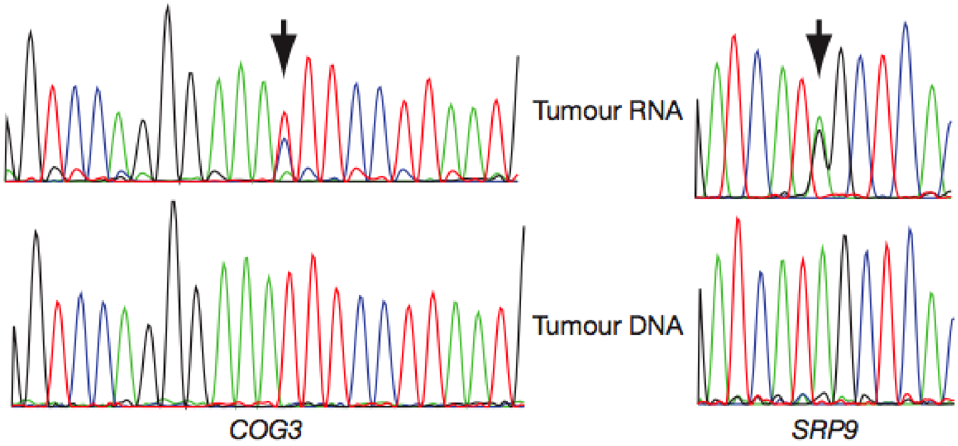


Figure 2.5: RNA editing in *COG3* and *SRP3*. Sanger sequencing chromatograms validating the RNA editing for each gene. The arrow marks the editing site. Figure from [125].

An important A-to-I editing event in Acute Myeloid Leukemia (AML) was identified in the protein tyrosine phosphatase *PTPN6* gene [14]. The *PTPN6* gene is recognized as a tumor suppressor gene and is important for the down-regulation of growth-promoting receptors. The A-to-I conversion causes the splicing mechanism to ignore a splice junction, leading to a non-functional PTPN6 protein via the inclusion of an intron in the mature RNA transcript. In addition, A-to-I editing at this position was lower in patients who were in remission [14]. The authors emphasized that this abnormal A-to-I editing that was only found in tumor samples could be at the origin of AML [14], but this is still an open question.

Two A-to-I editing events that alter the amino acid sequence in the *SRP3* and *COG3* genes were identified by next-generation sequencing of both the genome and transcriptome of an estrogen-receptor-α-positive metastatic lobular breast cancer [125]. *SRP3* and *COG3*, were highly edited and resulted in non-synonymous changes that created variant protein sequences as shown in Figure 2.5. In addition to identifying the specific editing events and sites, one of the ADAR enzymes was in the top 5% of genes that were expressed.

We are just beginning to gain insight on the role of A-to-I editing in cancer. An understanding of how A-to-I editing is regulated *in vivo* has begun, and it is important to determine the link between ADARs and cancer and other diseases. It has been shown that both ADAR1 and ADAR2 are co-expressed in a cell, and ADAR homo- and hetro-dimers are found in different cell types [39]. While it seems that there is a general decrease in ADAR activity for brain tumors, a complex pattern of hypo- and hyper-editing of individual genes is emerging in other cancers [21, 39, 106, 90]. This thesis will focus on hyperediting of the *AZIN1* gene in multiple cancers.

2.3.4.3 Hyperediting of *AZIN1* in Liver Cancer

Over-editing of the *AZIN1* gene in hepatocellular carcinoma (HCC) was associated with gain-of-function phenotypes such as tumor initiation, accelerated growth rates, and increased invasive capabilities both *in vitro* and *in vivo* [24]. The edited AZIN1 protein neutralizes antizyme mediated degradation of orinthine decarboxylase (ODC) and cyclin D1 (CCND1) leading to increased cell proliferation [24]. Until now, a thorough investigation of *AZIN1* editing in other cancers has not been performed.

2.3.4.4 Role of Cyclin D1 in Cancer

The retinoblastoma gene (*RB*) is a tumor suppressor gene and a key regulator of cell cycle progression [140]. The Rb protein is responsible for preventing excessive cell growth by inhibiting progression of the cell cycle from the G1 (first gap phase) to S (synthesis phase) phase until a cell is ready to divide [140]. Rb binds to and inhibits the E2F transcription factor family that can trigger a cell's entry into S phase [140]. When a cell is ready to divide, Rb is phosphorylated by cyclins and cyclin-dependent kinases (CDKs) [64, 88, 140]. The initial phosphorylation is done by the cyclin D-CDK4-CDK6 complex [64, 88, 140]. Rb is inactivate when it is phosphorylated, allowing E2F to activate other cyclins and CDKs to continue the cell cycle [64, 88, 140].

Given that many cancers occur due to errors in cell cycle regulation and the role of cyclin D in activating cell cycle progression by inactivating Rb, cyclin D is considered to be a possible oncogene. Uncontrolled production of cyclin D allows more cyclinD-CDK4-CDK6 complexes to form, driving cell cycle progression.

2.4 Conclusion

Identifying and characterizing germline variants, somatic mutations, and RNA editing events is important to cancer research in many ways. It provides critical information about cancer development and progression. It also determines potential cancer genes that can be used for drug development, diagnosis, and treatment. Many cancer genes that have been discovered thus far have been successfully targeted by anticancer drugs, making the identification of new cancer genes one of the most important contributions to cancer research. With the enormous costs involved in drug development, it is important to characterize cancer genes as thoroughly and accurately as possible. Characterizing both the DNA and RNA of a mutation is crucial in the identification of good candidate genes for drug targets. With the influx of high-throughput sequencing data for both the DNA and RNA across multiple cancers, it is now possible to characterize genetic variants, somatic mutations, and RNA editing events on a genome-wide scale.

Chapter 3

RADIA: A Method to Identify and Characterize SNVs in Cancer Using Both DNA and RNA

Somatic mutation calling is traditionally performed on patient-matched pairs of tumor and normal genomes/exomes [25, 44, 74, 75, 79, 123]. The ability to accurately detect somatic mutations is hindered by both biological and technical artifacts that make it difficult to obtain both high sensitivity and high specificity. Different mutation calling algorithms often disagree about putative mutations in the same source data, and frequently have discernible systematic differences due to the trade-off between sensitivity and specificity [116]. This is especially true for somatic mutations with low variant allele frequencies (VAFs). By creating an algorithm that utilizes both DNA and RNA, the power to detect somatic mutations is greater, especially at low variant allele frequencies.

RADIA combines patient-matched tumor and normal DNA with the tumor RNA to detect somatic mutations. The DNA Only Method (DOM) (Figure 3.1) uses just the tumor/normal pairs of DNA (ignoring the RNA), while the Triple BAM Method (TBM) (Figure 3.1) uses all three datasets from the same patient to detect somatic mutations. The mutations from the TBM are further categorized into two sub-groups: RNA Confirmation and RNA Rescue mutations (Figure A.1). RNA Confirmation mutations are those that are made by both the DOM and the TBM due to the strong variant read support in both the DNA and RNA. RNA Rescue mutations are those that had very little DNA support, hence not called by the DOM, but strong RNA support, and thus called by the TBM. RNA Rescue mutations are typically missed by traditional methods that only interrogate the DNA.

RADIA operates on two or more BAM files, producing somatic mutation calls through a series of steps outlined in Figure 3.1. Each step in this process is described in detail, beginning with the initial selection of sites for further processing and ending with a description of filters used to eliminate false positives while maintaining true positives.

3.1 Variant Detection with RADIA

RADIA is typically run on three BAM [84] files consisting of a pair of patient-matched tumor and normal genomes and a tumor transcriptome and outputs germline (inherited) variants, somatic mutations, and RNA editing events. The DOM is run on the pairs of tumor and matched-normal DNA while the TBM is applied to the DNA and RNA triplets. After the DOM and TBM specific filters, the results are merged and run through a final read support filter (Figure 3.1). If RNA-Seq data is not available, RADIA can utilize paired tumor and normal DNA genomes using the DOM to detect germline variants and somatic mutations.

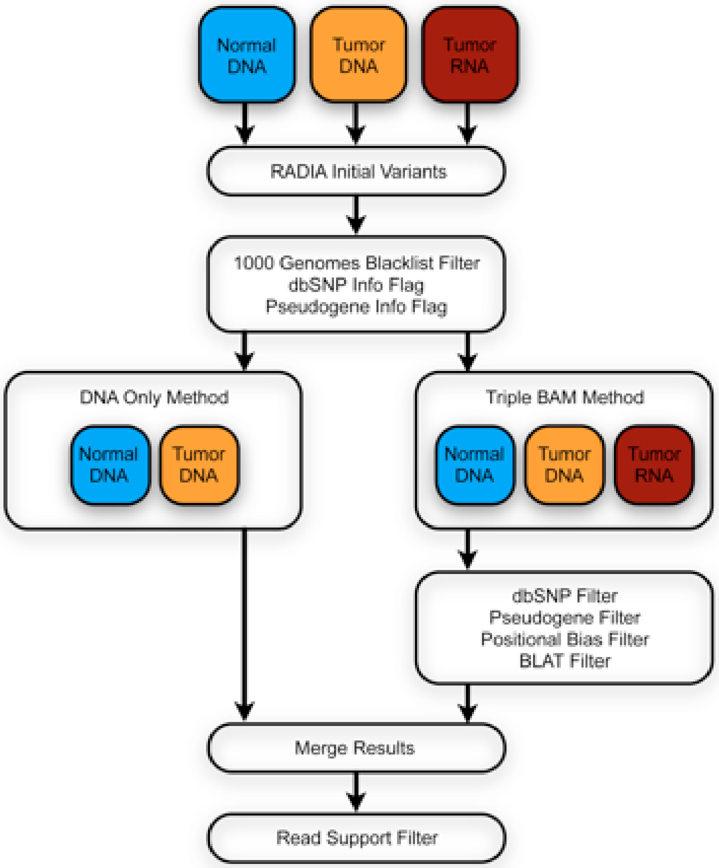


Figure 3.1: Overview of the RADIA work-flow for identifying SNVs. The normal DNA, tumor DNA, and tumor RNA BAMs are processed in parallel and initial low-level variants are identified. The variants are filtered by the DNA Only Method using the pairs of normal and tumor DNA and by the Triple BAM Method using all three datasets. The variants from the two methods are merged and output in VCF format.

Internally, RADIA uses the samtools [84] mpileup command (version 0.1.18) to examine the pileups of bases in each sample in parallel. A heuristic algorithm determines the existence and type of variant at any given position based on the user-configurable minimum thresholds for overall depth, variant depth, Base Alignment Quality (BAQ) [82], and mapping quality. Initially, RADIA requires a minimum overall depth of four bases, minimum variant depth of two bases, minimum phred BAQ of 10, and minimum phred mapping quality of 10. These initial calls are lenient in coverage and provide a good baseline set of calls for further filtering.

RADIA scans pileups of reads across the reference genome and outputs variants in Variant Call Format (VCF) (https://github.com/samtools/hts-specs). For each position, summary information such as the overall depth, allele specific depth and frequency, average BAQ base quality, average mapping quality, and the fraction of reads on the plus strand are calculated for both the DNA and RNA. All of this information is used during the filtering process.

3.2 Variant Filtering

After the initial variants are detected, a number of filters are applied to remove false positive variants that result from biological and technical artifacts. Each filter is described here in detail.

3.2.1 Filtering Around INDELs

Many current mutation calling algorithms have a pre-processing step to account for misaligned reads around INDELs. This realignment step is computationally expensive and relies on accurately predicting the location of INDELs which itself is not a trivial problem. Base Alignment Quality (BAQ) is an alternative option for dealing with alignment ambiguity around INDELs. It calculates the probability that a base has been misaligned and returns the minimum of the original base quality and the base alignment quality. BAQ is run by default when executing a samtools mpileup command and has been shown to improve SNP calling accuracy [82]. The extended version of BAQ (option –E) that is activated by default in the latest version of samtools (0.1.19) for increased sensitivity and slightly lower specificity is used [84].

3.2.2 1000 Genomes Blacklist Filter

The 1000 Genomes Project coined the term "accessible genome" to be the part of the reference genome that is reliable for accurate variant calling after removing ambiguous or highly repetitive regions [30]. Since the reference genome is incomplete, repetitive in places, and does not represent human genetic variation comprehensively, reads often get mapped incorrectly in locations outside the accessible genome (inaccessible sites), leading to false positive variant calls. Over 97% of inaccessible sites are due to high copy repeats or segmental duplications. In the pilot, the 1000 Genomes Project determined that 85% of the reference sequence and 93% of the coding region was accessible. Due to longer read lengths (75-100 bp) and improvements to both paired end protocols and sequence alignment algorithms, the accessible genome increased in Phase I to 94% of the reference and 98% of the coding region [31]. Variants that are not in the accessible genome using the Phase I mapping quality and depth blacklists (ftp://ftptrace.ncbi.nih.gov/1000genomes/ftp/phase1/analysis_results/supporting/accessible_genome_masks/) are filtered.

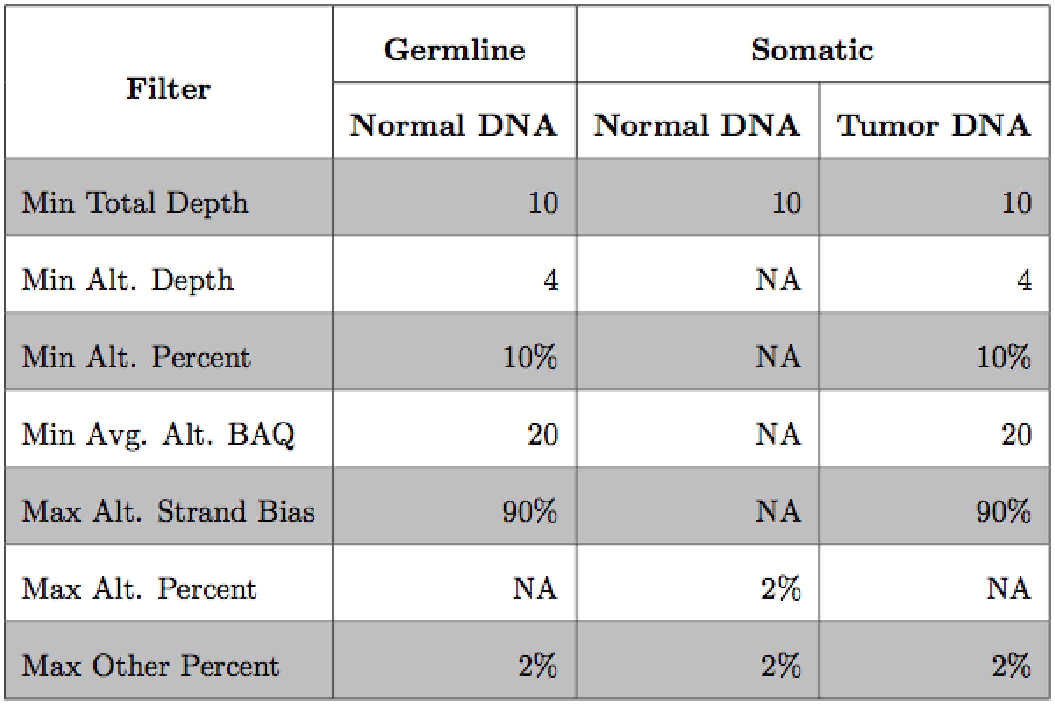
3.2.3 Strand-Bias Filter

It has been shown that variant allele reads that occur exclusively on one strand are largely associated with false positives [79]. In order to account for this technical artifact, RADIA filters based on the variant allele strand bias. If there are at least four total reads supporting the variant allele, then the strand bias filter is applied if more than 90% of the reads are on the forward strand or more than 90% are on the reverse strand.

3.2.4 Filtering by mpileup Support

RADIA can be executed on patient-matched pairs of tumor and normal DNA samples using the DOM to identify germline variants and somatic mutations. The matched normal DNA is first compared to the human reference genome. The normal DNA must pass the mpileup support filters described in Table 3.1 for all germline variants.

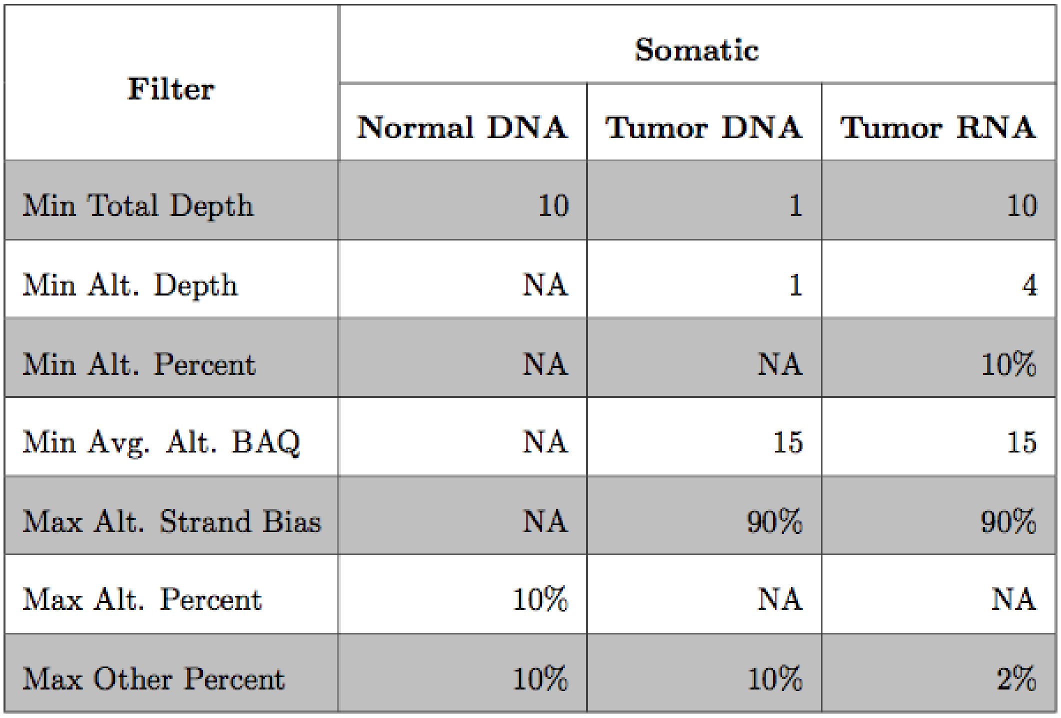
If no germline variant is found, the tumor DNA is compared to the matched normal DNA and the reference genome to search for somatic mutations. The normal DNA and tumor DNA must pass the mpileup support filters shown in Table 3.1 for all somatic variants. To ensure that there is adequate power to detect a possible germline variant at this site, the germline DNA depth must be 10 or more.



Filter	Germline	Somatic	
	Normal DNA	Normal DNA	Tumor DNA
Min Total Depth	10	10	10
Min Alt. Depth	4	NA	4
Min Alt. Percent	10%	NA	10%
Min Avg. Alt. BAQ	20	NA	20
Max Alt. Strand Bias	90%	NA	90%
Max Alt. Percent	NA	2%	NA
Max Other Percent	2%	2%	2%

Table 3.1: DNA Only Method (DOM) mpileup Support Filters. The germline variants and somatic mutations from the DOM are filtered according to the parameters described here. The minimum average alternative read BAQ filter uses the phred scale. The maximum other percent restricts the percentage of reads that are allowed to support an additional alternative allele.

The Triple BAM Method is used to augment the somatic mutation calls using both the pairs of DNA and the RNA-Seq data. The normal DNA, tumor DNA, and tumor RNA must pass the mpileup support filters shown in Table 3.2 for all somatic mutations. At least one read with a minimum BAQ phred score of 15 in the tumor DNA is required. To rule out possible germline variants, the normal DNA depth must again be 10 or more. In addition, calls that overlap with common SNPs that are not flagged as clinically relevant and found in at least one percent of the samples in dbSNP [127] are filtered out. This subset of dbSNP was downloaded from the “Common SNPs” track on the UCSC human genome browser [63, 68]. Many false positive variants overlapped with earlier versions of dbSNP. These variants were due to technical artifacts and were removed from subsequent versions of dbSNP [98]. Therefore, all variants that overlap with dbSNP versions 130, 132 or 135 (ftp://ftp.ncbi.nih.gov/snp/) are filtered out. The TBM calls are subjected to further filtering procedures as shown in Figure 3.1 and described below.



Filter	**Somatic**		
	Normal DNA	**Tumor DNA**	**Tumor RNA**
Min Total Depth	10	1	10
Min Alt. Depth	NA	1	4
Min Alt. Percent	NA	NA	10%
Min Avg. Alt. BAQ	NA	15	15
Max Alt. Strand Bias	NA	90%	90%
Max Alt. Percent	10%	NA	NA
Max Other Percent	10%	10%	2%

Table 3.2: Triple BAM Method (TBM) mpileup Support Filters. The somatic mutations from the TBM are filtered according to the parameters shown here. The minimum average alternative read BAQ filter uses the phred scale. The maximum other percent restricts the percentage of reads that are allowed to support an additional alternative allele.

3.2.5 Pseudogene Filter

An observation was made that many of the putative TBM mutations overlapped with predicted pseudogenes. Although expressed pseudogenes have recently been reported to be significant contributors to the transcriptional landscape and shown to play a role in cancer progression [60], mutations that overlap with predicted pseudogenes have a high false positive rate. Sequence similarity of pseudogene copies to their parent genes leads to uncertainty in alignment within these regions. Because of these technical artifacts, TBM mutations that overlap with pseudogenes annotated in GENCODE by the ENCODE project (version 19) [48] and predicted by RetroFinder (version 5) [48, 13] are removed. The pseudogene annotations were downloaded from the following tracks on the UCSC human genome browser [68, 118]: Gene Annotations from ENCODE/GENCODE and Retroposed Genes. The predicted pseudogenes occupy 1.5% of the total genome.

3.2.6 Highly Variable Genes Filter

TBM mutations that overlap with families of genes that have high sequence similarity are removed. Some examples of these gene families are Human Leukocyte Antigens (HLAs), Ribosomal Proteins (RPLs), and immunoglobulins. While mutations in these genes may exist, special processing would be needed to distinguish them from false positive calls due to misaligned reads. The mutations are annotated using SnpEff [26], and mutations landing in the following five gene families are removed: RPLs, RP11s, HLAs, IGHVs and IGHCs.

3.2.7 Positional Bias Filter

False positive calls are associated with misaligned reads where the alternative allele is consistently within a certain distance from the start or end of the read. The positional bias filter is applied when 95% or more of the reads that have an alternative allele are such that the alternate allele falls in the first third or last third of the read.

3.2.8 BLAT Filter

Multiple instances were observed where RNA-Seq reads appeared to be incorrectly mapped due to the added difficulties in aligning RNA-Seq data, such as dealing with hard to identify splice junctions and multiple gene isoforms. To guarantee that the RNA-Seq reads that support a variant do not map better to another location in the genome, a BLAT filter was created. All of the RNA-Seq reads that support a variant are extracted from the BAM file and aligned to the human genome using BLAT [67]. If the read maps to another location with a better score, the read is rejected. After using BLAT on each read, at least four valid reads that support the variant and a minimum of 10% or more of the reads that support the variant are required.

3.2.9 Read Support Filter

The calls from the DOM and the TBM are merged and one final filter is applied. Each somatic mutation must be supported by at least four “perfect” reads. A perfect read is defined as follows:

1. Minimum mapping quality of read is 10

2. Minimum base quality of alternative allele in read is 10
3. Minimum base qualities of the five bases up- and down-stream of the alternative allele are 10
4. Read is properly paired
5. Read has fewer than four mismatches across its entirety when compared to the reference
6. Read does not require an insertion or deletion to be mapped

After determining the number of perfect reads that support the reference and the alternative alleles at a coordinate, the strand bias filter is re-applied to guarantee that no more than 90% of the total perfect reads are from one strand.

Chapter 4

Somatic Mutations

RADIA has been applied to data derived from nearly 4,500 patients representing 22 different cancer types from TCGA (Table A.9). Overall, the RNA Rescue mutations that are made possible by the incorporation of the RNA-Seq data provide a two to seven percent increase in somatic mutations compared to the DOM (Table A.9). Many of these mutations were new discoveries that were not previously found by other mutation calling algorithms in TCGA. Of these new discoveries, some mutations were found in well-known cancer genes that were heavily mutated in a specific cohort. Mutations in new samples where the same gene had already been identified as harboring mutations in other samples from the cohort were also found. When these RNA Rescue mutations are added to the DNA Only mutations, these genes achieve a statistically significant overall mutation rate for the cohort.

The primary focus here is on results from 177 endometrial carcinoma [62] and 230 lung adenocarcinoma [11] patients from TCGA. To demonstrate the increase in sensitivity from including the tumor RNA-Seq dataset, artificial mutations were spiked into the tumor DNA and tumor RNA of a breast cancer patient using bamsurgeon (https://github.com/adamewing/bamsurgeon). Sensitivity and precision was evaluated on the endometrial carcinoma and lung adenocarcinoma data using validation data that was generated by TCGA. All patients in this study provided written informed consent to genomic studies in accordance with local Institutional Review Boards and the policies and guidelines outlined by the Ethics, Law and Policy Group from TCGA. All patient data is anonymous and was originally collected for routine therapeutic purposes. RNA Rescue mutations found by the TBM in tumor suppressor genes such as *TP53*, *STK11*, and *CDKN2A* in lung adenocarcinoma are highlighted.

4.0.10 Sensitivity on Simulation Data

In order to evaluate sensitivity and demonstrate the increase in power from including the RNA-Seq data, somatic mutations were simulated starting from patient data. Mutations were spiked into a pair of breast cancer tumor DNA and tumor RNA samples using bamsurgeon (https://github.com/adamewing/bamsurgeon), a tool developed to generate simulation data that closely mimics actual experimental data from high-throughput sequencing datasets. Bamsurgeon first determines the loci that have an appropriate DNA and RNA depth to spike in mutations. It then extracts the reads at the loci, adjusts the VAF according to the user-defined VAF distribution, and then re-maps the reads (Figure A.2). This simulation strategy is more sophisticated than simply generating simulated reads from a reference genome, as it retains the biological and technical artifacts that are inherently present in next generation sequencing data. Two spike-in experiments were performed: one varying the DNA VAF while holding the RNA VAF constant, and one varying the RNA VAF while holding the DNA VAF to 10% or less.

4.0.10.1 Sensitivity on Variable DNA-Constant RNA Simulation Data

To evaluate the sensitivity of RADIA, 1,594 mutations were spiked into the tumor DNA sequence with a variant allele frequency ranging from 1-50% and to the tumor RNA sequence at a constant frequency of 25%. The overall sensitivity rate averaged across all VAFs is 85% consisting of 1,351 out of 1,594 spiked in mutations (Figure 4.1A). Of the 243 calls that were filtered out, over 50% were removed because they failed to meet the minimum variant allele frequency, more than 20% landed in blacklist regions that the method ignored, and nearly 20% were discarded due to the BLAT filter. The number of mutations that were rejected by the full list of filters can be found in Figure A.3.

4.0.10.2 Sensitivity on Low Frequency DNA-Variable RNA Simulation Data

To demonstrate the ability of the TBM to rescue calls at low DNA VAFs, 1,761 mutations were spiked into the tumor RNA sequence with a variant allele frequency ranging from 1-50% and to the tumor DNA sequence at a frequency of 10% or less. Most of the mutations by the DOM were filtered out due to the low allelic frequency in the DNA (Figure A.4). For the mutations that had sufficient read support in the RNA,

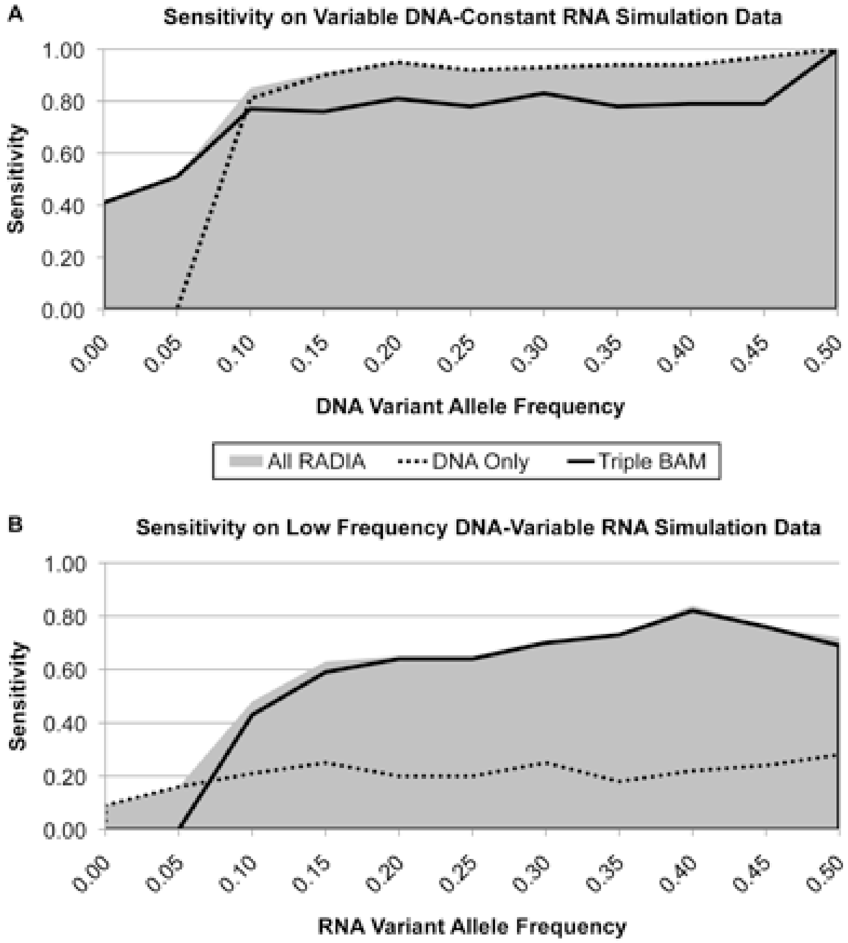


Figure 4.1: Sensitivity of RADIA on simulation data. Artificial mutations were spiked into the tumor DNA and RNA BAM files of a breast cancer patient using bamsurgeon. (A) Mutations were spiked into the DNA at variant allele frequencies distributed from 1-50% and into the RNA at a constant 25%. The overall sensitivity of RADIA was 85%. RNA Rescue calls from the Triple BAM method detected the mutations that had a DNA VAF less than 10%. (B) Mutations were spiked into the DNA at 10% or less and into the RNA distributed from 1-50%. Most of the DOM mutations were filtered due to the low DNA allelic frequency. The mutations that had adequate RNA read support were rescued back at these low DNA allelic frequencies.

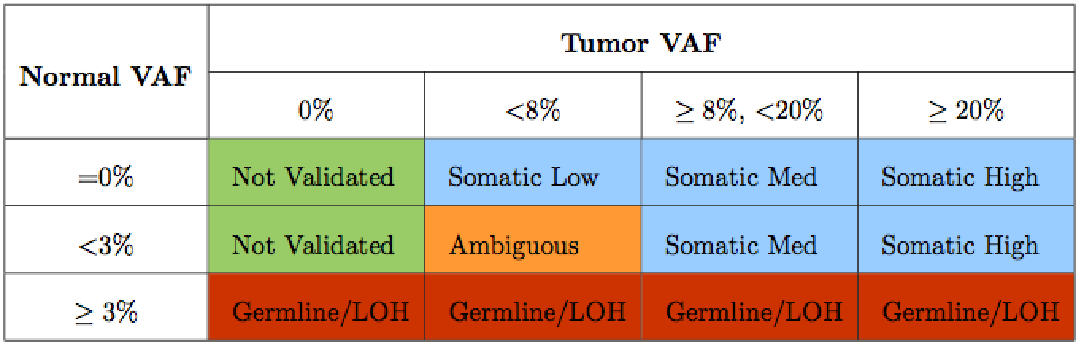
these low DNA VAFs were rescued back (Figure 4.1B).

4.0.11 Precision and Sensitivity on Patient Data

Somatic mutation calls were made on 177 non-hypermutated TCGA endometrial carcinoma samples [62]. All 177 tumor and matched normal whole exome sequencing and RNA-Seq alignments in BAM [84] format were downloaded from TCGA at the Cancer Genomics Hub (CGHub, https://cghub.ucsc.edu). The exomes were sequenced using the Illumina Genome Analyzer II, and the paired-end sequencing reads were aligned by BWA [83]. The RNA was sequenced using the Illumina Genome Analyzer II, and the single-end sequencing reads were aligned by MapSplice (V2) [136].

4.0.11.1 RADIA Precision on Endometrial Carcinoma Patient Data

For the study on endometrial carcinoma by TCGA [62], mutations were submitted by three independent TCGA Genomic Data Analysis Centers (GDACs). These mutations were merged and targeted for custom recapture and resequencing using new cDNA libraries from the tumor and normal DNA samples [62]. The validation BAMs containing the results of the hybrid capture and resequencing of targeted mutations were downloaded from CGHub (https://cghub.ucsc.edu). The identical validation criteria used by the TCGA Endometrial Analysis Working Group was utilized to validate the somatic mutations detected by RADIA [62]. For each somatic mutation, the patient-matched tumor and normal validation data was queried. At least 10 reads in both the tumor and normal data were required in order to determine if a call validated, otherwise it was classified as ambiguous. If the variant was present at low levels in both datasets, it was also classified as ambiguous. Otherwise, a mutation validated as germline/LOH, somatic, or neither according to Table 4.1. In addition, any RNA Rescue call in the "Not Validated" group that overlapped with a COSMIC (Catalogue of Somatic Mutations in Cancer) somatic mutation that was confirmed in another study was considered as validated.



Normal VAF	Tumor VAF			
	0%	<8%	≥ 8%, <20%	≥ 20%
=0%	Not Validated	Somatic Low	Somatic Med	Somatic High
<3%	Not Validated	Ambiguous	Somatic Med	Somatic High
≥ 3%	Germline/LOH	Germline/LOH	Germline/LOH	Germline/LOH

Table 4.1: Validation criteria for endometrial carcinoma data. Validation BAMs were used to determine the validation status for somatic mutations as shown here. A mutation is considered validated in the Somatic Low, Med, or High groups (blue), not validated in the Not Validated (green) and Germline/LOH groups (red), and Ambiguous when there was low read depth (<10 reads) or low VAFs in both the normal (<3%) and tumor (<8%) validation BAMs (orange).

A total of 27,900 somatic mutation calls over 177 endometrial samples were detected, of which the DOM and TBM made 27,390 and 6,325 calls respectively. Of the 6,325 TBM calls, there were 5,815 RNA Confirmation mutations that were made by both the DOM and TBM signifying high DNA and RNA support, and importantly, a total of 510 RNA Rescue mutations that were missed by the DOM.

Using the validation strategy described above, the overall precision for RADIA

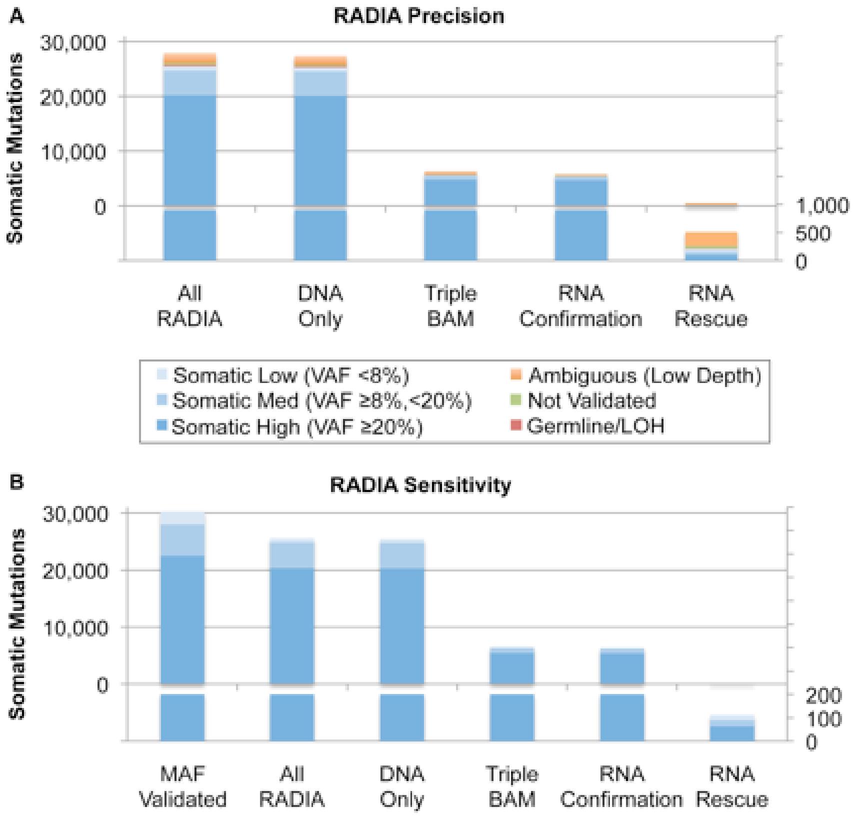


Figure 4.2: Precision and sensitivity of RADIA on 177 non-hypermutated endometrial carcinoma samples. Mutations were considered validated in the Somatic Low, Med, or High groups (blue), not validated in the Not Validated (green) and Germline/LOH (red) groups, and Ambiguous (orange) when there was low read depth (<10 reads) or ambiguity in the validation data. (A) An overall precision of 98% was demonstrated. RNA Confirmation mutations with strong DNA and RNA support validated over 99%. RNA Rescue mutations validated at 74%. (B) The union of all mutations submitted by TCGA GDACs that validated as somatic was considered as the truth set. RADIA demonstrated an overall sensitivity rate of 84%. Of the mutations that were missed, 33% occurred at low variant allele frequencies (<8%) and 23% occurred in blacklist regions that were ignored.

is 98% (Figure 4.2A). Due to lack of coverage or uncertainty in the tumor and normal validation BAMs, a total of 1,825 calls were considered to be ambiguous. Of the remaining 26,075 mutations called by RADIA, 25,520 validated as somatic, 271 validated as germline/LOH variants and 284 did not validate. The precision of calls made by the DOM and the TBM was 98% and 98.5% respectively. For the RNA Confirmation mutations made by both the DOM and the TBM, the precision was 99.3%. There were 510 RNA Rescue mutations made only by the TBM, and even though most of these calls were not targeted for validation, the precision was 74%. For the 510 RNA Rescue calls, 251 were classified as ambiguous, 6 validated as Germline/LOH, and 61 did not validate. Of the remaining 192 RNA Rescue mutations that validated, 178 (93%) were verified using the validation BAMs and 14 (7%) were confirmed as somatic mutations in COSMIC.

The precision of the DOM with varying RNA-Seq reads supporting the variant allele was examined as well as the precision of RNA Rescue mutations with differing levels of DNA supporting reads. Sixty-two percent of the DOM mutations were covered by reads in the RNA-Seq data, and 29% had at least 10 RNA-Seq reads covering the mutation. Nearly half (44%) had at least one RNA read supporting the DNA variant allele, while 25% of the DOM mutations had at least four supporting RNA reads. The precision of the DOM is lowest (92%) with no RNA-Seq support, increases to 95% with weak RNA-Seq support (at least one but less than five supporting reads), and increases to 99.3% for RNA Confirmation mutations. Overall, mutations that are detected by the DOM validate above 92%, regardless of the RNA-Seq support, and the precision increases as the RNA-Seq support increases.

On the other hand, RNA Rescue mutations weakly supported by the DNA validate at low levels. For RNA Rescue mutations, at least one variant supporting read in the DNA is required in order to distinguish between RNA Rescue mutations and possible RNA editing events. The precision of RNA Rescue mutations with only one read supporting the variant in the DNA was 11%, with two supporting reads in the DNA 23%, with three supporting reads in the DNA 43%, and with four or more supporting reads in the DNA 94%.

4.0.11.2 RADIA Sensitivity on Endometrial Carcinoma Patient Data

In order to measure the sensitivity of RADIA, the union of all mutations submitted by TCGA GDACs that validated as somatic were considered as our truth set. There were 30,239 mutations that validated as somatic from TCGA. A comparison of RADIA somatic mutations and this truth set demonstrated an overall sensitivity of 84% (Figure 4.2B, Figure A.5). Of the 4,751 calls that were missed, 1,539 (33%) were filtered by RADIA because they had a variant allele frequency less than 8% (Figure A.6). In addition, 1,072 (23%) landed in blacklist regions that were not considered (Figure A.6).

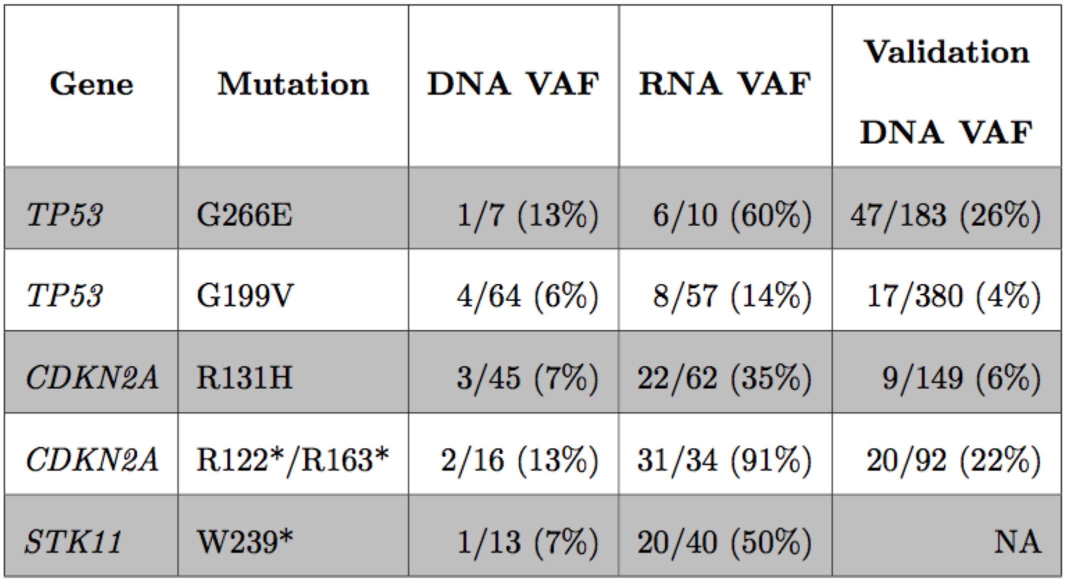
4.0.11.3 RADIA Precision on Lung Adenocarcinoma Patient Data

Finally, RADIA somatic mutations were analyzed during the course of our group's participation in the TCGA Lung Adenocarcinoma Analysis Working Group [11]. RADIA was executed on 230 TCGA lung adenocarcinoma triplets that were downloaded from CGHub (https://cghub.ucsc.edu). The exomes were sequenced using the Illumina HiSeq platform, and the paired-end sequencing reads were aligned by BWA [83]. The RNA was sequenced using the Illumina HiSeq platform, and the paired-end sequencing reads were aligned by MapSplice (V2) [136]. Validation was performed by the Broad Institute on 74 genes of interest along with an additional 1,150 somatic SNVs. Validation was attempted on 2,404 RADIA somatic mutations and 2,395 (99.63%) were verified. From the DOM, 2,336 of the 2,345 mutations (99.62%) validated. Importantly, 469/469 (100%) of the TBM mutations consisting of 410 RNA Confirmation and 59 RNA Rescue mutations validated.

4.0.12 Somatic Mutations in Specific Lung Adenocarcinoma Genes

Mutations in the tumor suppressor gene *TP53* are common in the majority of human cancers. Most of the mutations occur in the DNA-Binding Domain (DBD) and are considered change-of-function mutations that alter activity of *TP53*, sometimes acting in a dominant negative manner to sequester wildtype tp53 protein *in trans* [38]. As such, many p53 mutant proteins endow cells with oncogenic characteristics by promoting cell proliferation, survival, and metastasis [96].

RADIA was executed on the 230 TCGA lung adenocarcinoma triplets [11] and two non-synonymous *TP53* mutations that were below the detection threshold for other mutation calling algorithms used by TCGA were discovered (Table 4.2). Both of the mutations were validated by the deep-sequencing validation data and confirmed as somatic in COSMIC by other studies. One of the mutations (G266E) was confirmed



Gene	Mutation	DNA VAF	RNA VAF	Validation DNA VAF
TP53	G266E	1/7 (13%)	6/10 (60%)	47/183 (26%)
TP53	G199V	4/64 (6%)	8/57 (14%)	17/380 (4%)
CDKN2A	R131H	3/45 (7%)	22/62 (35%)	9/149 (6%)
CDKN2A	R122*/R163*	2/16 (13%)	31/34 (91%)	20/92 (22%)
STK11	W239*	1/13 (7%)	20/40 (50%)	NA

Table 4.2: RNA Rescue mutations in lung adenocarcinoma not detected by other methods in TCGA. These mutations were below the detection threshold for other mutation calling algorithms used by TCGA. The ratio of reads supporting the mutations along with the variant allele frequencies (VAFs) are shown for both the DNA and RNA. Validation was done on four of the mutations, and the resulting validation DNA variant allele frequencies are shown.

as somatic in another lung cancer study [61] as well as in prostate [86], pancreas [16], urinary tract [45], and hematopoietic and lymphoid [1] cancer studies. The G266E mutation occurs in the *TP53* DBD mutation hotspot frequently resulting in pathological effects [3, 37, 108]. This mutation has also been described as a gain-of-function mutation in a melanoma cell line [41]. The other *TP53* mutation (G199V) was confirmed as somatic in breast [9], ovarian [59], and medulloblastoma [117] studies. It is a known anti-apoptotic gain-of-function mutation that promotes cell survival through the Signal Transducer and Activator of Transcription-3 (STAT3) pathway [72]. Knockdown experiments of G199V p53 mutants demonstrated a level of anti-tumor activity similar to high doses of chemotherapeutic agents, suggesting that inhibition of G199V p53 mutants may be beneficial for cancer treatment [72].

Additionally, mutations were found in other well-known tumor suppressor genes such as *STK11* and *CDKN2A*. In the lung adenocarcinoma manuscript from TCGA, mutations in *STK11* and *CDKN2A* were reported in 17% and 4% of all patients, respectively [11]. *STK11* was the fourth most mutated gene and *CDKN2A* was the sixteenth [11]. The proximal-proliferative subtype in lung adenocarcinoma is characterized by an enrichment of mutations in *KRAS* along with inactivation mutations in *STK11* [11]. In the *STK11* gene, a nonsense mutation was discovered at W239* in the structurally conserved protein kinase domain that was below the detection threshold for other mutation algorithms used by TCGA. This mutation introduces an early stop codon in exon five (of ten) leading to a truncated protein. This site is in COSMIC and was previously reported to be part of a 398 nucleotide deletion in a lung cancer study [33].

In the *CDKN2A* gene, one nonsense mutation was found at R122*, R163* and one missense mutation was found at R131H, R80H. Both mutations were validated by TCGA and found in COSMIC. *CDKN2A* is silenced in many CpG island methylator phenotype-high (CIMP-High) tumors by DNA methylation [11], but mutations and deletions in *CDKN2A* also result in loss of function. The nonsense mutation at R122*, R163* results in an early stop codon in exon two (of three or four, isoform dependent) leading to a truncated protein. Previous lung cancer studies [6, 17, 53] have reported frameshifts and deletions at this site. The missense mutation at R131H was also found in colon cancer [8], clear cell sarcoma [132], and chronic myeloid leukemia [100] and confirmed as somatic in biliary tract cancer [133].

4.1 Discussion

Identifying single nucleotide variants is a key step in characterizing the cancer genome. Until now, algorithms for SNV detection have concentrated on comparing just the normal and tumor genomes within the same individual. In the past few years, it has become common to also sequence the tumor transcriptome using RNA-Seq technologies. Large genomics studies, such as those conducted by TCGA, primarily use the RNA-Seq data for gene expression, gene fusion, and splicing analyses. With the cost of sequencing steadily decreasing and the wealth of information that can by obtained from RNA-Seq data, the sequencing of the tumor RNA will continue to be routine in large cancer profiling projects. A novel method called RADIA that combines the normal DNA, tumor DNA, and tumor RNA from the same individual has been developed to increase sensitivity when detecting somatic mutations without compromising specificity. The primary focus so far has been on the ability of RADIA to detect germline variants, somatic single nucleotide variants, and RNA editing events. In the future, other classes of somatic mutations such as small insertions and deletions (INDELs) and loss of heterozygosity events (LOHs) will be included.

The accurate detection of SNVs is complicated by biological and technical artifacts such as tumor purity and subclonality, varying allele frequencies, sequencing depths, and copy-number variation. There is a trade-off between high sensitivity and high specificity, such that it is difficult to achieve both. By including an additional dataset, there is an increase in the ability to reliably detect mutations, especially at low variant allele frequencies (Figure A.7) where the signal to noise ratio becomes unfavorable.

Many widely used mutation calling algorithms see a large decrease in precision as the DNA variant allele frequency declines [25, 74, 75, 79, 116]. For RADIA, a DNA VAF of 10% provides the best balance between sensitivity and precision. To demonstrate this point, the DNA VAF was lowered to 5% and RADIA was rerun on the endometrial carcinoma data from Section 4.0.11. The same validation strategy as described in Section 4.0.11 was utilized and the results were compared to the ones with a DNA VAF of 10%. A slight 1% increase in overall sensitivity from 84% (at 10% VAF) to 85% (at 5% VAF) was gained but an 8% decrease in overall precision from 97% (at 10% VAF) to 89% (at 5% VAF) was lost.

By combining the RNA with the DNA, the expression of a mutation can be confirmed, providing insight into its likely functional effect. Confirming mutations through RNA-Seq is also advantageous for large genomic studies in providing a means for weak validation for mutations without costly resequencing for validation (Figure A.8). Over 99% of mutations that have both strong DNA and RNA support validate upon resequencing, suggesting that if one is not using mutations in clinical practice but rather estimating overall frequencies of specific mutations in a research cohort, the extreme expense in validating every mutation may not be warranted. While the integration of RNA and DNA provides an important but limited use as a DNA variant validation technique, studying the impacts on gene expression levels may lead to a deeper understanding of the functional impact of DNA-originating variants.

Some of the strengths of RADIA have been outlined here, but approaches that use RNA-Seq for detecting variants have clear limitations [27, 76]. Only expressed alleles can be evaluated, which reduces the number of genes that can be assessed. In addition, several classes of mutations, such as the introduction of premature stop codons that lead to nonsense mediated decay, cannot be verified. Expression levels can also confound the ability to detect an imbalance in the genomic VAF as influences due to feedback control to rebalance gene dosage are currently unknown.

RADIA is able to detect mutations in important cancer genes such as *TP53* that were previously not identified by other algorithms because the signal was lost in the noise. Somatic mutations are commonly used to group patients into subtypes that are critical for diagnosis and treatment of the disease. The ability to rescue back mutations for individual patients will assist in correctly identifying each patient's specific subtype and consequently their treatment options.

Chapter 5

RNA Editing

RNA editing of the *AZIN1* gene is a driver in the pathogenesis of hepatocellular carcinoma (HCC) and may be a potential driver for other human cancers as well. Over-editing of the *AZIN1* gene leads to the overexpression of cyclin D1 protein and an increase in cell proliferation. An investigation of *AZIN1* RNA editing in data collected from nearly 5,000 TCGA patients across 12 cancers has been performed. A particular focus on the luminal subtypes of breast cancer, known for overexpression of cyclin D1 is given. Increased editing of *AZIN1* appears to be an alternative to cyclin D1 gene amplification for increased cyclin D1 protein expression in the breast cancer luminal subtypes. In 44 samples with high cyclin D1 protein levels, devoid of cyclin D1 amplification, 19 (30%) were over-edited. Over-editing was significant in 10 out of the 12 cancers studied. Higher editing frequencies significantly correlated with clinical data such as larger tumor sizes, lymph node involvement, presence of metastases, and higher tumor grades. They were also associated with subtypes that often have the worst prognosis. In addition, over-editing in many cancers correlated with a poor overall and recurrence free survival. These results establish that increased *AZIN1* editing is a general mechanism for promoting cell proliferation in human cancers.

5.1 Introduction

An introduction to the fundamentals of RNA editing is given, along with specifics about the tumorigenic consequences of RNA editing in the *AZIN1* gene. The importance of cyclin D1 protein overexpression in cancer is illuminated, and a possible mechanism for this overexpression is provided. A computational investigation of this mechanism on data collected from nearly 5,000 patients across 12 different cancers is performed.

5.1.1 RNA Editing

RNA editing is a post-transcriptional modification of the precursor mRNA and microRNA. The most common type of RNA editing in mammals is A-to-I editing where an adenosine is deaminated into an inosine by the ADAR (adenosine deaminases acting on RNA) family of enzymes [77]. Since inosine preferentially base pairs with cytidine, it is interpreted as guanine during protein synthesis and reverse transcription. The ADAR family of enzymes binds to double-stranded RNA that occurs when single-stranded RNA folds back onto itself through perfect or imperfect base-pairing [135]. RNA editing in different regions of the gene can have various functional effects: non-synonymous protein coding substitutions, alternative splicing by either the creation of new splice sites or the read through of existing splice sites, an influence on translation from RNA editing in 5'UTRs, and an alteration to RNA stability, transport, and regulation from RNA editing in 3'UTRs and microRNAs [101]. RNA editing of some adenosine sites is highly selective while other sites are constitutively edited often occurring in clusters of edits [102, 110].

In mammals, ADAR1 is ubiquitously expressed, ADAR2 is primarily active in the brain, and ADAR3 is exclusively found in the brain. Well known functional consequences of A-to-I editing include amino acid substitutions in the glutamate and serotonin receptors [20, 124], mRNA retention in the nucleus [113], changes in RNA stability [2], modifications to both microRNA and microRNA target sequences in 3'UTRs [87], heterochromatin formation [138], and protection against viral RNA [144]. Although the most well-studied A-to-I editing sites are in coding sequences that lead to amino acid changes, the majority of known A-to-I editing sites occur within *Alu* elements primarily located in introns and 3'UTRs [7, 70, 81].

ADAR knock-out mice are embryonic lethal, suggesting that A-to-I editing is required for normal development [137]. A-to-I editing is essential for normal brain function and normal central nervous system function. Abnormal RNA editing has been linked to many diseases such as epilepsy, amyotrophic lateral sclerosis (ALS), schizophrenia, depression, inflammation, and cancer. There have been isolated cases of RNA editing events reported in oncogenes and tumor suppressor genes [125]. In Acute Myeloid Leukemia (AML), an A-to-I editing event was found in the branch site of an intron between the third and fourth exons in the *PTPN6* tumor suppressor gene causing the splicing machinery to ignore the splice site leading to an elongated, non-functional

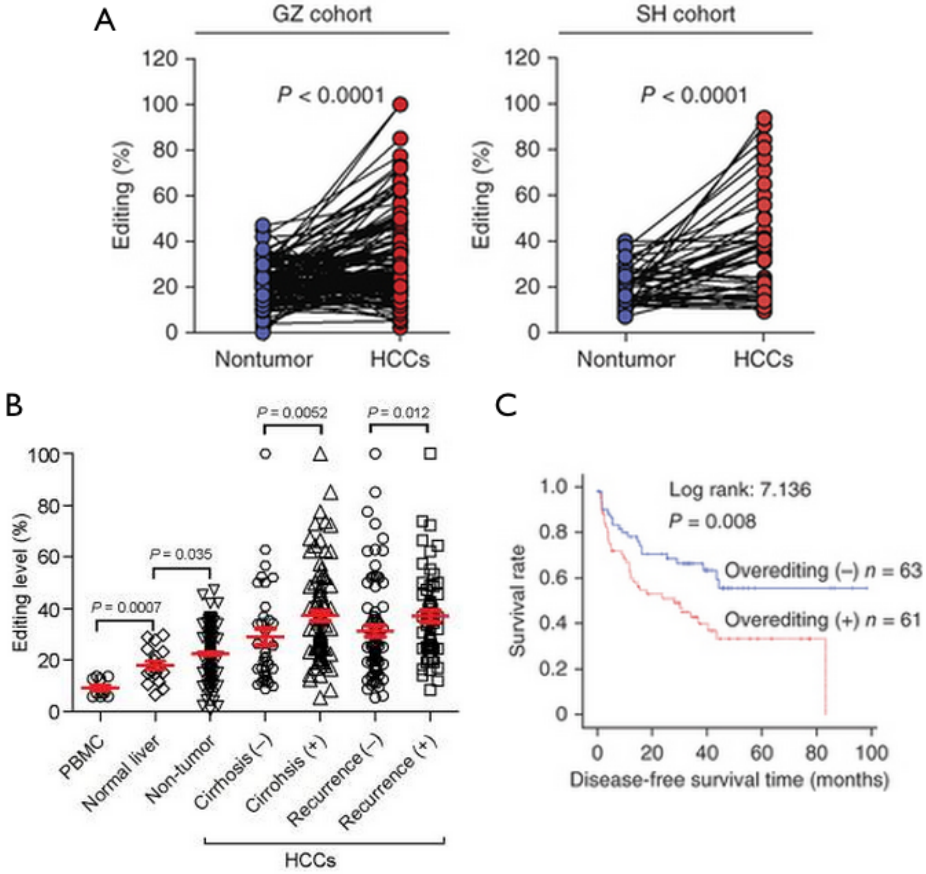


Figure 5.1: *AZIN1* editing is associated with HCC pathogenesis. (A) *AZIN1* editing differences between normal and tumor RNA from 135 and 46 HCC patients from the Guangzhou (GZ) and Shanghai (SH) cohorts, respectively (paired Students *t*-test). (B) A dot plot of *AZIN1* editing levels in PBMCs (n=10), healthy liver tissues (n=20), and adjacent normal liver samples (n=135). Tumor samples were divided into groups according to the presence or absence of cirrhosis and tumor recurrence. (C) Kaplan-Meier plot for disease free survival of HCC patients with or without over-editing. Figure from [24].

PTPN6 protein [15]. In brain cancer, A-to-I hypo-editing and a correlation between a decrease in ADAR1 and ADAR2 expression levels with the grade of tumor malignancy were discovered [107]. A 99% decrease in ADAR2 expression in glioblastoma multiforne (GBM) tumors was found compared to normal expression levels [107]. Recently, an A-to-I editing event in the antizyme inhibitor 1 (*AZIN1*) gene has been linked to hepatocellular carcinoma tumor initiation and development [24]. This thesis focuses on the A-to-I RNA editing modification reported in the *AZIN1* gene that causes an amino acid change in the final protein and the implications of this editing event in other cancers.

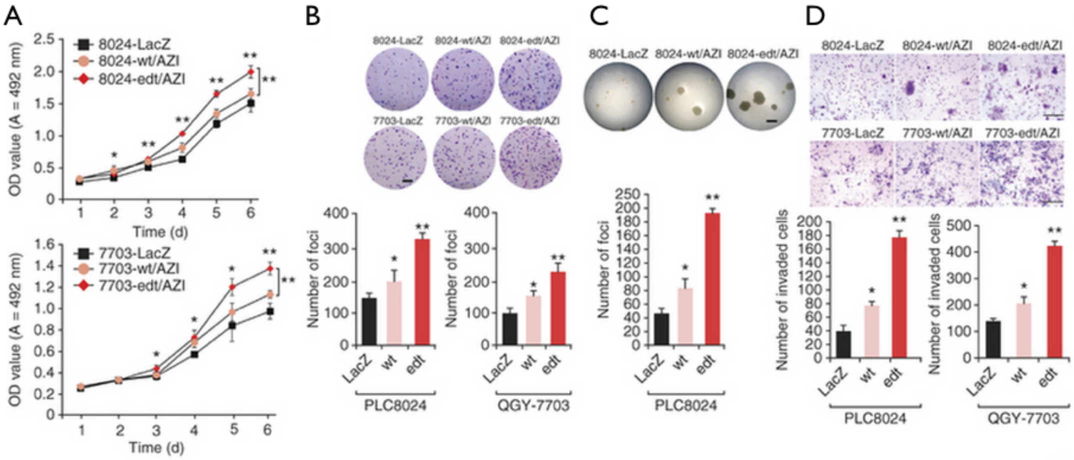
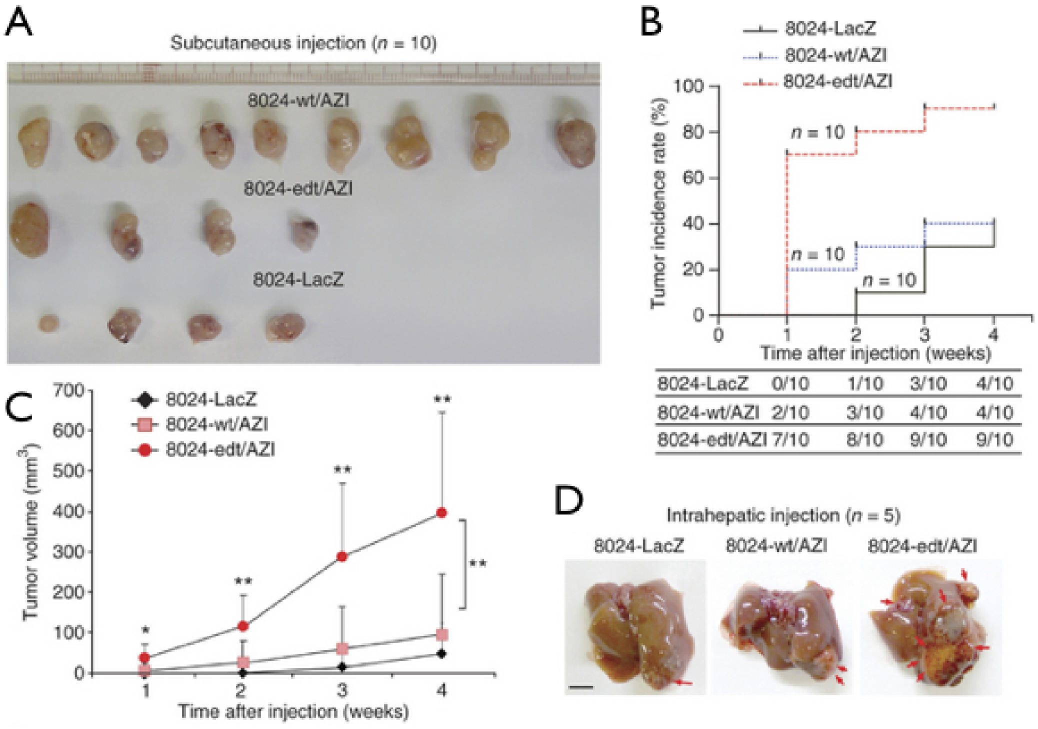


Figure 5.2: *AZIN1* RNA editing causes enhanced tumorigenicity in liver cancer cell lines. The growth rates of liver cancer cell lines with the edited form of *AZIN1* and the wild-type (A). Quantification of foci formation (B). Quantification of soft agar colonies induced by cell lines (C). Quantification of cells that invaded through Matrigel-coated membrane (D). Figure from [24].

5.1.2 A-to-I Editing of *AZIN1* in Liver Cancer

An A-to-I RNA editing event at chr8:103,841,636 in the *AZIN1* gene that leads to a Ser->Gly amino acid change (S367G) was identified in two separate hepa-



8024-LacZ	0/10	1/10	3/10	4/10
8024-wt/AZI	2/10	3/10	4/10	4/10
8024-edt/AZI	7/10	8/10	9/10	9/10

Figure 5.3: *AZIN1* RNA editing contributes to augmented tumor-initiating potential and enhanced *in vivo* tumorigenic ability. Xenograft studies in mice showed that the incidence of tumors from *AZIN1* edited cells was higher than wild-type or control cells (A). Seventy percent of mice injected with edt/AZI cells formed tumors within 1 week while 10% of mice injected with wt/AZI cells formed tumors after 4 weeks (A-B). Tumors induced by edt/AZI cells grew significantly faster (C). Mice injected with edt/AZI cells formed more liver nodules (D). Figure from [24].

tocellular carcinoma (HCC) cohorts [24]. Approximately 50% of the patients in each cohort had *AZIN1* over-editing which was defined as an increase of at least 10% (Figure 5.1A) [24]. There is a close link between *AZIN1* over-editing and HCC pathogenesis. The mean editing frequencies (EFs) were lowest in peripheral blood mononuclear cells (PBMCs) and normal liver samples and slightly increased in adjacent non-tumor tissues (Figure 5.1B). The EFs were significantly higher in HCC patients and highest in HCC patients with cirrhosis or recurrent tumors (Figure 5.1B). In addition, patients with over-editing had a significantly worse disease free survival (Figure 5.1C). ADAR1 is responsible for *AZIN1* RNA editing in human cancers [24]. Cell lines with expression constructs of the edited version of *AZIN1* showed accelerated growth rates (Figure 5.2A), higher frequencies of focus and colony formation (Figure 5.2B-C) and increased invasive capability (Figure 5.2D) compared to the wild-type [24]. Xenograft studies showed that mice that were injected with the edited form of *AZIN1* grew more tumors than the wild-type or control (Figure 5.3A-B). In addition, the tumors grew significantly faster and were larger compared to the wild-type (Figure 5.3A-D) [24].

Antizyme (AZ1 or OAZ1) is a tumor suppressor that regulates cell growth by binding to and inducing degradation of growth-promoting proteins such as ornithine decarboxylase (ODC) and cyclin D1 (CCND1) [24]. The antizyme-ODC/antizyme-CCND1 interaction leads to a conformational change of ODC/CCND1, inducing an ubiquitin-independent proteasomal degradation of ODC/CCND1 [24]. AZIN1 is a homolog to ODC that binds to antizyme with greater affinity than ODC, and the edited form of AZIN1 has an even higher binding affinity to antizyme than the wild-type

[24]. AZIN1 sequesters antizyme and prevents the degradation of ODC and CCND1, thus stimulating Retinoblastoma (Rb) phosphorylation and cell-cycle entry, leading to increased cell proliferation (Figure 5.4) [24].

5.1.3 Role of Cyclin D1 in Cancer

Several oncogenic activities due to the overexpression of cyclin D1 such as a decrease in the duration of the G1 phase, perturbation of the expression of other cellular growth-related genes including c-myc, c-jun, and cyclin A, and abnormalities in growth control have been described [58, 97]. The overexpression of cyclin D1 has been associated with aggressive tumorigenic characteristics such as a greater extent of lymph node involvement, metastasis, and a poorer prognosis for many cancers [69].

Overexpression of cyclin D can occur in many ways: gene amplification, chromosomal translocation, or impaired protein degradation. The human cyclin D1 gene is located on chromosome 11q13, an area known for DNA amplification and rearrangement in many human cancers including esophageal, breast, lung, larynx, head and neck, thyroid, bladder, and hepatocellular carcinoma (HCC) [129, 143]. However, for many tumors, cyclin D1 protein overexpression occurs even in cases where no CCND1 amplification or rearrangement is present [35, 69, 129, 143].

An investigation of RNA editing in the *AZIN1* gene was performed on data collected from a total of 12 cancers and nearly 5,000 patients by TCGA. Significant over-editing was evaluated for each cancer and correlations with clinical features and overall and recurrence free survival were examined. Significant associations with clinical

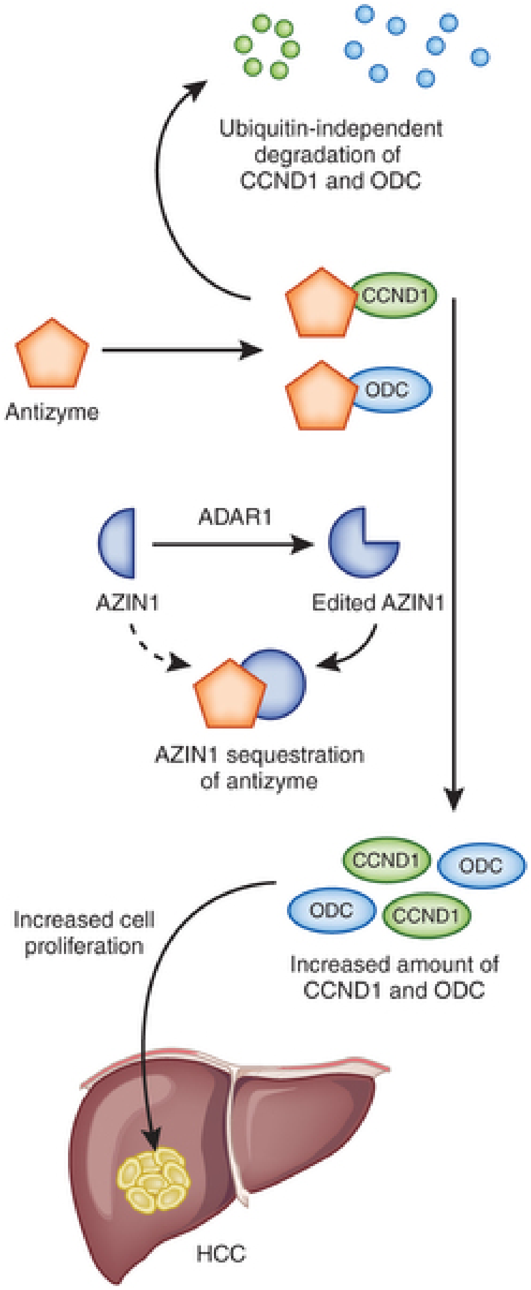


Figure 5.4: Edited *AZIN1* sequesters the tumor suppressor, antizyme, preventing the proteasomal degradation of ODC and CCND1, leading to increased cell proliferation. Figure from [24].

data such as the tumor size, lymph node involvement, presence of metastases, tumor stage and tumor grade were calculated. The tumor EFs also correlated with cancer specific subtypes.

5.1.4 A-to-I Editing of *AZIN1* in Breast Cancer

Cylin D1 amplification and high protein expression are common oncogenic events in breast cancer, especially in the Estrogen Receptor (ER) positive (i.e. luminal) subtype, most-notably in the luminal B subtype [9, 35, 51]. Patients with the basal subtype showed infrequent CCND1 amplification and cyclin D1 protein overexpression [35]. The overexpression of cyclin D1 mRNA is associated with increased risk of relapse, local recurrence, metastasis, and death in patients with ER positive tumors [66]. Although the luminal subtype is prognostically favorable compared to other breast cancer subtypes, overexpression of cyclin D1 and CCND1 amplification are associated with poorer prognosis and treatment failure [51].

In the TCGA comprehensive molecular study of breast cancer, 53% of luminal A and 54% of luminal B tumors had higher than average cyclin D1 protein expression levels. Many, but not all, of the high protein levels are associated with amplification of cyclin D1. This is consistent with findings in previous breast cancer studies [35]. A hypothesis for the high cyclin D1 protein levels for the remaining tumors is that edited *AZIN1* sequesters the tumor suppressor gene, antizyme, and interferes with the proteasomal degradation of cyclin D1.

An investigation of RNA editing in the *AZIN1* gene was performed on data collected from 1,129 breast cancer patients by TCGA. Significant correlations between the normal and tumor EFs and the luminal A and luminal B subtypes were found. As was the case in previous studies [24], *AZIN1* editing was strongly associated with *ADAR1* expression and cyclin D1 protein expression.

5.2 Materials and Methods

5.2.1 TCGA Samples

The sequencing of tumor RNA has become common in large cancer projects such as those conducted by TCGA. The retrieval of adjacent normal tissue, on the other hand, is less common, and only a subset of samples has adjacent normal RNA-Seq data available. Patients with DNA and RNA from both the tumor and matched normal samples were termed quadruplets. Patients with DNA from the tumor and matched normal samples and RNA only from the tumor were called triplets. All patients provided written informed consent for genomic studies in accordance with local institutional review boards. All samples were approved by the policies and guidelines outlined by the Ethics, Law and Policy Group from TCGA.

5.2.1.1 Quadruplets

DNA whole-exome sequencing (DNA-WES) and RNA-Seq BAM [84] files from the tumor and matched normal samples from 568 patients across 12 different cancer types were downloaded from the Cancer Genomics Hub (CGHub, https://cghub.ucsc.edu).

Each cancer had a minimum of 20 quadruplets with at least 10 quadruplets having an editing frequency difference of 5% or more. The 568 quadruplets were distributed across the 12 cancers ranging from 21 quadruplets in bladder cancer to 114 in breast cancer. The full table with the number of quadruplets for each cancer type is outlined in Supplementary Table B.1. RNA editing events in the tumor and normal RNA were identified by RADIA (RNA and DNA Integrated Analysis) [114], a method that interrogates high-throughput sequencing data to detect germline variants, somatic mutations, and RNA editing events. The frequencies and raw read counts were quantified for each sample.

5.2.1.2 Triplets

DNA whole-exome sequencing (DNA-WES) BAM [84] files from the tumor and matched normal samples and RNA-Seq from the tumor samples were downloaded from CGHub. Adding the triplets to the quadruplets resulted in a distribution of total patient samples from 32 in colon adenocarcinoma to 1,129 in breast cancer (Supplementary Table B.1). RNA editing events in the tumor RNA were identified by RADIA [114].

5.2.1.3 Clinical Data

All clinical data was downloaded from the TCGA Data Access Portal at https://tcga-data.nci.nih.gov/tcga/. All clinical features were analyzed for statistical significance. All follow-up data was used for recurrence-free and overall survival analysis.

5.2.2 Computational Analyses

5.2.2.1 RNA Editing Detection

RNA editing events in the *AZIN1* gene at position chr8:103,841,636 were identified by RADIA (https://github.com/aradenbaugh/radia/, v1.1.1) [114]. RADIA processes the four BAM files from the quadruplets or the three BAM files from the triplets in parallel to detect germline variants, somatic mutations, and RNA editing events. RNA editing in the normal samples was determined by comparing the normal DNA and RNA to the reference genome. RNA editing in the tumor samples was determined by comparing the normal and tumor DNA along with the tumor RNA to the reference genome. The RNA editing event must be supported by at least 10 total reads with base and mapping qualities of 10 or more (phred score) where a minimum of 4 reads are required to support the editing event. In addition, the RNA editing alternative allele can only occur in less than two percent of the total DNA reads.

5.2.2.2 Statistical Analyses

The R project for statistical computing (version 3.1.2) and Microsoft Excel 2000 were used for data analysis. The *AZIN1* editing frequencies in tumors and matched normals were compared using the paired Student's t test. The statistical significance of the clinical and pathological data features and the editing frequencies were calculated using the Analysis Of Variance (ANOVA) test. The recurrence-free-survival and overall-survival analyses were performed using the Kaplan-Meier method. A p-value less than

0.05 was considered statistically significant.

5.3 Results

5.3.1 *AZIN1* Editing in Human Cancers

The *AZIN1* chr8:103,841,636 site is significantly over-edited in multiple cancers. An analysis of RNA editing in *AZIN1* was performed on the DNA whole-exome sequencing (DNA-WES) and RNA-Seq from the tumor and matched normal (when available) samples from 4,741 patients (568 quadruplets and 4,173 triplets) across 12 different cancers. A comparison of the editing frequency differences between the normal and tumor samples showed a statistical significance ($p < 0.05$) in ten (breast, head and neck, lung adenocarcinoma, colon adenocarcinoma, thyroid, bladder, endometrial, stomach, lung squamous, and liver) of the 12 cancer types (Figure 5.5A). The increase in *AZIN1* RNA editing was most significant ($p < 0.0003$) in six cancers: breast, lung adenocarcinoma, head and neck squamous cell, colon adenocarcinoma, thyroid, and uterine corpus endometriod carcinoma. The mean editing differences between the tumor RNA EFs ranges from six percent in kidney renal papillary to 18% in liver cancer (Figure 5.5B).

High *AZIN1* EFs are often significantly correlated with pathological and clinical data such as larger tumor sizes, lymph node involvement, presence of metastases, and advanced staging classifications. They are also associated with higher tumor grades, cancer specific subtypes, and poor overall and recurrence free survival.

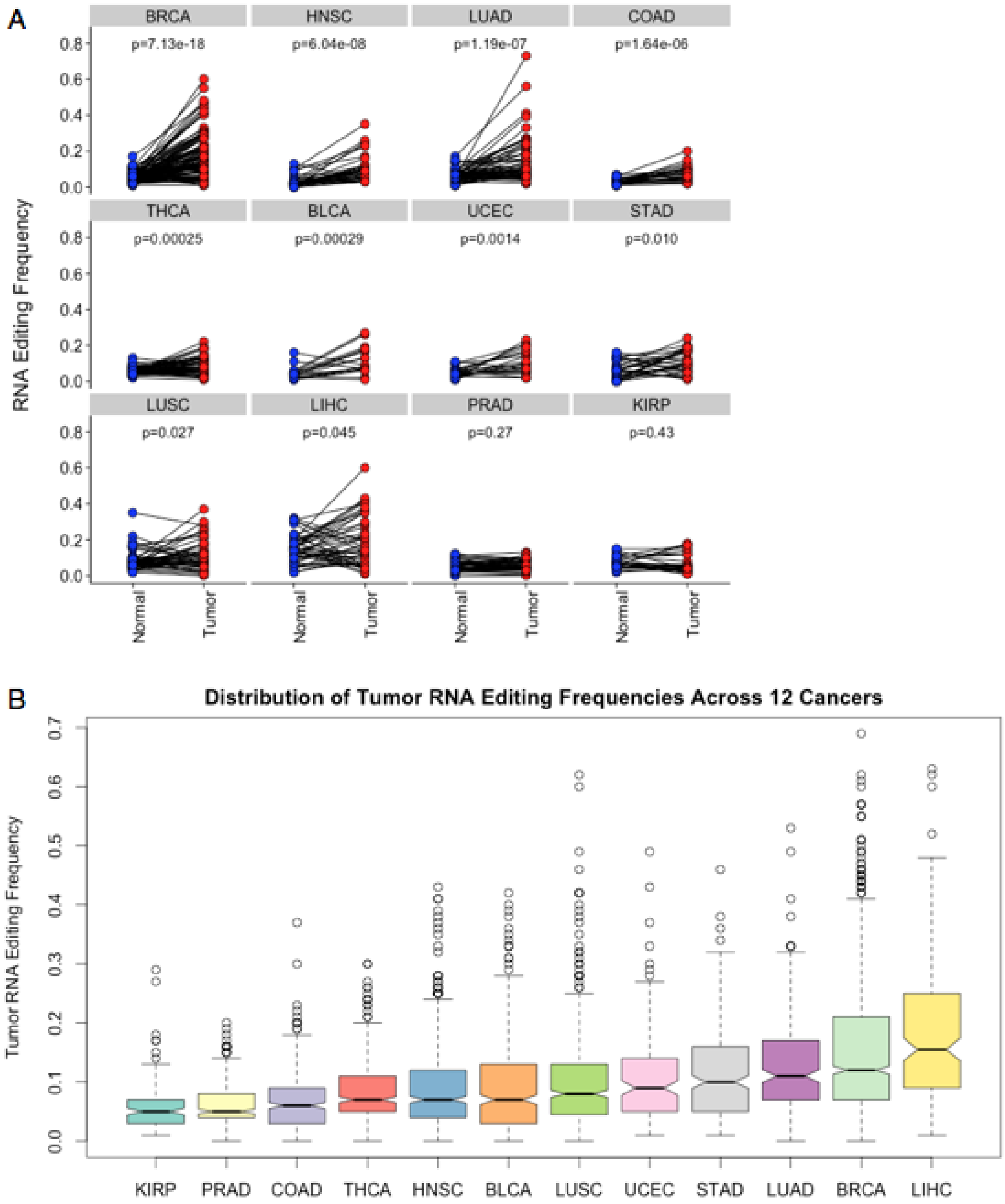


Figure 5.5: *AZIN1* editing frequencies for all 12 cancers. (A) A comparison of *AZIN1* editing in tumor and matched normals for 568 quadruplets from 12 cancers. The P values were calculated using the paired Students t-test. (B) The distribution of tumor RNA editing frequencies from 4,741 patients (quadruplets plus triplets) from 12 cancers. KIRP=Kidney Renal Papillary, PRAD=Prostate, COAD=Colon Adenocarcinoma, THCA=Thyroid, HNSC=Head and Neck Squamous, BLCA=Bladder, LUSC=Lung Squamous, UCEC=Endometrial, STAD=Stomach, LUAD=Lung Adenocarcinoma, BRCA=Breast, LIHC=Liver.

5.3.1.1 *AZIN1* Editing Associations with Staging Classifications

Pathologists and clinicians categorize tumors using the TNM Classification of Malignant Tumors staging system [54]. The staging classifications describe the severity of a person's cancer based on the size of their tumor (T), the degree of regional lymph node involvement (N), and the presence of distant metastases (M). TX, NX, and MX are used when no evaluation can be made. T0, N0, and M0 denote no tumor, no lymph node involvement, and no metastases present respectively. T1-T4 describe the size or extent of the primary tumor. The N1 classification is used when a tumor has spread to a small number or nearby lymph nodes, N3 designates a tumor has spread to numerous lymph nodes or more distant lymph nodes, and N2 is used when the lymph node involvement is between N1 and N3. The M1 classification is denoted when metastasis to distant organs has occurred. The TNM classifications are used to describe the overall stage of the cancer. The statistical correlation of *AZIN1* editing frequencies and the TNM classifications for each cancer was performed.

In thyroid cancer, the pathologist tumor size (T) classification was significantly correlated (p=5.27e^{-8}) with higher *AZIN1* EFs for both T3 and T4 (Figure 5.6A). In kidney renal papillary cell carcinomas, the pathologist T3 score was significantly associated (p=0.004) with higher EFs (Supplementary Figure B.2A). In bladder cancer, the pathologist T2 and T3 classifications had elevated *AZIN1* EFs (p=0.006, Supplementary Figure B.2B). In head and neck cancer, the pathologist T scores and clinical T scores steadily increased from T1 to T4 (p=0.03, p=0.03, Supplementary Figure B.2C-

D). In prostate cancer, both the pathologist and clinical tumor size (T) classification significantly correlated to higher *AZIN1* EFs (p=0.03, p=0.03, Supplementary Figure B.2E-F).

AZIN1 EFs significantly correlate with higher regional lymph node involvement classifications. In kidney renal papillary cancer, the *AZIN1* EFs steadily increased from no lymph node involvement (N0) to N1 and N2 (p=0.03, Figure 5.6B). In head and neck cancer, the EFs for N2 and N3 were higher than in N1 (p=0.03, Supplementary Figure B.3A). In addition, lymph nodes were examined to determine if they had only a few cancer cells in them (microscopic), many cancer cells (gross or macroscopic), or if the cancer had spread outside the wall of the node (extracapsular). Higher *AZIN1* EFs were associated with microscopic and gross extension (p=0.03, Supplementary Figure B.3B). In prostate cancer, higher *AZIN1* EFs correlated with the presence of lymph node involvement (p=0.02, Supplementary Figure B.3C).

High *AZIN1* EFs significantly correlate with distant metastases classifications. In kidney renal papillary cancer, the mean editing frequencies nearly doubled from no distant metastases (MO) to the presence of distant metastases (M1) (p=0.03, Figure 5.6C). Perineural invasion is when cancer cells are seen surrounding or tracking a nerve fiber and can be an indication that cancer has spread outside the tissue of origin [36]. In colon adenocarcinoma and head and neck cancer, higher EFs correlated with the presence of perineural invasion (p=0.0015 and p=0.0068 respectively, Supplementary Figure B.4A-B).

In thyroid cancer, the overall pathologist tumor stages significantly correlated

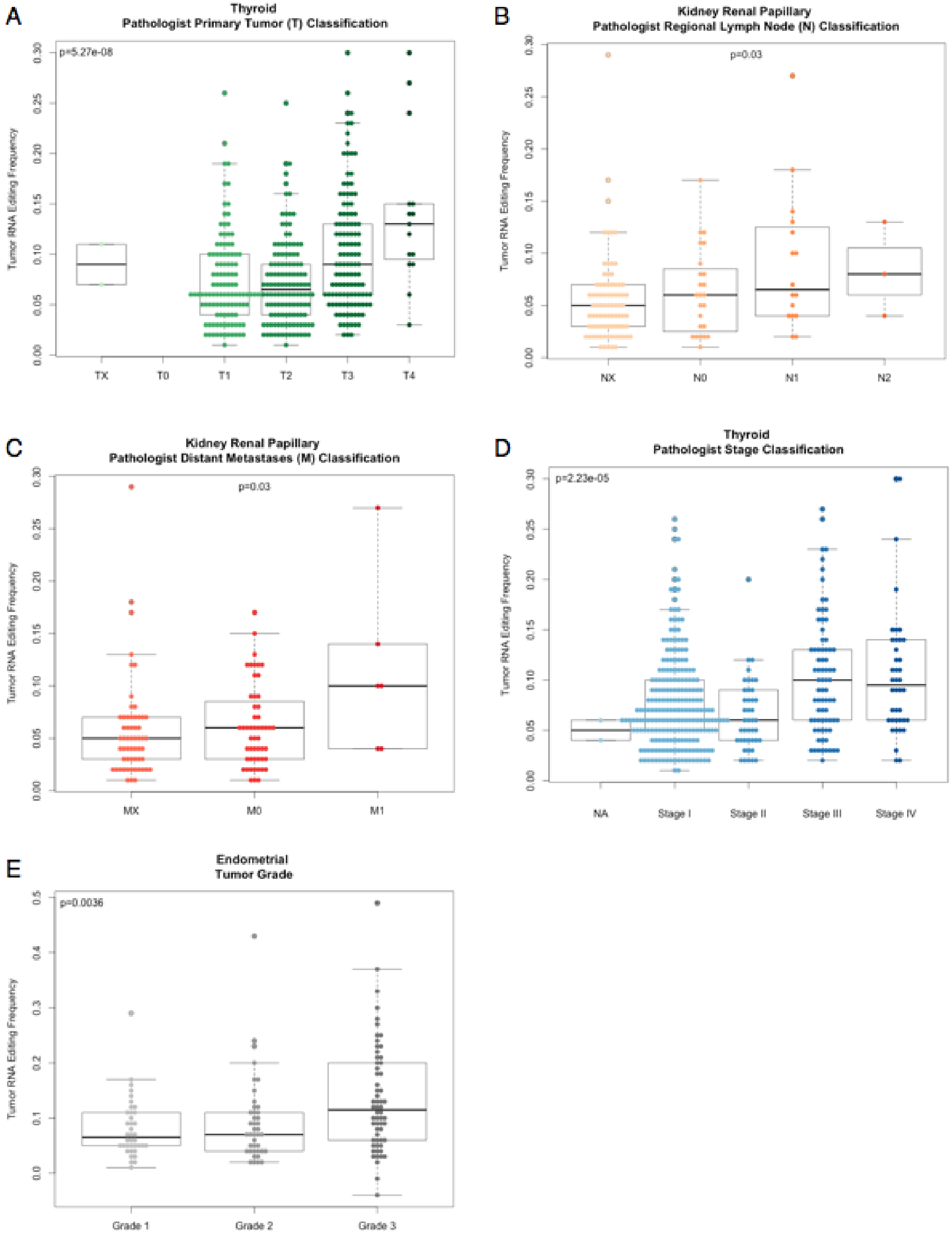


Figure 5.6: Higher *AZIN1* editing frequencies correlate with larger tumor sizes (A), a larger degree of lymph node involvement (B), and the presence of distant metastases (C). Higher *AZIN1* editing frequencies therefore also correlate to later tumor stages (D). Higher *AZIN1* editing frequencies also significantly correlate to the grade of the tumor (E).

($p=2.23e^{-5}$) with higher *AZIN1* EFs (Figure 5.6D). In head and neck cancer, the EFs steadily increased from stage 1 to stage 4 (p=0.001, Supplementary Figure B.5A). In kidney renal papillary cancer, the overall clinical stages 3 and especially 4 were associated with higher EFs (p=0.01, Supplementary Figure B.5B). Finally, in bladder cancer, the *AZIN1* EFs were elevated in stages 2-4 (p=0.04, Supplementary Figure B.5C).

5.3.1.2 Tumor Grade

The tumor grade describes how abnormal the tumor cells look under a microscope and is an indicator for how fast the tumor may grow and spread to other organs [55]. A pathologist examines the tissue from a biopsy to determine if it is benign or malignant. A grade of GX means that the grade could not be assessed. Grades 1-4 describe how differentiated the tumor cells are. Lower grades describe well-differentiated tumor cells that look similar to normal cells and will grow or spread slowly. Higher grades denote poorly differentiated or undifferentiated abnormal cells that will grow or spread faster than lower grades.

In prostate cancer, the gleason score is used to grade tumors. Pathologists assign a grade for both the primary and secondary patterns of tissue organization. Each pattern is given a grade from 1-5 where 1 denotes the tissue looking similar to normal prostate tissue and 5 denotes the most abnormal tissue [54]. The primary and secondary grades are added together to determine the gleason score. The Gleason X score means that a gleason score could not be determined. Gleason scores 2-6 describe normal (well-differentiated) looking tumor tissues. The gleason 7 score is used for moderately differentiated tissues while gleason scores 8-10 are for tumors that are poorly differentiated or undifferentiated.

In endometrial cancer, higher *AZIN1* EFs significantly correlated with Grade 3 tumors (p=0.0036, Figure 5.6E). In bladder cancer, higher *AZIN1* EFs were associated with "High Grade" histological grades (p=0.045, Supplementary Figure B.6A) In prostate cancer, the *AZIN1* EFs were highest for the worst overall gleason scores (8-10) (p=1.88e^{-5}, Supplementary Figure B.6B). Both the primary and secondary patterns were also significantly correlated (p=8.73e^{-4} and p=0.01 respectively, Figure B.6C-D) where the most poorly differentiated or undifferentiated tumors had the highest *AZIN1* EFs.

5.3.1.3 Other Clinical Associations

Extra-thyroidal extension is the spread of the primary tumor beyond the tumor to areas such as the trachea, larynx, vasculature, and esophagus [28]. It is a well-known prognostic factor for patients with thyroid cancer [28]. It is associated with high mortality and high tumor recurrence [28]. The moderate/advanced category of extra-thyroidal extension significantly correlated with the highest *AZIN1* EFs (p=2.23e^{-5}, Supplementary Figure B.7A) in the TCGA thyroid patients.

In bladder cancer, the T1 category indicates that the tumor has grown into the surrounding connective tissue, T2 represents that the tumor has grown into the muscle tissue, T3 illustrates that the tumor has grown into the fatty tissue that surrounds it, and T4 specifies that the tumor has spread to nearby organs such as the prostate, uterus, vagina, pelvic wall, or abdominal wall [128]. In the study on bladder cancer from TCGA [10], concomitant prostate tumor classification was given. Patients with bladder cancer that had spread to the prostate with size T3 had the highest *AZIN1* EFs (p=0.006, Supplementary Figure B.7B).

5.3.1.4 Subtypes

AZIN1 editing frequencies significantly correlate with cancer specific subtypes, especially those with poorer prognosis. An exception to this is the association of the luminal breast cancer subtypes with high EFs discussed below. Although the breast cancer luminal subtype is prognostically favorable compared to other subtypes, the overexpression of cyclin D1 and CCND1 amplification are associated with poorer prognosis [51, 66].

In endometrial cancer, high *AZIN1* EFs were mostly associated with the copy-number high (serous-like) integrative cluster subtype (p=0.023, Figure 5.7A) as defined by the integrated genomic characterization of endometrial cancer from TCGA [62]. The copy-number high integrative cluster had the worst Progressive Free Survival (PFS) of all of the integrated clusters and consisted entirely of patients from the copy number cluster 4, and primarily of patients with serous histology types and grade 3 tumors, all which were independently significantly correlated with high *AZIN1* EFs (p=9.28e^{-06}, Supplementary Figure B.8A; p=2.1e^{-04}, Supplementary Figure B.8B; p=0.0036, Figure 5.6E). Copy-number cluster 4 also contains most of the serous and a subset of the grade 3 endometrioid tumors and has the worst PFS of the copy-number clusters [62].

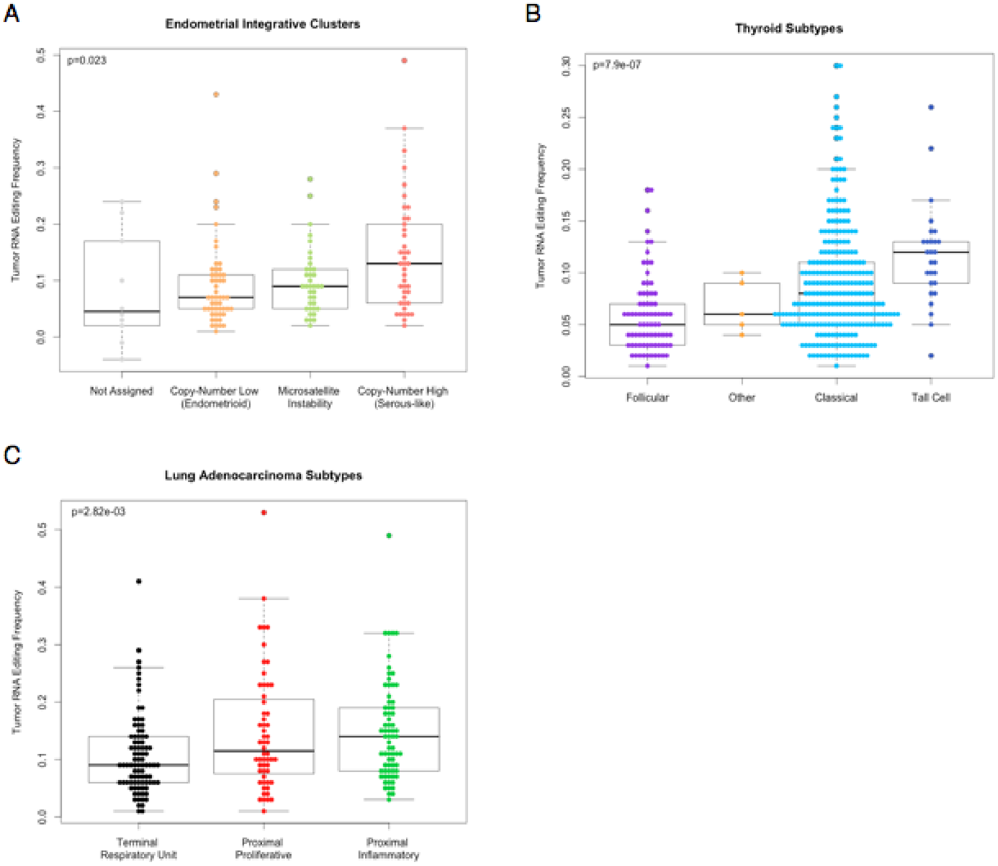


Figure 5.7: Higher *AZIN1* editing frequencies correlate with specific cancer subtypes, often those with a poor prognosis.

In thyroid cancer, the tall cell subtype significantly correlated with high *AZIN1* EFs (p=7.88e^{-07}, Figure 5.7B). The tall cell subtype had the highest frequency of BRAF (V600E) mutations and the least differentiated tumors [12]. It was deemed clinically relevant due to its association with more advanced stage and high risk tumors [12].

In lung adenocarcinoma, high *AZIN1* EFs (p=2.82e^{-03}, Figure 5.7C) significantly correlated mainly to the proximal inflammatory (PI) subtype and partially to the proximal proliferative (PP) subtype. In comparison to the terminal respiratory subtype (TRU), the PP and PI subtypes had worse overall survival with PI having the worst overall survival [11].

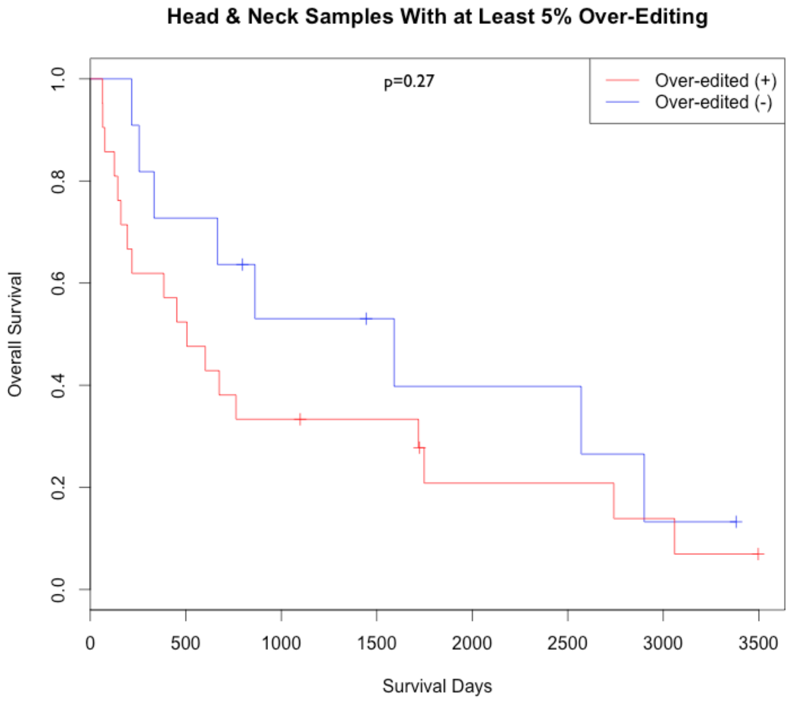


Figure 5.8: A minimum of 5% over-editing of *AZIN1* leads to a poorer Overall Survival (OS) in head and neck cancer.

5.3.1.5 Survival

Patients with over-editing of the *AZIN1* gene have a poorer overall and recurrence free survival. The poorer survival may only be applicable for a particular subtype, and the amount of over-editing necessary may vary from cancer to cancer. An attempt at calculating survival curves for all cancers was made, but due to the small number of samples with both normal RNA and survival data for some cancers, the p-values are often not significant. The Overall Survival (OS) curves were calculated for samples with more than 5% over-editing for head and neck (Figure 5.8), endometrial, kidney papillary, and thyroid (Supplementary Figure B.9A-C) and with a minimum of 10% over-editing for lung adenocarcinoma (Supplementary Figure B.9D). The Recurrence Free Survival (RFS) curves were calculated for samples with more than 5% over-editing for lung squamous, colon, and bladder (Supplementary Figure B.10A-C) and a minimum of 18% over-editing for lung adenocarcinoma and liver (Supplementary Figure B.10D-E). In all cases, samples with over-editing had a worse overall and recurrence free survival.

5.3.2 *AZIN1* Editing in TCGA Primary Breast Cancer Tumors

The most significant *AZIN1* editing difference between tumors and normals in the 12 cancers studied here occurred in breast cancer (p=7.13e-18). A detailed analysis on 114 quadruplets (WES and RNA-Seq data for tumor and adjacent normal tissue) and an additional 1015 triplets (lacking RNA-Seq from adjacent normal tissue) in breast cancer was performed. For the 114 quadruplets, the editing difference between the normal RNA and tumor RNA were analyzed (Figure 5.9A). The editing differences were statistically significant in the luminal A, luminal B, and basal subtypes (Figure 5.9A). For an additional 1015 breast cancer patients, tumor RNA was available. RNA editing frequencies in the tumor RNA were calculated on the entire cohort of 1129 breast cancer patients. The tumor RNA editing frequencies significantly correlated with the luminal A and luminal B subtypes (Figure 5.9B).

In a study of *AZIN1* RNA editing in liver cancer [24], the term "over-editing" was defined as a 10% increase in editing from the normal to the tumor. In order to include some of the 1015 TCGA patients where no adjacent normal data was available in the analysis, the normal EFs from the 114 TCGA patients with adjacent normal data were used as a proxy. The normal RNA EFs in the 114 patients ranged from $<$ 1% to 17% with a mean normal editing frequency of 5% (the maximum normal RNA editing frequency of 17% was an outlier by 5%, with the second highest maximum at 12%). To be conservative, all samples with no normal RNA-Seq data available and a tumor editing frequency $\geq$15% and $<$27% were ignored. This resulted in 812 samples that were further analyzed.

Similar to the 114 quadruplets, the editing differences for all 812 patients were highest in the luminal A and luminal B subtypes. As expected, *AZIN1* over-editing in breast cancer is significantly correlated with *ADAR1* expression (Figure 5.10A-B), especially for the luminal A and luminal B subtypes (Figure 5.10C-D).

AZIN1 over-editing in luminal A and especially luminal B is significantly correlated with higher cyclin D1 protein levels (Figure 5.11A-C). For the luminal B subtype, there were 60 samples that had cyclin D1 RPPA data, SNP6 copy number data, and

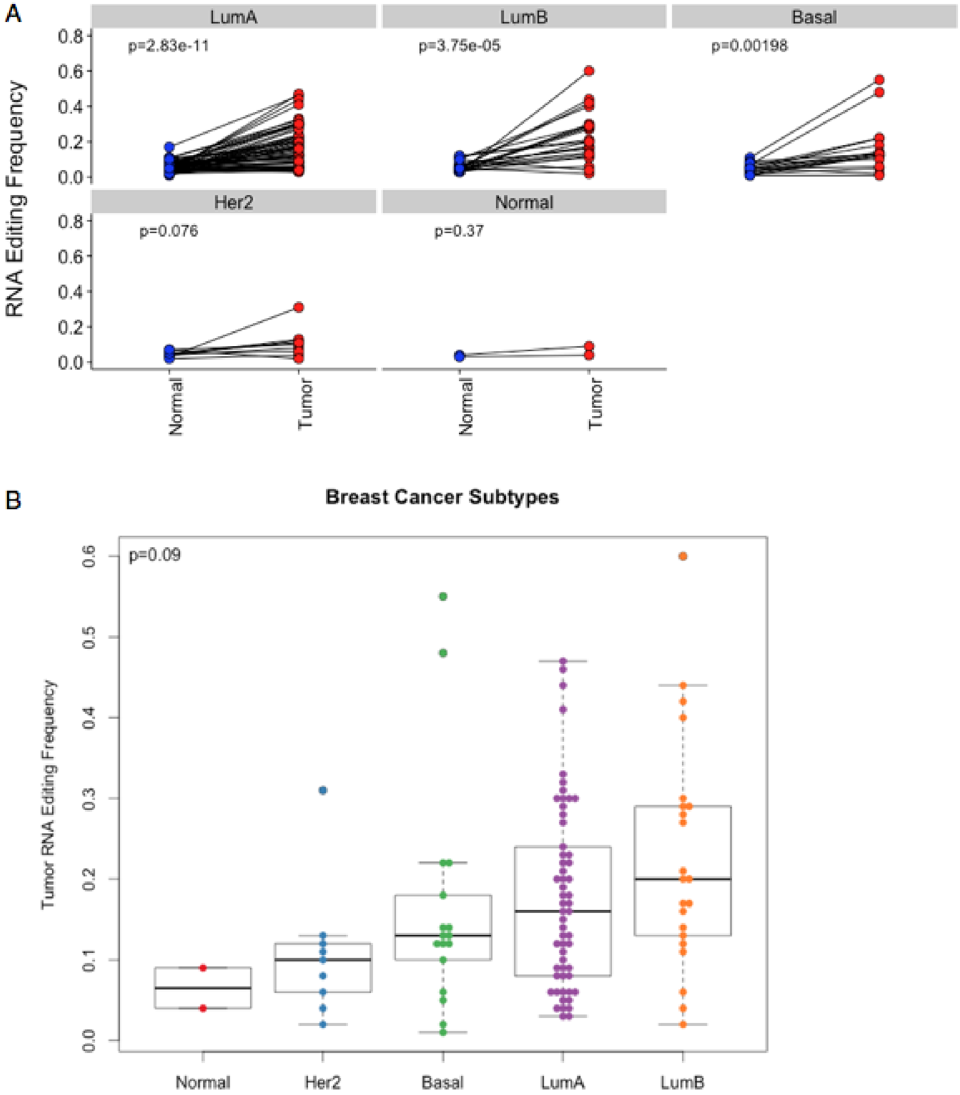


Figure 5.9: Elevated *AZIN1* RNA editing frequencies significantly correlated to the luminal A and luminal B breast cancer subtypes. (A) A comparison of *AZIN1* editing in tumor and matched normals for 114 quadruplets from breast cancer. The p-values were calculated using the paired Students *t*-test. (B) The distribution of tumor RNA editing frequencies from 1129 breast cancer patients.

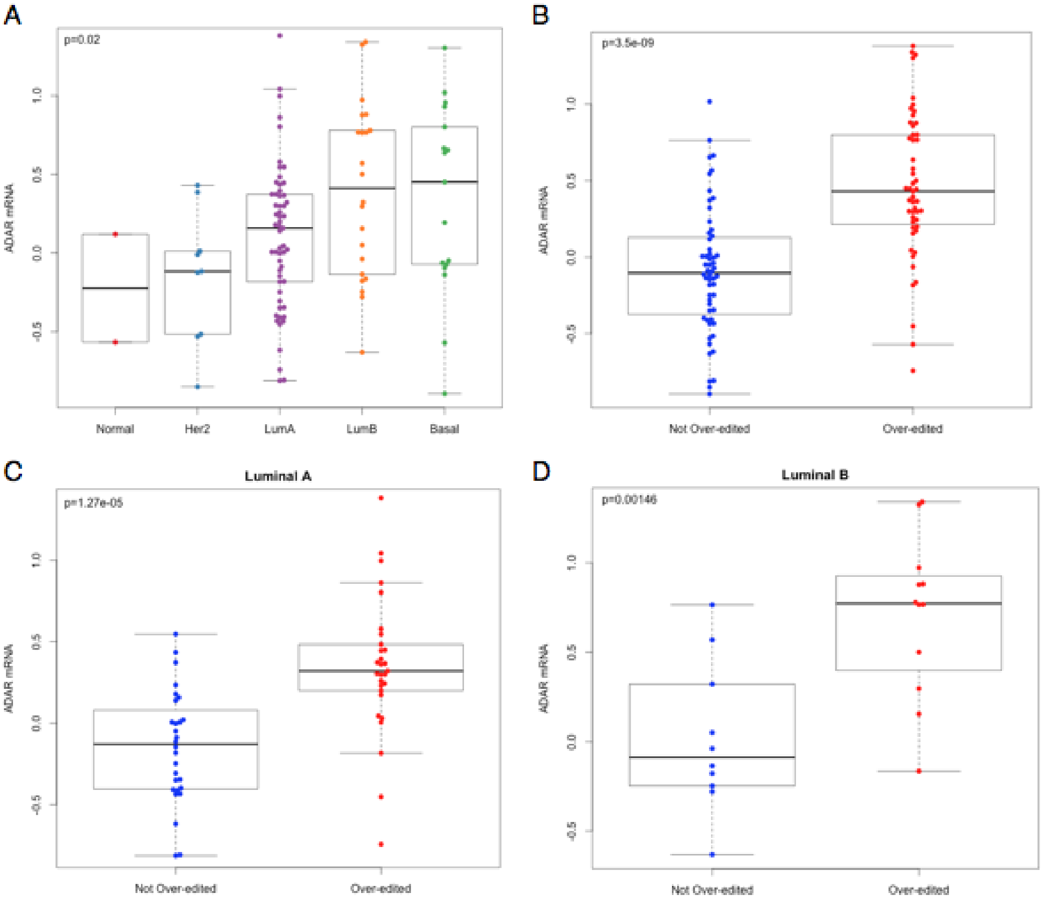


Figure 5.10: *AZIN1* over-editing is strongly associated with higher ADAR mRNA expression. (A) Correlation between *AZIN1* editing and ADAR mRNA expression by subtype. (B) Over-editing of *AZIN1* and ADAR mRNA expression for all subtypes. Over-editing of *AZIN1* and ADAR mRNA for the luminal A (C) and luminal B (D) subtypes.

RNA editing of *AZIN1*. Thirty-seven samples have high cyclin D1 protein levels, and 28 of them have cyclin D1 amplification. For the remaining 10 samples with high cyclin D1 protein levels and no cyclin D1 amplification, four (40%) are over-edited (Table 5.1). For the luminal A subtype, there were 73 samples with high cyclin D1 protein expression, of which 39 have cyclin D1 amplification. Of the remaining 34 samples, nine (26%) are over-edited (Table 5.1).

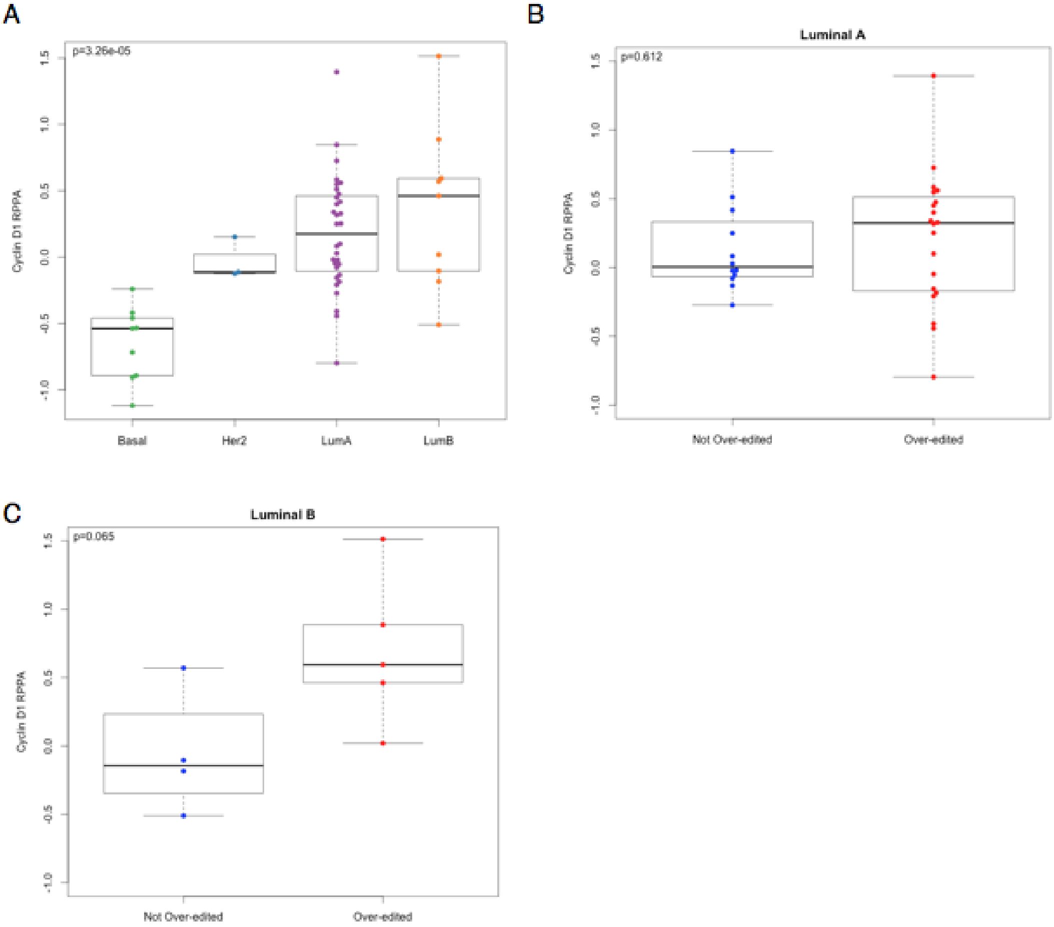
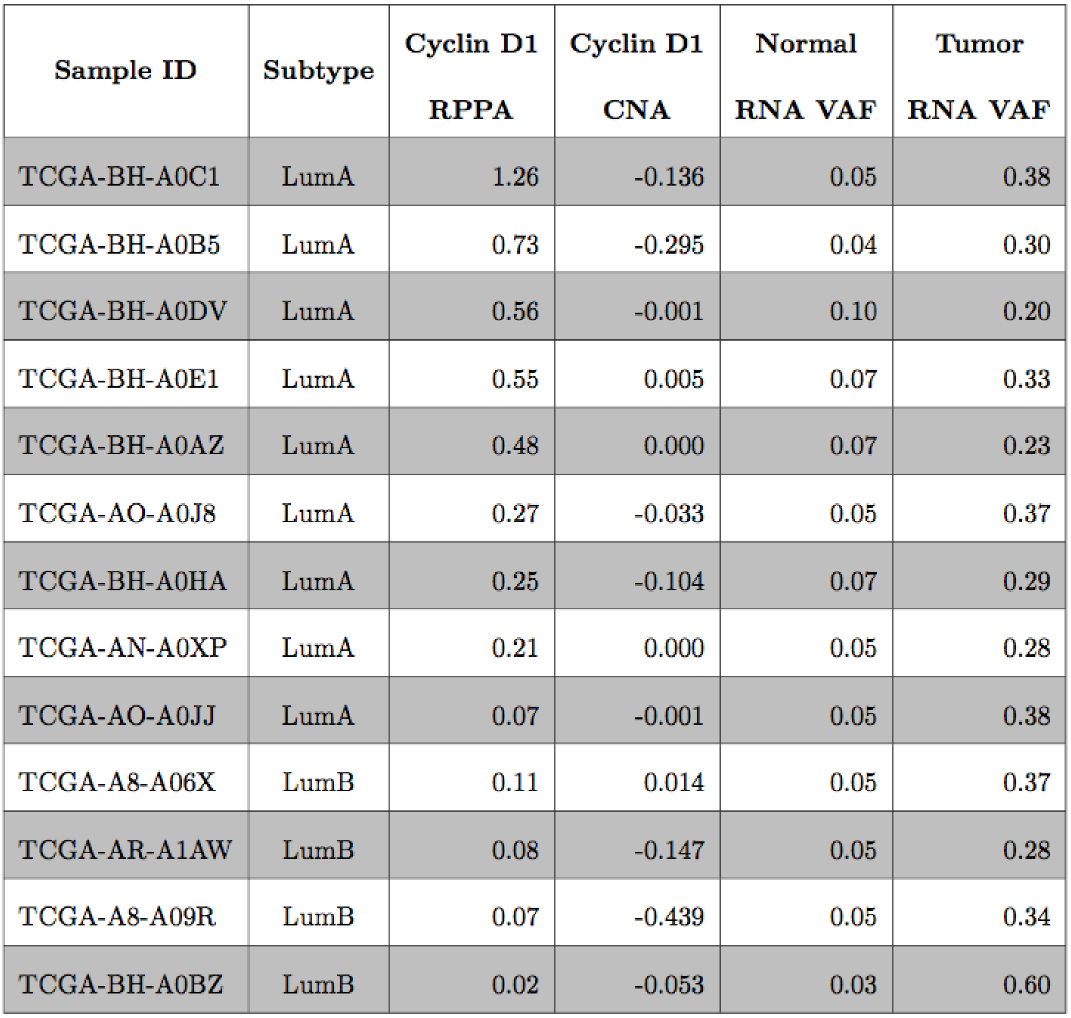


Figure 5.11: *AZIN1* over-editing in the luminal A and luminal B subtypes is associated with higher cyclin D1 protein expression. (A) Correlation between *AZIN1* editing and cyclin D1 protein expression by subtype. Over-editing of *AZIN1* and cyclin D1 protein expression for the luminal A (B) and luminal B (C) subtypes.



Sample ID	Subtype	Cyclin D1 RPPA	Cyclin D1 CNA	Normal RNA VAF	Tumor RNA VAF
TCGA-BH-A0C1	LumA	1.26	-0.136	0.05	0.38
TCGA-BH-A0B5	LumA	0.73	-0.295	0.04	0.30
TCGA-BH-A0DV	LumA	0.56	-0.001	0.10	0.20
TCGA-BH-A0E1	LumA	0.55	0.005	0.07	0.33
TCGA-BH-A0AZ	LumA	0.48	0.000	0.07	0.23
TCGA-AO-A0J8	LumA	0.27	-0.033	0.05	0.37
TCGA-BH-A0HA	LumA	0.25	-0.104	0.07	0.29
TCGA-AN-A0XP	LumA	0.21	0.000	0.05	0.28
TCGA-AO-A0JJ	LumA	0.07	-0.001	0.05	0.38
TCGA-A8-A06X	LumB	0.11	0.014	0.05	0.37
TCGA-AR-A1AW	LumB	0.08	-0.147	0.05	0.28
TCGA-A8-A09R	LumB	0.07	-0.439	0.05	0.34
TCGA-BH-A0BZ	LumB	0.02	-0.053	0.03	0.60

Table 5.1: TCGA breast cancer samples where *AZIN1* editing may be driving higher cyclin D1 protein levels. These samples have higher than average cyclin D1 protein expression, no cyclin D1 amplification, and are over-edited (the editing difference between the normal and tumor samples is at least 10%).

In summary, *ADAR1* over-expression results in an increase in A-to-I RNA editing in *AZIN1*. The edited AZIN1 protein sequesters the tumor suppressor antizyme leading to an increase in cyclin D1 protein expression. Over-editing occurs in 40% of luminal B and 26% of luminal A breast cancer patients with high cyclin D1 protein expression and little to no cyclin D1 amplification.

5.4 Discussion

RNA editing of *AZIN1* has been shown to be a potential driver in tumor initiation and progression in the pathogenesis of hepatocellular carcinoma (HCC). An important target of *AZIN1* is the cyclin D1 oncogene. Over-editing of *AZIN1* leads to a higher binding affinity to antizyme, a tumor suppressor that regulates cell growth. The edited form of *AZIN1* sequesters antizyme and inhibits the degradation of cyclin D1. This leads to overexpression of cyclin D1, stimulating Rb phosphorylation and resulting in a deregulation of cell cycle progression.

Here, an investigation of *AZIN1* RNA editing in 12 cancers with extensive genomic characterization by TCGA was performed. Strong correlations between increased *AZIN1* editing and clinical features associated with more advanced disease were found. Ten of the twelve cancers examined showed significantly higher tumor *AZIN1* EFs than the matched normal tissues.

Breast cancer had the highest editing differences, particularly in the luminal subtypes. As CCND1 is a known oncogenic driver of breast cancers, additional analysis of the relationship between *AZIN1* editing, *ADAR1* expression, and cyclin D1 protein expression levels was performed. Over-editing of the *AZIN1* gene by the ADAR1 enzyme was found in luminal breast cancers. Cyclin D1 protein levels were highest in the luminal subtypes and significantly correlated with over-editing of *AZIN1*. In 4 out of 10 (40%)

luminal B samples and 9 out of 34 (26%) luminal A samples with high cyclin D1 protein levels and no cyclin D1 amplification, over-editing of *AZIN1* was present, suggesting that *AZIN1* editing is a possible mechanism for the increase in cyclin D1 protein expression.

AZIN1 has been shown to be a driver for liver cancer and has been proposed as a driver for other human cancers. An analysis of *AZIN1* editing in 12 cancers showed that *AZIN1* over-editing is indeed prevalent in many cancers. The highest tumor EFs were significantly associated with the most severe clinical variables such as larger tumor sizes, greater lymph node involvement, the presence of metastases, and more advanced tumor stages and tumor grades. Cancer specific subtypes, often those with the worst survival, correlated with the highest EFs. Finally, a poor overall and recurrence free survival correlated with over-editing in many cancers. In summary, *AZIN1* editing is wide-spread in many cancers and associated with increased cyclin D1 protein levels. These results suggest that *AZIN1* editing promotes cancer by acting as a driver of cell proliferation through the deregulation of cell cycle progression.

Chapter 6

Conclusion

Cancer is the second most common cause of death in the US, with heart disease being the leading cause (American Cancer Society (ACS) [4]). Nearly a quarter of all deaths in the US are due to cancer. There are a number of factors that influence whether an individual will develop cancer including the germline variants that they are born with and somatic mutations that occur randomly in normal cells during the course of a lifetime. Several important germline variants such as those in *BRCA1* and *BRCA2* and somatic mutations found in *BRAF*, *EGFR*, *ERBB2*, *VEGF*, and *TP53* that are implicated in cancer have already been discovered.

The accurate and comprehensive identification and characterization of single nucleotide variants is crucial to cancer research in many ways, including determining potential cancer genes for drug development, diagnosis, and prognosis. With the accumulation of high-throughput sequencing datasets for both the DNA and RNA from the same patients across multiple cancers, it is possible to thoroughly characterize single nucleotide variants on a genome-wide scale for many cancers. The inclusion of the RNA in variant detection allows one to assess whether a variant is being expressed and also provides a boost in power for variant detection. RADIA has been used to detect mutations for nearly 4,500 patients across 22 different cancers, and the inclusion of the RNA provided a 2-7% increase in sensitivity. TCGA researchers have focused on identifying the most significantly mutated genes for each cancer studied. These mutational patterns help to divide patients into subtypes with the same molecular characteristics. These subtypes are important for determining the best treatment options for a particular patient.

RNA editing has been identified as an additional mechanism relevant to cancer development and progression. Alterations to RNA editing have been linked to diseases such as depression, epilepsy, schizophrenia, ALS, and various types of cancer. Previous studies [106] discovered a global hypoediting of *Alu* elements in brain tumors, and a correlation between editing levels, $ADAR$ expression, and the grade of tumor malignancy [106]. Other studies identified a more complex pattern of hypo- and hyper-editing of specific genes that are relevant to various types of cancer such as an event discovered in the *AZIN1* gene linked to tumor initiation and progression in liver cancer. This thesis focused on characterizing RNA editing of the *AZIN1* gene in nearly 5,000 patients across 12 different cancers. An over-editing of *AZIN1* was discovered in 10 of the 12 cancers, and this over-editing was often significantly correlated to advanced and aggressive characteristics of tumorigenesis.

With projects like TCGA providing sequencing data for both DNA and RNA from the same patients across multiple cancers, it is now possible to characterize germline variants, somatic mutations, and RNA editing events on a genome-wide scale. The identification of single nucleotide variants that occur in specific genes across multiple cancers provides a powerful way to discover genes that are important to these diseases.

Appendix A

RADIA Supplementary Figures

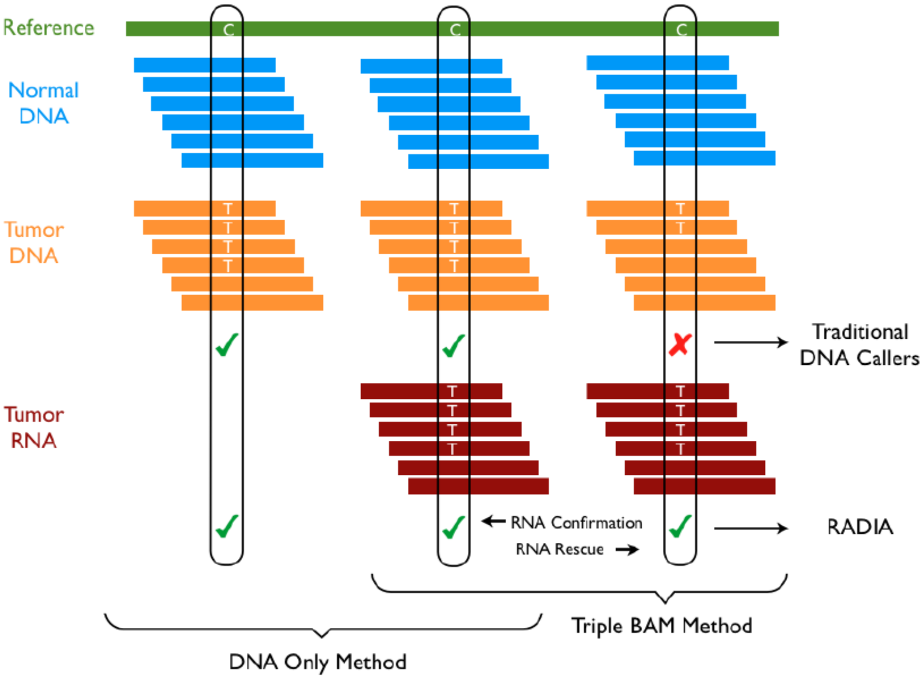


Figure A.1: Schematic of the types of calls made by the RADIA DNA Only Method (DOM) and Triple BAM Method (TBM). In the first and middle columns, there is enough DNA read support for the DOM and other algorithms examining DNA pairs to make a call. In the middle and last columns, there is enough RNA read support for the TBM to make a call. The middle column illustrates "RNA Confirmation" calls that are detected by both the DOM and the TBM due to high read support in both the DNA and RNA. The last column represents the "RNA Rescue" calls that have some support in the DNA and strong evidence in the RNA.

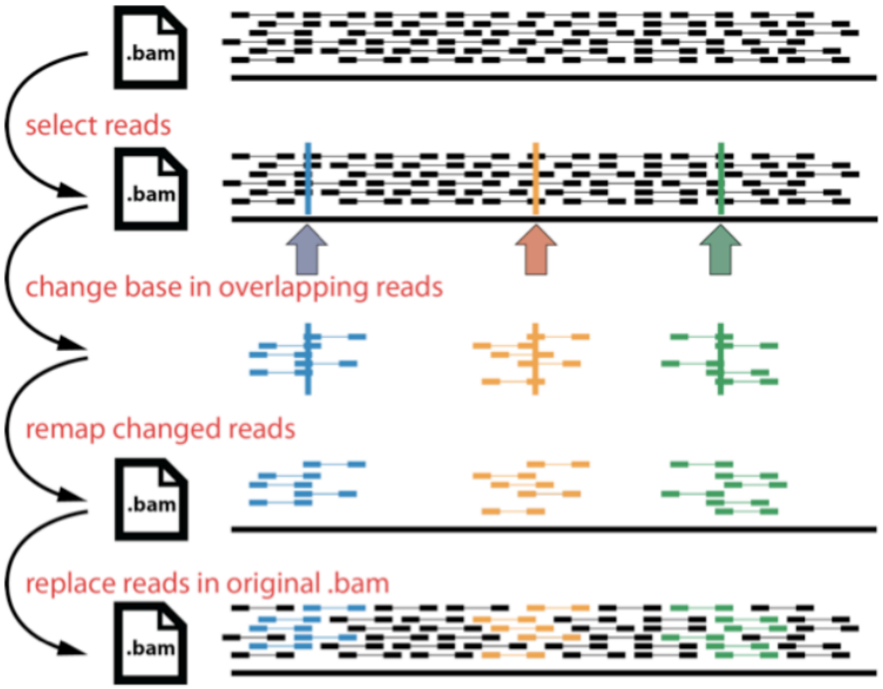

.bam
select reads
.bam
change base in overlapping reads
remap changed reads
.bam
replace reads in original .bam
.bam

Figure A.2: Diagram of bamsurgeon methodology. Mutations are spiked into BAM files by selecting locations with adequate coverage, extracting the reads, and adjusting the VAF according to the desirable VAF distribution. Once the bases in the reads are changed, the reads are remapped to the genome, replacing the reads in the original BAM file.

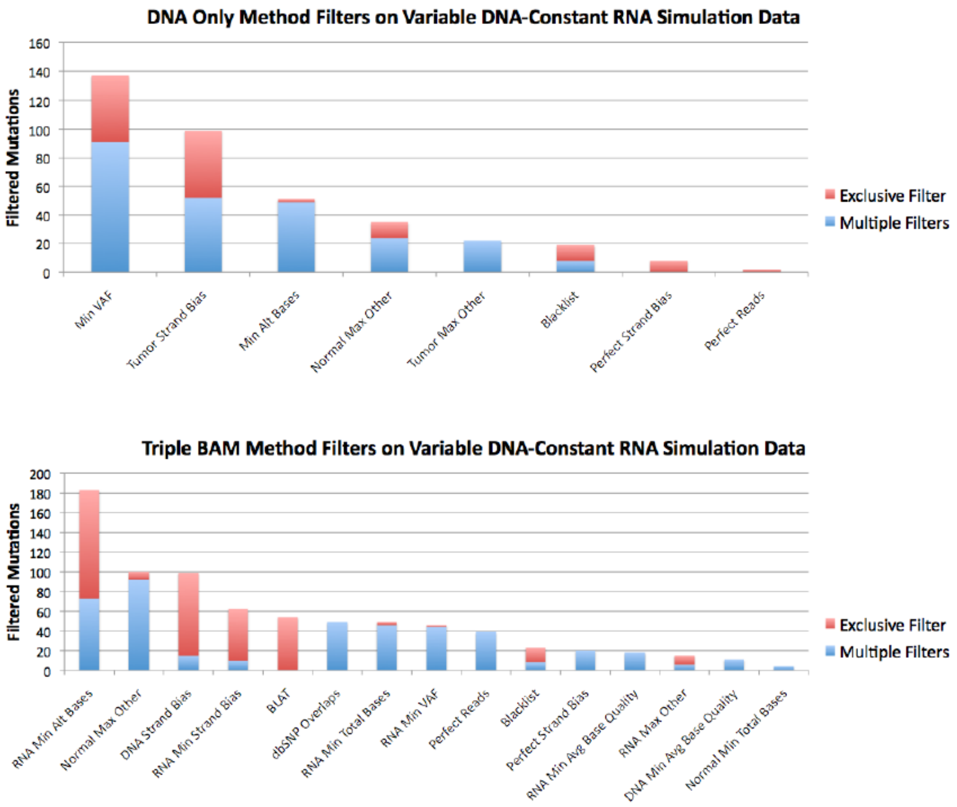


Figure A.3: Filters applied in the bamsurgeon simulation experiment where the DNA variant allele frequencies were distributed from 1-50% and the RNA was held constant at 25%. Most of the DOM mutations were filtered because of the low variant allele frequency and tumor strand bias. In the TBM, most of the mutations were filtered due to the minimum number of alternative alleles required to make a call (n=4) and strand bias in the tumor DNA and RNA.

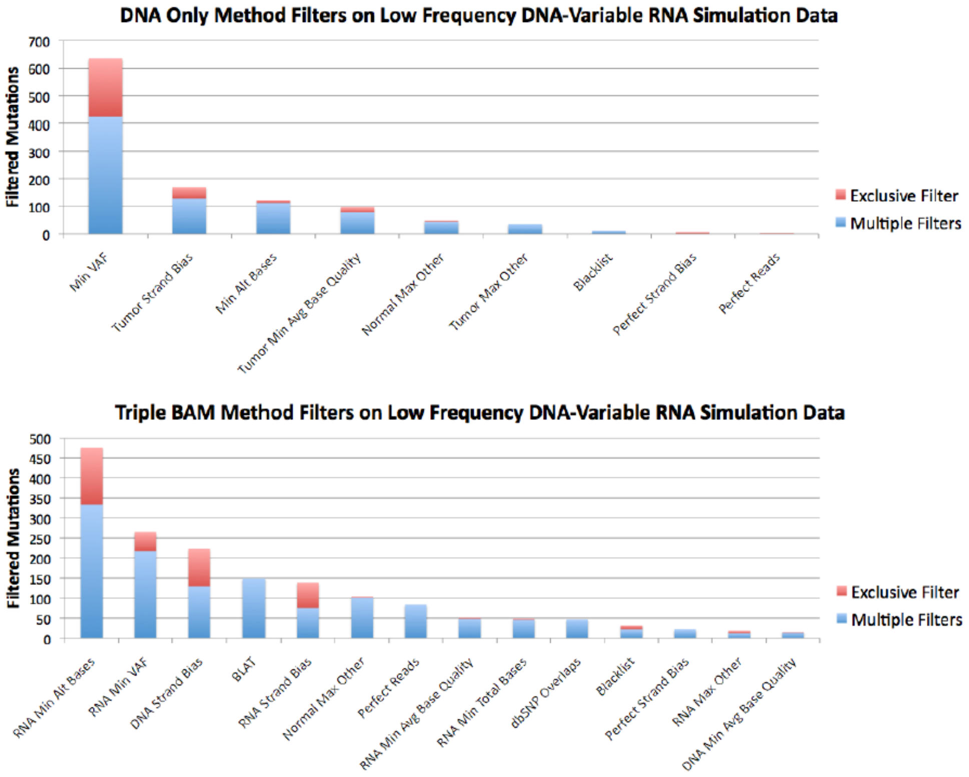


Figure A.4: Filters applied in the bamsurgeon simulation experiment where the RNA variant allele frequencies were distributed from 1-50% and the DNA was held at 10% or less. Most of the DOM mutations were filtered because of the low DNA variant allele frequency and tumor strand bias. In the TBM, most of the mutations were filtered due to the minimum number of alternative alleles required to make a call (n=4) and the low RNA variant allele frequency.

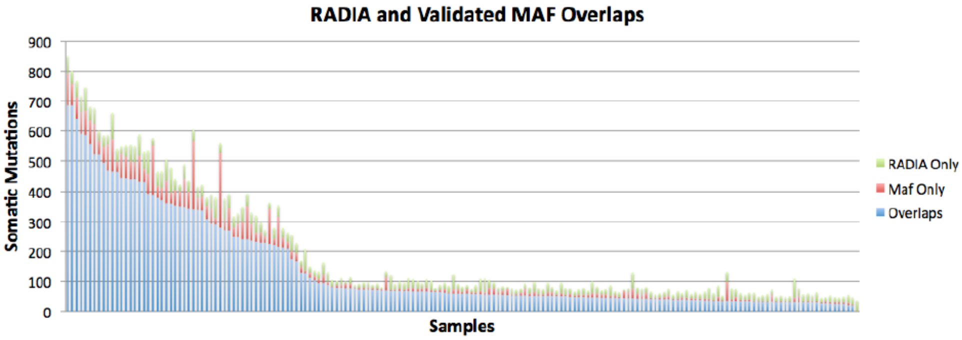


Figure A.5: The distribution of the overlaps between RADIA and the validated somatic mutations from the TCGA endometrial MAF file.

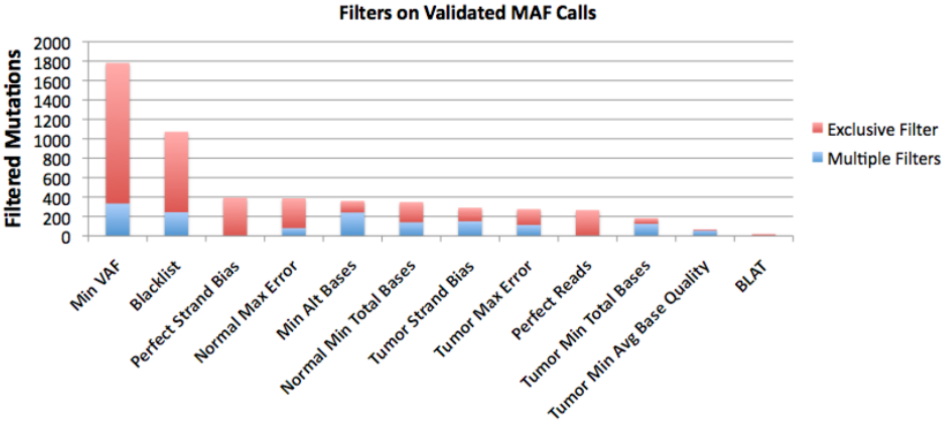


Figure A.6: Filters applied to the RADIA mutations that validated as somatic in the endometrial MAF file. Thirty-three percent of the mutations had a DNA VAF of 8% or less while 23% landed in blacklist regions that were not considered.

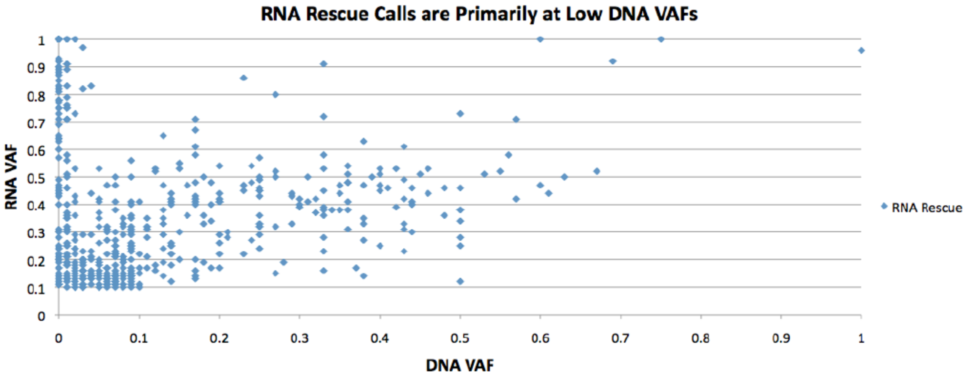


Figure A.7: RNA Rescue calls are primarily found at low DNA variant allele frequencies, but they are also found at higher frequencies where the call was originally filtered due to non-depth related artifacts (e.g. strand-bias).

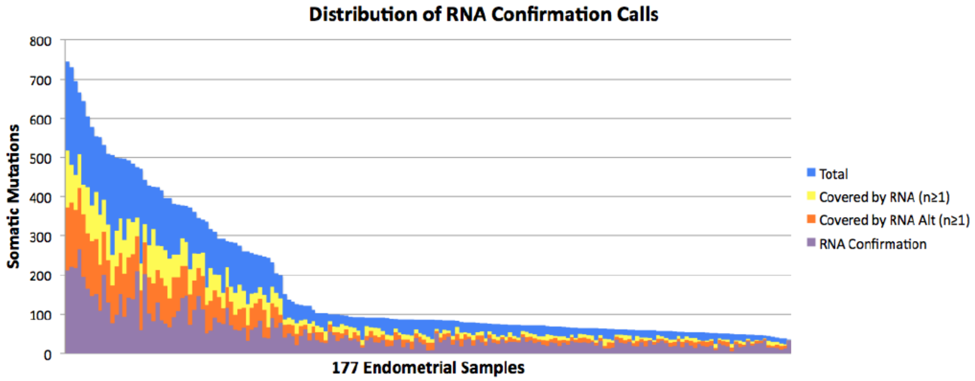
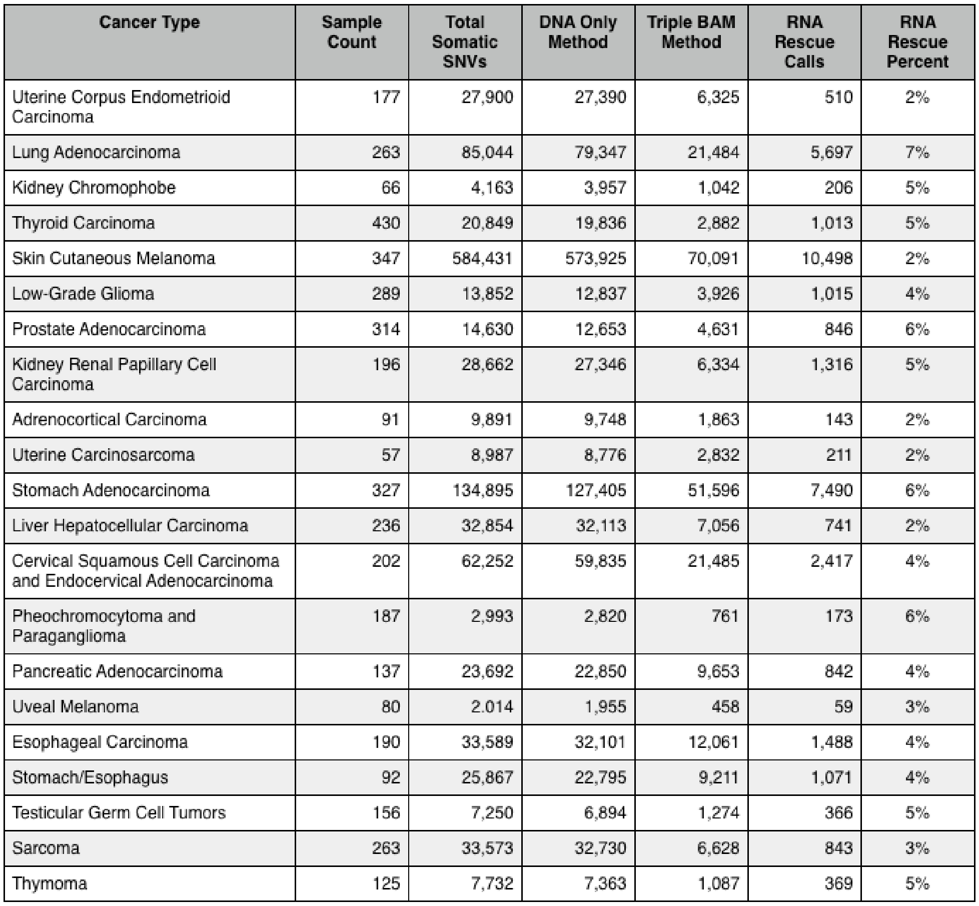


Figure A.8: The total number of mutations (blue) that are covered by at least one RNA read (yellow), one RNA read supporting the mutant allele (orange), and RNA Confirmation mutations with high support in both the DNA and RNA (purple).

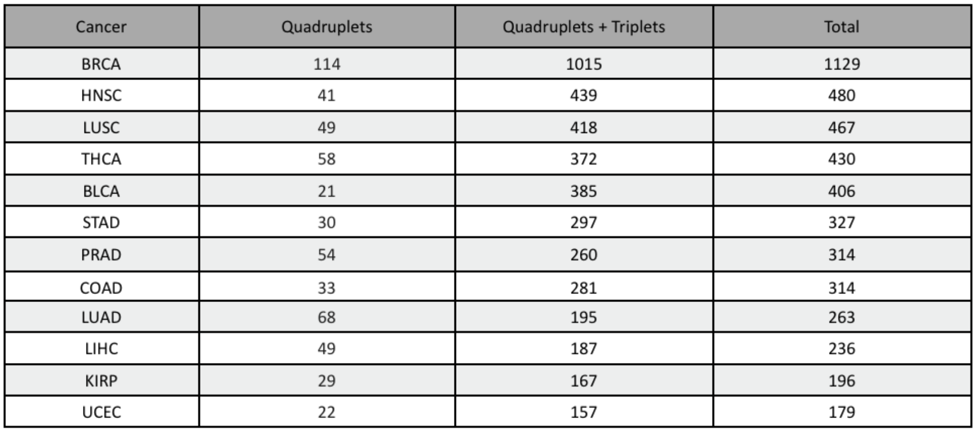


Cancer Type	Sample Count	Total Somatic SNVs	DNA Only Method	Triple BAM Method	RNA Rescue Calls	RNA Rescue Percent
Uterine Corpus Endometrioid Carcinoma	177	27,900	27,390	6,325	510	2%
Lung Adenocarcinoma	263	85,044	79,347	21,484	5,697	7%
Kidney Chromophobe	66	4,163	3,957	1,042	206	5%
Thyroid Carcinoma	430	20,849	19,836	2,882	1,013	5%
Skin Cutaneous Melanoma	347	584,431	573,925	70,091	10,498	2%
Low-Grade Glioma	289	13,852	12,837	3,926	1,015	4%
Prostate Adenocarcinoma	314	14,630	12,653	4,631	846	6%
Kidney Renal Papillary Cell Carcinoma	196	28,662	27,346	6,334	1,316	5%
Adrenocortical Carcinoma	91	9,891	9,748	1,863	143	2%
Uterine Carcinosarcoma	57	8,987	8,776	2,832	211	2%
Stomach Adenocarcinoma	327	134,895	127,405	51,596	7,490	6%
Liver Hepatocellular Carcinoma	236	32,854	32,113	7,056	741	2%
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	202	62,252	59,835	21,485	2,417	4%
Pheochromocytoma and Paraganglioma	187	2,993	2,820	761	173	6%
Pancreatic Adenocarcinoma	137	23,692	22,850	9,653	842	4%
Uveal Melanoma	80	2.014	1,955	458	59	3%
Esophageal Carcinoma	190	33,589	32,101	12,061	1,488	4%
Stomach/Esophagus	92	25,867	22,795	9,211	1,071	4%
Testicular Germ Cell Tumors	156	7,250	6,894	1,274	366	5%
Sarcoma	263	33,573	32,730	6,628	843	3%
Thymoma	125	7,732	7,363	1,087	369	5%

Figure A.9: Summary of TCGA samples analyzed for somatic mutation detection. RADIA has been run on nearly 4,500 TCGA patients across 22 different cancer types. The RNA Rescue calls make up 2-7% of the total somatic mutation calls across the 22 types of cancer. Variant Call Format (VCF) and Mutation Annotation Format (MAF) files can be downloaded from the TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/). Open-access somatic MAFs can be visualized and downloaded via the UCSC Cancer Browser (https://genomecancer.ucsc.edu/).

Appendix B

RNA Editing Supplementary Figures



Cancer	Quadruplets	Quadruplets + Triplets	Total
BRCA	114	1015	1129
HNSC	41	439	480
LUSC	49	418	467
THCA	58	372	430
BLCA	21	385	406
STAD	30	297	327
PRAD	54	260	314
COAD	33	281	314
LUAD	68	195	263
LIHC	49	187	236
KIRP	29	167	196
UCEC	22	157	179

Figure B.1: Summary information on the number of quadruplets, triplets, and total number of samples included in this study by cancer. BRCA=Breast, HNSC=Head and Neck Squamous, LUSC=Lung Squamous, THCA=Thyroid, BLCA=Bladder, STAD=Stomach, PRAD=Prostate, COAD=Colon Adenocarcinoma, LUAD=Lung Adenocarcinoma, LIHC=Liver, KIRP=Kidney Renal Papillary, UCEC=Endometrial.

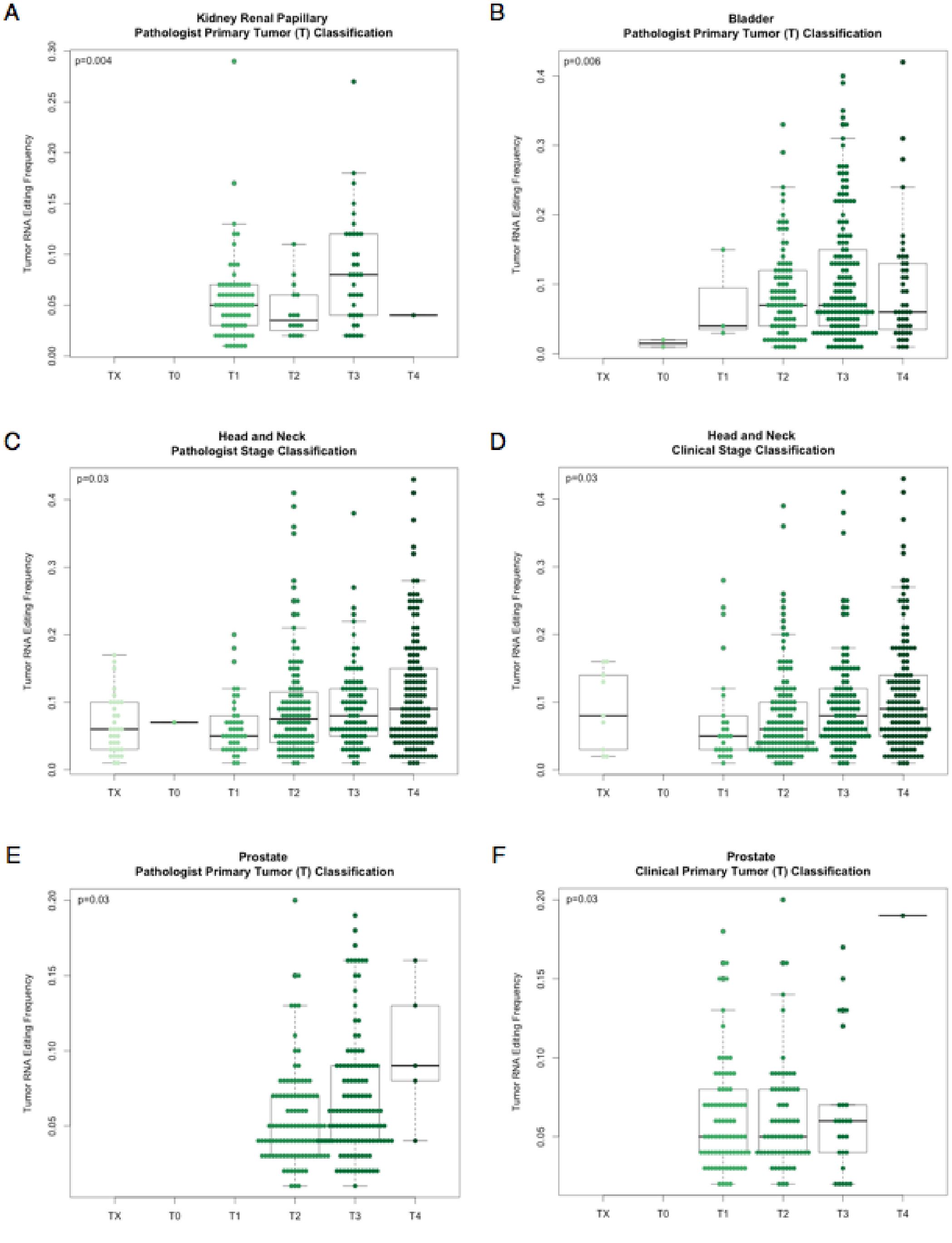


Figure B.2: Higher *AZIN1* EFs are associated with larger pathologist and/or clinical primary tumor size (T) classifications in (A) kidney renal papillary, (B) bladder urothelial carcinoma, (C-D) head and neck, and (E-F) prostate.

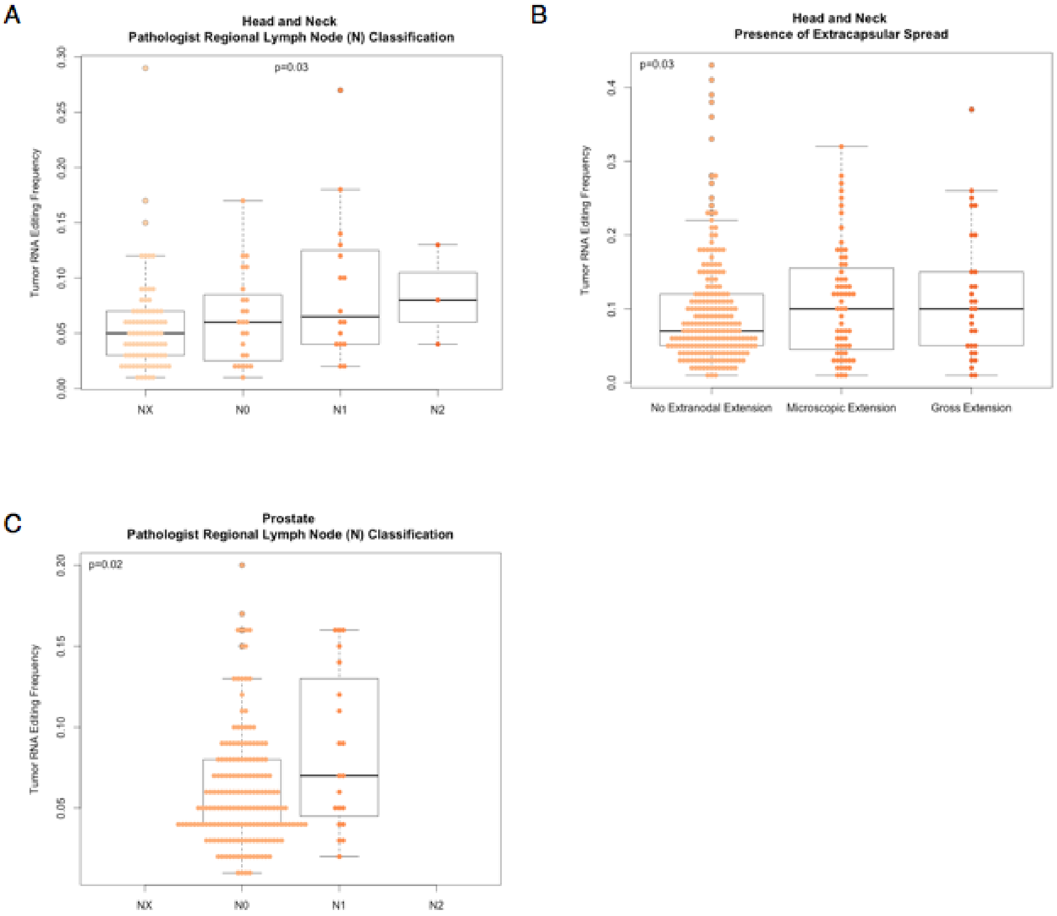


Figure B.3: In head and neck cancer, higher *AZIN1* EFs correlate with (A) greater pathologist regional lymph node involvement (N) classifications and (B) the presence of extracapsular spread. In addition, higher *AZIN1* EFs are associated with (C) greater pathologist regional lymph node involvement in prostate.

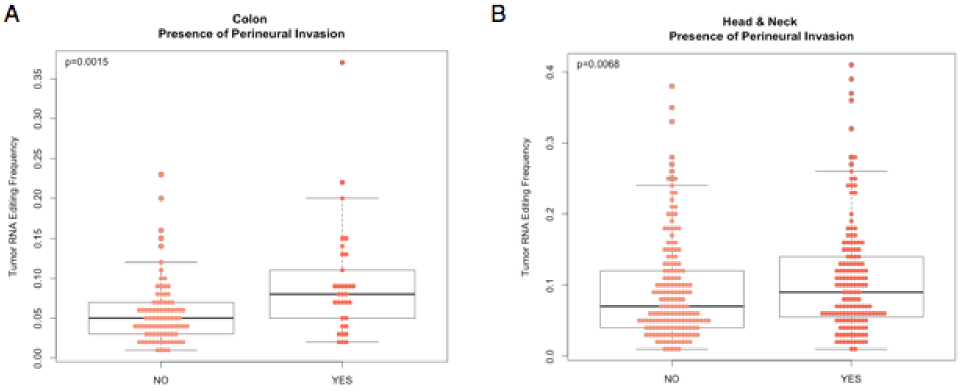


Figure B.4: Higher *AZIN1* EFs correlate with the presence of perineural invasion in (A) colon and (B) head and neck cancers.

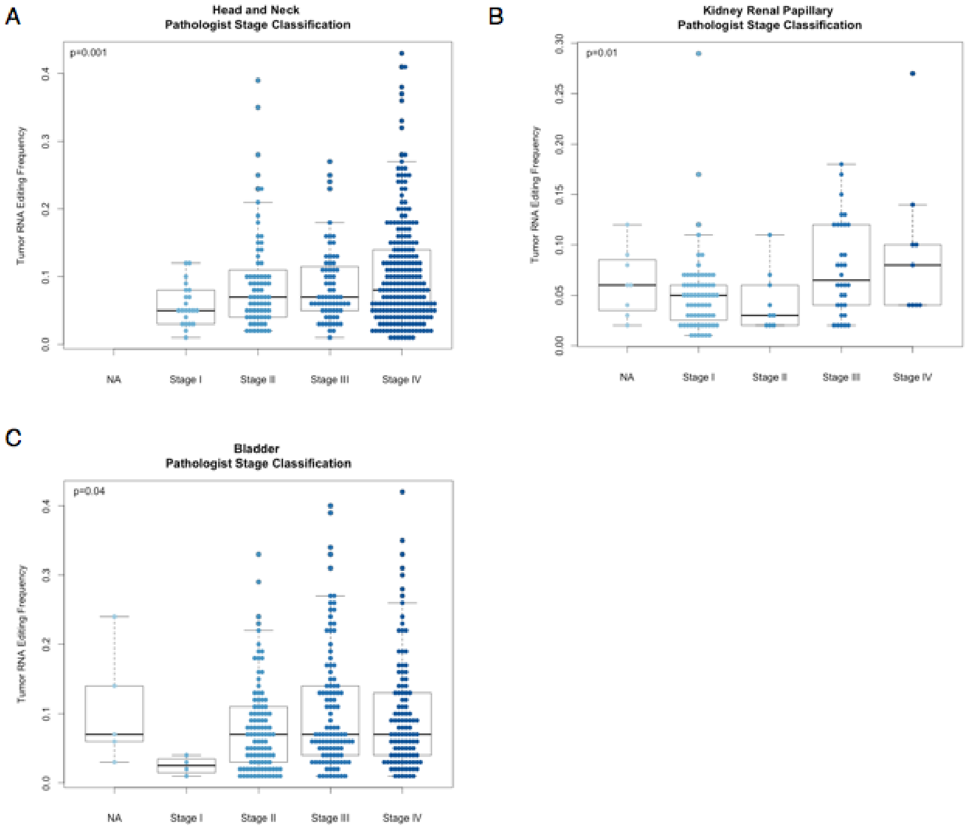


Figure B.5: Higher *AZIN1* EFs are associated with more advanced pathological and clinical stage classifications in (A) head and neck, (B) kidney renal papillary, and (C) bladder cancers.

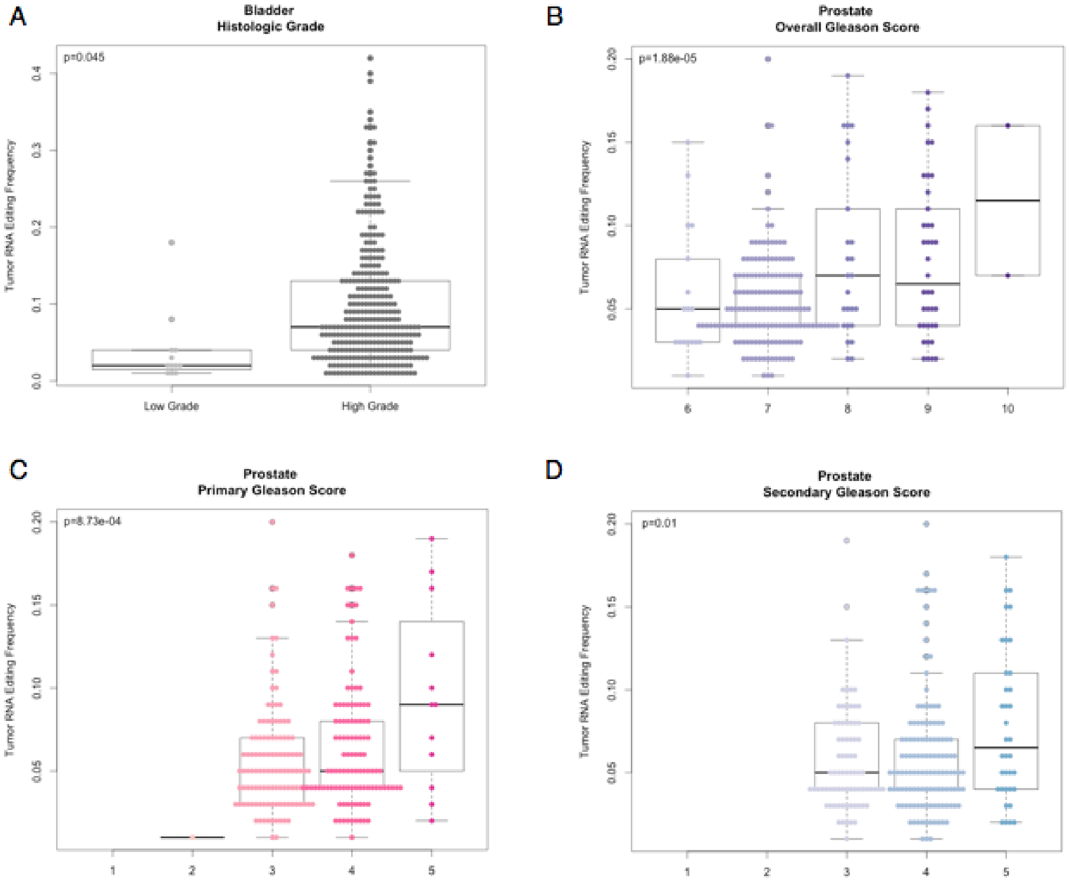


Figure B.6: Higher *AZIN1* EFs correlate with the “High Grade” histological grade in bladder cancer (A). In prostate cancer, higher *AZIN1* EFs are associated with the worst overall Gleason score (B) along with the largest primary (C) and secondary pattern scores (D).

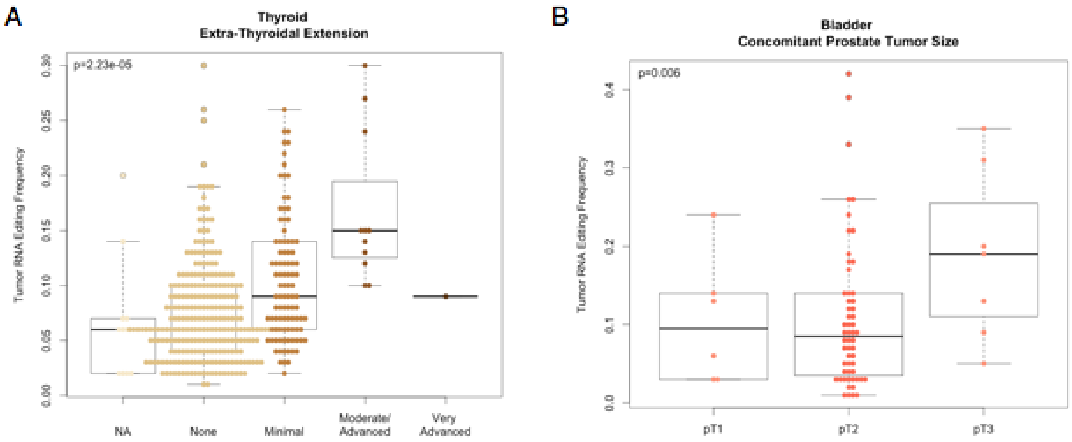


Figure B.7: Higher *AZIN1* EFs correlate with extra-thyroidal extension in thyroid cancer (A) and larger concomitant prostate tumor (T) size classifications in bladder cancer (B).

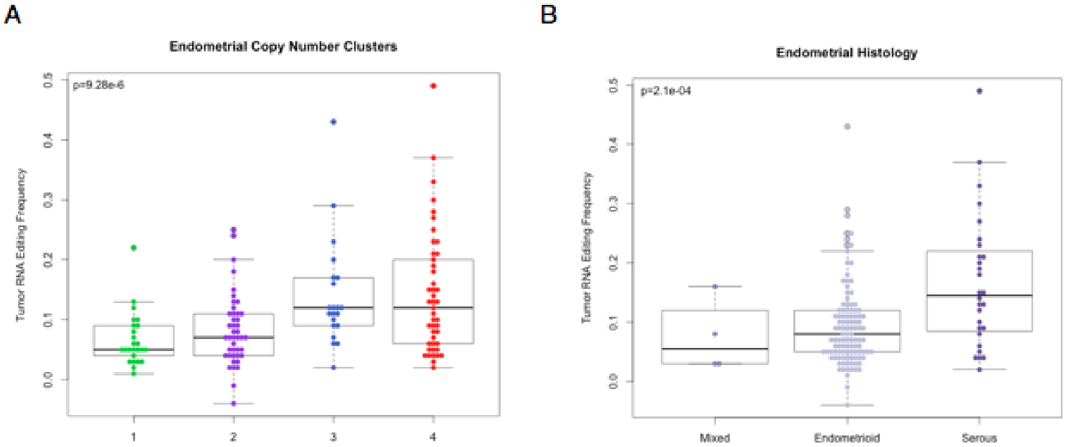


Figure B.8: Higher *AZIN1* EFs are associated with copy number aberration subtypes (A) and the serous histology types (B) in endometrial cancer.

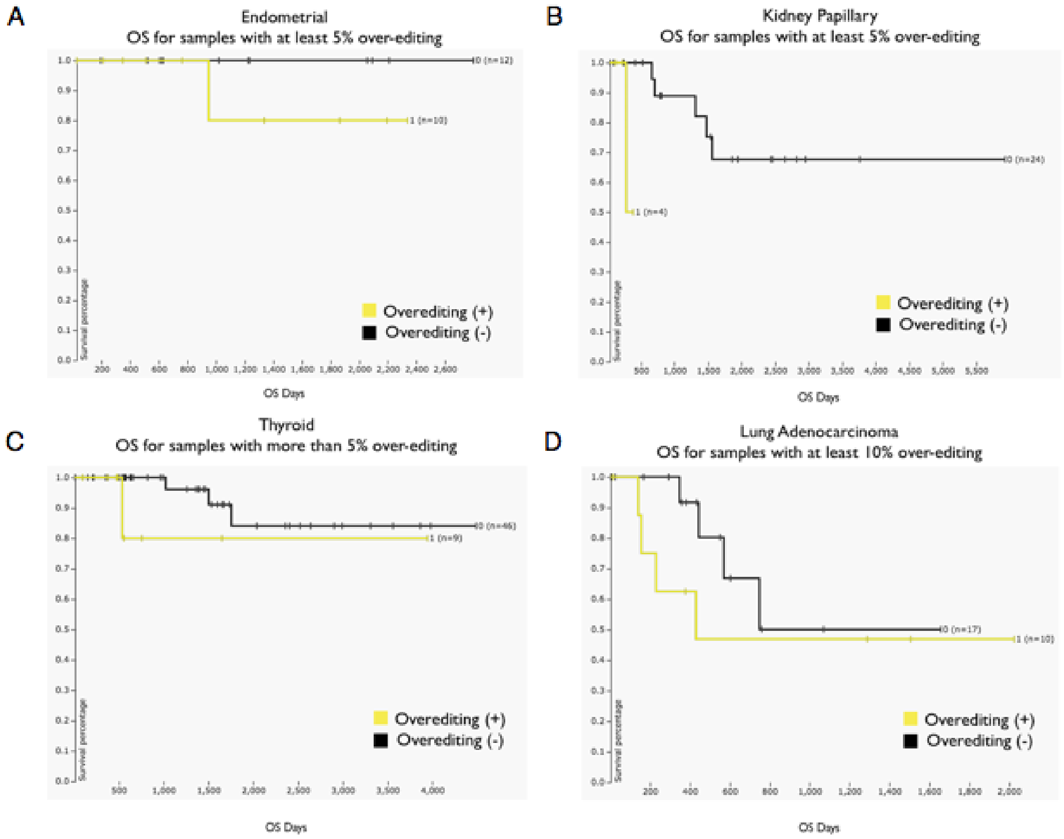


Figure B.9: Patients with at least 5% over-editing in (A) endometrial and (B) kidney, and more than 5% over-editing in (C) thyroid, and a minimum of 10% over-editing in (D) lung adenocarcinoma have a worse Overall Survival (OS).

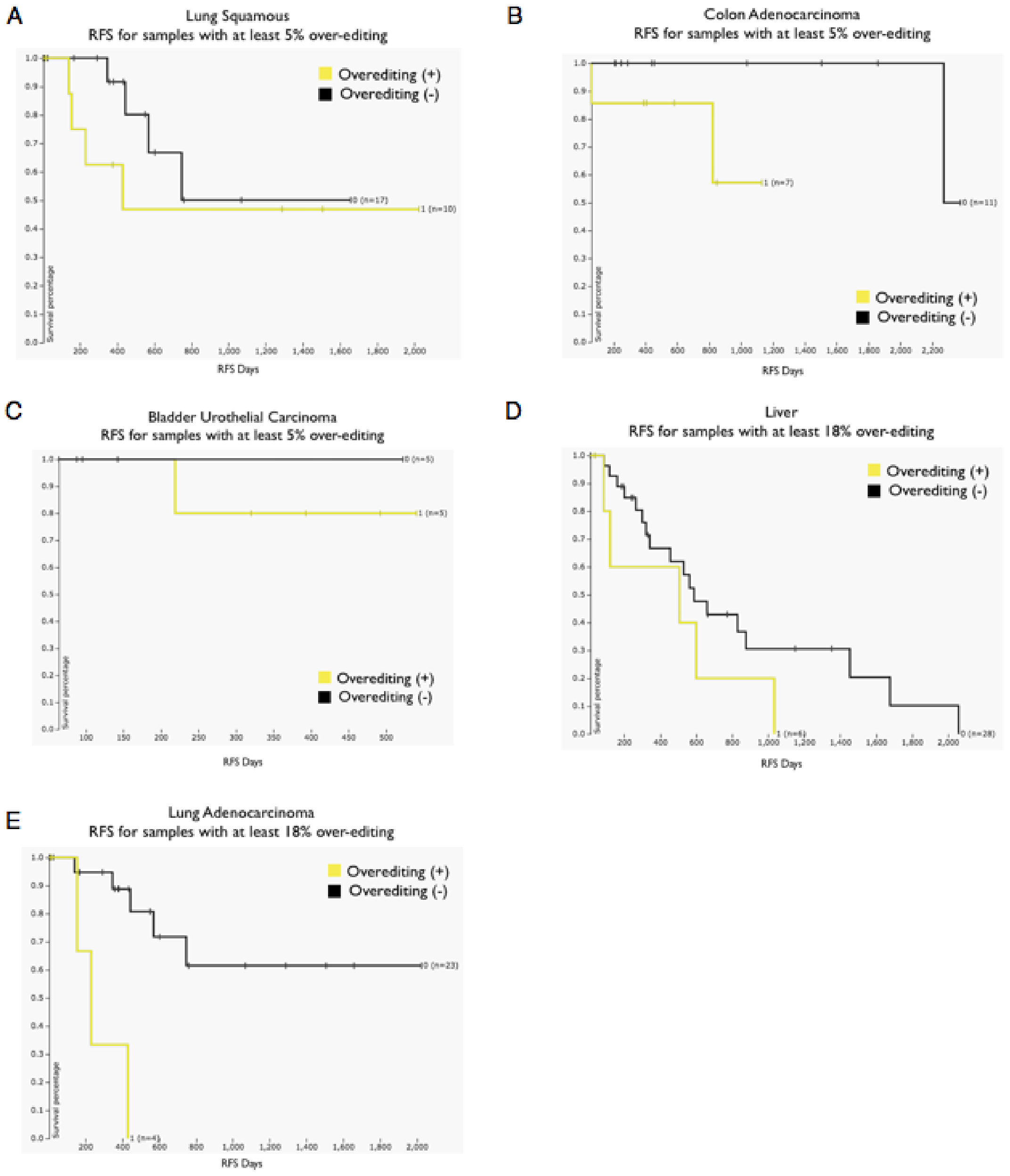


Figure B.10: Patients with at least 5% over-editing in (A) lung squamous, (B) colon, and (C) bladder, and a minimum of 18% over-editing in (D) liver, and (E) lung adenocarcinoma have a worse Recurrence Free Survival (RFS).

Bibliography

[1] O. D. Abaan, E. C. Polley, S. R. Davis, Y. J. Zhu, S. Bilke, R. L. Walker, M. Pineda, Y. Gindin, Y. Jiang, W. C. Reinhold, S. L. Holbeck, R. M. Simon, J. H. Doroshow, Y. Pommier, and P. S. Meltzer. The exomes of the nci-60 panel: a genomic resource for cancer biology and systems pharmacology. *Cancer Res*, 73(14):4372–82, 2013.

[2] L. Agranat, O. Raitskin, J. Sperling, and R. Sperling. The editing enzyme adar1 and the mrna surveillance protein hupf1 interact in the cell nucleus. *Proc Natl Acad Sci U S A*, 105(13):5028–33, 2008.

[3] J. Alsner, M. Yilmaz, P. Guldberg, L. L. Hansen, and J. Overgaard. Heterogeneity in the clinical phenotype of tp53 mutations in breast cancer patients. *Clin Cancer Res*, 6(10):3923–31, 2000.

[4] American Cancer Society. Cancer Facts & Figures 2011. Technical report, American Cancer Society, Atlanta, Georgia, 2011.

[5] American Cancer Society. Cancer Facts & Figures 2015. Technical report, American Cancer Society, Atlanta, Georgia, 2015.

[6] P. Andujar, J. Wang, A. Descatha, F. Galateau-Salle, I. Abd-Alsamad, M. A. Billon-Galland, H. Blons, B. Clin, C. Danel, B. Housset, P. Laurent-Puig, F. Le Pimpec-Barthes, M. Letourneux, I. Monnet, J. F. Regnard, A. Renier, J. Zucman-Rossi, J. C. Pairon, and M. C. Jaurand. p16ink4a inactivation mechanisms in non-small-cell lung cancer patients occupationally exposed to asbestos. *Lung Cancer*, 67(1):23–30, 2010.

[7] A. Athanasiadis, A. Rich, and S. Maas. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol.*, 2:e391, Dec 2004.

[8] The Cancer Genome Atlas. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–7, 2012.

[9] The Cancer Genome Atlas. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[10] The Cancer Genome Atlas. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492):315–22, 2014.

[11] The Cancer Genome Atlas. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–50, 2014.

[12] The Cancer Genome Atlas. Integrated genomic characterization of papillary thyroid carcinoma. *Cell*, 159(3):676–90, 2014.

[13] R. Baertsch, M. Diekhans, W. J. Kent, D. Haussler, and J. Brosius. Retrocopy contributions to the evolution of the human genome. *BMC Genomics*, 9:466, 2008.

[14] A. Beghini, C. B. Ripamonti, P. Peterlongo, G. Roversi, R. Cairoli, E. Morra, and L. Larizza. RNA hyperediting and alternative splicing of hematopoietic cell phosphatase (PTPN6) gene in acute myeloid leukemia. *Human Molecular Genetics*, 9(15):2297–2304, September 22 2000.

[15] A. Beghini, C. B. Ripamonti, P. Peterlongo, G. Roversi, R. Cairoli, E. Morra, and L. Larizza. Rna hyperediting and alternative splicing of hematopoietic cell phosphatase (ptpn6) gene in acute myeloid leukemia. *Hum Mol Genet*, 9(15):2297–304, 2000.

[16] A. V. Biankin, N. Waddell, K. S. Kassahn, M. C. Gingras, L. B. Muthuswamy, A. L. Johns, D. K. Miller, P. J. Wilson, A. M. Patch, J. Wu, D. K. Chang, M. J. Cowley, B. B. Gardiner, S. Song, I. Harliwong, S. Idrisoglu, C. Nourse, E. Nourbakhsh, S. Manning, S. Wani, M. Gongora, M. Pajic, C. J. Scarlett, A. J. Gill, A. V. Pinho, I. Rooman, M. Anderson, O. Holmes, C. Leonard, D. Taylor, S. Wood, Q. Xu, K. Nones, J. L. Fink, A. Christ, T. Bruxner, N. Cloonan, G. Kolle, F. Newell, M. Pinese, R. S. Mead, J. L. Humphris, W. Kaplan, M. D. Jones, E. K. Colvin, A. M. Nagrial, E. S. Humphrey, A. Chou, V. T. Chin, L. A. Chantrill, A. Mawson, J. S. Samra, J. G. Kench, J. A. Lovell, R. J. Daly, N. D. Merrett, C. Toon, K. Epari, N. Q. Nguyen, A. Barbour, N. Zeps, N. Kakkar, F. Zhao, Y. Q. Wu, M. Wang, D. M. Muzny, W. E. Fisher, F. C. Brunicardi, S. E. Hodges, J. G. Reid, J. Drummond, K. Chang, Y. Han, L. R. Lewis, H. Dinh, C. J. Buhay, T. Beck, L. Timms, M. Sam, K. Begley, A. Brown, D. Pai, A. Panchal, N. Buchner, R. De Borja, R. E. Denroche, C. K. Yung, S. Serra, N. Onetto, D. Mukhopadhyay, M. S. Tsao, P. A. Shaw, G. M. Petersen, S. Gallinger, R. H. Hruban, A. Maitra, C. A. Iacobuzio-Donahue, R. D. Schulick, C. L. Wolfgang, R. A. Morgan, et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature*, 491(7424):399–405, 2012.

[17] H. Blons, K. Pallier, D. Le Corre, C. Danel, M. Tremblay-Gravel, C. Houdayer, E. Fabre-Guillevin, M. Riquet, P. Dessen, and P. Laurent-Puig. Genome wide snp comparative analysis between egfr and kras mutated nsclc and characterization of two models of oncogenic cooperation in non-small cell lung carcinoma. *BMC Med Genomics*, 1:25, 2008.

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

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

[20] C. M. Burns, H. Chu, S. M. Rueter, L. K. Hutchinson, H. Canton, E. Sanders-Bush, and R. B. Emeson. Regulation of serotonin-2c receptor g-protein coupling by rna editing. *Nature*, 387(6630):303–8, 1997.

[21] C. Cenci, R. Barzotti, F. Galeano, S. Corbelli, R. Rota, L. Massimi, C. Di Rocco, M. A. O'Connell, and A. Gallo. Down-regulation of RNA editing in pediatric astrocytomas: ADAR2 editing activity inhibits cell migration and proliferation. *J. Biol. Chem.*, 283:7251–7260, Mar 2008.

[22] B. J. Chang, P. Lau, and L. Chan. Apolipoprotein b mrna editing. In Henri Grosjean and Rob Benne, editors, *Modification and Editing of RNA*, pages 325–342. ASM Press, 1998.

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

[24] L. Chen, Y. Li, C. H. Lin, T. H. Chan, R. K. Chow, Y. Song, M. Liu, Y. F. Yuan, L. Fu, K. L. Kong, L. Qi, N. Zhang, A. H. Tong, D. L. Kwong, K. Man, C. M. Lo, S. Lok, D. G. Tenen, and X. Y. Guan. Recoding rna editing of azin1 predisposes to hepatocellular carcinoma. *Nat Med*, 19(2):209–16, 2013.

[25] K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander, and G. Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*, 31(3):213–9, 2013.

[26] P. Cingolani, A. Platts, L. Wang le, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, 6(2):80–92, 2012.

[27] E. T. Cirulli, A. Singh, K. V. Shianna, D. Ge, J. P. Smith, J. M. Maia, E. L. Heinzen, J. J. Goedert, and D. B. Goldstein. Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *Genome Biol*, 11(5):R57, 2010.

[28] J. B. Clain, S. Scherl, L. Dos Reis, A. Turk, B. M. Wenig, S. Mehra, W. E. Karle, and M. L. Urken. Extrathyroidal extension predicts extranodal extension in patients with positive lymph nodes: an important association that may affect clinical management. *Thyroid*, 24(6):951–7, 2014.

[29] Graham A Colditz, Thomas A Sellers, and Edward Trapido. Epidemiology - identifying the causes and preventability of cancer? *Nat Rev Cancer*, 6(1):75–83, Jan 2006.

[30] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, 2010.

[31] 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[32] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, August 8 1970.

[33] H. Davies, C. Hunter, R. Smith, P. Stephens, C. Greenman, G. Bignell, J. Teague, A. Butler, S. Edkins, C. Stevens, A. Parker, S. O'Meara, T. Avis, S. Barthorpe, L. Brackenbury, G. Buck, J. Clements, J. Cole, E. Dicks, K. Edwards, S. Forbes, M. Gorton, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, D. Jones, V. Kosmidou, R. Laman, R. Lugg, A. Menzies, J. Perry, R. Petty, K. Raine, R. Shepherd, A. Small, H. Solomon, Y. Stephens, C. Tofts, J. Varian, A. Webb, S. West, S. Widaa, A. Yates, F. Brasseur, C. S. Cooper, A. M. Flanagan, A. Green, M. Knowles, S. Y. Leung, L. H. Looijenga, B. Malkowicz, M. A. Pierotti, B. T. Teh, S. T. Yuen, S. R. Lakhani, D. F. Easton, B. L. Weber, P. Goldstraw, A. G. Nicholson, R. Wooster, M. R. Stratton, and P. A. Futreal. Somatic mutations of the protein kinase gene family in human lung cancer. *Cancer Res*, 65(17):7591–5, 2005.

[34] J. Egebjerg, V. Kukekov, and S. F. Heinemann. Intron sequence directs RNA editing of the glutamate receptor subunit GluR2 coding sequence. *Proc. Natl. Acad. Sci. U.S.A.*, 91:10270–10274, Oct 1994.

[35] S. Elsheikh, A. R. Green, M. A. Aleskandarany, M. Grainge, C. E. Paish, M. B. Lambros, J. S. Reis-Filho, and I. O. Ellis. Ccnd1 amplification and cyclin d1 expression in breast cancer and their relation with proteomic subgroups and patient outcome. *Breast Cancer Res Treat*, 109(2):325–35, 2008.

[36] J. J. Fagan, B. Collins, L. Barnes, F. D'Amico, E. N. Myers, and J. T. Johnson. Perineural invasion in squamous cell carcinoma of the head and neck. *Arch Otolaryngol Head Neck Surg*, 124(6):637–40, 1998.

[37] L. Fernandez-Cuesta, C. Oakman, P. Falagan-Lotsch, K. S. Smoth, E. Quinaux, M. Buyse, M. S. Dolci, E. D. Azambuja, P. Hainaut, P. Dell'orto, D. Larsimont,

P. A. Francis, J. Crown, M. Piccart-Gebhart, G. Viale, A. D. Leo, and M. Olivier. Prognostic and predictive value of tp53 mutations in node-positive breast cancer patients treated with anthracycline- or anthracycline/taxane-based adjuvant therapy: results from the big 02-98 phase iii trial. *Breast Cancer Res*, 14(3):R70, 2012.

[38] P. N. Friedman, X. Chen, J. Bargonetti, and C. Prives. The p53 protein is an unusually shaped tetramer that binds directly to dna. *Proc Natl Acad Sci U S A*, 90(8):3319–23, 1993.

[39] A. Gallo and S. Galardi. A-to-I RNA editing and cancer: from pathology to basic science. *RNA Biol*, 5:135–139, 2008.

[40] A. Gallo, L. P. Keegan, G. M. Ring, and M. A. O'Connell. An ADAR that edits transcripts encoding ion channel subunits functions as a dimer. *EMBO J.*, 22:3421–3430, Jul 2003.

[41] A. L. Gartel, C. Feliciano, and A. L. Tyner. A new method for determining the status of p53 in tumor cell lines of different origin. *Oncol Res*, 13(6-10):405–8, 2003.

[42] A. Gerber, H. Grosjean, T. Melcher, and W. Keller. Tad1p, a yeast tRNA-specific adenosine deaminase, is related to the mammalian pre-mRNA editing enzymes ADAR1 and ADAR2. *EMBO J.*, 17:4780–4789, Aug 1998.

[43] J. M. Gott and R. B. Emeson. Functions and mechanisms of RNA editing. *AnnuRevGenet*, 34:499–531, 2000.

[44] R. Goya, M. G. Sun, R. D. Morin, G. Leung, G. Ha, K. C. Wiegand, J. Senz, A. Crisan, M. A. Marra, M. Hirst, D. Huntsman, K. P. Murphy, S. Aparicio, and S. P. Shah. Snvmix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, 26(6):730–6, 2010.

[45] Y. Gui, G. Guo, Y. Huang, X. Hu, A. Tang, S. Gao, R. Wu, C. Chen, X. Li, L. Zhou, M. He, Z. Li, X. Sun, W. Jia, J. Chen, S. Yang, F. Zhou, X. Zhao, S. Wan, R. Ye, C. Liang, Z. Liu, P. Huang, C. Liu, H. Jiang, Y. Wang, H. Zheng, L. Sun, X. Liu, Z. Jiang, D. Feng, S. Wu, J. Zou, Z. Zhang, R. Yang, J. Zhao, C. Xu, W. Yin, Z. Guan, J. Ye, H. Zhang, J. Li, K. Kristiansen, M. L. Nickerson, D. Theodorescu, Y. Li, X. Zhang, S. Li, J. Wang, H. Yang, and Z. Cai. Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. *Nat Genet*, 43(9):875–8, 2011.

[46] D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100:57–70, Jan 2000.

[47] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74, 2011.

[48] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo, and T. J. Hubbard. Gencode: the reference human genome annotation for the encode project. *Genome Res*, 22(9):1760–74, 2012.

[49] M. Higuchi, S. Maas, F. N. Single, J. Hartner, A. Rozov, N. Burnashev, D. Feldmeyer, R. Sprengel, and P. H. Seeburg. Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature*, 406:78–81, Jul 2000.

[50] M. Higuchi, F. N. Single, M. Kohler, B. Sommer, R. Sprengel, and P. H. Seeburg. RNA editing of AMPA receptor subunit GluR-B: a base-paired intron-exon structure determines position and efficiency. *Cell*, 75:1361–1370, Dec 1993.

[51] K. Holm, J. Staaf, G. Jonsson, J. Vallon-Christersson, H. Gunnarsson, A. Arason, L. Magnusson, R. B. Barkardottir, C. Hegardt, M. Ringner, and A. Borg. Characterisation of amplification patterns and target genes at chromosome 11q13 in ccnd1-amplified sporadic and familial breast tumours. *Breast Cancer Res Treat*, 133(2):583–94, 2012.

[52] R. F. Hough and B. L. Bass. Purification of the Xenopus laevis double-stranded RNA adenosine deaminase. *J. Biol. Chem.*, 269:9933–9939, Apr 1994.

[53] M. Imielinski, A. H. Berger, P. S. Hammerman, B. Hernandez, T. J. Pugh, E. Hodis, J. Cho, J. Suh, M. Capelletti, A. Sivachenko, C. Sougnez, D. Auclair, M. S. Lawrence, P. Stojanov, K. Cibulskis, K. Choi, L. de Waal, T. Sharifnia, A. Brooks, H. Greulich, S. Banerji, T. Zander, D. Seidel, F. Leenders, S. Ansen, C. Ludwig, W. Engel-Riedel, E. Stoelben, J. Wolf, C. Goparju, K. Thompson, W. Winckler, D. Kwiatkowski, B. E. Johnson, P. A. Janne, V. A. Miller, W. Pao, W. D. Travis, H. I. Pass, S. B. Gabriel, E. S. Lander, R. K. Thomas, L. A. Garraway, G. Getz, and M. Meyerson. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*, 150(6):1107–20, 2012.

[54] National Cancer Institute. Cancer staging, 01/06/2015 2015.

[55] National Cancer Institute. Tumor grade, 05/03/2013 2015.

[56] S. Ishiuchi, K. Tsuzuki, Y. Yoshida, N. Yamada, N. Hagimura, H. Okado, A. Miwa, H. Kurihara, Y. Nakazato, M. Tamura, T. Sasaki, and S. Ozawa. Blockage of Ca(2+)-permeable AMPA receptors suppresses migration and induces apoptosis in human glioblastoma cells. *Nat. Med.*, 8:971–978, Sep 2002.

[57] S. Ishiuchi, Y. Yoshida, K. Sugawara, M. Aihara, T. Ohtani, T. Watanabe, N. Saito, K. Tsuzuki, H. Okado, A. Miwa, Y. Nakazato, and S. Ozawa. Ca2+-permeable AMPA receptors regulate growth of human glioblastoma via Akt activation. *J. Neurosci.*, 27:7987–8001, Jul 2007.

[58] W. Jiang, S. M. Kahn, P. Zhou, Y. J. Zhang, A. M. Cacace, A. S. Infante, S. Doi, R. M. Santella, and I. B. Weinstein. Overexpression of cyclin d1 in rat fibroblasts causes abnormalities in growth control, cell cycle progression and gene expression. *Oncogene*, 8(12):3447–57, 1993.

[59] S. Jones, T. L. Wang, M. Shih Ie, T. L. Mao, K. Nakayama, R. Roden, R. Glas, D. Slamon, Jr. Diaz, L. A., B. Vogelstein, K. W. Kinzler, V. E. Velculescu, and N. Papadopoulos. Frequent mutations of chromatin remodeling gene arid1a in ovarian clear cell carcinoma. *Science*, 330(6001):228–31, 2010.

[60] S. Kalyana-Sundaram, C. Kumar-Sinha, S. Shankar, D. R. Robinson, Y. M. Wu, X. Cao, I. A. Asangani, V. Kothari, J. R. Prensner, R. J. Lonigro, M. K. Iyer, T. Barrette, A. Shanmugam, S. M. Dhanasekaran, N. Palanisamy, and A. M. Chinnaiyan. Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell*, 149(7):1622–34, 2012.

[61] Z. Kan, B. S. Jaiswal, J. Stinson, V. Janakiraman, D. Bhatt, H. M. Stern, P. Yue, P. M. Haverty, R. Bourgon, J. Zheng, M. Moorhead, S. Chaudhuri, L. P. Tomsho, B. A. Peters, K. Pujara, S. Cordes, D. P. Davis, V. E. Carlton, W. Yuan, L. Li, W. Wang, C. Eigenbrot, J. S. Kaminker, D. A. Eberhard, P. Waring, S. C. Schuster, Z. Modrusan, Z. Zhang, D. Stokoe, F. J. de Sauvage, M. Faham, and S. Seshagiri. Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature*, 466(7308):869–73, 2010.

[62] C. Kandoth, N. Schultz, A. D. Cherniack, R. Akbani, Y. Liu, H. Shen, A. G. Robertson, I. Pashtan, R. Shen, C. C. Benz, C. Yau, P. W. Laird, L. Ding, W. Zhang, G. B. Mills, R. Kucherlapati, E. R. Mardis, and D. A. Levine. Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73, 2013.

[63] D. Karolchik, G. P. Barber, J. Casper, H. Clawson, M. S. Cline, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, R. A. Harte, S. Heitner, A. S. Hinrichs, K. Learned, B. T. Lee, C. H. Li, B. J. Raney, B. Rhead, K. R. Rosenbloom, C. A. Sloan, M. L. Speir, A. S. Zweig, D. Haussler, R. M. Kuhn, and W. J. Kent. The ucsc genome browser database: 2014 update. *Nucleic Acids Res*, 42(Database issue):D764–70, 2014.

[64] J. Kato, H. Matsushime, S. W. Hiebert, M. E. Ewen, and C. J. Sherr. Direct binding of cyclin d to the retinoblastoma gene product (prb) and prb phosphorylation by the cyclin d-dependent kinase cdk4. *Genes Dev*, 7(3):331–42, 1993.

[65] L. P. Keegan, A. Gallo, and M. A. O'Connell. The many roles of an RNA editor. *Nat. Rev. Genet.*, 2:869–878, Nov 2001.

[66] F. S. Kenny, R. Hui, E. A. Musgrove, J. M. Gee, R. W. Blamey, R. I. Nicholson, R. L. Sutherland, and J. F. Robertson. Overexpression of cyclin d1 messenger rna predicts for poor prognosis in estrogen receptor-positive breast cancer. *Clin Cancer Res*, 5(8):2069–76, 1999.

[67] W. J. Kent. BLAT–the BLAST-like alignment tool. *Genome Res.*, 12:656–664, Apr 2002.

[68] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at UCSC. *Genome Res.*, 12:996–1006, Jun 2002.

[69] M. L. Khoo, S. Ezzat, J. L. Freeman, and S. L. Asa. Cyclin d1 protein expression predicts metastatic behavior in thyroid papillary microcarcinomas but is not associated with gene amplification. *J Clin Endocrinol Metab*, 87(4):1810–3, 2002.

[70] D. D. Kim, T. T. Kim, T. Walsh, Y. Kobayashi, T. C. Matise, S. Buyske, and A. Gabriel. Widespread rna editing of embedded alu elements in the human transcriptome. *Genome Res*, 14(9):1719–25, 2004.

[71] Dennis D. Y. Kim, Thomas T. Y. Kim, Thomas Walsh, Yoshifumi Kobayashi, Tara C. Matise, Steven Buyske, and Abram Gabriel. Widespread RNA editing of embedded *Alu* elements in the human transcriptome. *Genome Research*, 14(9):1719–1725, September 2004.

[72] T. H. Kim, S. Y. Lee, J. H. Rho, N. Y. Jeong, Y. H. Soung, W. S. Jo, D. Y. Kang, S. H. Kim, and Y. H. Yoo. Mutant p53 (g199v) gains antiapoptotic function through signal transducer and activator of transcription 3 in anaplastic thyroid cancer cells. *Mol Cancer Res*, 7(10):1645–54, 2009.

[73] U. Kim, Y. Wang, T. Sanford, Y. Zeng, and K. Nishikura. Molecular cloning of cDNA for double-stranded RNA adenosine deaminase, a candidate enzyme for nuclear RNA editing. *Proc. Natl. Acad. Sci. U.S.A.*, 91:11457–11461, Nov 1994.

[74] D. C. Koboldt, K. Chen, T. Wylie, D. E. Larson, M. D. McLellan, E. R. Mardis, G. M. Weinstock, R. K. Wilson, and L. Ding. Varscan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17):2283–5, 2009.

[75] D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson. Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*, 22(3):568–76, 2012.

[76] C. S. Ku, M. Wu, D. N. Cooper, N. Naidoo, Y. Pawitan, B. Pang, B. Iacopetta, and R. Soong. Exome versus transcriptome sequencing in identifying coding region variants. *Expert Rev Mol Diagn*, 12(3):241–51, 2012.

[77] M. Kumar and G. G. Carmichael. Nuclear antisense rna induces extensive adenosine modifications and nuclear retention of target transcripts. *Proc Natl Acad Sci U S A*, 94(8):3542–7, 1997.

[78] F. Lai, C. X. Chen, V. M. Lee, and K. Nishikura. Dramatic increase of the RNA editing for glutamate receptor subunits during terminal differentiation of clonal human neurons. *J. Neurochem.*, 69:43–52, Jul 1997.

[79] D. E. Larson, C. C. Harris, K. Chen, D. C. Koboldt, T. E. Abbott, D. J. Dooling, T. J. Ley, E. R. Mardis, R. K. Wilson, and L. Ding. Somaticsniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–7, 2012.

[80] E. Y. Levanon, E. Eisenberg, R. Yelin, S. Nemzer, M. Hallegger, R. Shemesh, Z. Y. Fligelman, A. Shoshan, S. R. Pollock, D. Sztybel, M. Olshansky, G. Rechavi, and M. F. Jantsch. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat. Biotechnol.*, 22:1001–1005, Aug 2004.

[81] E. Y. Levanon, E. Eisenberg, R. Yelin, S. Nemzer, M. Hallegger, R. Shemesh, Z. Y. Fligelman, A. Shoshan, S. R. Pollock, D. Sztybel, M. Olshansky, G. Rechavi, and M. F. Jantsch. Systematic identification of abundant a-to-i editing sites in the human transcriptome. *Nat Biotechnol*, 22(8):1001–5, 2004.

[82] H. Li. Improving snp discovery by base alignment quality. *Bioinformatics*, 27(8):1157–8, 2011.

[83] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–60, 2009.

[84] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 15 2009.

[85] Jin Billy Li, Erez Y. Levanon, Jung-Ki Yoon, John Aach, Bin Xie, Emily Leproust, Kun Zhang, Yuan Gao, and George M. Church. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science*, 324(5931):1210–1213, May 29 2009.

[86] J. Lindberg, I. G. Mills, D. Klevebring, W. Liu, M. Neiman, J. Xu, P. Wikstrom, P. Wiklund, F. Wiklund, L. Egevad, and H. Gronberg. The mitochondrial and autosomal mutation landscapes of prostate cancer. *Eur Urol*, 63(4):702–8, 2013.

[87] D. J. Luciano, H. Mirsky, N. J. Vendetti, and S. Maas. Rna editing of a mirna precursor. *RNA*, 10(8):1174–7, 2004.

[88] A. S. Lundberg and R. A. Weinberg. Functional inactivation of the retinoblastoma protein requires sequential modification by at least two distinct cyclin-cdk complexes. *Mol Cell Biol*, 18(2):753–61, 1998.

[89] S. Maas, Y. Kawahara, K. M. Tamburro, and K. Nishikura. A-to-I RNA editing and human disease. *RNA Biol*, 3:1–9, 2006.

[90] S. Maas, S. Patt, M. Schrey, and A. Rich. Underediting of glutamate receptor GluR-B mRNA in malignant gliomas. *Proc. Natl. Acad. Sci. U.S.A.*, 98:14687–14692, Dec 2001.

[91] T. Melcher, S. Maas, A. Herb, R. Sprengel, M. Higuchi, and P. H. Seeburg. RED2, a brain-specific member of the RNA-specific adenosine deaminase family. *J. Biol. Chem.*, 271:31795–31798, Dec 1996.

[92] T. Melcher, S. Maas, A. Herb, R. Sprengel, P. H. Seeburg, and M. Higuchi. A mammalian RNA editing enzyme. *Nature*, 379:460–464, Feb 1996.

[93] T. Melcher, S. Maas, M. Higuchi, W. Keller, and P. H. Seeburg. Editing of alpha-amino-3-hydroxy-5-methylisoxazole-4-propionic acid receptor GluR-B pre-mRNA in vitro reveals site-selective adenosine to inosine conversion. *J. Biol. Chem.*, 270:8566–8570, Apr 1995.

[94] D. P. Morse, P. J. Aruscavage, and B. L. Bass. RNA hairpins in noncoding regions of human brain and Caenorhabditis elegans mRNA are edited by adenosine deaminases that act on RNA. *Proc. Natl. Acad. Sci. U.S.A.*, 99:7906–7911, Jun 2002.

[95] D. P. Morse and B. L. Bass. Long RNA hairpins that contain inosine are present in Caenorhabditis elegans poly(A)+ RNA. *Proc. Natl. Acad. Sci. U.S.A.*, 96:6048–6053, May 1999.

[96] P. A. Muller and K. H. Vousden. p53 mutations in cancer. *Nat Cell Biol*, 15(1):2–8, 2012.

[97] E. A. Musgrove, C. S. Lee, M. F. Buckley, and R. L. Sutherland. Cyclin d1 induction in breast cancer cells shortens g1 and is sufficient for cells arrested in g1 to complete the cell cycle. *Proc Natl Acad Sci U S A*, 91(17):8022–6, 1994.

[98] L. Musumeci, J. W. Arthur, F. S. Cheung, A. Hoque, S. Lippman, and J. K. Reichardt. Single nucleotide differences (snds) in the dbsnp database may lead to errors in genotyping and haplotyping studies. *Hum Mutat*, 31(1):67–73, 2010.

[99] S. R. Nagalla, B. J. Barry, and E. R. Spindel. Cloning of complementary DNAs encoding the amphibian bombesin-like peptides Phe8 and Leu8 phyllolitorin from Phyllomedusa sauvagei: potential role of U to C RNA editing in generating neuropeptide diversity. *Mol. Endocrinol.*, 8:943–951, Aug 1994.

[100] E. Nagy, Z. Beck, A. Kiss, E. Csoma, B. Telek, J. Konya, E. Olah, K. Rak, and F. D. Toth. Frequent methylation of p16ink4a and p14arf genes implicated in the evolution of chronic myeloid leukaemia from its chronic to accelerated phase. *Eur J Cancer*, 39(16):2298–305, 2003.

[101] K. Nishikura. Functions and regulation of rna editing by adar deaminases. *Annu Rev Biochem*, 79:321–49, 2010.

[102] K. Nishikura, C. Yoo, U. Kim, J. M. Murray, P. A. Estes, F. E. Cash, and S. A. Liebhaber. Substrate specificity of the dsrna unwinding/modifying activity. *EMBO J*, 10(11):3523–32, 1991.

[103] M. A. O'Connell and W. Keller. Purification and properties of double-stranded RNA-specific adenosine deaminase from calf thymus. *Proc. Natl. Acad. Sci. U.S.A.*, 91:10596–10600, Oct 1994.

[104] M. A. O'Connell, S. Krause, M. Higuchi, J. J. Hsuan, N. F. Totty, A. Jenny, and W. Keller. Cloning of cDNAs encoding mammalian double-stranded RNA-specific adenosine deaminase. *Mol. Cell. Biol.*, 15:1389–1397, Mar 1995.

[105] H. Okada and T. W. Mak. Pathways of apoptotic and non-apoptotic death in tumour cells. *Nat. Rev. Cancer*, 4:592–603, Aug 2004.

[106] N. Paz, E. Y. Levanon, N. Amariglio, A. B. Heimberger, Z. Ram, S. Constantini, Z. S. Barbash, K. Adamsky, M. Safran, A. Hirschberg, M. Krupsky, I. Ben-Dov, S. Cazacu, T. Mikkelsen, C. Brodie, E. Eisenberg, and G. Rechavi. Altered adenosine-to-inosine RNA editing in human cancer. *Genome Res.*, 17:1586–1595, Nov 2007.

[107] N. Paz, E. Y. Levanon, N. Amariglio, A. B. Heimberger, Z. Ram, S. Constantini, Z. S. Barbash, K. Adamsky, M. Safran, A. Hirschberg, M. Krupsky, I. Ben-Dov, S. Cazacu, T. Mikkelsen, C. Brodie, E. Eisenberg, and G. Rechavi. Altered adenosine-to-inosine rna editing in human cancer. *Genome Res*, 17(11):1586–95, 2007.

[108] E. Pfaff, M. Remke, D. Sturm, A. Benner, H. Witt, T. Milde, A. O. von Bueren, A. Wittmann, A. Schottler, N. Jorch, N. Graf, A. E. Kulozik, O. Witt, W. Scheurlen, A. von Deimling, S. Rutkowski, M. D. Taylor, U. Tabori, P. Lichter, A. Korshunov, and S. M. Pfister. Tp53 mutation is frequently associated with ctnnb1 mutation or mycn amplification and is compatible with long-term survival in medulloblastoma. *J Clin Oncol*, 28(35):5188–96, 2010.

[109] A. G. Polson and B. L. Bass. Preferential selection of adenosines for modification by double-stranded RNA adenosine deaminase. *EMBO J.*, 13:5701–5711, Dec 1994.

[110] A. G. Polson and B. L. Bass. Preferential selection of adenosines for modification by double-stranded rna adenosine deaminase. *EMBO J*, 13(23):5701–11, 1994.

[111] A. G. Polson, P. F. Crain, S. C. Pomerantz, J. A. McCloskey, and B. L. Bass. The mechanism of adenosine to inosine conversion by the double-stranded RNA unwinding/modifying activity: a high-performance liquid chromatography-mass spectrometry analysis. *Biochemistry*, 30:11507–11514, Dec 1991.

[112] H. Poulsen, R. Jorgensen, A. Heding, F. C. Nielsen, B. Bonven, and J. Egebjerg. Dimerization of ADAR2 is mediated by the double-stranded RNA binding domain. *RNA*, 12:1350–1360, Jul 2006.

[113] K. V. Prasanth, S. G. Prasanth, Z. Xuan, S. Hearn, S. M. Freier, C. F. Bennett, M. Q. Zhang, and D. L. Spector. Regulating gene expression through rna nuclear retention. *Cell*, 123(2):249–63, 2005.

[114] A. J. Radenbaugh, S. Ma, A. Ewing, J. M. Stuart, E. A. Collisson, J. Zhu, and D. Haussler. Radia: Rna and dna integrated analysis for somatic mutation detection. *PLoS One*, 9(11):e111516, 2014.

[115] E. M. Riedmann, S. Schopoff, J. C. Hartner, and M. F. Jantsch. Specificity of ADAR-mediated RNA editing in newly identified targets. *RNA*, 14:1110–1118, Jun 2008.

[116] N. D. Roberts, R. D. Kortschak, W. T. Parker, A. W. Schreiber, S. Branford, H. S. Scott, G. Glonek, and D. L. Adelson. A comparative analysis of algorithms for somatic snv detection in cancer. *Bioinformatics*, 29(18):2223–30, 2013.

[117] G. Robinson, M. Parker, T. A. Kranenburg, C. Lu, X. Chen, L. Ding, T. N. Phoenix, E. Hedlund, L. Wei, X. Zhu, N. Chalhoub, S. J. Baker, R. Huether, R. Kriwacki, N. Curley, R. Thiruvenkatam, J. Wang, G. Wu, M. Rusch, X. Hong, J. Becksfort, P. Gupta, J. Ma, J. Easton, B. Vadodaria, A. Onar-Thomas, T. Lin, S. Li, S. Pounds, S. Paugh, D. Zhao, D. Kawauchi, M. F. Roussel, D. Finkelstein, D. W. Ellison, C. C. Lau, E. Bouffet, T. Hassall, S. Gururangan, R. Cohn, R. S. Fulton, L. L. Fulton, D. J. Dooling, K. Ochoa, A. Gajjar, E. R. Mardis, R. K. Wilson, J. R. Downing, J. Zhang, and R. J. Gilbertson. Novel mutations target distinct subgroups of medulloblastoma. *Nature*, 488(7409):43–8, 2012.

[118] K. R. Rosenbloom, C. A. Sloan, V. S. Malladi, T. R. Dreszer, K. Learned, V. M. Kirkup, M. C. Wong, M. Maddren, R. Fang, S. G. Heitner, B. T. Lee, G. P. Barber, R. A. Harte, M. Diekhans, J. C. Long, S. P. Wilder, A. S. Zweig, D. Karolchik, R. M. Kuhn, D. Haussler, and W. J. Kent. Encode data in the ucsc genome browser: year 5 update. *Nucleic Acids Res*, 41(Database issue):D56–63, 2013.

[119] S. M. Rueter, C. M. Burns, S. A. Coode, P. Mookherjee, and R. B. Emeson. Glutamate receptor RNA editing in vitro by enzymatic conversion of adenosine to inosine. *Science*, 267:1491–1494, Mar 1995.

[120] S. M. Rueter, T. R. Dawson, and R. B. Emeson. Regulation of alternative splicing by RNA editing. *Nature*, 399:75–80, May 1999.

[121] S.M. Rueter and R.B. Emeson. Adenosine-to-inosine conversion in mrna. In Henri Grosjean and Rob Benne, editors, *Modification and Editing of RNA*, pages 343–361. ASM Press, 1998.

[122] L. H. Saal, S. K. Gruvberger-Saal, C. Persson, K. Lovgren, M. Jumppanen, J. Staaf, G. Jonsson, M. M. Pires, M. Maurer, K. Holm, S. Koujak, S. Subramaniyam, J. Vallon-Christersson, H. Olsson, T. Su, L. Memeo, T. Ludwig, S. P. Ethier, M. Krogh, M. Szabolcs, V. V. Murty, J. Isola, H. Hibshoosh, R. Parsons, and A. Borg. Recurrent gross mutations of the PTEN tumor suppressor gene in breast cancers with deficient DSB repair. *Nat. Genet.*, 40:102–107, Jan 2008.

[123] C. T. Saunders, W. S. Wong, S. Swamy, J. Becq, L. J. Murray, and R. K. Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28(14):1811–7, 2012.

[124] P. H. Seeburg, M. Higuchi, and R. Sprengel. Rna editing of brain glutamate receptor channels: mechanism and physiology. *Brain Res Brain Res Rev*, 26(2-3):217–29, 1998.

[125] S. P. Shah, R. D. Morin, J. Khattra, L. Prentice, T. Pugh, A. Burleigh, A. Delaney, K. Gelmon, R. Guliany, J. Senz, C. Steidl, R. A. Holt, S. Jones, M. Sun, G. Leung, R. Moore, T. Severson, G. A. Taylor, A. E. Teschendorff, K. Tse, G. Turashvili, R. Varhol, R. L. Warren, P. Watson, Y. Zhao, C. Caldas, D. Huntsman, M. Hirst, M. A. Marra, and S. Aparicio. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, 461:809–813, Oct 2009.

[126] P. M. Sharma, M. Bowman, S. L. Madden, F. J. Rauscher, and S. Sukumar. RNA editing in the Wilms' tumor susceptibility gene, WT1. *Genes Dev.*, 8:720–731, Mar 1994.

[127] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res*, 29(1):308–11, 2001.

[128] American Cancer Society. How is bladder cancer diagnosed?, 2015.

[129] S. Sporny, D. Slowinska-Klencka, and M. Ratynska. Cyclin d1 expression in primary thyroid carcinomas. *Neuro Endocrinol Lett*, 26(6):815–8, 2005.

[130] M. R. Stratton. Exploring the genomes of cancer cells: progress and promise. *Science*, 331:1553–1558, Mar 2011.

[131] J. P. Struewing, P. Hartge, S. Wacholder, S. M. Baker, M. Berlin, M. McAdams, M. M. Timmerman, L. C. Brody, and M. A. Tucker. The risk of cancer associated

with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *N. Engl. J. Med.*, 336:1401–1408, May 1997.

[132] T. Takahira, Y. Oda, S. Tamiya, H. Yamamoto, K. Kawaguchi, C. Kobayashi, Y. Iwamoto, and M. Tsuneyoshi. Alterations of the p16ink4a/p14arf pathway in clear cell sarcoma. *Cancer Sci*, 95(8):651–5, 2004.

[133] T. Ueki, A. W. Hsing, Y. T. Gao, B. S. Wang, M. C. Shen, J. Cheng, J. Deng, Jr. Fraumeni, J. F., and A. Rashid. Alterations of p16 and prognosis in biliary tract cancers from a population-based study in china. *Clin Cancer Res*, 10(5):1717–25, 2004.

[134] L. Valente and K. Nishikura. RNA binding-independent dimerization of adenosine deaminases acting on RNA and dominant negative effects of nonfunctional subunits on dimer functions. *J. Biol. Chem.*, 282:16054–16061, Jun 2007.

[135] R. W. Wagner, J. E. Smith, B. S. Cooperman, and K. Nishikura. A double-stranded rna unwinding activity introduces structural alterations by means of adenosine to inosine conversions in mammalian cells and xenopus eggs. *Proc Natl Acad Sci U S A*, 86(8):2647–51, 1989.

[136] K. Wang, D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich, X. He, P. Mieczkowski, S. A. Grimm, C. M. Perou, J. N. MacLeod, D. Y. Chiang, J. F. Prins, and J. Liu. Mapsplice: accurate mapping of rna-seq reads for splice junction discovery. *Nucleic Acids Res*, 38(18):e178, 2010.

[137] Q. Wang, M. Miyakoda, W. Yang, J. Khillan, D. L. Stachura, M. J. Weiss, and K. Nishikura. Stress-induced apoptosis associated with null mutation of adar1 rna editing deaminase gene. *J Biol Chem*, 279(6):4952–61, 2004.

[138] Q. Wang, Z. Zhang, K. Blackwell, and G. G. Carmichael. Vigilins bind to promiscuously a-to-i-edited rnas and are involved in the formation of heterochromatin. *Curr Biol*, 15(4):384–91, 2005.

[139] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, Jan 2009.

[140] R. A. Weinberg. The retinoblastoma protein and cell cycle control. *Cell*, 81(3):323–30, 1995.

[141] J. H. Yang, P. Sklar, R. Axel, and T. Maniatis. Editing of glutamate receptor subunit B pre-mRNA in vitro by site-specific deamination of adenosine. *Nature*, 374:77–81, Mar 1995.

[142] S. Yano, Y. Nishioka, H. Goto, and S. Sone. Molecular mechanisms of angiogenesis in non-small cell lung cancer, and therapeutics targeting related molecules. *Cancer Sci.*, 94:479–485, Jun 2003.

[143] Y. J. Zhang, W. Jiang, C. J. Chen, C. S. Lee, S. M. Kahn, R. M. Santella, and I. B. Weinstein. Amplification and overexpression of cyclin d1 in human hepatocellular carcinoma. *Biochem Biophys Res Commun*, 196(2):1010–6, 1993.

[144] H. Zheng, T. B. Fu, D. Lazinski, and J. Taylor. Editing on the genomic rna of human hepatitis delta virus. *J Virol*, 66(8):4693–7, 1992.