A Novel Statistical Framework for the Accurate Identification of RNA-edits with Application to Human Cancers

by

Ryan S. Giuliany

B.Sc, The University of British Columbia, 2010

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE STUDIES
(Bioinformatics)

The University Of British Columbia
(Vancouver)

July 2012

Abstract

RNA-editing is the post-transcriptional, enzymatic modification of RNA molecules resulting in an altered nucleotide sequence. These modifications play a critical role in mammalian tissues and are essential for proper function of liver and neuronal development, among other processes. The advent of high-throughput sequencing (HTS) technologies (e.g. Illumina HiSeq) has renewed interest in RNA-editing discovery due to unprecedented opportunities for simultaneous interrogation of whole genome and transcriptome sequences. In the past several months a number of studies have been published describing methods and results of RNA-editing discovery in HTS data. These methods have been ad hoc approaches based on repurposing SNP calling tools designed for genome-based variant detection. However, the statistical properties of RNA-editing warrant specialized analytical strategies that leverage the non-uniform substitution distributions inherent in RNA-editing processes.

A novel statistical framework, called Auditor, that simultaneously analyzes the genomic and transcriptomic base-counts and infers the likelihood of an RNA-edit at each position in the transcriptome is reported. This model leverages the inherent correlation present in the RNA and DNA sequence while encoding the non-uniform substitution distributions induced by RNA-editing, conferring increased sensitivity. Further, a Random-Forest based technical artifact removal tool that accurately identifies sequencing and alignment errors has been implemented, greatly increasing the specificity of the method. The combination of these approaches leads to a robust, principled method that accurately detects RNA-edits in the presence of both biological and technical noise.

It is systematically shown, in both a simulation study and on real matched whole genome and transcriptome data generated from 11 lymphoma samples, that Auditor significantly outperforms similar, but simpler statistical frameworks, including a Samtools/bcftools based approach that is similar to a recently published study. Finally by profiling 11 diffuse large B-cell lymphomas and 16 triple negative breast cancers with Auditor, it is shown that RNA-editing is an active process in human malignancies. Surprisingly, consistent patterns of nucleotide substitutions and regional enrichment of RNA-edits in 3 UTRs suggests that RNA-editing processes are invariant between cell lineages and between tumours of similar histological subtypes and even cancers from distinct tissues of origin.

Table of Contents

List of Tables

List of Figures

Acknowledgments

My thanks...

Thank you to David, for believing in me from day one.

Thank you to Sohrab, for giving me that job years ago...and for being my mentor ever since.

Thank you to Mom and Dad, for 25 years of unending support and encouragement.

Thank you to Sarah for being there for me. I can never thank you enough.

Chapter 1

Introduction

The molecular flow of information in biology is well-established: DNA is transcribed into RNA which in turn is translated into protein products. However, over the past several decades mechanisms that violate this so-called central dogma have emerged. The discovery of reverse transcriptase, allowing DNA sequence to be created from an RNA template, in 1970 (1, 2) was initially met with scepticism and resistance by the scientific community; however, a massive body of research led to deep understanding of the mechanisms and fundamental importance of the process. In the 1980's reports began to emerge of other enzymes that violated the central dogma, breaking the one-to-one mapping of RNA sequence from a given DNA template, by post-transcriptionally modifying the primary RNA sequence (3, 4, 5). Unlike reverse transcription, these processes, known as RNA-editing, were not the subject of wide-spread, coordinated, research efforts and are still comparatively understudied.

Despite the lack of comprehensive interest, the studies that were conducted led to understanding of some mechanisms and to identification of the enzymes responsible for the most common forms of RNA-editing, as well as some of the consequences of sub- or abnormal activity of these processes (6, 7). However, this research is far from exhaustive, and many aspects of RNA-editing, including the potential role in tumour pathogenesis, remain unclear or unknown.

Until recently, attempting to identify RNA-editing transcriptome wide has been impractical due to the cost and labour involved in sequencing the matched whole-genome and whole-transcriptome of the same sample. However, with the advent of high-throughput sequencing technologies, such procedures are now practical for not only single samples, but are also possible on the scale of thousands of samples. The emergence of this data has led to the rapid development of computational tools designed to align this data to reference genomes and to accurately analyze this data in order to infer the nucleotide sequence, and individual nucleotide positions that vary from the reference genome.

Unsurprisingly, there has been a corresponding attempt to use these variant calling methods to detect RNA-edits from matched DNA/RNA samples. Unfortunately, these attempts have led in some cases to questionable results, and it has become clear that RNA-editing analyses are affected by a substantial false positive burden due to technical artifacts. To date, no principled method has

emerged that effectively considers the known properties of RNA-editing and appropriately identifies technical artifacts.

This thesis seeks to address two primary goals: to demonstrate a method of RNA-editing detection that substantially improves over previous efforts and to apply this method to a set of human cancer samples to explore the landscape of RNA-editing in human malignancy. To this effect, this thesis describes the following contributions:

1. A statistical framework that systematically addresses the shortcomings of previous methods
2. The implementation of this framework in a distributable software package
3. The analysis of twenty-seven human cancer samples, resulting in the first landscape of RNA-editing in diffuse large B-cell lymphoma and triple negative breast cancer

1.1 Mechanisms of RNA-editing

The RNA-editing process is thought to be wide-spread in eukaryotes (8) and universal among metazoans (9). However, among different kingdoms, RNA-editing can result in different modifications. For example, in *Trypanosoma brucei*, RNA-editing most commonly appears as the insertion or deletion of uracil (10). In mammals however, RNA-editing has only been observed in the form of base substitutions achieved by enzymatic deamination. Despite the potential benefits of better understanding indel-based RNA-editing, this thesis focuses on the specific problem of base-substitution RNA-editing in the context of mammalian samples.

1.1.1 Adenosine Deaminase Acting on RNA (ADAR)

Members of the ADAR gene family have been shown to enzymatically deaminate adenosine (A) to inosine (I), and are thought to be responsible for more than 90% of RNA-editing events (11, 12). ADARs require a double-stranded RNA (dsRNA) substrate and are thought to be highly promiscuous with respect to primary sequence motif (5), though there have been reports of weak preferences for certain bases surrounding the edit site (13, 14, 15).

The patterns of A→I RNA-editing are known to be stochastic (11), further reflecting the promiscuity of the ADAR gene family. In general, ADARs are thought to bind to any appropriate dsRNA substrate, complete the deamination and release. However, proximity to the same dsRNA substrate results in a high probability of re-binding the same dsRNA molecule at a new adenosine and repeating the deamination process. This can repeat until the dsRNA molecule is sufficiently deaminated to break the double stranded structure (as inosine does not pair with uracil), or until either the enzyme or the RNA is removed from proximity (5). Repeat structures such as *Alu* elements are common targets of RNA-editing due to their propensity to form secondary structures and RNA-edits within *Alu* repeats account for a substantial proportion of known events (16).

Inosine, despite frequently being referred to as a ”wobble-base” (because it is incorporated into tRNA molecules in the third base of the anti-codon), is interpreted by the majority of cellular

machinery as a guanine (G). Inosine is also observed as a guanine by nucleotide sequencing technologies, and for the purpose of consistency with experimental observations A→G will be used in place of A→I hereafter. The effect of inosine substitution depends heavily on the type of RNA molecule as well as the functional region containing the RNA-edit, but can range in severity from irrelevant to an amino-acid replaced by a stop codon.

1.1.2 Apolipoprotein B mRNA Editing Enzyme (APOBEC)

The APOBEC family of RNA-editing enzymes, like the ADAR family, are deaminases. Unlike the ADAR family however, APOBECs act on single-stranded RNA molecules, deaminating cytosine to uracil. Also, while ADARs are promiscuous it is thought that APOBECs are accompanied by motif recognition subunits, leading to a more specific substrate and more control over specific bases edited (17).

Overall the APOBECs are considered to be responsible for a relatively small number of RNA-edits across the human transcriptome, but the effects of APOBEC RNA-edits are sometimes more easily determined (see Section 1.2.1).

1.1.3 Unknown Mechanisms of RNA-editing

Recent studies have presented evidence of the existence of all twelve possible nucleotide substitutions (18). However, others have found little evidence of non-canonical RNA-editing (12, 19), and peer-reviewed correspondence has refuted the evidence of non-canonical RNA-editing presented in some cases (20, 21, 22). Currently there is significant debate regarding the existence of these non-canonical RNA-edits and further research is needed before a suitably confident conclusion can be drawn. This need for further research is one of the key sources of demand for principled methods for RNA-editing detection and analysis, and is one of the main needs that the method described here seeks to address (see Section 1.5).

1.2 Biological Relevance of RNA-editing

In general, the biological relevance of RNA-editing in mammalian contexts is not well understood. However, it is known that ADAR mediated RNA-editing frequently affects 3' UTR regions (23), which could affect mRNA stability (and indirectly mRNA abundance) and/or miRNA or protein binding sites. While it is difficult to quantify such effects, the degree to which many 3' UTRs are subject to RNA-editing leads to speculation as to the impact of these modifications. Additionally, RNA-edits in exonic regions, while less frequent than 3' UTR events, have been shown to result in amino-acid substitutions (3, 4, 19, 23, 24), thus increasing proteomic diversity in the cell. Further, in any region, RNA-editing could potentially disrupt RNA secondary structure, leading to changes in protein binding affinity, translation efficiency and other consequences.

1.2.1 RNA-editing in Healthy Tissues

RNA-editing is known to occur in diverse cell types and across a wide range of developmental stages (11). Unlike well studied polymorphisms, mutations and splice variants, little is known about the effects of specific RNA-editing events in the context of healthy tissues. However, despite the lack of comprehensive study, the effects of double knockouts of ADAR and APOBEC in mice have been relatively well studied, granting some insight into the importance of RNA-editing.

ADAR1 and ADAR2 -/- knockout mouse models have been constructed to explore the effects of ADAR deficiency. The ADAR1 knockout is embryonic lethal, resulting in severe damage and/or developmental impairment to embryonic hepatic tissue (25). The mechanisms by which ADAR1 deletion causes this effect are unknown, but ADAR1 is clearly required for normal cell development and function. ADAR2 knockout results in lack of critical RNA-editing of GluR-B receptor, causing strong neuronal abnormalities resulting in epileptic seizures and short life-span (26). Further Hoopengardner *et. al.* (27) and Levanon *et. al.* (28) have reported RNA-editing sites in genes in human, mouse and chicken that further support a role of RNA-editing in neuronal development and function.

In some cases, such as the canonical ApoB C→U at position 6666 (mediated by APOBEC1), RNA-edits recode the codon to result in a STOP (3). The truncated ApoB protein is only expressed in the small intestine (the full length protein is expressed in the liver), and possesses specialized roles in lipid absorption and metabolism (29).

While these examples do not represent all known effects of RNA-editing, they are presented as illustrative examples to demonstrate the importance of RNA-editing for normal cellular processes.

1.2.2 RNA-editing in Pathological Contexts

In the general context of disease, very little is known about the role of RNA-editing. Given the severe ramifications of double knockouts of ADAR, it is easy to hypothesize as to how reduced expression of ADAR could lead to acute cellular dysfunction. Similarly, significant overabundance of ADAR could lead to transcriptome-wide hyper-editing leading to wide-spread stochastic modification of the protein ensemble of the affected tissue.

In the context of cancer, both coding and non-coding changes have been reported in a lobular breast cancer (24). Additionally, hypo-editing of known ADAR targets in tumour samples has been found (30). Paz *et al.* (31) explored the expression and targets of the ADAR gene family in brain tumours. They reported that expression of ADARB1 correlated with grade of malignancy in glioblastoma multiforme and that there were differences in the degree of RNA-editing at coding positions in CYFIP2, FLNA and BLCAP. The functional role of differential expression of ADAR or variable RNA-editing of these genes is unknown, but the existence of these patterns demonstrates that further exploration of RNA-editing in cancer is merited.

Until now, no study has attempted to define the transcriptome-wide landscape of RNA-editing across multiple tumours of the same type, or between tumours of differing tissue of origin. This thesis describes the first whole-trancriptome survey of the RNA-editing landscape in multiple cancers

of the same type as well as the first comparison of RNA-editing between cancer types.

1.3 High Throughput Nucleotide Sequencing Data

Generating full genome and transcriptome data has become a standard practice in the study of heritable genomic disorders as well as diseases with acquired genomic aberrations such as cancer. Such data is a powerful tool due to the wide array of analyses that can be performed on a single data set to interrogate many aspects of the sample. For example, copy number, chromosomal rearrangements, polymorphisms and somatic mutations can be inferred from a matched healthy/diseased genome pair from the same individual, providing a comprehensive view of landscape for that particular disorder. Similarly, from a whole transcriptome sequencing assay, gene expression, gene fusions, alternative splicing, polymorphisms and expressed mutations can be discovered.

In the past decade a number of technologies, collectively known as high-throughput sequencing (HTS), have emerged that allow relatively inexpensive nucleotide sequencing on a gigabase scale in a matter of days (32). The proliferation of HTS has made practical the sequencing of full genomes and transcriptomes of thousands of individuals and samples in sequencing centers around the world, and has led to the launch of substantial sequencing and analysis consortia such as the 1000 Genomes Project (33) and The Cancer Genome Atlas (34).

However, despite the clear potential of HTS, any analysis relies on sophisticated bioinformatics tools due to the properties of the data. The primary result of an HTS assay is a collection of hundreds of millions of short (*e.g.* 50-100 bases, dependent on sequencing chemistry) sequences, each derived from an unknown position in the sequenced sample. Each of these so-called reads must be either assembled *de-novo* or aligned to a known reference sequence (*e.g.* the reference human genome). Given the number of reads and the vast search space, alignment remains an open problem with many aligners available with different benefits and limitations. The alignment of RNA sequences (RNA-Seq) is particularly challenging as sequences that contain exon-exon junctions are not present in the reference genome. A common method to address this issue is to align to not only the reference genome, but also a set of known exon-exon junctions, and post-process the results to map the junction aligned sequences back onto genomic coordinates (35, 36).

Regardless of aligner choice, the result of the alignment process is the digital allelic count at each position in the genome, including zero-counts when the position is not present in the sequence data (Figure 1.1). Typically, whole genome sequencing experiments aim to achieve average depths (where depth refers to the sum of allelic counts at a given position) of thirty or greater across the genome. The purpose of this redundant sequence is to allow for the accurate typing of each base since HTS sequence data is known to be error prone (see Section 1.4.4). Each sequenced base from each read is accompanied by a number of features recorded or computed by the sequencing apparatus as well as those computed by the alignment software. These features include the confidence in the base call, the confidence in the alignment, the direction of the sequence, the number of mismatches in the read and more. Several software tools exist for the computation of these features including the popular Samtools (37) and GATK (38) packages. These features, along with the al-

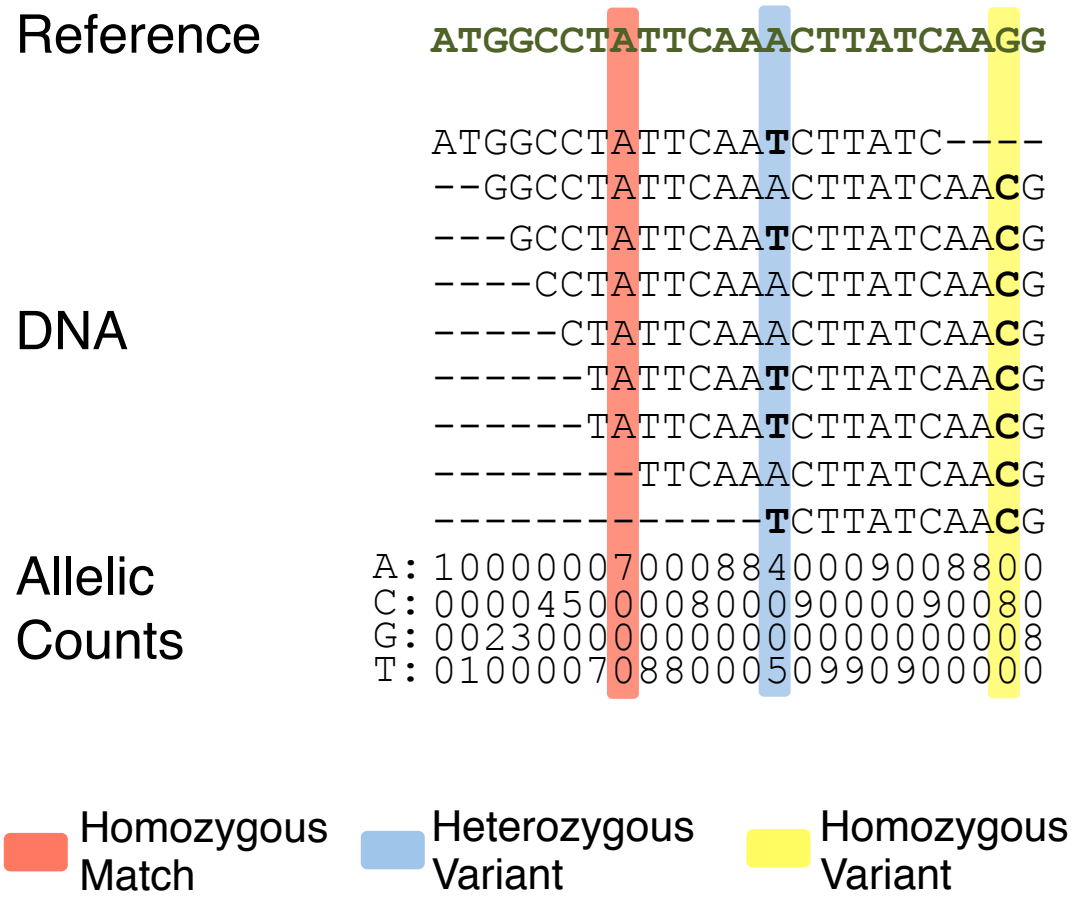


Figure 1.1: A constructed example of aligned HTS data. For each position, the allelic count of the reads overlapping a given position are recorded. These counts can be used to infer whether each position is a match to the reference genome (red), a heterozygous variant (blue) or a homozygous variant (yellow). Many methods for variant calling exist (see Section 1.4) and technical artifacts can interfere with accurate classification of some positions (see Section 1.4.4).

lelic counts at a given position are used to determine the nature of each position in the genome or transcriptome sample (*e.g.* matches the reference, mutation/SNP, indel etc).

1.4 Nucleotide Variant Calling from High Throughput Sequencing Data

Calling single nucleotide differences between the reference sequence and the sample of interest has become an area of active research and is a critical step in utilizing HTS data for disease research. Typically variants are determined on the basis of counts of the number of bases at a given position that match the reference and the count of those that do not. Frequently these counts are scaled

or modified by the base and mapping qualities provided by the sequencer and aligner respectively, granting additional weight to confident bases and penalizing poor quality data.

The methods by which the determination of variant or non-variant are diverse, and with differing degrees of accuracy and ease of interpretation. The most straightforward method, also typically the least accurate, is based on sequential thresholding of various features related to the position in the HTS data. Due to the lack of principled justification for these methods, statistical models have emerged as the accepted approach for accurate nucleotide typing.

1.4.1 Thresholding Methods

Thresholding methods are appealing due to their simplicity and ease of implementation. While the features used vary from method to method, a simple example, Algorithm 1, serves well to illustrate the concept.

Algorithm 1 Example Thresholding Method

```
1:  function THRESHOLD(bases, bq_thresh, mq_thresh, var_thresh)
2:      ref_match ← 0
3:      ref_mismatch ← 0
4:      for base ∈ bases do
5:          bq ← base_quality(base)                    ▷ Get base quality
6:          mq ← map_quality(base)                     ▷ Get mapping quality
7:          if bq < base_qual_thresh then
8:              goto next base                         ▷ Failed bq threshold
9:          end if
10:         if mq < map_qual_thresh then
11:             goto next base                         ▷ Failed mq threshold
12:         end if
13:         if matches_ref(base) == TRUE then
14:             ref_match ← ref_match + 1              ▷ Increment match count
15:         else
16:             ref_mismatch ← ref_mismatch + 1        ▷ Increment mismatch count
17:         end if
18:     end for
19:     var_ratio ← ref_mismatch/(ref_match + ref_mismatch)
20:     if var_ratio > var_thresh then
21:         return TRUE                                ▷ Variant ratio exceeds threshold
22:     else
23:         return FALSE                               ▷ Variant ratio fails threshold
24:     end if
25: end function
```

In this case, only base quality and mapping quality are subject to thresholds. Other common features include total depth and mappability of the region (for various definitions of mappability).

Thresholding methods, at best, are based on empirically determined appropriate threshold values for the relevant available features, and on selecting positions that meet these criteria. The fundamen-

tal concern with this approach lies in the fact that surpassing a threshold by a wide margin confers no advantage over passing minimally. This leads to a lack of associated confidence measures (*i.e.* a position is either variant or not, there is no uncertainty), and also leads to sensitivity and specificity issues with positions that pass or fail a threshold by a small margin. Additionally, these methods are often used with thresholds set to "intuitive values" or similar non-quantitative criteria, leading to further lack of confidence in the effectiveness of thresholding-based methods. See Goya *et al* (39) a more complete description of the deficiencies of thresholding methods.

Despite the known shortcomings of thresholding methods, several studies investigating RNA-editing from HTS data have implemented thresholding-based systems for RNA-editing detection (12, 18, 40). However, the study by Li *et al* has been questioned in published responses and, in general, these thresholding methods are being considered as naive, early solutions to this problem, as shown by studies using less *ad-hoc* methods (19, 41).

1.4.2 Likelihood Models

An early statistical approach to nucleotide typing was to use log-likelihood ratios, testing a variant model against that of the non-variant, null, model. This approach was utilized by the popular Samtools variant caller (37), prior to the implementation of bcftools (42), and as such has inspired a number of methods since. Recently such a likelihood model was used in Bahn *et al* (19) to classify RNA-edits from paired DNA/RNA data. However, in this method the genotype was taken as perfect knowledge, and not associated with any confidence measure. This lack of explicit uncertainty in the null distribution is a primary limitation of likelihood ratio methods for RNA-editing discovery, and a significant motivation for more sophisticated statistical methods.

1.4.3 Mixture Models

Mixture models, specifically mixtures of Binomial distributions, have gained popularity and achieved success in the nucleotide typing field. Examples of this approach include SNVMix2 (39), SOAPSnp (43), the maq SNP caller (44) and JointSNVMix (45).

SNVMix2 is a mixture of three Binomial distributions, parameterized to model homozygous reference, heterozygous reference and homozygous non-reference genotypes. The posterior probability of each genotype is inferred from the model given the base counts, base and mapping qualities. This approach gives a confidence for each genotype allowing the user to choose cutoffs depending on the specific tolerances for false positive and false negatives. An SNVMix2 based approach was attempted as a pilot study prior to the development of the framwork described in Chapter 2. However, this approach proved to be ill-suited to the detection of RNA-editing and was abandoned in favor of the creation of a novel method.

JointSNVMix adds the ability to jointly model two related sequences (*e.g.* a tumour and a healthy sample from the same individual). This expands the state space to nine joint genotypes, and the model borrows statistical strength across samples to increase the specificity of the classification. This approach has proved effective in the discovery of somatic mutation in cancer cells (36).

While this model has not been utilized for RNA-editing detection, the joint modelling approach used by JointSNVMix served as inspiration for the Auditor model described in Chapter 2.

1.4.4 Sequence and Alignment Artifacts

The throughput of HTS technologies is coupled with the cost of high error rates, relative to Sanger sequencing (46). The extreme density of the nucleotides within the sequencing equipment leads to errors in interpreting the optical signal of each base during sequencing, and various PCR biases during library preparation can lead to repeated or underrepresented fragments (47). These artifacts are largely randomly distributed, but some systematic patterns have been observed, such as a bias for base-calling errors toward the 3' end of each read. During alignment, these sequencing errors can lead to false positioning of the read with respect to the reference, causing wildtype positions to appear as variant, and leading false positive classifications of these positions by many methods.

Fortunately, many of these artifacts display detectable signatures. A common feature of false positive variant calls is for all reads containing a variant to be sequenced in a single direction when the expectation is a mixture of both directions (47). Similarly, a common RNA-Seq alignment artifact arises due to alignment to exon-exon junctions, and by accounting for transcript structure, these artifacts can be detected. Further, more general confidence scores, such as base and mapping quality, can be integrated into the detection of these more elaborate signatures. The artifacts described here represent only a small portion of the known artifacts found in HTS data, but serve to provide insight into the more general problem of artifact detection. These artifacts can, as described above, lead to the generation of false signals in the data resulting in the erroneous classification of variants leading to prohibitive false positive rates if the artifacts are not effectively identified and removed.

However, many methods seeking to exploit these signatures rely on thresholding or other heuristics to identify artifacts. In general, many of the shortcomings of thresholding-based nucleotide typing methods affect thresholding methods for artifact detection and a more principled approach is desirable. Discriminative classifiers are emerging as a powerful tool for this purpose (48) and provide the quantitative rigour that thresholding methods lack.

1.5 Meeting the Need for Principled RNA-Editing Detection

No cohesive principled method has emerged to appropriately address the following known properties of RNA-editing:

- Base-specific substitution likelihoods
- High-variance in RNA-seq base count evidence (due to stochastic properties)
- RNA dependence on DNA

Additionally, and of equal importance, no published method has yet taken a formal, principled approach to the detection and removal of sequencing and alignment artifacts that arise in RNA-editing detection.

By utilizing these properties when modelling RNA-editing, as well as taking a more principled approach to artifact detection, substantial improvements over existing methods are possible.

1.5.1 Specific Challenges in RNA-Editing Detection

The detection of RNA-edits is subject to an array of challenges that have not arisen in other nucleotide typing problems. First, the base-specific properties of RNA-editing require a shift from reference/non-reference counts to counts of each base. This requirement leads to a need for strand correction as a pre-processing step (in order to count the correct base), as well as a larger state space to accommodate a larger set of possible substitutions. To date, no method has addressed this challenge. Second, due to the stochastic nature of RNA-editing, RNA-edits may present in RNA-seq data with allelic ratios that deviate significantly from the expected 50:50 of a heterozygous polymorphism. This increase in variance invalidates critical assumptions made by many existing DNA genotyping methods (*e.g.* Binomial mixture based methods), and has yet to be accounted for in published studies. Third, the dependence of RNA on DNA has been largely ignored as a tool for increasing statistical power when classifying RNA-edits. In cases where a variant allele has been undersampled, but is still present, in the DNA data, calling a DNA variant on that evidence alone may not be possible. However, by simultaneously examining the RNA, evidence of the same variant can be taken into account, and used to appropriately classify the position as a variant and not an RNA-edit. This approach leads to a theoretical increase in specificity and has not been applied by published methods. Finally, and possibly the most significant challenge, is the substantial false positive rate inherent in RNA-edit classification due to sequencing and alignment artifacts. While accounting for artifacts is standard practice in RNA-editing detection, to date published methods can all be categorized as thresholding and/or filtering methods. Recent work (48) has demonstrated the power of discriminative classifiers to address the need for principled artifact detection in single nucleotide variation classification, but these approaches have not been utilized in RNA-editing detection to date.

1.5.2 Proposed Serial Generative and Discriminative Classification of RNA-edits

To address the challenges described in Section 1.5.1, a new computational approach to the detection of RNA-editing sites from an HTS whole genome/whole transcriptome paired dataset is proposed. This new framework, called Auditor, is tailored specifically to the intrinsic properties of RNA-editing biology and the systematic artifacts that are expected from comparison of whole genome sequencing to RNA-seq data. Auditor, is a generative probabilistic model that at once models the correlation expected between the WGS and RNA based allelic counts due to expressed SNPs and the non-random substitution distribution over nucleotide-encoded alleles we expect from RNA-editing enzymatic processes. The allelic counts are modelled using a Polya distribution accounting for base-specific properties rather than the standard Binomial (reference vs non-reference) allelic counts traditionally used in SNP detection. The substitution distributions are represented with a transition matrix encoding the expected probability of emitting a particular substitution in

the transcriptome given the observed alleles in the genome. Finally, to account for systematic artifacts, a previously described machine learning approach, called MutationSeq, for somatic mutation detection (48) tailored to the specifics of the RNA-editing detection problem is used.

By applying this combination of generative and discriminative classifiers, the proposed method addresses the specific challenges of RNA-editing detection in a principled manner, and represents a substantial advance over previous methods. In order to determine the benefits of each individual advance, the proposed method is compared to several simpler methods of similar design, as well as a modified version of a recently published Samtools/bcftools based detection method. The methods are compared using a combination of simulated data, Illumina HTS data and SOLiD HTS data to provide comprehensive analysis of each feature.

Section 2.1 describes the specification of the Auditor generative model, the variants used to evaluate the advances of Auditor and the simulation of relevant data sets. Section 2.3 demonstrates that Auditor outperforms simpler variants on simulated data. Chapter 3 reports the application of Auditor, and MutationSeq, to eleven diffuse large B-cell lymphoma and sixteen triple negative breast cancer samples. Section 3.2 shows that Auditor achieves superior sensitivity and specificity on HTS data than simpler variants and a Samtools approach modified from Denecek *et al* (41) and Section 3.3. In Section 3.4 the first landscape of RNA-editing in human cancers is presented and shows RNA-editing to be remarkably stable despite malignancy. Finally Chapter 4 reflects on the advances and limitations of the Auditor model and the results of the first large scale RNA-editing study in human cancers. Also in Chapter 4, possible avenues of future research are briefly explored to put this thesis in context with potential advances to come.

Chapter 2

The Auditor Model

Statistical approaches for RNA-editing afford many desirable properties. Perhaps of greatest benefit is that statistical methods allow for the assignment of principled confidence scores to each classification. This confidence score is a powerful tool when designing validation experiments as well as when performing functional analyses as it grants an explicit ranking to the classifications. Further, Bayesian methods allow the principled encoding of prior belief, allowing users to encourage expected/known patterns without preventing novel discovery. In the case of generative models, the parameters can be trained from unlabelled data allowing novel discovery in the absence of ground-truth. In a field as poorly understood as RNA-editing these properties are critical. Results will require secondary validation and priors encouraging common substitutions (*e.g.* A→G) discourage false positive calls, increasing confidence in computational predictions.

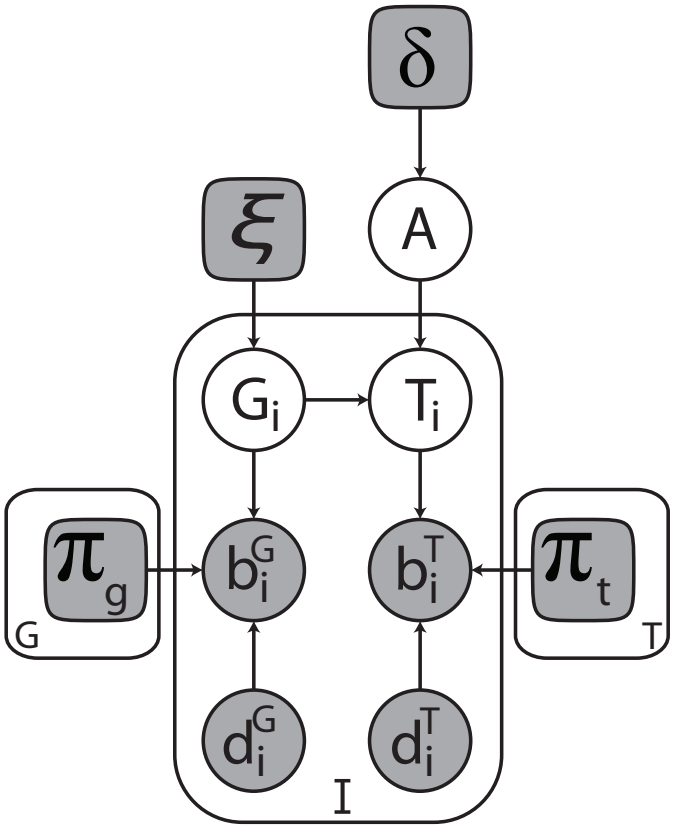
This chapter describes a statistical framework, called Auditor, that provides these benefits with sound mathematical foundations.

2.1 Joint Polya Mixture Modelling with a Substitution Specific Transition Matrix

2.1.1 Description of the Auditor Model

The Auditor model is shown as a probabilistic graphical model in Figure 2.1, with parameters as described in Table 2.1. The goal of the Auditor model is to identify positions for which the majority of probability mass (computed as described below) indicates the presence of an RNA-edit. The input to Auditor is two $4\times N$ matrices (where N is the number of positions), the first matrix containing the base counts of A, C, T, and G at each position in the genome and the second matrix containing the base counts for the transcriptome. The model outputs $p(Edit)$ for each position where $p(Edit) > 0.5$ ($p(Edit)$ is defined below).

It is assumed that both the genotype, G_i, and transcriptotype (mRNA genotype), T_i, consist of a single paternal and single maternal allele, each drawn from $\{A,C,G,T\}$. The cross-product of alleles then gives 16 possible types for each position in the genome and transcriptome. Palindromic



$$
\begin{aligned}
A_g \mid \delta_g^A &\sim Dirichlet(A_g \mid \delta_g) \\
G_i \mid \xi &\sim Multinomial(G_i \mid 1, \xi) \\
T_i \mid A, G_i = g &\sim Multinomial(T_i \mid 1, A_g) \\
b_i^G \mid \pi_G, d_i^G, G_i = g &\sim Polya(b_i^G \mid d_i^G, \pi_g) \\
b_i^T \mid \pi_T, d_i^T, T_i = t &\sim Polya(b_i^T \mid d_i^T, \pi_t)
\end{aligned}
$$

Figure 2.1: Auditor graphical model. Shaded boxes are fixed parameters, shaded circles are given data and open circles are unknown values modelled as random variables. Directed arrows indicate conditional dependencies. Boxes surrounding a group of variables represents an independent set of random variables for each element in the set denoted by the plate (*e.g.* $i \in$I). A is a transition matrix encoding the probability of transitioning from a given genotype to a given transcriptotype. The rows are Multinomially distributed with Dirichlet prior δ_g. G_i and T_i are 1-of-11 Multinomial distributions over the possible genotypes and transcriptotypes, respectively. ξ is the prior over the possible genotypes $g \in G$. b_i^G, b_i^T, d_i^G, and dT_i are the base counts and total depth of the genome and transcriptome at position i. π_g and π_t are the mixture parameters for the mixture of components g and t of the Polya mixture over G and T.

duplicates are disambiguated (*e.g.* AG, GA) reducing to a state space of size 10. Further, a state denoted ZZ, representing an unknown type, is added to both the genotype and transcriptotype to allow the model to avoid making a canonical type call in the presence of multi-allelic (>2) data. This results in an 11-state space: $G_i, T_i \in \{AA, AC, AG, AT, CA, CG, CT, GG, GT, TT, ZZ\}$ for each position $i \in (1, \ldots, N)$ in the input data. A constructed example of input and resultant genotype-transcriptotype outputs is shown in (Figure 2.2).

Polya distributions, also known as Dirichlet-Multinomials, are used to model the state-dependent base counts at each position in the genome, π_G, and transcriptome, π_T. The parameterizations demonstrated below are set such that $\pi_G = \pi_T$, but these could be independently tuned if necessary. Additional variance of the Polya distribution leads to significant advantages over the, arguably, more intuitive Multinomial model as the overdispersion properties of the Polya distributions more closely model the high variance observed in RNA-Seq data (Figure 2.3). It is the nature of a Multinomial mixture to move the vast majority of probability mass between classes (*e.g.* AA→AG) with the addition or subtraction of only a single base count, resulting in little uncertainty in the Multinomial mixture and unduly peaked distributions, whereas Polya mixtures shift probability mass more gradually (Figure 2.4).

To capture the known process of RNA-editing (DNA is transcribed to RNA, which is subsequently enzymatically modified), the transcriptotype is explicitly modelled as a function of the genotype. To capture the base-modification specific properties of RNA-editing (e.g. A→G represents the vast majority of known RNA-edits), an 11×11 transition matrix, $A(g,t)$, is used, with each entry containing $p(T_i = t | G_i = g)$. Thus, information from the genome is leveraged when calling the transcriptotype, and visa versa, and the non-uniform substitutions inherent in RNA-editing due to enzymatic processes can be encoded. Furthermore $A(g,t)$ takes advantage of the inherent correlation present in the genotype and transcriptotype by simultaneous inference and renders the model less susceptible to calling expressed germline polymorphisms as RNA-edits. This is not possible using methods that call variation in the genome and transcriptome independently as has been previously shown for somatic mutation calling from tumour-normal DNA datasets in cancer (45). The parameters of $A(g,.)$ are assumed to be Multinomial such that $\sum_{T=t} A(g,t) = 1$ and can, in theory, be estimated with maximum a posteriori estimation with conjugate Dirichlet priors (2.1). For the purposes of this study, literature-informed estimates were used due to the intractability of parameter estimation (this is further explained in Section 2.1.2).

To determine the probability of an RNA-edit at a given position, first the joint genotype and transcriptotype is inferred according to the conditional probabilities of the model:

$$\begin{aligned} p(G_i = g, T_i = t \mid b_i^G, d_i^G, b_i^T, d_i^T, \theta) &\propto p(b_i^G \mid \pi_G, d_i^G, G_i = g) \\ &\times p(G_i = g \mid \xi) p(T_i = t \mid A_{gt}) \\ &\times p(b_i^T \mid \pi_T, d_i^T, T_i = t) \end{aligned}$$

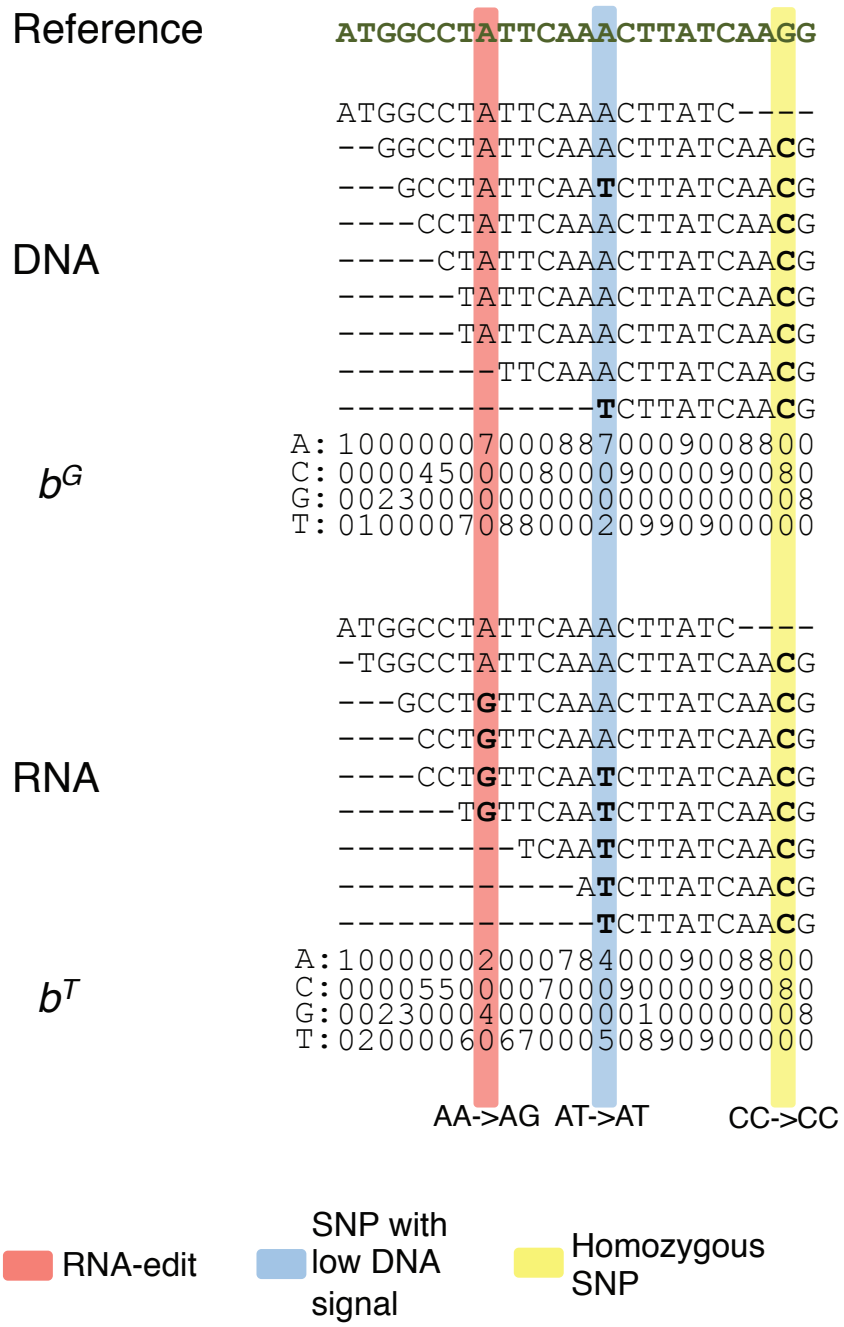


Figure 2.2: A constructed example of three joint geno-transcriptotypes. In red, an example RNA-edit is shown, where the variant is present only in the RNA. In blue, a SNP that would likely require joint modelling to identify as the DNA evidence alone may not be enough to legitimately classify a SNP; however with the inclusion of the RNA data the SNP can be identified. In yellow, a homozygous SNP is shown; without the DNA data (*i.e* with reference genome alone) this would be indistinguishable from a genuine RNA-edit.

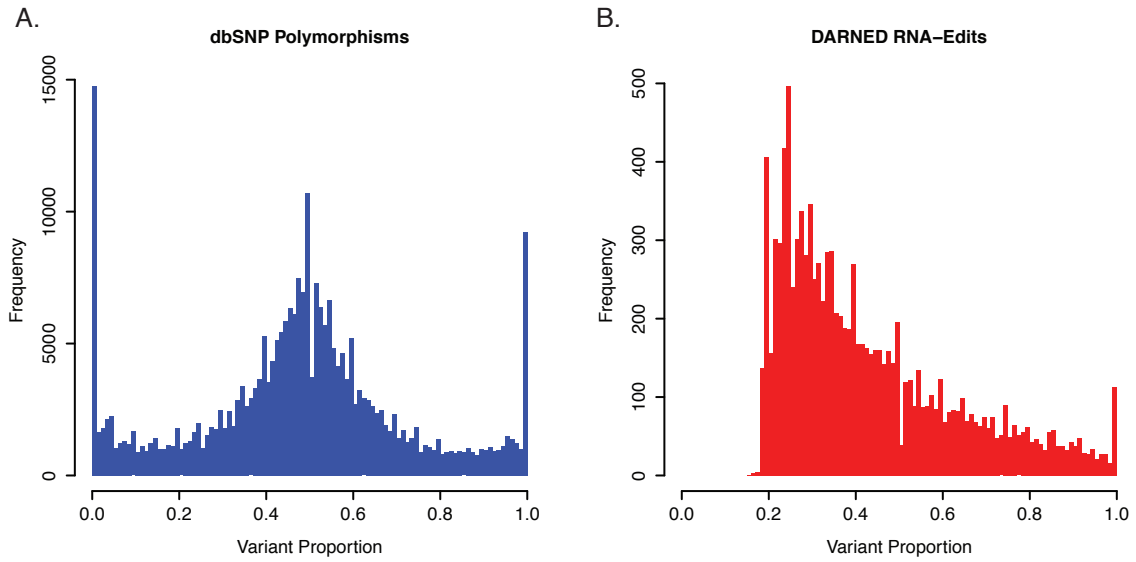


Figure 2.3: (A.) The distribution of RNA-Seq variant proportion of known SNPs present in the DNA of 27 cancer samples with depth of at least 20 in the RNA-Seq. The distribution is trimodal with peaks at 0, 0.5 and 1.0. **(B.)** The distribution of RNA-Seq variant proportion of known RNA-edits detected in 27 cancer samples with depth of at least 20. The distribution is highly variant with probability mass broadly distributed.

where $\theta = \{\pi_g, \pi_t, \xi, A_{gt}\}$ and

$$p(b_i^G \mid \pi_G, d_i^G, G_i = g) = Polya(b_i^G \mid d_i^G, \pi_g) \quad (2.1)$$

$$p(b_i^T \mid \pi_T, d_i^T, T_i = t) = Polya(b_i^T \mid d_i^T, \pi_t) \quad (2.2)$$

where

$$Polya(\mathbf{x} \mid N, \alpha) = \frac{N!}{\prod_{k=1}^{K}(x_k!)} \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(N + \sum_k \alpha_k)} \prod_{k=1}^{K} \frac{\Gamma(x_k + \alpha_k)}{\Gamma(\alpha_k)} \quad (2.3)$$

where $\mathbf{x} = (x_1, \ldots, x_K)$ discreet counts for each of K classes, $N = \sum_k x_k$, Γ is the gamma function and $\alpha = (\alpha_1, \ldots, \alpha_K)$ is the parameterization of the Polya distribution.

After computing and normalizing the posterior joint probabilities are marginalized over the set of RNA-edit states, defined as $R = \{g \in G, t \in T : g \neq t \wedge g_i \neq ZZ \wedge t_i \neq ZZ\}$:

$$p(Edit) = \sum_{g,t \in R} p(G_i = g, T_i = t \mid b_i^G, d_i^G, b_i^T, d_i^T, \theta) \quad (2.4)$$

2.1.2 Parameter Learning

The equations and distributions necessary for inference are shown below the graphical model in Figure 2.1.

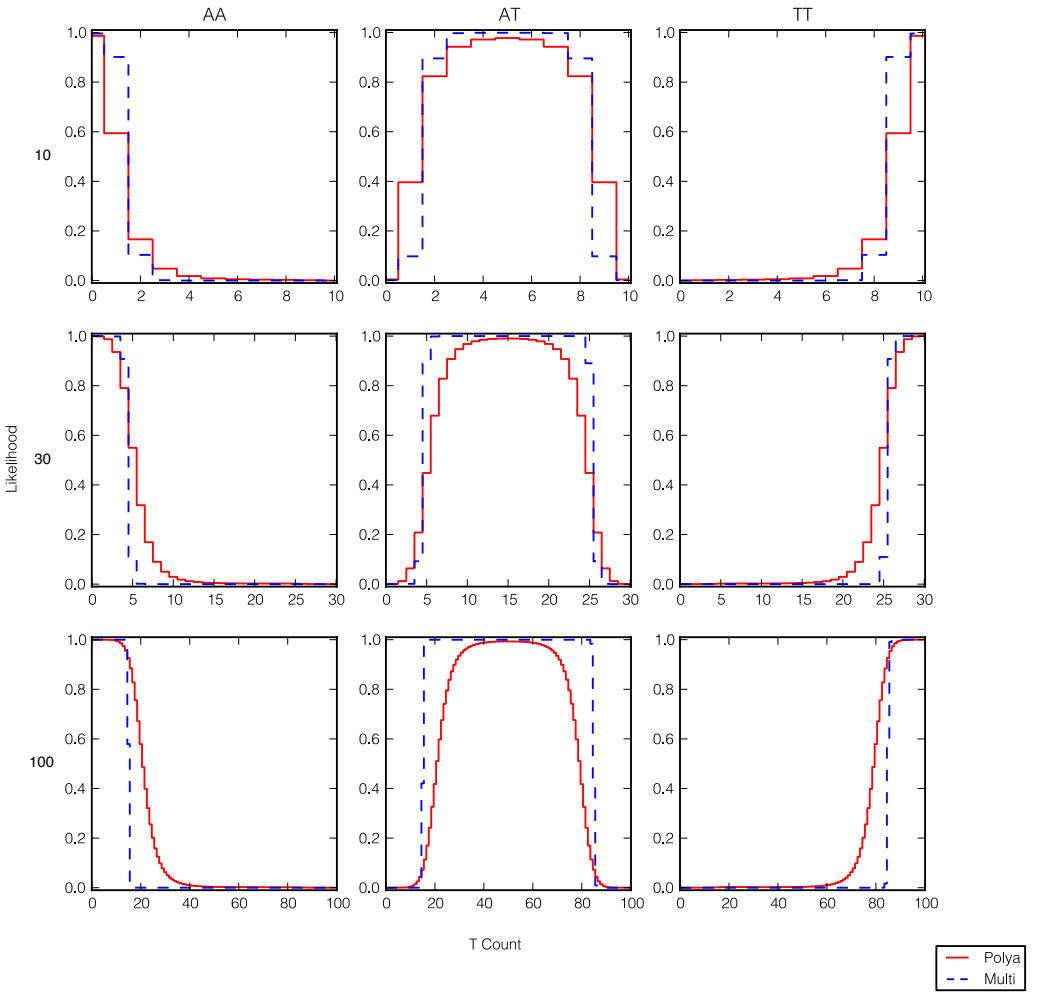


Figure 2.4: The Polya (red solid lines) distribution allows for state transitions over a greater range of variation in counts, compared to the Multinomial (blue dashed lines). The three rows of figures are demonstrations of this property at depths of 10, 30, and 100 respectively, top to bottom. In this case, the reference is declared to be A, and the variant to be T. Across the x-axis of each plot, the number of T's (out of the total depth of 10, 30 or 100) is increased. From left to right, each plot shows the likelihood of the counts having been drawn from the AA, AT, or TT distributions of the Polya or Multinomial mixture. It is clear that in all cases, the Multinomial mixture shifts the majority of the probability mass between classes with the addition, or subtraction, of a single base, where as the Polya mixture shift more gradually.

The transition matrix can be learned via expectation maximization (EM) using the standard Dirichlet-Multinomial update formula. Specifically:

$$A_{g,t}^{new} = \frac{\sum_{i \in N} \left[p(T_i = t) p(G_i = g) + \delta_{g,t}^{A} - 1 \right]}{\sum_{i \in N} \left[\sum_{t \in T} \left[p(T_i = t) p(G_i = x) + \delta_{g}^{A} - 1 \right] \right]} \tag{2.5}$$

EM is allowed to to iterate until the complete data log-likelihood (CLL) converges. The CLL for the Audtior model is expressed as:

$$\begin{aligned} CLL &= \sum_{i \in N} \left[\sum_{g \in G} \left[\sum_{t \in T} \left[log(p(G_i = g \mid \pi_G, d_i^G, b_i^G, \xi)) \right. \right. \right. \\ &+ \left. \left. \left. log(p(T_i = t \mid \pi_T, d_i^T, A_{g,t}, G_i = g)) \right] \right] \right] \\ &+ log(p(A \mid \delta)) \end{aligned}$$

In this framework EM proceeds with monotonically increasing CLL, and converges on a locally optimal value for A. An example of the values of the transition matrix through iterations of EM on simulated data is shown in Figure 2.5.

Despite the success of training on simulated data, EM has proved impractically slow to run on entire HTS data sets. Randomly subsampling from the data to form a training set is an appealing approach, but it is unlikely that a random subsample small enough to make EM practical would contain a representative sample of RNA-edits. In the future, when more is known regarding RNA-editing and commonly edited positions, targeted subsampling may be a feasible approach to training Auditor in a sample-specific manner.

2.1.3 Parameter Settings

In place of fitting A to each sample, the substitution proportions from Bahn *et al* (19) is used to generate the transition matrix. It is demonstrated below that processing putative RNA-edits with MutationSeq effectively overpowers the transition matrix, and the results support that these settings are suitable until per-sample training is practical.

The parameters for the Polya mixtures, π_G and π_T were set by empirically examining the nature of various parameterizations. Figure 2.6 shows the behaviour of the three Polya mixtures to illustrate the rationale behind the selection of the parameters described in Table 2.1. This parameterization was selected because it demonstrated appropriate class-switching from homozygous to heterozygous types (*i.e.* switching from AA to AT was gradual, but complete). Other parameters led to mixtures that would shift too abruptly between classes, or to mixtures that never shifted completely to the heterozygous type. See Section 4.5 for a discussion on training the Polya mixture parameters from labelled data and why this approach was not used in this thesis and Section 2.2 for experiments demonstrating the robustness of the Polya distribution in general and the parameters specified in Table 2.1 specifically.

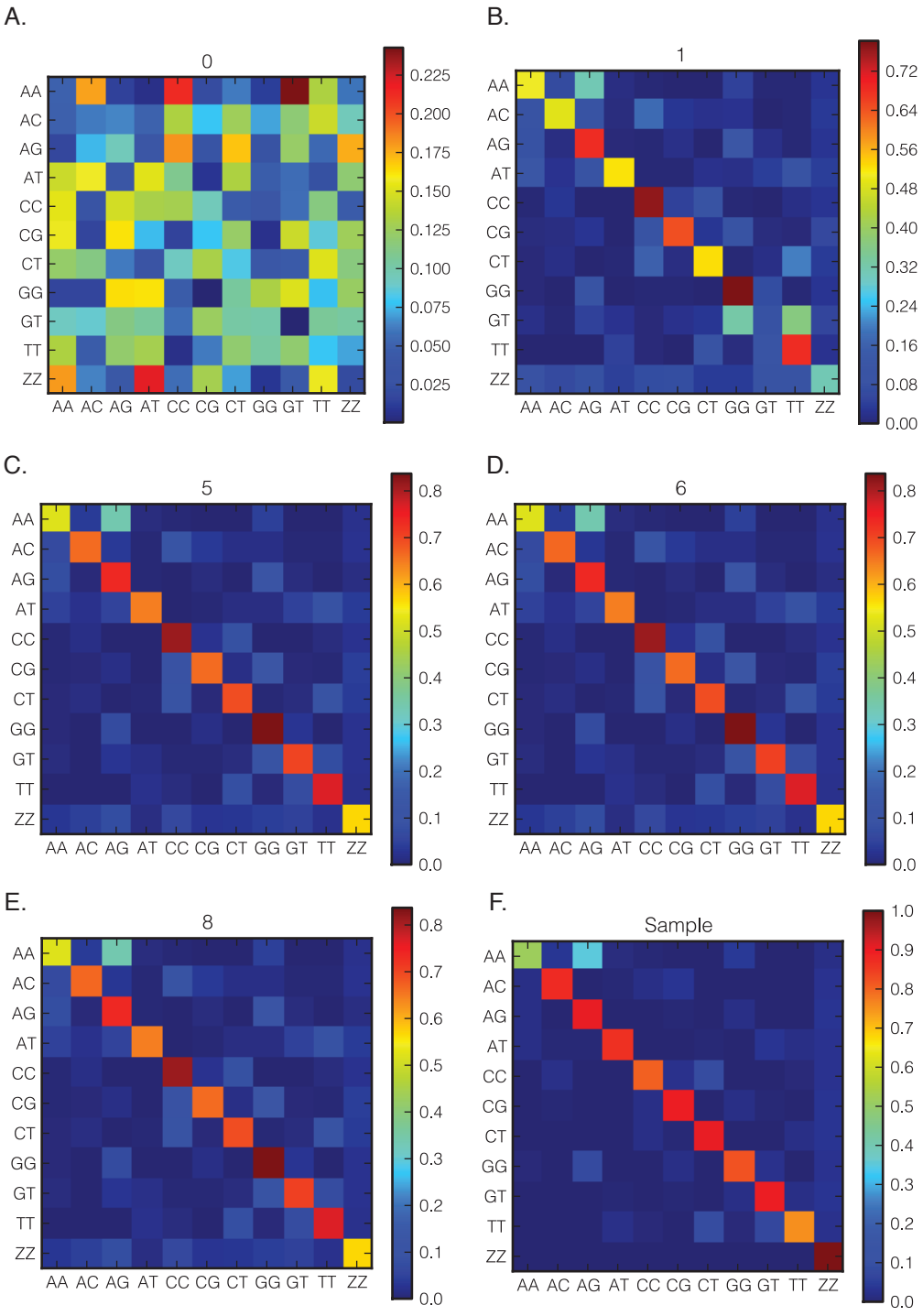


Figure 2.5: Example of progression of the Auditor transition matrix during EM iterations. Iteration 0 **(A)** shows the starting parameters which were generated uniformly at random. The 10,000 data points used for this simulation were drawn from the Auditor model parameterized as in Table 2.1. The transition matrix parameters are shown as a heat map in **(F)**. Iterations 5 and 6 **(C,D)** represent intermediate iterations, and iteration 8 **(E)** is the final values after EM. EM converges quickly, in 8 iterations, and results in values representative of the true sampling distribution.

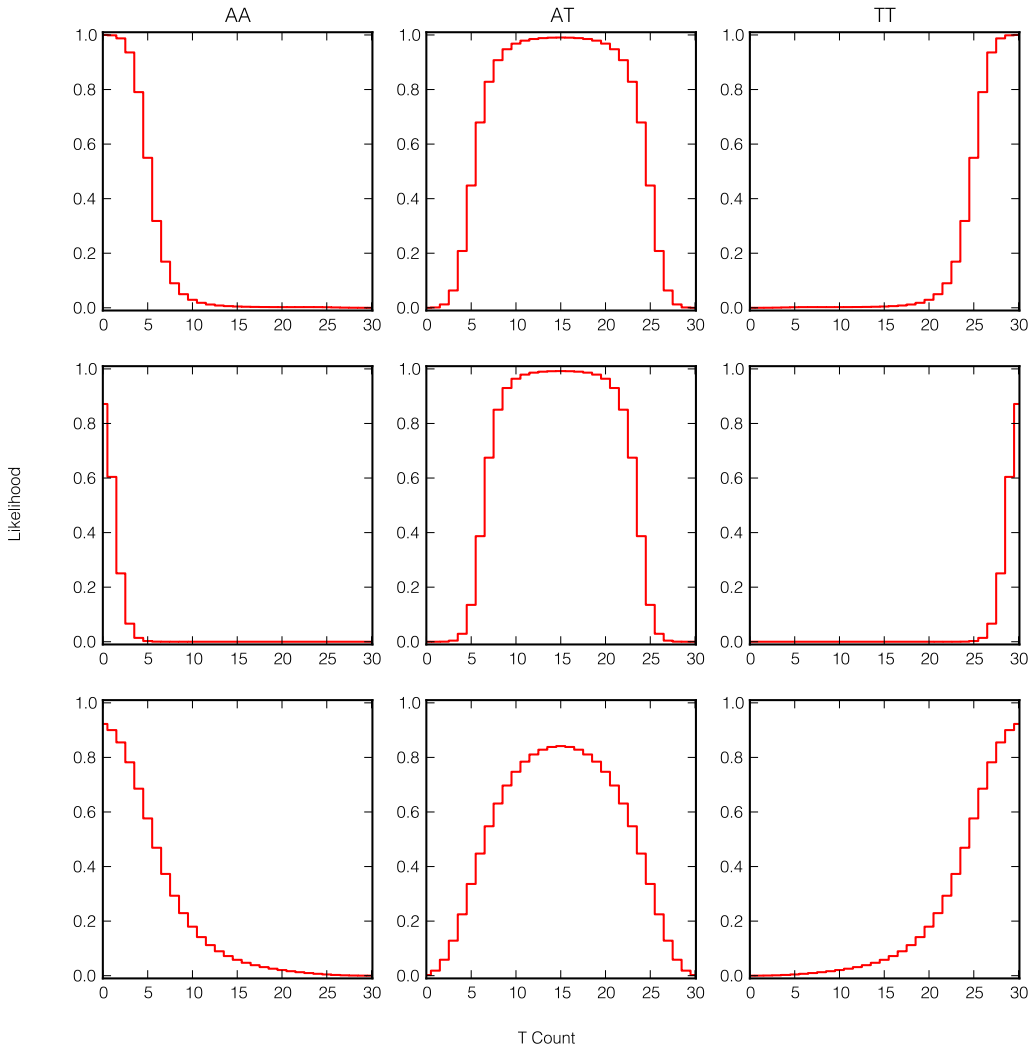


Figure 2.6: Demonstration of three different parameterizations of a four-count Polya distribution. The reference is declared to be A, and the variant to be T. Across the x-axis of each plot, the number of T's (out of the total depth of 30) is increased. The top row of plots shows the behaviour of the Polya mixture as parameterized in the Auditor model (Table 2.1). Transition to heterozygous types begins with the addition of of only one non-reference base, and the AT type retains substantial probability mass until the number of T's is strongly dominant. The second row demonstrates a Polya mixture with more pronounced peaks, which effectively delays the transition from AA to AT and AT to TT. The third row shows a Polya mixture with significantly greater uncertainty, resulting in the AT class never being fully dominant. Given the broadly distributed variant ratios observed in RNA-editing (see Figure 2.3) the behaviour of the top Polya mixture was deemed most suitable.

2.1.4 Variants of the Auditor Model

In addition to the fully featured Auditor model, two simpler variants were implemented in order to systematically evaluate the properties of the proposed advances.

The first variant independently computes the posterior probability of each genotype and transcriptotype and then generates the final transition posterior by taking the cross-product of the independent results. This model is referred to as IndepPolya, and is shown, with relevant conditional probability distributions in Figure 2.7. IndepPolya allows for analysis without the benefit of the transition matrix and transcriptomic dependence on the genome, and is used for evaluation of these properties.

In the second variant,the Polya distributions in the π_G and π_T mixtures are replaced with Multinomial distributions, η_G and η_T, keeping the transition matrix intact and computation of the transition $p(T_i = t|G_i = g)$ is equivalent to Auditor. This model is referred to as JointMulti and is shown, with relevant conditional probability distributions in Figure 2.8. This model was implemented in order to demonstrate the benefits of the Polya distribution over the Multinomial distribution in isolation.

These models are demonstrated on simulated data in Section 2.3.

2.1.5 Simulated Data Generation

We simulated two-hundred datasets of ten-thousand positions, divided into two groups of one-hundred, for use in comparing the performance of Auditor and related variants. The first group was simulated from Auditor itself (hereafter AuditorSet), and the second was simulated from the JointMulti variant (hereafter MultiSet). In both cases the transition matrix used to simulate the data was set so that self-transitions (*i.e*. no RNA-edit) were twenty times more likely than any other transition. All other transitions were equally likely. For the Polya mixture models, base-counts were set as shown in Table 2.1. These counts were normalized to generate the parameters for JointMulti.

Depth was simulated using a hyper-variable Poisson distribution:

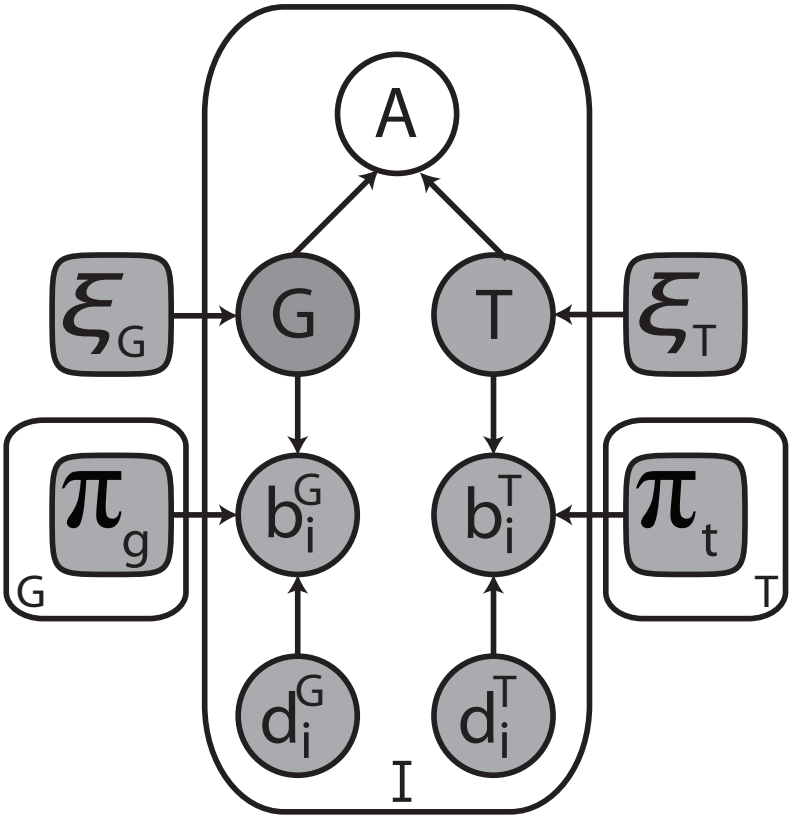
$$D \sim Poisson(\lambda) + Uniform(-0.5 \times \lambda, 0.5 \times \lambda) \tag{2.6}$$

where $\lambda = 40$ for the genomic counts and $\lambda = 50$ for the transcriptomic counts.

Sampling from both Auditor and JointMulti was intended to reduce bias that would be inherent to one approach over the other.

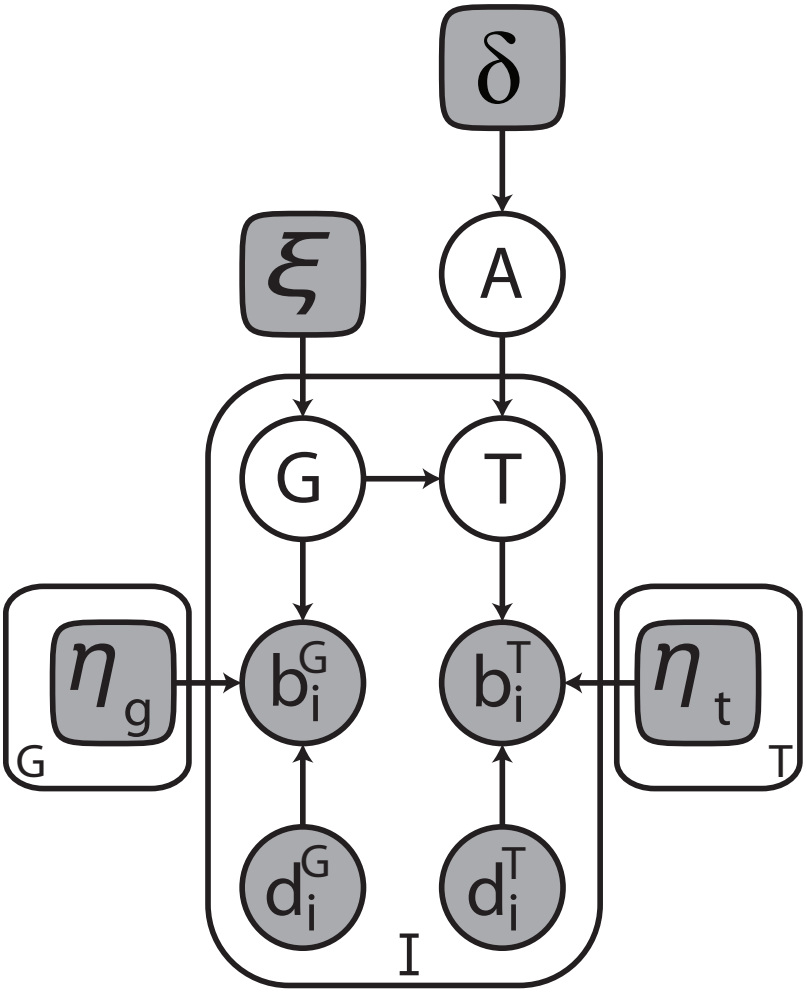
2.1.6 Implementation.

Auditor is implemented in Python, and has dependencies for Numpy, SciPy, bam-counter (and consequently pysam), and Matplotlib. Auditor provides three primary commands: syncnt, for converting matched DNA-RNA BAM files into the synchronized count format; classify, to compute the probability of an RNA-edit for each position in a syncnt file; and train, to train the transition matrix from a syncnt file using EM (see Section 2.1.2). MutationSeq is implemented in Matlab with



$$
\begin{aligned}
p(b_i^T \mid \pi_T, d_i^T, T_i = t) &= Polya(b_i^T \mid d_i^T, \pi_t) \\
p(b_i^G \mid \pi_G, d_i^G, G_i = g) &= Polya(b_i^G \mid d_i^G, \pi_g) \\
p(G_i \mid \xi) &= Multinomial(G_i \mid 1, \xi) \\
p(T_i \mid \xi) &= Multinomial(T_i \mid 1, \xi) \\
p(G_i = g \mid \pi_G, d_i^G, b_i^G, \xi) &\propto p(b_i^G \mid \pi_G, d_i^G, G_i = g) p(G_i = g \mid \xi_G) \\
p(T_i = t \mid \pi_T, d_i^T, b_i^T, \xi) &\propto p(b_i^T \mid \pi_T, d_i^T, T_i = t) p(T_i = t \mid \xi_T) \\
p(A_{gt} \mid G_i, T_i) &= p(G_i = g) p(T_i = t)
\end{aligned}
$$

Figure 2.7: Graphical model for the IndepPolya variant of Auditor and the relevant conditional probability distributions. IndepPolya breaks the dependence of the transcriptome on the genome, but all other aspects of the model are help invariant compared to Auditor.



$$
\begin{aligned}
p(b_i^T \mid \eta T, d_i^T, T_i = t) &= Multinomial(b_i^T \mid d_i^T, \eta_t) \\
p(b_i^G \mid \eta_G, d_i^G, G_i = g) &= Multinomial(b_i^G \mid d_i^G, \eta_g) \\
p(G_i \mid \xi) &= Multinomial(G_i \mid 1, \xi) \\
p(A_g \mid \delta_g^A) &= Dirichlet(A_g \mid \delta_g^A) \\
p(G_i = g \mid \eta_G, d_i^G, b_i^G, \xi) &\propto p(b_i^G \mid \eta_G, d_i^G, G_i = g) p(G_i = g \mid \xi) \\
p(T_i \mid A, G_i = g) &= Multinomial(T_i \mid 1, A_g) \\
p(T_i = t \mid \eta_T, d_i^T, b_i^T, A, G_i = g) &\propto p(b_i^T \mid \eta_T, d_i^T, T_i = t) p(T_i = t \mid A, G_i = g)
\end{aligned}
$$

Figure 2.8: Graphical model for the JointMulti variant of Auditor, and the relevant conditional probability distributions. JointMulti is identical to Auditor for the purposes of inference, except that the mixture components, η_G and η_T, are Multinomial distributions rather than Polya distributions.

Parameter	Description	Value

δ — Pseudo counts in Dirichlet prior on A

	AA	AC	AG	AT	CC	CG	CT	GG	GT	TT	ZZ
AA	500.0	10.0	100.0	10.0	10.0	10.0	10.0	30.0	10.0	10.0	20.0
AC	70.0	500.0	10.0	10.0	70.0	10.0	10.0	10.0	10.0	10.0	20.0
AG	70.0	10.0	500.0	10.0	10.0	10.0	10.0	70.0	10.0	10.0	20.0
AT	70.0	10.0	10.0	500.0	10.0	10.0	10.0	10.0	10.0	10.0	20.0
CC	10.0	10.0	10.0	10.0	500.0	10.0	70.0	10.0	10.0	20.0	20.0
CG	10.0	10.0	10.0	10.0	70.0	500.0	10.0	70.0	10.0	10.0	20.0
CT	10.0	10.0	10.0	10.0	70.0	10.0	500.0	10.0	10.0	70.0	20.0
GG	10.0	10.0	10.0	10.0	10.0	10.0	10.0	500.0	10.0	10.0	20.0
GT	10.0	10.0	10.0	10.0	10.0	10.0	10.0	70.0	500.0	70.0	20.0
TT	10.0	10.0	10.0	10.0	10.0	10.0	100.0	10.0	10.0	500.0	20.0
ZZ	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	500.0

A — Multinomial substitution-specific transition matrix

	AA	AC	AG	AT	CC	CG	CT	GG	GT	TT	ZZ
AA	0.5208	0.0417	0.3542	0.0130	0.0052	0.0000	0.0000	0.0443	0.0000	0.0000	0.0208
AC	0.0220	0.8811	0.0000	0.0044	0.0220	0.0352	0.0000	0.0000	0.0000	0.0000	0.0352
AG	0.0228	0.0000	0.9132	0.0000	0.0000	0.0046	0.0000	0.0228	0.0000	0.0000	0.0365
AT	0.0218	0.0044	0.0044	0.8734	0.0000	0.0000	0.0044	0.0000	0.0349	0.0218	0.0349
CC	0.0000	0.0247	0.0000	0.0000	0.8230	0.0247	0.0864	0.0000	0.0000	0.0082	0.0329
CG	0.0000	0.0045	0.0000	0.0000	0.0227	0.9091	0.0000	0.0227	0.0045	0.0000	0.0364
CT	0.0000	0.0000	0.0000	0.0000	0.0228	0.0046	0.9132	0.0000	0.0000	0.0228	0.0365
GG	0.0083	0.0000	0.0792	0.0000	0.0000	0.0167	0.0000	0.8333	0.0292	0.0000	0.0333
GT	0.0000	0.0000	0.0000	0.0045	0.0000	0.0045	0.0000	0.0227	0.9091	0.0227	0.0364
TT	0.0000	0.0000	0.0000	0.0268	0.0077	0.0000	0.0881	0.0077	0.0728	0.7663	0.0307
ZZ	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

ξ	Multinomial parameters for genotype states	AA, CC, GG, TT: 0.21; AC, AG, AT, CG, CT, GT: 0.021; ZZ: 0.0021
G^i	Genotype at position i	Inferred
T^i	Transcriptotype at position i	Inferred
b_i^G	Count of each base (A,C,G,T) at position i in genome	Observed
b_i^G	Count of each base (A,C,G,T) at position i in transcriptome	Observed
d_i^G	Depth of coverage at position i in genome	Observed
d_i^T	Depth of coverage at position i in transcriptome	Observed

π_G — Count parameters for genotype Polya mixture

	AA	AC	AG	AT	CC	CG	CT	GG	GT	TT	ZZ
A	4	12	12	12	0.05	0.05	0.05	0.05	0.05	0.05	4
C	0.05	12	0.05	0.05	4	12	12	0.05	0.05	0.05	4
G	0.05	0.05	12	0.05	0.05	12	0.05	4	12	0.05	4
T	0.05	0.05	0.05	12	0.05	0.05	12	0.05	12	4	4

π_T	Count parameters for transcriptotype Polya mixture	Same as π_G

Table 2.1: Description of the parameters and random variables of the Auditor model. Fixed values for parameters are given, and inferred or observed values are noted.

convenience wrappers in Perl.

Additionally a Perl pipeline that takes as input paired DNA/RNA single chromosome BAM files and executes conversion to counts, classification and MutationSeq post-processing, has been implemented. The output of this pipeline contains the probability of each substitution, the total $p(Edit)$, the MutationSeq confidence score and the product of $p(Edit)$ and the MutationSeq score for all positions where $p(Edit) > 0.5$.

The entire suite will be publicly available at `http://compbio.bccrc.ca/software/` in the near future.

2.2 Assessment of the Robustness of Audtior with Respect to Polya Parameters

The Polya distribution, as an infinite mixture of Multinomials, is expected to be robust with respect to parameter settings. However, the robustness of Auditor specifically was examined to ensure that the model was not overly sensitive to the Polya mixture parameterization.

To determine the sensitivity of Auditor to various parameterizations of the Polya mixtures two experiments were performed. In the first experiment, ten sets of one-thousand positions were simulated from a single parameterization of Auditor as described in Table 2.1. These positions were then classified using an array of different Polya parameters as follows: the four count parameters of each Polya were reduced to strength and skew, such that for a given class (*e.g.* AG) nucleotides present in that class (in this case A and G) were assigned the skew value, and all other elements were assigned 1. The set of parameters were then scaled by the strength value. For example, if $skew = 50$ and $strength = 0.1$ the AG vector would be $\{5, 0.1, 5, 0.1\}$. Skew values ranged between 2 and 200 in steps of 5, while the strength values ranged from 0.01 to 676.4 in 40 steps such that $strength = (0.01) \times (1.33)^{(step)}$. This range of strengths was used to adequately explore extreme values in the parameter space. For each (strength, skew) tuple, the median AUC over the ten datasets was recorded and is shown in Figure 2.9. The results indicate that any combination of strength and skew is effective at classifying the data generated by the Auditor model, demonstrating that many differing parameterizations of the Polya mixture would likely be effective in practice.

Further, using the RNA base counts from the same positions that were used to generate the RNA-editing variant ratio distribution in Figure 2.3, the parameters for the A→G Polya distribution were trained with second order gradient descent. The resulting parameters, [2.6903, 0.0357, 1.9304, 0.0239] were converted to skew and strength. Strength was computed by averaging the C and T components of the vector (strength = 0.0298) and skew was computed by taking the product of the strength and the average of the A and G components (skew = 77.529). This value is similar to the strength and skew used for the Auditor parameters (strength = 0.05, skew = 80).

The second experiment covered the same range of strength and skew, but simulated ten sets of one-thousand positions from an instantiation of Auditor for *each* strength-skew tuple. For each data set Auditor, parameterized as shown in Table 2.1, was used to classify each position. The median AUCs of the ten sets simulated from each parameter setting was recorded and is shown in

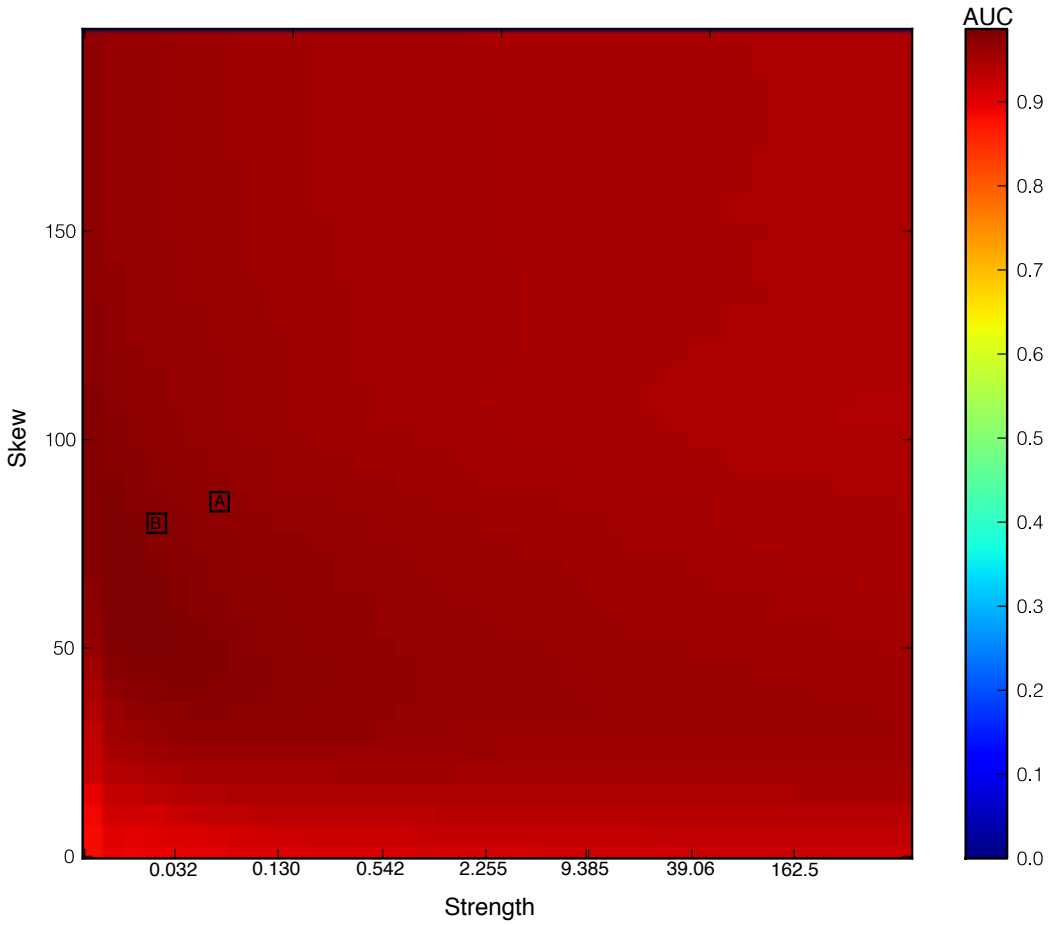


Figure 2.9: The results of classifying ten sets of one-thousand positions, simulated from Auditor as parameterized in Table 2.1 **[A]**, with instantiations of Auditor with varying Polya parameters (the trained value of the A→G DARNED edits is labelled as **[B]**). The Polya parameterization was reduced to strength and skew, such that for a given class (*e.g.* AG) nucleotides present in that class (in this case A and G) were assigned the skew value, and all other elements were assigned 1. The set of parameters were then scaled by the strength value. Skew values explored were between 2 and 200 in steps of 5, while the strength values ranged from 0.01 to 676.4 in 40 steps such that $strength = (0.01) \times (1.33)^{(step)}$. Across all values of strength and skew, the various parameterization of the Polya mixture are effective at classifying positions simulated from the Auditor parameters in Table 2.1.

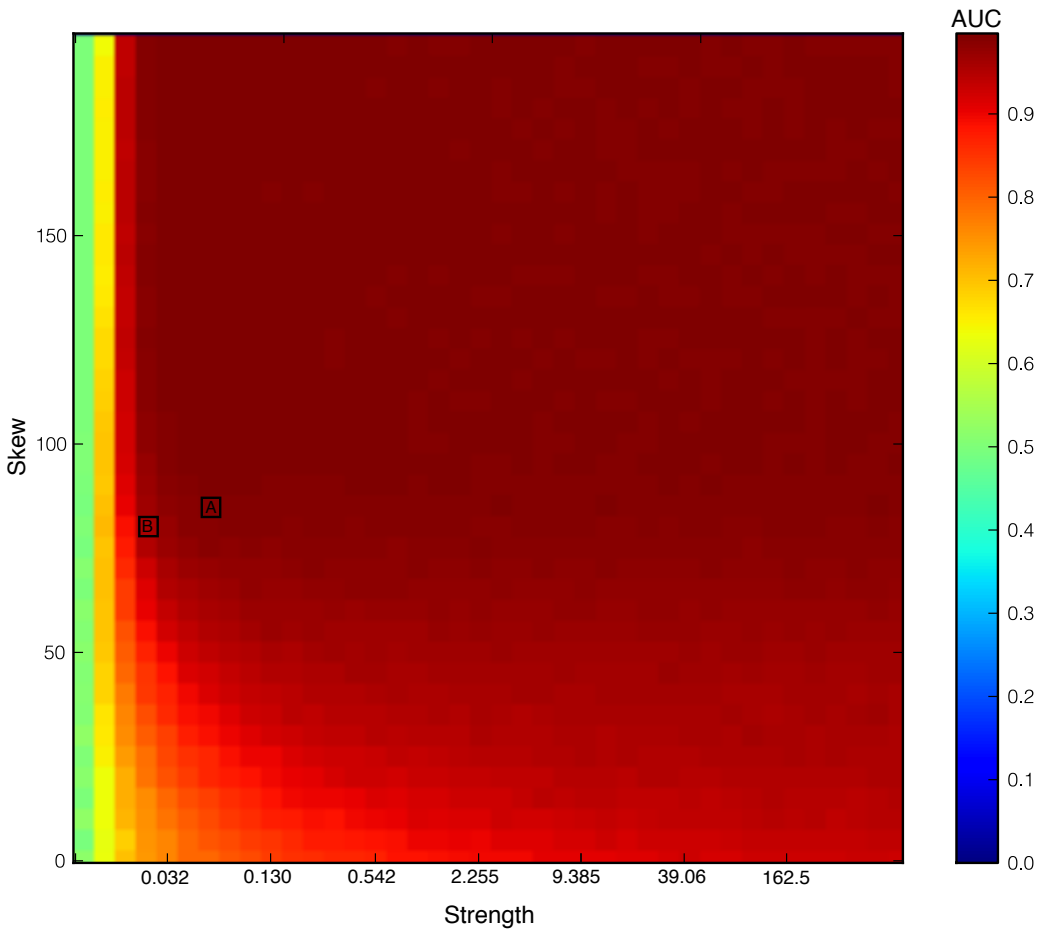


Figure 2.10: The results of classifying ten sets of one-thousand positions, simulated from Auditor with a wide range of Polya mixture parameterizations, with an instantiation of Auditor using parameters as described in Table 2.1 **[A]** (the trained value of the A→G DARNED edits is labelled as **[B]**). The Polya parameterization was reduced to strength and skew, such that for a given class (*e.g.* AG) nucleotides present in that class (in this case A and G) were assigned the skew value, and all other elements were assigned 1. The set of parameters were then scaled by the strength value. Skew values explored were between 2 and 200 in steps of 5, while the strength values ranged from 0.01 to 676.4 in 40 steps such that $strength = (0.01) \times (1.33)^{(step)}$. Auditor, using the parameters in Table 2.1 is effective in classifying positions simulated from a wide range of Polya distributions. This suggests that Auditor is robust to differences between the Polya mixture parameters and the unknown, true distribution of RNA-edits in HTS data.

Figure 2.10. Auditor performs well on a wide range of strength and skew parameters and shows weak performance only on highly variable (low strength and skew) data. This suggests that the parameterization described for Auditor is robust to differences between the parameter settings and the unknown, true distribution of RNA-edits in HTS data.

2.3 Systematic Evaluation of Auditor on Simulated Data

The performance of Auditor was evaluated against the variants described in Section 2.1.4 using simulated data as described in Section 2.1.5 . Posterior marginal probabilities $p(Edit)$ were computed to obtain the probability of an RNA-edit at each site in the data as described above. The performance of Auditor, IndepPolya and JointMulti was evaluated for each of the 200 runs with area under the receiver operator characteristic curve (AUC) statistics. (Figure 2.11B).

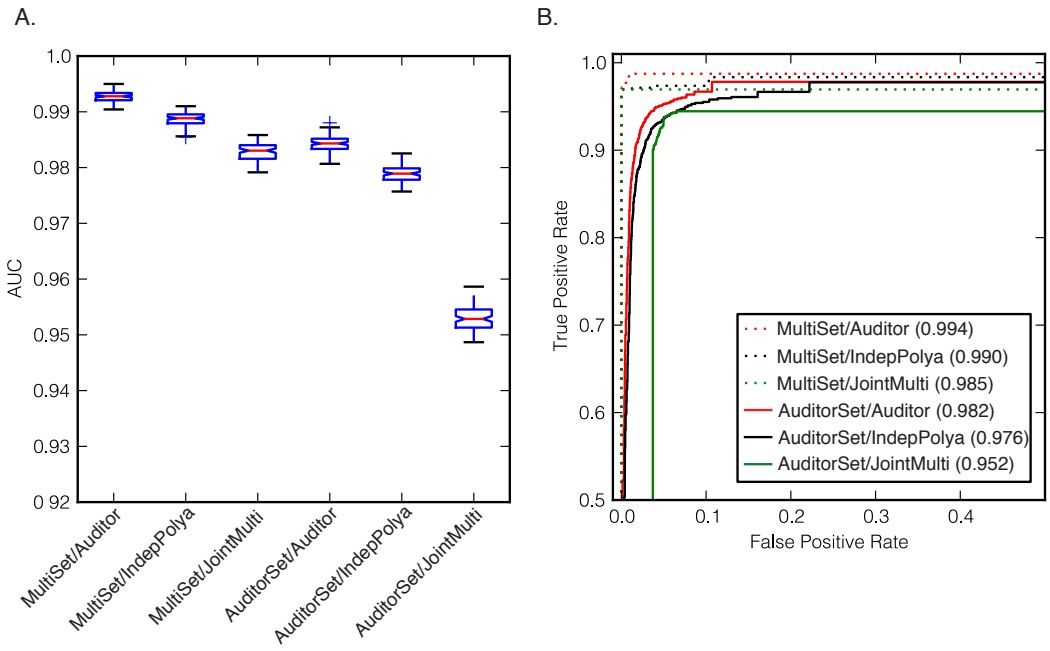


Figure 2.11: (A.) Distribution of area under the curve (AUC) for each method on 200 sets of 10,000 (100 simulated from Auditor, 100 simulated from JointMulti) simulated data points. Classification was performed with three methods: Auditor, JointMulti , and IndepPolya. On both simulation sets, Auditor is superior to both JointMulti and IndepPolya. **(B.)** A representative example of ROC curves resulting from classification of two sets of 10,000 simulated data points. The AUC for each method (Auditor in red, IndepPolya in black and JointMulti in green) is shown in brackets in the legend. The first set, shown with solid lines, was simulated from Auditor. The second set, shown with dotted lines, was simulated from JointMulti (see Section 2.1.4).

As expected, Auditor achieved the highest accuracy on both simulation datasets (Figure 2.11A). Performance was consistent across runs for both datasets despite random generation of the data. A representative run is shown in Figure 2.11B with full ROC curves. For MultiSet, the median AUC was was 0.993 for Auditor followed by IndepPolya and JointMulti (median AUC of 0.989 and 0.983, respectively). Similar results were obtained for the AuditorSet where Auditor had a median AUC of 0.984, followed by IndepPolya and JointMulti (0.980 and 0.953, respectively). These results indicate that if base counts are generated according to a Multinomial distribution, the Polya

Simulation Model	Classification Model	Mean AUC	Median AUC	AUC Variance
JointMulti	Auditor	0.9927	0.9928	9.41×10^{-7}
JointMulti	IndepPolya	0.9888	0.9888	1.51×10^{-6}
JointMulti	JointMulti	0.9828	0.9830	2.63×10^{-6}
Auditor	Auditor	0.9843	0.9843	1.76×10^{-6}
Auditor	IndepPolya	0.9790	0.9790	2.14×10^{-6}
Auditor	JointMulti	0.9530	0.9529	4.75×10^{-6}

Table 2.2: Summary statistics of AUCs for classification of 100 data sets per simulation method/classification method pair.

distributions, encoding greater variance that the Multinomial distributions, confer an advantage over using the Multinomial itself. Observed differences between all 6 AUC distributions (Table 2.2) were statistically significant (ANOVA $p < 0.0001$, Tukey HSD p-values <0.001). This suggests that Polya distributions are significantly more effective than Multinomials, and adding joint modelling achieves a similarly significant performance gain. It is notable that all three methods achieve higher median AUCs on the MultiSet model (0.993 vs. 0.984, 0.989 vs 0.979, 0.983 vs. 0.953, Auditor, IndepPolya and JointMulti, respectively), likely due to the lower variance generated by Multinomial distributions. Therefore in the presence of higher variance in the allelic distributions (as is typical of RNA-edits (Figure 2.3), it is clear that the models that employ Polya distributions will be more robust than Multinomial based models.

Chapter 3

Application of Auditor to Human Cancers

The performance of Auditor, IndepPolya, JointMulti and Samtools/bcftools were compared by classifying RNA-edits from eleven diffuse large B-cell lymphoma samples (DLBCL), and determining relative sensitivity and specificity for the four methods. Relative sensitivity and specificity is a semi-quantitative analysis to determine the rank order of the methods, and has emerged as an acceptable approach when a suitable, gold-standard data set does not exist (38). The substitution profiles resulting from Auditor and Samtools/bcftools classifications were then compared to known properties of RNA-editing as well as known polymorphic substitution distributions in order to determine which method generated distributions that most closely resembled the known RNA-editing patterns while diverging from the polymorphic distribution.

By these metrics, Auditor achieves the best performance and was subsequently applied to sixteen triple-negative breast cancer (TNBC) genome-transcriptome pairs. In contrast to the Illumina-based DLBCL data, the TNBC genome data was sequenced using the SOLiD platform. The substitution distribution of the TNBC samples was computed using Auditor and compared to that of the DLBCL samples and resulted in substantial similarities between the two cancer types. Further, the distribution of RNA-edits over six genomic regions (3'UTR, 5'UTR, introns, splice-sites, synonymous-exonic and non-synonymous-exonic) was computed for both cancer types. Analysis of this regional distribution both within and between cancer types suggests that RNA-editing processes are abundant and stable in both cancer types.

3.1 Feature-Based Classifiers for Modelling Technical Artifacts in RNA-Seq Data

MutationSeq (48) uses supervised learning techniques to leverage alignment properties and scores, such as mapping/base quality and strand bias to accurately identify true variants in high-throughput sequence data. A single-sample, 21 feature variant (Table 3.1) of the original method was implemented and trained for use in filtering putative RNA-edits. The importances of each feature are

1. number of reads covering or bridging the site
2. number of reference Q13 bases on the forward strand
3. number of reference Q13 bases on the reverse strand
4. number of non-reference Q13 bases on the forward strand
5. number of non-reference Q13 bases on the reverse strand
6. sum of reference base qualities
7. sum of squares of reference base qualities
8. sum of non-reference base qualities
9. sum of squares of non-reference base qualities
10. sum of reference mapping qualities
11. sum of squares of reference mapping qualities
12. sum of non-reference mapping qualities
13. sum of squares of non-reference mapping qualities
14. sum of tail distances for reference bases
15. sum of squares of tail distance for reference bases
16. sum of tail distances for non-reference bases
17. sum of squares of tail distance for non-reference bases
18. $P(D \mid G_i = aa)$, phred-scaled, i.e., x is transformed to $-10\log(x)$
19. $\max_{G_i \neq aa}(P(D \mid G_i))$, phred-scaled
20. $\sum_{G_i \neq aa}(P(D \mid G_i))$, phred-scaled
21. Distance to nearest known splice site

Table 3.1: The definitions of MutationSeq features x_1 to x_{21}. Q13 is defined as base quality greater or equal to Phred score 13; D represents the three dimensional vector (depth, number of reference bases and number of non-reference bases) at the current site; $G_i \in \{aa, ab, bb\}$ means the genotype at site i, where $a, b \in \{A, C, T, G\}$ and a is the reference allele and b is the non-reference allele. These features are constructed from Samtools.

shown in Figure 3.1. MutationSeq was used to evaluate the presence or absence of a variant in the RNA-Seq data for each putative RNA-edit. After classification with Auditor, resulting putative RNA-edit positions were analyzed by MutationSeq. A final confidence score for each position was generated by taking the product of the score from the classifier method and score from MutationSeq. All variants of the model and the Samtools method were similarly post-processed in this way in the application to real datasets to ensure the comparability of results.

See Appendix A for a detailed description of the Auditor and Samtools/bcftools workflows (IndepPolya and JointMulti follow the same workflow as Auditor).

Examples of false positive RNA-edit classifications that require MutationSeq post-processing to identify are given in Appendix C.

3.2 Auditor Achieves Superior Sensitivity and Specificity than Variant Models and Samtools/bcftools

3.2.1 Description of Concordance Analyses For Evaluation of Sensitivity and Specificity

The BWA (49) aligned whole-genome shotgun sequence (WGSS) and whole-transcriptome shotgun sequence (WTSS, RNA-Seq) from eleven DLBCL samples reported in a recent sequencing project (35) was obtained. Sample acquisition and preparation are described in the orginal manuscript, and coverage statistics can be found in Appendix B.

Putative RNA-edits were classified in the DLBCL data using four methods: Auditor, IndepPolya, JointMulti and the Samtools/BCFTools variant caller (37). All putative RNA-edits from each

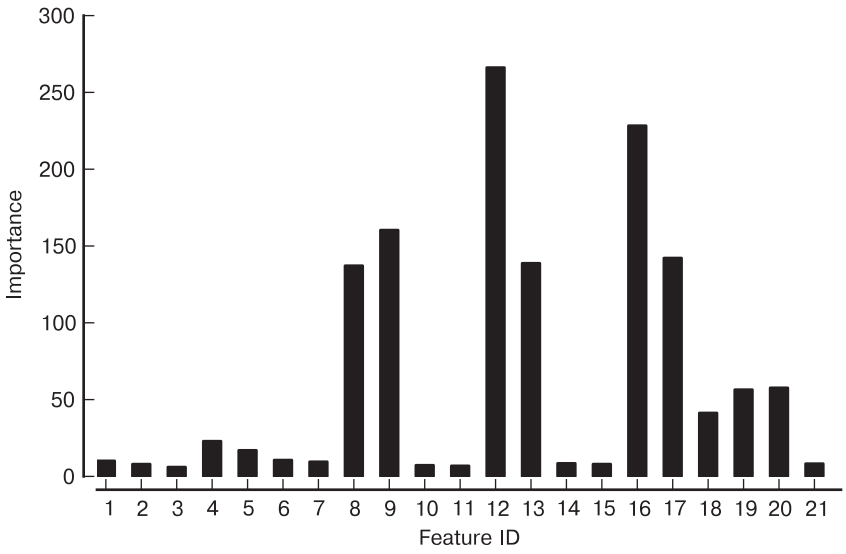


Figure 3.1: The importance score for each MutationSeq feature is shown. Features pertaining to non-reference base quality (features 8 and 9), non-reference mapping quality (features 12 and 13) and sum of tail distances of non-reference bases (features 16 and 17) are the most relevant to identifying technical artifacts. Base quality and mapping quality directly relate to the confidence in the base call, thus their importance is not surprising. Features 16 and 17 are effective in removing splicing-related artifacts since such errors frequently manifest as the beginning or end of a read overhanging a splice site due to misalignment. Feature 21, despite explicitly encoding the distance to the nearest splice site is of relatively low importance. This is likely due to insufficient splicing related artifacts in the training data.

method were post-processed with the modified MutationSeq algorithm to determine the likelihood that the RNA variant was due to an artifact. The final confidence score for each putative RNA-edit is the product of $p(Edit)$ from Auditor and the $p(variant)$ from MutationSeq (*i.e.* the position is an RNA-edit *and* there's a true variant in the RNA).

To assess sensitivity and specificity of each of the methods, the concordance of the putative RNA-edits predicted by these four methods on two datasets, dbSNP and the RNA-edit database DARNED (50), was computed. Given the lack of a comprehensively validated RNA-edit sample set for which there also exists deep DNA and RNA sequence data, these concordance comparisons were used to generate a comparison of *relative* sensitivity and specificity. For each method, putative edits were ranked by confidence and the percent of calls concordant with each external dataset was computed for the highest n scored RNA-edits, $1 \leq n \leq 10,000$. Lower concordance with the dbSNP positions indicates higher specificity (*i.e.* a smaller proportion of predictions are known SNP positions) under the assumption that a position known to be a polymorphic position is significantly more likely to be an undersampled SNP, causing a false-positive RNA-edit prediction, than a genuine RNA-editing event. Higher concordance with the DARNED RNA-edits indicates higher

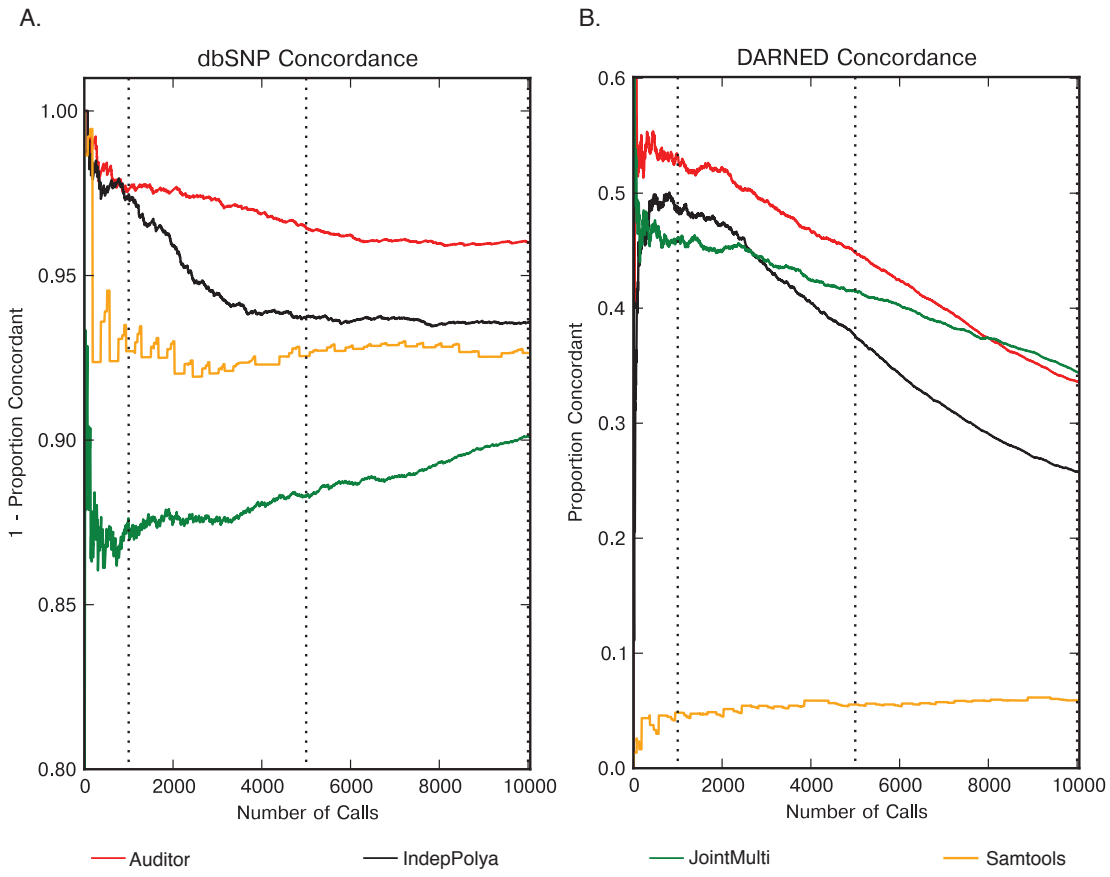


Figure 3.2: Concordance of the 10,000 highest confidence RNA-edits predicted by five methods to dbSNP **(A.)** and the DARNED RNA-edit set **(B.)**. Lower relative concordance to the dbSNP positions indicates higher specificity, while higher relative concordance to the DARNED set indicates higher sensitivity. Auditor achieves superior sensitivity and specificity.

sensitivity (*i.e.* a larger proportion of predictions are known RNA-edit sites) due to the relatively small size of the DARNED position set (42,055 positions). The probability of randomly distributed technical artifacts inducing a false positive at one of the 42,055 DARNED positions in a space of approximately 7×10^7 coding positions is remote and is unlikely to skew the results. For ease of interpretation, 1-(dbSNP concordance) is reported so that low concordance values indicate high specificity.

3.2.2 Comparison of Specificities

The relative specificity of each model was examined via the respective concordance to dbSNP at rank cut-offs of 1000, 5000 and 10,000 (Figure 3.2A; Table 3.2).

Auditor achieves the highest specificity by these metrics. At 1000 calls the Auditor displays very high specificity of 0.976. The specificity of the IndepPolya is similarly high at 0.974 and

Rank	Auditor	IndepPolya	JointMulti	Samtools
1000	**0.9760**	0.9740	0.8741	0.9270
5000	**0.9644**	0.9374	0.8830	0.9255
10,000	**0.9601**	0.9358	0.9011	0.9265

Table 3.2: 1-(dbSNP concordance) values of Auditor, IndepPolya, JointMulti and Samtools at top 1000, 5000, and 10000 most confident RNA-edit positions classified from 11 DLBCL samples.

Samtools also displays high specificity at 0.960. JointMulti, however, somewhat less specific at 0.8741. This is attributed to the peaked nature of the Multinomial mixtures described in Section 2.1.1 and shown in Figure 2.4. This results in false positives when the evidence for a SNP in the DNA is just under the threshold for the Multinomial to detect and in false negatives when an RNA-edit has sub-threshold variant proportion (see Section 3.2.3). It is noted that dbSNP is known to contain a small-number of RNA-editing events (51), and that it is likely that all methods compared are being penalized for a small portion of false negatives in the dbSNP data.

The specificity ranking of the four compared methods remains consistent at all benchmarks (Figure 3.2A; Table 3.2), but note that at 10,000 calls the specificity of IndepPolya reduces to 0.934 becoming similar to the specifity of Samtools, 0.926, whereas Auditor remains more consistent with specificity of 0.960 at the same benchmark. The improvement of Auditor over IndepPolya can be directly attributed to the simultaneous inference of the genotype and transcriptotype and suggests that such joint inference results in fewer false positive predictions of expressed polymorphisms.

3.2.3 Comparison of Sensitivities

Next, the relative sensitivities of the methods as determined by concordance to the DARNED data set are compared (Figure 3.2; Table 3.3).

Auditor, again, achieves the best performance with sensitivity of 0.531 at rank 1000. IndepPolya and JointMulti demonstrate lower sensitivities of 0.486 and 0.4615, respectively. Samtools displays the lowest sensitivity by a wide margin, with sensitivity of 0.048 at the rank 1000 benchmark. The sensitivity of Auditor, IndepPolya and JointMulti decay as calls of lower confidence are included (0.336, 0.258 and 0.3444 at rank 10,000 respectively), as expected, but consistently remain above the sensitivity achieved by Samtools. Samtools, despite low dbSNP concordance, achieves low sensitivity at all rankings suggesting that the calls made by Samtools are neither SNPs nor RNA-edits.

3.3 Substitution Profiles of Eleven DLBCLs as Classified by Auditor and Samtools

The emerging canonical substitution profile of RNA-edits is characterized by the overwhelming enrichment of A→G events at the expense of near-complete depletion of other substitutions (12). The

Rank	Auditor	IndepPolya	JointMulti	Samtools
1000	**0.5315**	0.4865	0.4615	0.0483
5000	**0.4483**	0.3756	0.4143	0.0554
10,000	0.3361	0.2578	**0.3444**	0.059

Table 3.3: Sensitivity of Auditor, IndepPolya, JointMulti and Samtools at top 1000, 5000, and 10000 most confident RNA-edit positions classified from 11 DLBCL samples.

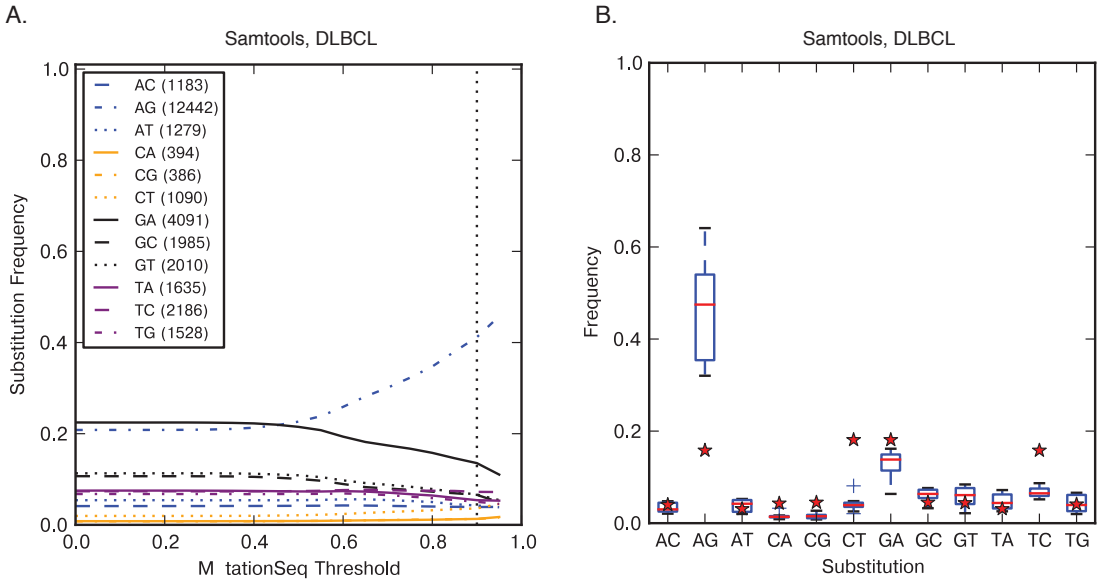


Figure 3.3: (A.) Substitution distribution of Samtools-classified RNA-edits of 11 DLBCL cases as a function of MutationSeq score. **(B.)** Substitution distribution across the 11 DLBCL cases at a MutationSeq threshold of 0.9. The distribution shows some similarity to that of known SNPs (red stars), particularly G→A.

substitution profiles generated from the Auditor and Samtools classifications, at a range of MutationSeq confidence thresholds (0.0 to 0.95 in steps of 0.05), were compared in order to determine to what extent this known pattern could be reproduced. A critical assumption made in this comparison is that the greater the proportion of A→G events, the more accurate the method. While this assumption may not be strictly valid, the majority of evidence to date suggests that it is not unreasonable (11, 12). The substitution profiles are also compared to the known polymorphism substitution distribution, and, in the case of Auditor, the model parameters. Substitution profiles differing from these two reference profiles further indicates that the phenomenon being observed is not polymorphic nor a result of parameter bias.

The base substitution distribution across the eleven DLBCL samples as classified by Auditor and by Samtools was compared. For each method, the substitution distribution of the pooled putative RNA-edits from all eleven cases was plotted as a function of MutationSeq score (Figures 3.3A

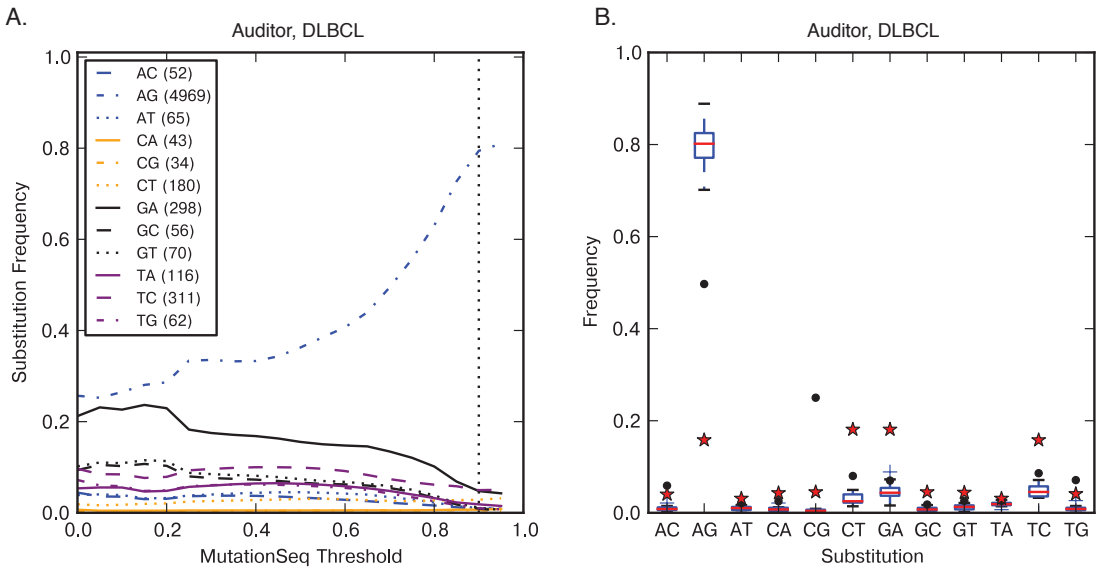


Figure 3.4: (A.) Substitution distribution of Auditor-classified RNA-edits of 11 DLBCL cases as a function of MutationSeq score. As the MutationSeq score increases A→G substitutions become dominant with 80% of substitutions being A→G at a MutationSeq threshold of 0.9. **(B.)** Substitution distribution across the 11 DLBCL cases at a MutationSeq threshold of 0.9. The distribution shows little similarity to that of known SNPs (red stars), and diverges strongly from the model parameterization (black dots).

and 3.4A). As the MutationSeq score increases, from 0 to 0.95 in steps of 0.05, the substitution distribution predicted by Auditor converges toward the known distribution of RNA-edits, dominated by A→G events (A→G proportion = 0.80) . However, in the Samtools predictions, while A→G events are enriched as MutationSeq score increases (A→G proportion = 0.45), the enrichment is weaker than observed in the Auditor calls and further, the G→A frequency remains relatively high (G→A proportion = 0.13). Indeed, Samtools retains a greater proportion of all non-A→G substitutions than Auditor at a MutationSeq score of 0.9 (Table 3.4). It is noted that while Auditor is biased by design toward calling A→G events due to the transition matrix parameterization, MutationSeq harbours no such bias. Therefore any enrichment of a given base substitution as a function of MutationSeq is the result of unbiased selection.

The substitution frequencies of the Auditor predictions, at a MutationSeq threshold of 0.9, are consistent with known RNA-editing distributions and diverge substantially from the original parameters (Figure 3.4B). There are strong examples (*e.g.* C→G) where the final substitution frequency differs from the initial parameters, and it is concluded that Auditor, post-processed by MutationSeq, is not inappropriately constrained by the prior knowledge encoded in the model. The substitution distribution of the Samtools classifications, at the same MutationSeq threshold, much more closely resemble that of the SNP distribution (Figure 3.3B). This supports the results from concordance analysis and suggests strongly that Samtools predictions are composed substantially of SNPs,

Substitution	Samtools Mean	Auditor Mean	Samtools Median	Auditor Median	Samtools Variance	Auditor Variance
AC	0.03	0.01	0.03	0.01	1.01×10^{-4}	2.84×10^{-5}
AG	0.45	0.80	0.47	0.80	0.01	2.97×10^{-3}
AT	0.04	0.01	0.04	0.01	1.50×10^{-4}	2.17×10^{-5}
CA	0.02	0.01	0.01	0.01	3.49×10^{-5}	2.97×10^{-5}
CG	0.02	0.00	0.02	0.00	2.82×10^{-5}	6.43×10^{-6}
CT	0.04	0.03	0.04	0.03	2.20×10^{-4}	1.30×10^{-4}
GA	0.13	0.05	0.14	0.04	8.18×10^{-4}	4.00×10^{-4}
GC	0.06	0.01	0.06	0.01	2.08×10^{-4}	2.34×10^{-5}
GT	0.06	0.01	0.06	0.01	4.36×10^{-4}	3.30×10^{-5}
TA	0.05	0.02	0.04	0.02	2.70×10^{-4}	2.72×10^{-5}
TC	0.07	0.04	0.06	0.05	1.20×10^{-4}	3.78×10^{-5}
TG	0.04	0.01	0.04	0.01	3.03×10^{-4}	4.91×10^{-5}

Table 3.4: Substitution profile statistics resulting from Samtools and Auditor RNA-edit classification of 11 DLBCL cases. Auditor generates a profile with higher proportion of A→G substitutions, and lower proportions of all other substutions, consistent with the canonical distribution.

whereas the more specific predictions of Auditor are more likely to be true RNA-edits.

3.4 Auditor Analysis Reveals RNA-Editing to Be an Abundant and Stable Process in Human Cancers

With accuracy metrics established on both simulated and real data, Auditor was used to profile two histologically distinct cancer types for global patterns of RNA-editing. In addition the the DLBCL data set described above, sixteen triple-negative breast cancer (TNBC) genome-transcriptome pairs from a recent study (36) were analyzed (sample acquisition and preparation are described in the original manuscript and coverage statistics can be found in Appendix B).

Using conservative thresholds (MutationSeq probability > 0.9), 6,256 RNA-edits were called in the 11 DLBCL cases, with a mean of 575.2 RNA-edits per case, affecting 1257 genes, with a mean of 308.0 genes per case (Figure 3.5A,C). The TNBC cases exhibited 8,023 RNA-edits, with a mean of 509.1 RNA-edits per case, affecting 2243 genes with a mean of 348.6 genes per case (Figure 3.5B,D).

The substitution distribution in both DLBCL and TNBC was dominated by A→G substitutions comprising 80% and 65% of events, respectively (Figures 3.4B and 3.6B). This suggests an overall similarity in RNA-editing processes between the two cancer types (*e.g.* canonical ADAR-mediated

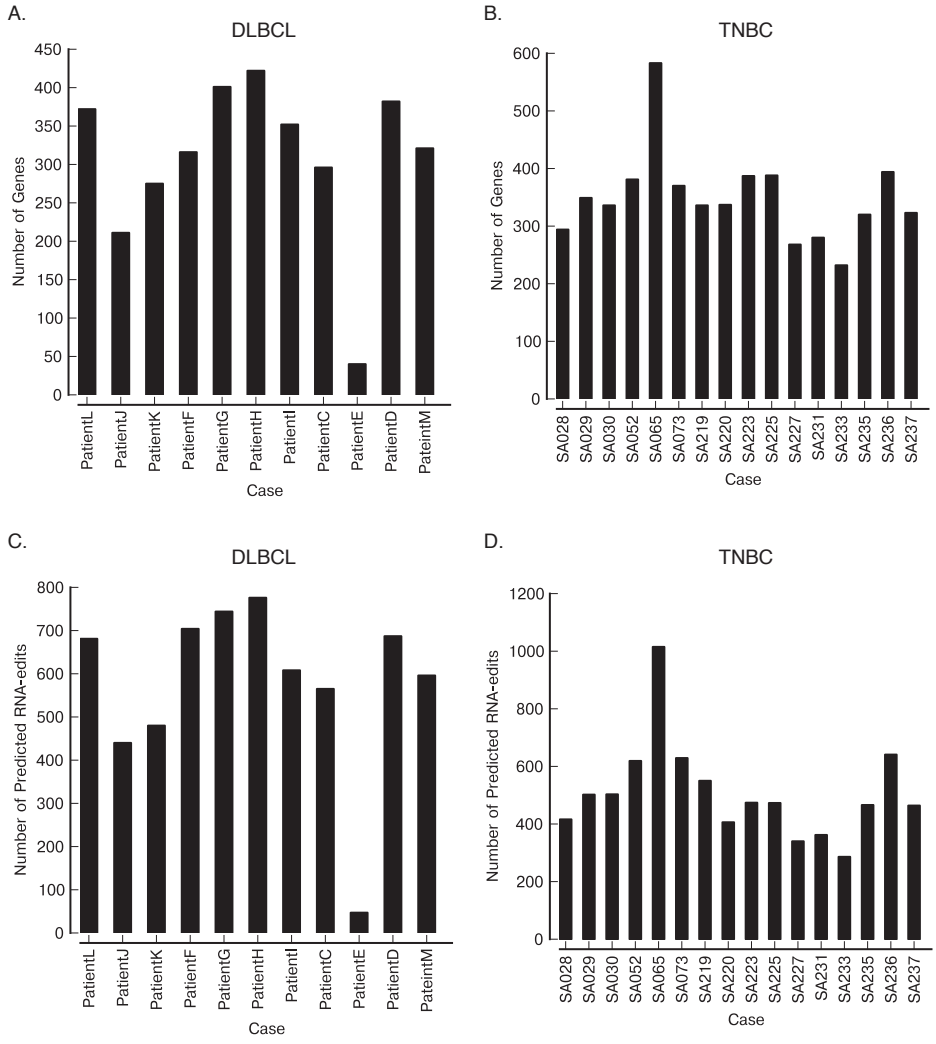


Figure 3.5: The distribution of the number RNA-edit affected genes per case for DLBCL **(A.)** and TNBC **(B.)**. The distribution of the number of RNA-edits per case for DLBCL **(C.)** and TNBC **(D.)**. As expected the two metrics are highly correlated. While there appears to be non-negligible variance between cases, it is correlated to sequence coverage (Pearson r = 0.47) suggesting that at least some of the variation can be attributed to differing amounts of RNA sequence (see Figure B.1).

Substitution	Mean	Median	Variance
AC	0.02	0.01	3.55×10^{-5}
AG	0.65	0.67	0.01
AT	0.02	0.02	6.75×10^{-5}
CA	0.02	0.02	5.35×10^{-5}
CG	0.02	0.02	2.13×10^{-5}
CT	0.06	0.06	3.44×10^{-4}
GA	0.07	0.07	4.12×10^{-4}
GC	0.01	0.01	2.05×10^{-5}
GT	0.01	0.01	6.85×10^{-5}
TA	0.03	0.02	1.23×10^{-4}
TC	0.08	0.07	3.99×10^{-4}
TG	0.02	0.02	6.84×10^{-5}

Table 3.5: Substitution profile statistics resulting from Auditor RNA-edit classification of 16 TNBC cases. Despite the A→G proportion being lower than that observed in the DLBCL cases, it is still dominant over all other substitutions, consistent with the canonical distribution.

RNA-editing is active in both), but does not discount the possibility of more subtle variations. Moreover, these results indicate that additional enzymatic processes (*i.e.* C→T deamination by APOBEC enzymes) play at most a minor role in the landscape of RNA-editing in human cancers. Notably, the observed substitution distributions were relatively consistent within cases in both DLBCL (Figure 3.4) and TNBC (Figure3.6). The variance for A→G substitutions was 0.3% and 1% in the DLBCL set and TNBC set, respectively (Tables 3.4 and 3.5).

Each putative RNA-edit called in the DLBCL and TNBC data sets was annotated with snpEff (52) as one of six classes: 3' UTR, 5' UTR, intronic, splice-site, synonymous-exonic, and non-synonymous-exonic. Next, the distribution of RNA-edits across these categories was examined as a function of MutationSeq score (Figure 3.7A,C), similarly to the procedure used above for computing substitution distributions. The DLBCL and TNBC data sets show remarkably similar class distributions with respect to both as collective cancer types (Figure 3.7C,D; Table 3.6) and as individual cases (Figure 3.8). Splice-sites, which are likely to be artifacts, are strongly over-represented with no MutationSeq post-processing, and 3' UTR events dominate the distribution of each cancer type with proportions of 0.838 and 0.793 at a MutationSeq threshold of 0.9, DLBCL and TNBC respectively. The modified MutationSeq algorithm is highly effective at removing splice-site artifacts. This is attributed largely to features 14 and 15 (sum of tail distances for non-reference bases and sum of squares of tail distances for non-reference bases, respectively) (Table 3.1). Features 14 and 15 are indirectly related as most splice-site artifacts are generated by short overhangs at the begin-

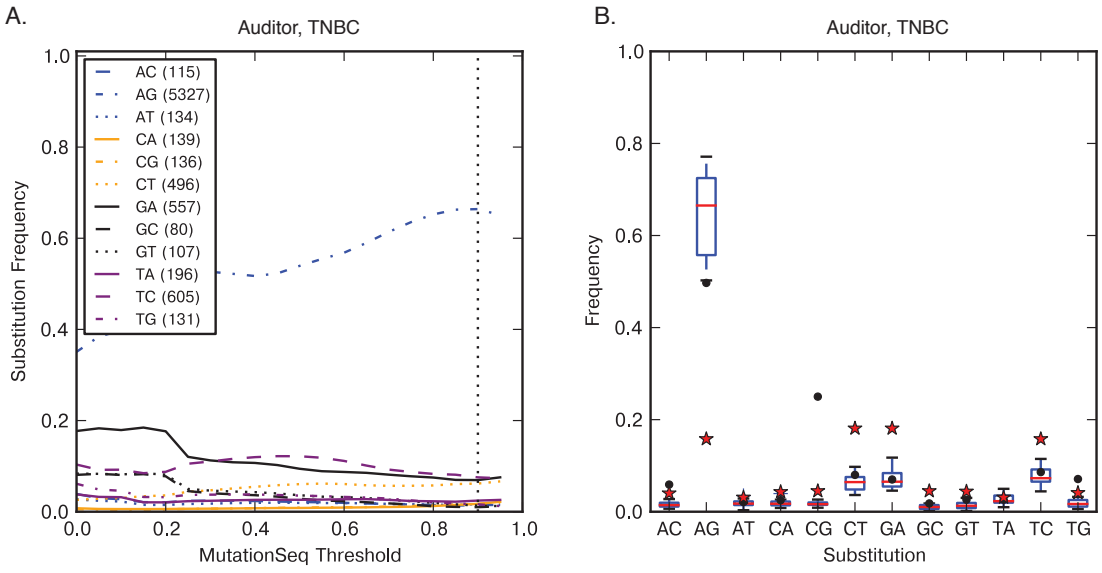


Figure 3.6: (A.) Substitution distribution of Auditor-classified RNA-edits of 16 TNBC cases as a function of MutationSeq score. Similar to the DLBCL substitution profile, A→G proportion increases significantly as MutationSeq score increases. **(B.)** Substitution distribution across the 16 TNBC cases at a MutationSeq threshold of 0.9. The distribution is remarkably similar to that of the DLBCL cases and shows little similarity to the known SNP distribution (red stars) and the model parameterization (black dots).

nings or ends of reads cause by misalignment to junction sequences and subsequent remapping to the reference genome. The use of MutationSeq for this purpose is emphasized here as a significant advancement over previous RNA-editing detection methods.

The remainder of a non-negligable number of non-synonymous RNA-edits is noted. Manual examination suggests that many of them are in fact unlikely to be genuine; many of these non-synonymous events lie only one or two nucelotides from a splice-site within an exon. However, a portion of these non-synonymous RNA-edits appear to be genuine and merit further validation.

The depletion of RNA-edits in both conding regions and 5'UTRs is unsurprising given the potential functional consequences of these modifications. Wide-spread stochastic modification of the protein complement of a cell would likely lead to challenges in regulation and isoform abundance control. Similarly, 5'UTRs are known to contain signals pertaining to protein binding and translation initiation (53), and random modification of these sites could lead to fluctuations of mRNA or protein abundance in the cell.

Overall, the most probable RNA-editing events are those that are A→G and lie within 3'UTRs. This suggests that the majority of RNA-editing activity is restricted to these properties and future, targeted analysis seems to be justified. Further, these patterns suggest that exome capture would be an ill-suited platform for RNA-editing detection due to the lack of 3'UTR sequences captured by the protocol by design.

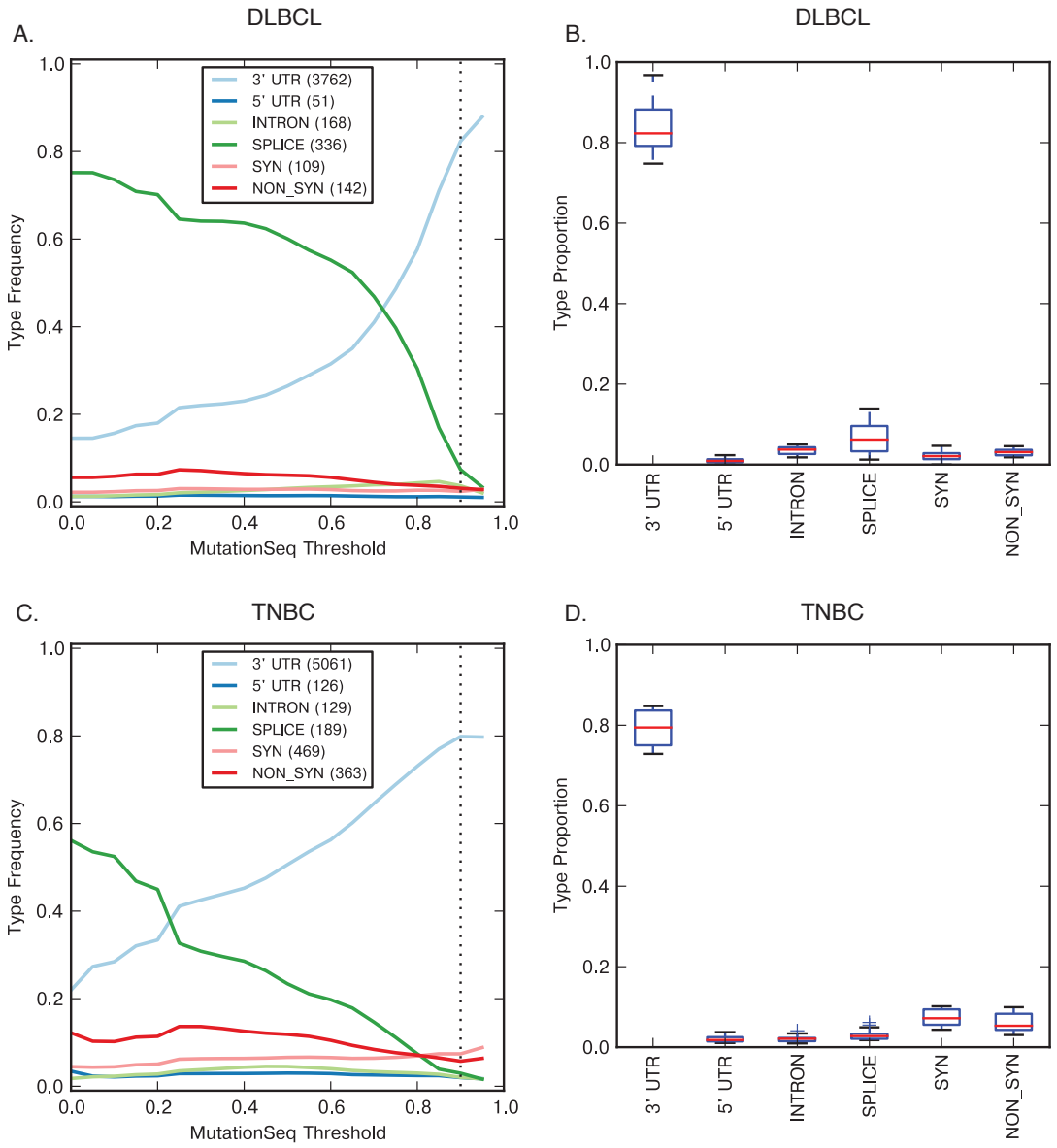


Figure 3.7: Distribution of RNA-edits across genomic regions as a function of MutationSeq threshold in DLBCL **(A.)** and TNBC **(B.)**. Splice site RNA-edits decrease sharply as MutationSeq threshold increases as expected given the artifact-prone nature of splice site alignments. The total count of events categorized into each region from each data set is shown in parentheses in the legend of each plot. The distribution of regional proportion across the each of the cancer types are shown in **(C.)** and **(D.)**, DLBCL and TNBC respectively. The overall distributions are highly similar and primarily differ in the proportion of remaining splice sites.

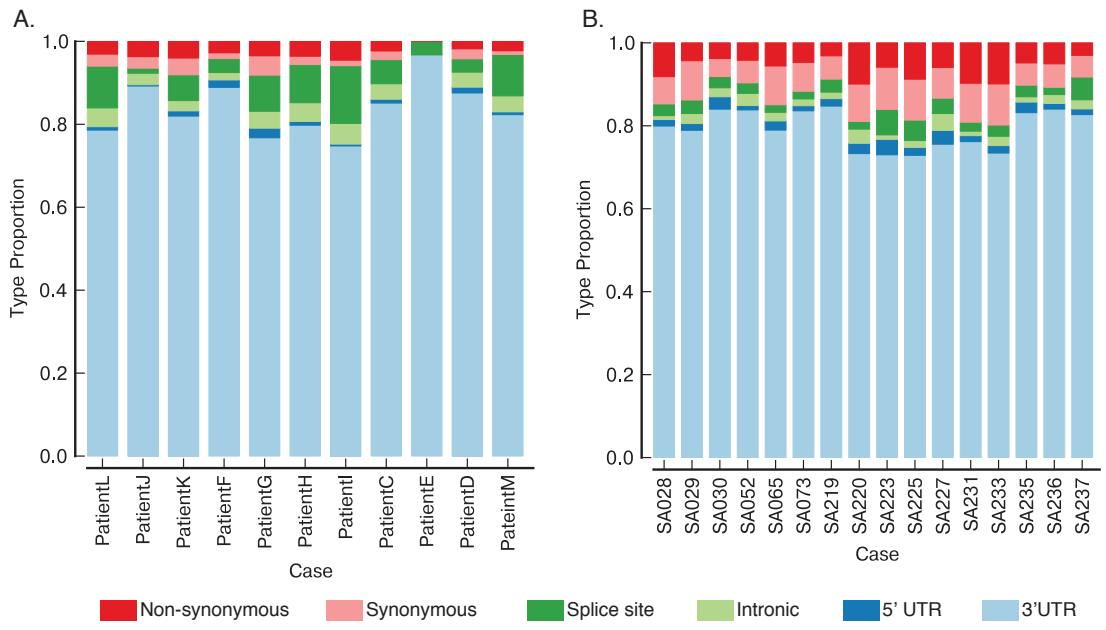


Figure 3.8: The regional distribution for each case in the DLBCL set **(A.)** and in the TNBC set **(B.)**. The distributions are highly similar both within and between the two tumour types.

Region	DLBCL Mean	TNBC Mean	DLBCL Median	TNBC Median	DLBCL Variance	TNBC Variance
3' UTR	0.84	0.79	0.82	0.79	3.81×10^{-3}	1.97×10^{-3}
5' UTR	0.01	0.02	0.01	0.02	4.12×10^{-5}	5.85×10^{-5}
INTRON	0.03	0.02	0.04	0.02	1.93×10^{-4}	7.05×10^{-5}
SPLICE	0.07	0.03	0.06	0.03	1.37×10^{-3}	1.69×10^{-4}
SYN	0.02	0.07	0.02	0.07	1.66×10^{-4}	3.98×10^{-4}
NON_SYN	0.03	0.06	0.03	0.05	1.45×10^{-4}	5.57×10^{-4}

Table 3.6: Regional Auditor classified RNA-edit distribution of 11 DLBCL and 16 TNBC cases. The abundance of RNA-edits falling in 3' UTRs is consistent between both cancer types as well as with previously reported patterns of RNA-editing.

Chapter 4

Discussion

This thesis has:

1. Demonstrated of the power of Polya distributions for genotyping in the presence of biological and technical noise
2. Shown that joint modelling, while beneficial, is of less effect than the use of Polya distributions for the discovery of RNA-edits
3. Described a novel generative model, Auditor, that, when coupled with sophisticated artifact detection accurately identifies RNA-edits from paired genome/transcriptome HTS data
4. Demonstrated the necessity of effective artifact removal in RNA-editing detection
5. Described the first landscape of RNA-editing in human cancer, and shown RNA-editing to be largely invariant in DLBCL and TNBC and that RNA-editing patterns in these tumours closely reflect the patterns seen in healthy tissues

The current state of the art knowledge regarding RNA-editing, while informative, also emphasizes significant shortcomings in the understanding of this process. The advent of high-throughput sequencing now provides a means by which the scientific community can begin to elaborate on underexplored aspects of RNA-editing as well as to potentially inspire novel insights into the mechanisms and effects of post-transcriptional modification of RNA.

Despite the power of HTS data for this purpose, previously published methods for RNA-edit discovery have lacked rigour and the results of some studies remain suspect. Auditor paves the way future studies and discoveries as the first mathematically principled method for RNA-edit classification from HTS data. Further, Auditor has provided the first large scale exploration of RNA-editing in human cancers and found RNA-editing to be remarkably stable inspite of malignancy and cellular lineage.

4.1 Auditor is a Principled Method for the Detection of RNA-Edits

Auditor, with the introduction of joint Polya-mixture modelling as well as the base-specific transition matrix, represents a significant improvement over existing methods. The use of Polya distributions explicitly encodes the count of each base as well as the increased variance of RNA-seq and RNA-editing. This approach results in significant improvements over both Multinomial distribution methods and the Samtools/bcftools software.

Additionally, the joint modelling of genotype and transcriptotype allows the model to borrow statistical strength between the DNA and RNA samples increasing the specificity of RNA-edit classifications. While the use of joint modelling does not result in performance gains on the same scale as the usage of Polya distributions, the improvements are measurable and significant.

4.2 Effective Artifact Removal is Critical to RNA-editing Detection

In this study, in both substitution profiling and regional analysis, only after artifact removal were the expected patterns generated, demonstrating that sophisticated artifact detection approaches are a necessary component of RNA-edit detection. It is clear that without such artifact removal, RNA-editing detection becomes overwhelmed with technical artifacts and no biological conclusions can be drawn in the face of such a severe false positive burden.

Discriminative modelling of technical artifacts in HTS data has previously proven to be effective in controlling false positive classifications of somatic mutations (48), and the twenty-one feature adaptation presented here is highly effective in the context of RNA-editing detection. Given the recent uncertainty in the literature (20, 21, 22), and the overall laxity of statistical rigour in current artifact detection approaches in published RNA-editing studies (12, 18, 19, 40, 41), use of principled approaches, such as the modified MutationSeq algorithm presented here, could provide much needed confidence in the field.

4.3 Preliminary Landscape of RNA-Editing in DLBCL and TNBC

This study represents the first examination of the landscape of RNA-editing in DLBCL and TNBC. A$\rightarrow$G RNA-editing dominates the substitution landscape, and the majority of high-confidence predictions lie within 3' UTR regions.

There is striking similarity within and between the DLBCL and TNBC samples, and further the patterns observed closely match the known patterns found in healthy tissues. There is little evidence in this study of what, if any, role RNA-editing may play in these cancers, but for these cancer types it seems that any effects will be the result of fine scale aberrations rather than gross, mechanistic divergences from normal activity. Further, the evidence presented here suggests the possibility that RNA-editing mechanisms, as an active and necessary cellular process, is generally invariant across cell types of differing lineages and even in the presence of malignancy.

4.4 Limitations

Currently Auditor admits several limitations. The most significant limitation is that the method requires matched DNA and RNA HTS data. However, it is noted that this limitation is shared by all existing RNA-edit discovery tools. Further, in light of the evidence presented here, and elsewhere (12), future large scale studies could alleviate the financial and technical burden by performing targeted sequencing of 3'UTR sequences only. While this approach may fail to elucidate novel properties of RNA-editing it would capture the significant majority of RNA-editing activity and potentially provide further insights into the consequences of RNA-editing. A second limitation is that sensitivity is currently reliant solely on count data, whereas specificity is controlled by count data and the features used by MutationSeq. Finally, it is noted that while the model is theoretically suited to parameter learning via EM, the process is too computationally intensive, due to the necessity of large matrix operations, to train on entire data sets.

4.5 Future Directions

Currently Auditor uses a single set of parameters regardless of the functional region currently being examined. This thesis has described significant biases to the prevalence of RNA-editing in various genomic regions and region-specific parameters appear to be justified. The structure of Auditor would easily allow for the creation of separate transition matrices and Polya mixtures for each functional region, and the annotated region would simply become a known-value indicator variable that determines which set of parameters are used for any given classification. In a similar sense, the necessity of RNA secondary structure for ADAR mediated RNA-editing has been established, and RNA secondary structure prediction could be overlaid as a prior on the parameters to further inform RNA-edit classification in regions of varying secondary structure.

In addition to secondary structure considerations, ADAR mediated events are known to be spatially correlated; a feature ignored by the current implementation of Auditor. A modelling approach, such as an HMM, that utilized these patterns could be a powerful tool for detecting RNA-edits arising from specific processes.

It was demonstrated throughout this study that discriminative classifier based artifact removal is a necessary and powerful element of the described method. However, the classifier used here is trained on RNA-Seq based variants only, due to the lack of a comprehensive set of validated RNA-edits. In the future, once such a set exists, a full MutationSeq framework can be constructed, simultaneously using features of both the DNA and RNA, and trained on RNA-edits. This would alleviate the shortcoming of sensitivity relying on base counts alone and further increase the performance of the model.

Addressing the final limitation noted above, as new data emerges, it should be possible to train on a carefully selected subset of positions for improved results. This would allow for discovery of substitution distributions from entire samples as well as more accurate measurements of RNA-editing rates on a per-sample basis, which are currently only available as parameter biased post-

processing steps. In addition to training the transition matrix, an attempt was also made to train the Polya mixture distributions from labelled RNA-seq data using second order gradient descent. However, all training attempts resulted in pseudo-counts of less than one for all parameters encoding the heterozygous types. When such parameterizations are used, the behaviour of the Polya distribution becomes ill-suited to nucleotide typing as the probability density becomes concentrated at the edges of the distributions. This would, for example, manifest as the CT distribution being most likely in the presence of all C's or all T's and unlikely in the presence of an equal mixture. Developing effective training for the Polya parameters would likely increase performance, but will require significant further development.

As a final research detection, the repurposing of Auditor for somatic mutation detection is proposed. Audior could be used to simultaneously infer mutational signatures (*e.g.* substitution profiles) and mutated positions from matched healthy and diseased genomic samples. This functionality is unavailble in any other tool to date and represents an unexplored area of somatic mutation research.

4.6 Conclusions

Auditor is the first principled method for RNA-editing detection, and is a clear improvement over previous methods. The patterns of RNA-editing discovered by applying Auditor to twenty-seven cancer samples closely match the canonical prototype, demonstrating the effectiveness of the model and providing the preliminary landscape for RNA-editing in DLBCL and TNBC. RNA-editing in DLBCL and TNBC was found to be largely invariant within and between the cancer types, suggesting significant stability of the RNA-editing processes in the contexts of both malignancy and distinct cellular lineages.

Additionally, Auditor will enable future studies of RNA-editing in both healthy and diseased tissue, and due to the extensible design of the model will only benefit and improve as new data and insights arise.

Bibliography

[1] D. Baltimore. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature*, 226(5252):1209–1211, Jun 1970. → pages 1

[2] H. M. Temin and S. Mizutani. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature*, 226(5252):1211–1213, Jun 1970. → pages 1

[3] L. M. Powell, S. C. Wallis, R. J. Pease, Y. H. Edwards, T. J. Knott, and J. Scott. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell*, 50(6): 831–840, Sep 1987. → pages 1, 3, 4

[4] S. H. Chen, G. Habib, C. Y. Yang, Z. W. Gu, B. R. Lee, S. A. Weng, S. R. Silberman, S. J. Cai, J. P. Deslypere, and M. Rosseneu. Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science*, 238(4825):363–366, Oct 1987. → pages 1, 3

[5] B. L. Bass and H. Weintraub. An unwinding activity that covalently modifies its double-stranded RNA substrate. *Cell*, 55(6):1089–1098, Dec 1988. → pages 1, 2

[6] T. Melcher, S. Maas, A. Herb, R. Sprengel, P. H. Seeburg, and M. Higuchi. A mammalian RNA editing enzyme. *Nature*, 379(6564):460–464, Feb 1996. doi:10.1038/379460a0. URL http://dx.doi.org/10.1038/379460a0. → pages 1

[7] 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):303–308, May 1997. doi:10.1038/387303a0. URL http://dx.doi.org/10.1038/387303a0. → pages 1

[8] L. F. Landweber and W. Gilbert. Phylogenetic analysis of RNA editing: a primitive genetic phenomenon. *Proc Natl Acad Sci U S A*, 91(3):918–921, Feb 1994. → pages 2

[9] Brenda L. Bass. RNA editing by adenosine deaminases that act on RNA. *Annu Rev Biochem*, 71:817–846, 2002. doi:10.1146/annurev.biochem.71.110601.135501. URL http://dx.doi.org/10.1146/annurev.biochem.71.110601.135501. → pages 2

[10] J. Cruz-Reyes, L. N. Rusch, K. J. Piller, and B. Sollner-Webb. T. brucei RNA editing: adenosine nucleotides inversely affect U-deletion and U-insertion reactions at mRNA cleavage. *Mol Cell*, 1(3):401–409, Feb 1998. → pages 2

[11] 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 2009. doi:10.1126/science.1170995. URL http://dx.doi.org/10.1126/science.1170995. → pages 2, 4, 35

[12] Zhiyu Peng, Yanbing Cheng, Bertrand Chin-Ming Tan, Lin Kang, Zhijian Tian, Yuankun Zhu, Wenwei Zhang, Yu Liang, Xueda Hu, Xuemei Tan, Jing Guo, Zirui Dong, Yan Liang, Li Bao, and Jun Wang. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat Biotechnol*, 30(3):253–60, 2012. doi:10.1038/nbt.2122. → pages 2, 3, 8, 34, 35, 44, 45

[13] Julie M. Eggington, Tom Greene, and Brenda L. Bass. Predicting sites of ADAR editing in double-stranded RNA. *Nat Commun*, 2:319, 2011. doi:10.1038/ncomms1324. URL http://dx.doi.org/10.1038/ncomms1324. → pages

[14] A. G. Polson and B. L. Bass. Preferential selection of adenosines for modification by double-stranded RNA adenosine deaminase. *EMBO J*, 13(23):5701–5711, Dec 1994. → pages 2

[15] K. A. Lehmann and B. L. Bass. Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. *Biochemistry*, 39(42):12875–12884, Oct 2000. → pages 2

[16] Matthew Blow, P Andrew Futreal, Richard Wooster, and Michael R. Stratton. A survey of RNA editing in human brain. *Genome Res*, 14(12):2379–2387, Dec 2004. doi:10.1101/gr.2951204. URL http://dx.doi.org/10.1101/gr.2951204. → pages

[17] A. Mehta, M. T. Kinter, N. E. Sherman, and D. M. Driscoll. Molecular cloning of apobec-1 complementation factor, a novel RNA-binding protein involved in the editing of apolipoprotein B mRNA. *Mol Cell Biol*, 20(5):1846–1854, Mar 2000. → pages 3

[18] Mingyao Li, Isabel X. Wang, Yun Li, Alan Bruzel, Allison L. Richards, Jonathan M. Toung, and Vivian G. Cheung. Widespread RNA and DNA sequence differences in the human transcriptome. *Science*, 333(6038):53–58, Jul 2011. doi:10.1126/science.1207018. URL http://dx.doi.org/10.1126/science.1207018. → pages 3, 8, 44

[19] J H Bahn, J H Lee, G Li, C Greer, G Peng, and X Xiao. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res*, 22(1):142–150, Jan 2012. doi:10.1101/gr.124107.111. URL http://www.hubmed.org/display.cgi?uids=21960545. → pages 3, 8, 18, 44

[20] Claudia L Kleinman and Jacek Majewski. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*, 335(6074):1302; author reply 1302, Mar 2012. doi:10.1126/science.1209658. → pages 3, 44

[21] Joseph K Pickrell, Yoav Gilad, and Jonathan K Pritchard. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*, 335(6074):1302; author reply 1302, Mar 2012. doi:10.1126/science.1210484. → pages 3, 44

[22] Wei Lin, Robert Piskol, Meng How Tan, and Jin Billy Li. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*, 335(6074):1302; author reply 1302, Mar 2012. doi:10.1126/science.1210624. → pages 3, 44

[23] Alekos Athanasiadis, Alexander Rich, and Stefan Maas. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol*, 2(12):e391, Dec 2004. doi:10.1371/journal.pbio.0020391. URL http://dx.doi.org/10.1371/journal.pbio.0020391. → pages 3

[24] Sohrab P Shah, Ryan D Morin, Jaswinder Khattra, Leah Prentice, Trevor Pugh, Angela Burleigh, Allen Delaney, Karen Gelmon, Ryan Guliany, Janine Senz, Christian Steidl, Robert A Holt, Steven Jones, Mark Sun, Gillian Leung, Richard Moore, Tesa Severson, Greg A Taylor, Andrew E Teschendorff, Kane Tse, Gulisa Turashvili, Richard Varhol, René L Warren, Peter Watson, Yongjun Zhao, Carlos Caldas, David Huntsman, Martin Hirst, Marco A Marra, and Samuel Aparicio. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, 461(7265):809–13, Oct 2009. doi:10.1038/nature08489. → pages 3, 4

[25] Jochen C. Hartner, Carolin Schmittwolf, Andreas Kispert, Albrecht M. Mller, Miyoko Higuchi, and Peter H. Seeburg. Liver disintegration in the mouse embryo caused by deficiency in the RNA-editing enzyme ADAR1. *J Biol Chem*, 279(6):4894–4902, Feb 2004. doi:10.1074/jbc.M311347200. URL http://dx.doi.org/10.1074/jbc.M311347200. → pages 4

[26] 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(6791):78–81, Jul 2000. doi:10.1038/35017558. URL http://dx.doi.org/10.1038/35017558. → pages 4

[27] Barry Hoopengardner, Tarun Bhalla, Cynthia Staber, and Robert Reenan. Nervous system targets of RNA editing identified by comparative genomics. *Science*, 301(5634):832–836, Aug 2003. doi:10.1126/science.1086763. URL http://dx.doi.org/10.1126/science.1086763. → pages

[28] Erez Y. Levanon, Martina Hallegger, Yaron Kinar, Ronen Shemesh, Kristina Djinovic-Carugo, Gideon Rechavi, Michael F. Jantsch, and Eli Eisenberg. Evolutionarily conserved human targets of adenosine to inosine RNA editing. *Nucleic Acids Res*, 33(4): 1162–1168, 2005. doi:10.1093/nar/gki239. URL http://dx.doi.org/10.1093/nar/gki239. → pages

[29] M. M. Vniant, C. H. Zlot, R. L. Walzem, V. Pierotti, R. Driscoll, D. Dichek, J. Herz, and S. G. Young. Lipoprotein clearance mechanisms in LDL receptor-deficient "Apo-B48-only" and "Apo-B100-only" mice. *J Clin Invest*, 102(8):1559–1568, Oct 1998. doi:10.1172/JCI4164. URL http://dx.doi.org/10.1172/JCI4164. → pages 4

[30] 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(25):14687–14692, Dec 2001. doi:10.1073/pnas.251531398. URL http://dx.doi.org/10.1073/pnas.251531398. → pages 4

[31] Nurit Paz, Erez Y. Levanon, Ninette Amariglio, Amy B. Heimberger, Zvi Ram, Shlomi Constantini, Zohar S. Barbash, Konstantin Adamsky, Michal Safran, Avi Hirschberg, Meir Krupsky, Issachar Ben-Dov, Simona Cazacu, Tom Mikkelsen, Chaya Brodie, Eli Eisenberg, and Gideon Rechavi. Altered adenosine-to-inosine RNA editing in human cancer. *Genome Res*, 17(11):1586–1595, Nov 2007. doi:10.1101/gr.6493107. URL http://dx.doi.org/10.1101/gr.6493107. → pages 4

[32] Olena Morozova and Marco A. Marra. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5):255–264, Nov 2008. doi:10.1016/j.ygeno.2008.07.001. URL http://dx.doi.org/10.1016/j.ygeno.2008.07.001. → pages 5

[33] Jocelyn Kaiser. DNA sequencing. A plan to capture human diversity in 1000 genomes. *Science*, 319(5862):395, Jan 2008. doi:10.1126/science.319.5862.395. URL http://dx.doi.org/10.1126/science.319.5862.395. → pages 5

[34] International Cancer Genome Consortium. International network of cancer genome projects. *Nature*, 464(7291):993–998, Apr 2010. doi:10.1038/nature08987. URL http://dx.doi.org/10.1038/nature08987. → pages 5

[35] Ryan D Morin, Maria Mendez-Lago, Andrew J Mungall, Rodrigo Goya, Karen L Mungall, Richard D Corbett, Nathalie A Johnson, Tesa M Severson, Readman Chiu, Matthew Field, Shaun Jackman, Martin Krzywinski, David W Scott, Diane L Trinh, Jessica Tamura-Wells, Sa Li, Marlo R Firme, Sanja Rogic, Malachi Griffith, Susanna Chan, Oleksandr Yakovenko, Irmtraud M Meyer, Eric Y Zhao, Duane Smailus, Michelle Moksa, Suganthi Chittaranjan, Lisa Rimsza, Angela Brooks-Wilson, John J Spinelli, Susana Ben-Neriah, Barbara Meissner, Bruce Woolcock, Merrill Boyle, Helen McDonald, Angela Tam, Yongjun Zhao, Allen Delaney, Thomas Zeng, Kane Tse, Yaron Butterfield, Inanç Birol, Rob Holt, Jacqueline Schein, Douglas E Horsman, Richard Moore, Steven J M Jones, Joseph M Connors, Martin Hirst, Randy D Gascoyne, and Marco A Marra. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature*, 476(7360):298–303, Aug 2011. doi:10.1038/nature10351. → pages 5, 31

[36] Sohrab P. Shah, Andrew Roth, Rodrigo Goya, Arusha Oloumi, Gavin Ha, Yongjun Zhao, Gulisa Turashvili, Jiarui Ding, Kane Tse, Gholamreza Haffari, Ali Bashashati, Leah M. Prentice, Jaswinder Khattra, Angela Burleigh, Damian Yap, Virginie Bernard, Andrew McPherson, Karey Shumansky, Anamaria Crisan, Ryan Giuliany, Alireza Heravi-Moussavi, Jamie Rosner, Daniel Lai, Inanc Birol, Richard Varhol, Angela Tam, Noreen Dhalla, Thomas Zeng, Kevin Ma, Simon K. Chan, Malachi Griffith, Annie Moradian, S-W Grace Cheng, Gregg B. Morin, Peter Watson, Karen Gelmon, Stephen Chia, Suet-Feung Chin, Christina Curtis, Oscar M. Rueda, Paul D. Pharoah, Sambasivarao Damaraju, John Mackey, Kelly Hoon, Timothy Harkins, Vasisht Tadigotla, Mahvash Sigaroudinia, Philippe Gascard, Thea Tlsty, Joseph F. Costello, Irmtraud M. Meyer, Connie J. Eaves, Wyeth W. Wasserman, Steven Jones, David Huntsman, Martin Hirst, Carlos Caldas, Marco A. Marra, and Samuel Aparicio. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, Apr 2012. doi:10.1038/nature10933. URL http://dx.doi.org/10.1038/nature10933. → pages 5, 8, 37

[37] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup . The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009. → pages 5, 8, 31

[38] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing

next-generation DNA sequencing data. *Genome Res*, 20(9):1297–1303, Sep 2010. doi:10.1101/gr.107524.110. URL http://dx.doi.org/10.1101/gr.107524.110. → pages 5, 30

[39] Rodrigo Goya, Mark G F. Sun, Ryan D. Morin, Gillian Leung, Gavin Ha, Kimberley C. Wiegand, Janine Senz, Anamaria Crisan, Marco A. Marra, Martin Hirst, David Huntsman, Kevin P. Murphy, Sam Aparicio, and Sohrab P. Shah. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, 26(6):730–736, Mar 2010. doi:10.1093/bioinformatics/btq040. URL http://dx.doi.org/10.1093/bioinformatics/btq040. → pages 8

[40] Gokul Ramaswami, Wei Lin, Robert Piskol, Meng How Tan, Carrie Davis, and Jin Billy Li. Accurate identification of human Alu and non-Alu RNA editing sites. *Nat Methods*, Apr 2012. doi:10.1038/nmeth.1982. URL http://dx.doi.org/10.1038/nmeth.1982. → pages 8, 44

[41] Petr Danecek, Christoffer Nellker, Rebecca E. McIntyre, Jorge E. Buendia-Buendia, Suzannah Bumpstead, Chris P. Ponting, Jonathan Flint, Richard Durbin, Thomas M. Keane, and David J. Adams. High levels of RNA-editing site conservation amongst 15 laboratory mouse strains. *Genome Biol*, 13(4):r26, Apr 2012. doi:10.1186/gb-2012-13-4-r26. URL http://dx.doi.org/10.1186/gb-2012-13-4-r26. → pages 8, 11, 44, 54

[42] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21): 2987–2993, Nov 2011. doi:10.1093/bioinformatics/btr509. URL http://dx.doi.org/10.1093/bioinformatics/btr509. → pages 8

[43] Ruiqiang Li, Yingrui Li, Xiaodong Fang, Huanming Yang, Jian Wang, Karsten Kristiansen, and Jun Wang. SNP detection for massively parallel whole-genome resequencing. *Genome Res*, 19(6):1124–1132, Jun 2009. doi:10.1101/gr.088013.108. URL http://dx.doi.org/10.1101/gr.088013.108. → pages 8

[44] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–1858, Nov 2008. doi:10.1101/gr.078212.108. URL http://dx.doi.org/10.1101/gr.078212.108. → pages 8

[45] Andrew Roth, Jiarui Ding, Ryan Morin, Anamaria Crisan, Gavin Ha, Ryan Giuliany, Ali Bashashati, Martin Hirst, Gulisa Turashvili, Arusha Oloumi, Marco A. Marra, Samuel Aparicio, and Sohrab P. Shah. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics*, 28(7):907–913, Apr 2012. doi:10.1093/bioinformatics/bts053. URL http://dx.doi.org/10.1093/bioinformatics/bts053. → pages 8, 14

[46] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nat Biotechnol*, 26(10): 1135–1145, Oct 2008. doi:10.1038/nbt1486. URL http://dx.doi.org/10.1038/nbt1486. → pages 9

[47] Michael A. Chapman, Michael S. Lawrence, Jonathan J. Keats, Kristian Cibulskis, Carrie Sougnez, Anna C. Schinzel, Christina L. Harview, Jean-Philippe Brunet, Gregory J. Ahmann, Mazhar Adli, Kenneth C. Anderson, Kristin G. Ardlie, Daniel Auclair, Angela Baker, P Leif Bergsagel, Bradley E. Bernstein, Yotam Drier, Rafael Fonseca, Stacey B. Gabriel, Craig C. Hofmeister, Sundar Jagannath, Andrzej J. Jakubowiak, Amrita Krishnan, Joan Levy, Ted Liefeld, Sagar Lonial, Scott Mahan, Bunmi Mfuko, Stefano Monti, Louise M. Perkins, Robb

Onofrio, Trevor J. Pugh, S Vincent Rajkumar, Alex H. Ramos, David S. Siegel, Andrey Sivachenko, A Keith Stewart, Suzanne Trudel, Ravi Vij, Douglas Voet, Wendy Winckler, Todd Zimmerman, John Carpten, Jeff Trent, William C. Hahn, Levi A. Garraway, Matthew Meyerson, Eric S. Lander, Gad Getz, and Todd R. Golub. Initial genome sequencing and analysis of multiple myeloma. *Nature*, 471(7339):467–472, Mar 2011. doi:10.1038/nature09837. URL http://dx.doi.org/10.1038/nature09837. → pages 9

[48] J Ding, A Bashashati, A Roth, A Oloumi, K Tse, T Zeng, G Haffari, M Hirst, M A Marra, A Condon, S Aparicio, and S P Shah. Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics*, 28(2):167–175, Jan 2012. doi:10.1093/bioinformatics/btr629. URL http://www.hubmed.org/display.cgi?uids=22084253. → pages 9, 10, 11, 30, 44

[49] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009. doi:10.1093/bioinformatics/btp324. URL http://dx.doi.org/10.1093/bioinformatics/btp324. → pages 31

[50] Anmol Kiran and Pavel V. Baranov. DARNED: a DAtabase of RNa EDiting in humans. *Bioinformatics*, 26(14):1772–1776, Jul 2010. doi:10.1093/bioinformatics/btq285. URL http://dx.doi.org/10.1093/bioinformatics/btq285. → pages 32

[51] Eli Eisenberg, Konstantin Adamsky, Lital Cohen, Ninette Amariglio, Abraham Hirshberg, Gideon Rechavi, and Erez Y. Levanon. Identification of RNA editing sites in the SNP database. *Nucleic Acids Res*, 33(14):4612–4617, 2005. doi:10.1093/nar/gki771. URL http://dx.doi.org/10.1093/nar/gki771. → pages 34

[52] P Cingolani. snpEff: Variant effect prediction, 2012. URL http://snpeff.sourceforge.net. → pages 39

[53] Can Cenik, Adnan Derti, Joseph C. Mellor, Gabriel F. Berriz, and Frederick P. Roth. Genome-wide functional analysis of human 5' untranslated region introns. *Genome Biol*, 11 (3):R29, 2010. doi:10.1186/gb-2010-11-3-r29. URL http://dx.doi.org/10.1186/gb-2010-11-3-r29. → pages 40

Appendix A

Method Workflows

A.1 Auditor Workflow

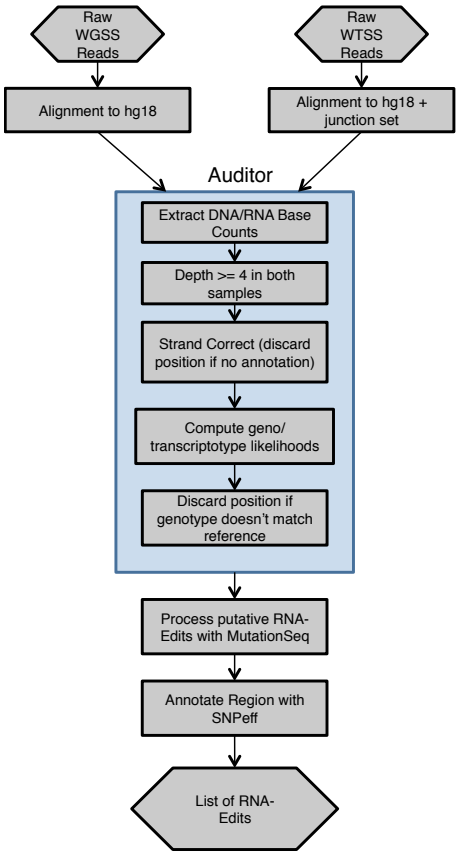


Figure A.1: Auditor workflow. DNA and RNA NGS data are converted to counts subject to user-specified constraints (e.g. depth and quality), strand corrected and then the likelihood of an RNA-edit at each position is computed. Positions where p(edit) > 0.5 are classified with MutationSeq to assess confidence that the RNA variant is genuine.

A.2 Samtools Workflow

- Published method (41):
  - Align DNA and RNA to reference genome with BWA
  - Use Samtools on DNA and RNA to call types
  - Identify putative RNA-edits by deterministically selecting positions where:
    - RNA contains a variant
    - DNA is homozygous for reference
  - Realign reads supporting RNA-edits with a junction aware aligner to remove junction artifacts
  - Filter on Samtools p-vals for various artifact categories
- The Samtools/bcftools procedure described in the algorithm is a similar approach but more principled:
  - Align DNA with BWA to reference genome
  - Align RNA with BWA to reference genome + known junctions
  - Use Samtools on DNA and RNA to call types
  - Identify putative RNA-edits by deterministically selecting positions where:
    - RNA contains a variant
    - DNA is homozygous for reference
  - Use modified MutationSeq using Samtools features, plus 'distance-to-nearest-junction'
  - Rank calls by MutationSeq score

The realignment step is replaced with a supplemented set of junction sequences during the alignment of the RNA-seq data. Additionally, the Samtools-based features used by MutationSeq are augmented with a 'distance-to-nearest-junction' feature. Similarly, heuristic filters on Samtools p-values (for features such as $p(strand-bias)$), are replaced by the MutationSeq platform, granting the ability to control the confidence threshold, rather than setting arbitrary cut-offs for these p-values.

Appendix B

Sequence Data Coverage Statistics and Correlation to RNA-edits

Patient ID	Genome Library	WTSS Library	Genome Cover (fold)	WTSS Mapped Bases
PatientL	A01413	HS0637	51.74	9115713150
PatientJ	A01415	HS0926	28.26	2734734350
PatientK	A01416	HS0928	24.9	5181143850
PatientF	A01418	HS0936	31.91	5369008300
PatientG	A01424	HS1462	31.81	6288448875
PatientH	A01434	HS2605	32.82	10830100200
PatientI	A01453	HS2048	26.42	12120724350
PatientC	A03291	HS3105	29.14	9127296000
PatientE	HS2702	HS0647	29.76	1828711740
PatientD	HS2706	HS1133	40.49	5081588150
PatientM	HS2974	HS2051	28.53	10768700700

Table B.1: DLBCL Sequence Coverage Statistics

Case	Genome Cover (fold)	WTSS Mapped Bases
SA028	26.04	4752835350
SA029	20.28	5867238050
SA030	19.69	6217164650
SA052	20.97	5907148250
SA065	15.88	6992404750
SA073	20.43	7361049850
SA219	25.38	6399383450
SA220	30.84	6832277300
SA223	35.28	6643575050
SA224	34.49	14510812300
SA225	42.2	10555014600
SA227	26.67	5676236200
SA231	21.42	3111736850
SA233	28.9	5712104050
SA235	32.89	5756271050
SA236	27.47	6232548250
SA237	27.07	6819030700

Table B.2: TNBC Sequence Coverage Statistics

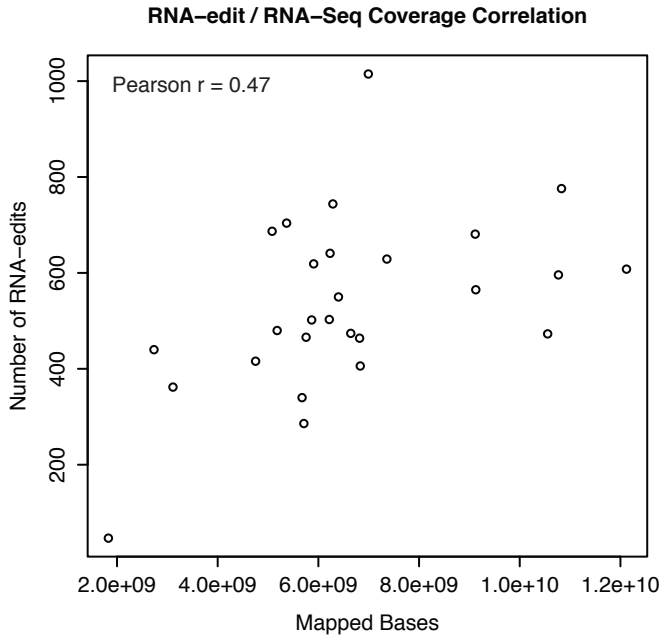


Figure B.1: Correlation of the number of putative RNA-edits and the RNA-Seq coverage per case. A positive correlation is found between the two values with Pearson r = 0.47. This suggests that at least some of the variance in number of RNA-edits per case is due to difference in sequence coverage.

Appendix C

Examples of Technical Artifacts Identified by MutationSeq

Here three examples of common technical artifacts are presented. In all cases, $p(Edit) > 0.9$ (computed by Auditor) and $p(variant) < 0.15$ (computed by MutationSeq).

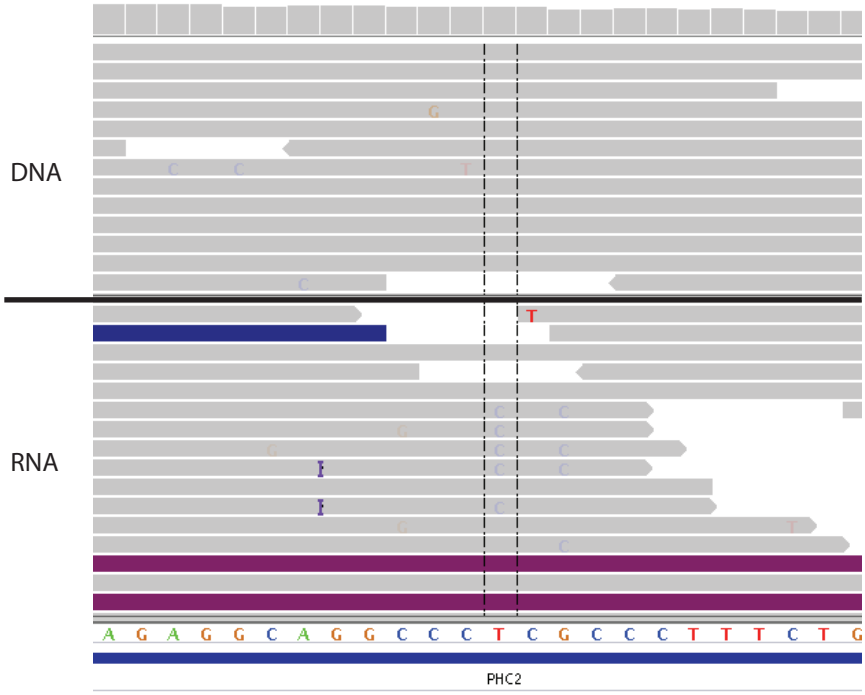


Figure C.1: A false positive RNA-edit caused by low base quality. All of the C's in this example have base quality < 10, which is visualized by high translucency. Without the MutationSeq classifier, this position would rank among the highest confidence putative RNA-edits.

Figure C.2: A false positive RNA-edit caused by low mapping quality. The reads containing a G at the position of interest all have mapping quality < 20. The transparent reads are non-uniquely mapped, suggesting that this region is repetitive and difficult to accurately map. Without the MutationSeq classifier this position would rank among the highest confidence putative RNA-edits.

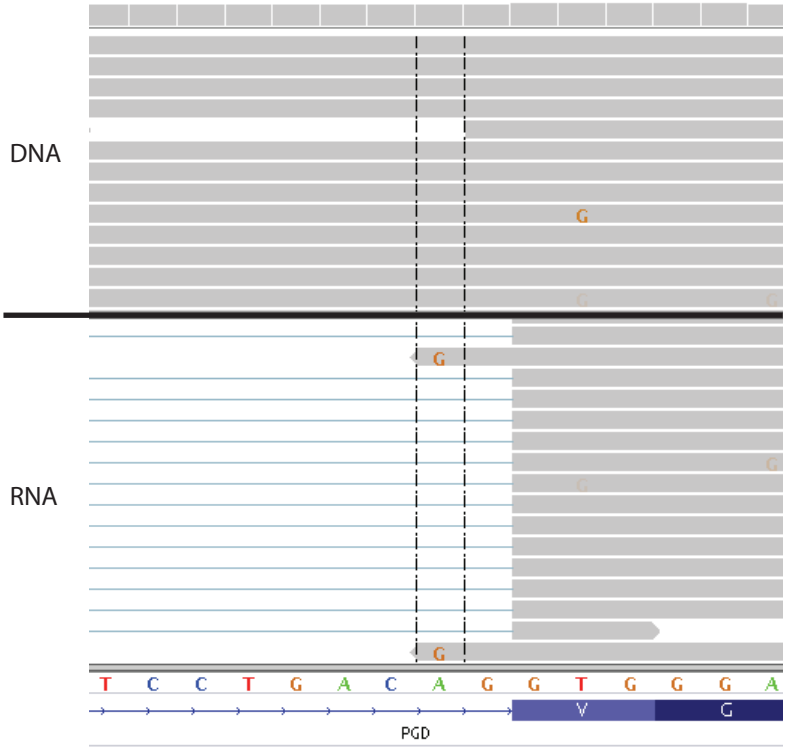


Figure C.3: A false positive RNA-edit caused by incorrect mapping proximal to an exon-exon junction. The reads containing a G at the position of interest are misaligned due to the splice site, and give rise to a high confidence RNA-edit classification. Without the MutationSeq classifier this position would rank among the highest confidence putative RNA-edits.