Identifying potential RNA editing sites from RNA-Sequencing data from human brain tissue

Francesca North

4276108

School of Life Sciences

University of Nottingham

Medical School, QMC

Nottingham

NG7 2UH

Supervisors: Professor Kevin Morgan, Dr Keeley Brookes, Dr Sally Chappell.

MRes in Human Molecular Genetics

March 2017

21,526 words

Abstract: 235 words

Number of tables: 11

Number of figures: 12

Contents

1 Abstract 4
2 Introduction 4
2.1 Types of RNA editing 5
2.2 RNA editing in the human genome 6
2.3 ADAR Enzymes 7
2.4 RNA editing in Humans, Gene Expression, and Disease 10
2.5 Identification of Potential Editing Sites using Sequencing technology 12
2.6 Aims and Objectives 14
3 Materials and Methods 14
3.1 Bioinformatics pipeline 15
3.2 Selection of Potential Editing Sites for experimental validation 16
3.3 Designing primers 16
3.4 Cryoprepping of brain samples 17
3.5 RNA Extraction 17
3.6 Checking for RNA Quality 18
3.7 PCR to check for gDNA contamination 19
3.8 cDNA synthesis 19
3.9 Checking for successful cDNA synthesis 20
3.10 Polymerase Chain Reaction amplification of PES 20
3.11 Sanger Sequencing 21
4 Results 22
4.1 Bioinformatics Pipeline 22
4.1.1 Quality control graphs 22
4.1.2 Number of paired reads vs. number of mapped reads for all samples 26
4.1.3 Comparison of mapping tool efficiency 26
4.2 Identification of Potential Editing Sites 27
4.2.1 Filtering for AG polymorphisms 28

4.2.2 Filtering for novel PES using the “Extracted PES” file 33
4.3 Validating PES in the laboratory 38
4.3.1 RNA Quantity and Quality 38
4.3.2 Sanger Sequencing 38
4.3.3 Summary of Validation Results 43
5 Discussion 44
5.1 Identification of PES 44
5.2 Evaluation of PES called from pipeline 47
5.3 MEG3 non-coding RNA Editing site 49
5.4 Evaluation of Bioinformatics Pipeline 50
5.5 Ranking of PES for validation 52
5.6 Future work 53
6 Acknowledgments 56
7 References 56

1 Abstract

RNA A-to-I editing; carried out by the family of adenosine deaminase acting on RNA (ADAR) proteins, is the most common type of RNA editing in vertebrates. It can lead to the disruption of protein synthesis or structure, as well as changes in both the coding and non-coding sequence of the RNA. This can lead to the development of many diseases, and when present in the brain, contribute to the onset of conditions such as depression, epilepsy, schizophrenia, and amyotrophic lateral sclerosis (ALS) (Kwak and Kawahara, 2005, Maas *et al*., 2006). The aim of this project was to develop a bioinformatics pipeline to identify potential A-to-I RNA editing sites from RNA sequencing data. A number of identified sites were then verified using PCR and Sanger Sequencing to test the stringency of the pipeline. A site identified as likely to be a false positive due to sequencing error was confirmed as such. Two other sites that were likely to be false positive were also confirmed to be false positive and one true editing site was verified. This project shows that RNA-Sequencing (RNA-Seq) data provides a good basis for the prediction of potential editing sites from the transcriptome with a good bioinformatics pipeline. However, it seems using reference genomic sequence data lacks the specificity of using exome or whole genome data generated from the same samples as the RNA sequence, therefore leading to a number of false positives.

2 Introduction

DNA (deoxyribonucleic acid) is the fundamental genetic material of all organisms, and allows the expression of genes and consequent production of proteins necessary for life. To allow this production of proteins, DNA must first be transcribed into RNA (ribonucleic acid) as an intermediate step. However, modifications in both DNA and RNA can be introduced during or after transcription, changing the genetic sequence and how subsequent genes and proteins are expressed. The production of abnormal proteins or a change in protein expression can lead to the development of disorders. One such modification that can bring about disease is RNA editing.

RNA editing is classified as the process by which the nucleotide sequence of an RNA transcript is altered relative to that of the encoding DNA (Gray, 2012, Huang *et al*., 2013). This can take place through the insertion, deletion

or modification of nucleotides but does not include events such as splicing or polyadenylation (Farajollahi and Maas, 2010). These changes have been observed in the cell nucleus, cytosol, mitochondria and the plasmids (Tang et al., 2012) and occur commonly in transfer RNA (tRNA), ribosomal RNA (rRNA), and messenger RNA (mRNA) molecules of eukaryotes (Nishikura, 2010). The first incidence of RNA editing was reported in the mitochondrial mRNA of two species of trypanosomes, *Trypanosoma brucei* and *Crithidia fasciculate* in 1986 by Benne *et al*., (Benne et al., 1986, Tang et al., 2012). In this case, four extra nucleotides were found to be inserted in the cytochrome oxidase subunit II (*coxII*) gene at amino acid 170 when comparing the mRNA to the DNA sequence (Benne *et al*., 1986). Despite this editing, the gene remains highly conserved between different species of trypanosomes and appears to be functional (Benne *et al*., 1986).

2.1 Types of RNA editing

There are currently three distinct types of post-transcriptional RNA editing (Moreira *et al*., 2016). The first main category of RNA editing involves the insertion of new nucleotides or deletion of existing ones to create changes in the RNA sequence. The addition/deletion of uridine residues in pre-mRNA transcripts commonly takes place in kinetoplastid mitochondria and is mediated by short guide RNAs (gRNAs) (Farajollahi and Maas, 2010). Another example of this type of RNA editing occurs in the slime mould Physarum polycephalum through the insertion of cytosine (Gott and Emeson, 2000). This type of editing does not take place in humans. The second type of RNA editing is the substitution of cytosine residues to uracil, uracil to cytosine and adenosine to inosine through deamination and transamination (Moreira *et al*., 2016). A-to-I editing was first discovered in yeast transfer RNAs by Holley *et al*., in 1965 when digesting Alanine-RNA (Holley et al., 1965, Gott and Emeson, 2000) and is the most common type of editing in vertebrates. C-to-U editing is most common in plant mitochondria, however, it also occurs in human apolipoprotein B gene (APOB100), which plays a major role in the clearance of low density lipoproteins (LDLs) (Farajollahi and Maas, 2010). The C-to-U mutation generates an in-frame stop codon which prematurely truncates the protein translated (ApoB48) (Farajollahi and Maas, 2010). This shortened protein has a different functionality causing disorders characterised by the inability to absorb and

transport fats, a deficiency in apolipoprotein B, and high cholesterol ('Database Resources of the National Center for Biotechnology Information,' 2017). The third type of RNA editing is also a nucleotide substitution but this occurs only in the 5′ and 3′ ends of the acceptor stem of mitochondrial transfer RNAs (tRNAs) (Moreira *et al*., 2016). It encompasses the removal of mispaired nucleotides from one side of the helix and the insertion of those complimentary to the other side of the helix (Moreira *et al*., 2016).

RNA editing in humans primarily takes place within *Alu* regions; primate-specific repetitive elements that constitute over 10% of the genome (Lander *et al*., 2001, Chen *et al*., 2008). There are up to 1.4 million copies of the *Alu* transposable element within the genome, occurring in gene-rich, non-coding regions such as introns and untranslated regions (UTRs) (Kim et al., 2004, Tang et al., 2012). Due to their abundance, many *Alu* elements form intramolecular long duplexes with other closely located inverted *Alu* sequences, creating a double-stranded RNA substrate key for A-to-I RNA editing (Chen *et al*., 2008). A-to-I editing within introns and UTRs can lead to the creation of new splice sites, which in turn permits the inclusion of intronic *Alu* sequences in the mature mRNAs (*Alu* exonisation) (Nishikura, 2016). This generally can cause decreased expression of the appropriate transcript. Contrary to this, the development of cryptic *Alu* splice sites to create a more functional transcript, can disrupt gene expression sufficiently to lead to conditions such as Alport syndrome or chorioretinal degeneration (Deininger, 2011). A large proportion of *Alu* insertions reside within genes that are involved in neuronal function and have hence been implicated in the development of the nervous system (Chen et al., 2008, Tang et al., 2012).

2.2 RNA editing in the human genome

The conversion of adenosine to inosine (A-to-I) in RNA, post-transcriptionally is a widespread editing mechanism and is highly conserved in the human genome (Slotkin and Nishikura, 2013). Of the various types of RNA editing, A-to-I base modification is the most common in higher eukaryotes (Bass, 2002, Nishikura, 2010, Savva et al., 2012, Li and Church, 2013, Khermesh et al., 2016). To date, it has been estimated that there are over one hundred million editing sites in the human genome (Bazak *et al*., 2014a).

It is a prerequisite that for A-to-I editing to take place the target RNA must be partially or completely double-stranded (Tang et al., 2012). A-to-I editing takes place in three different types of sequence; protein-coding exons in pre-mRNA, repetitive sequence elements in untranslated exons and introns,

and microRNA (miRNA) precursor transcripts (Farajollahi and Maas, 2010). Due to the differing secondary structures of these sequences the RNA editing machinery interacts and modifies them in alternate ways; however, in all cases no cofactors are required (Tang et al., 2012).

RNA A-to-I editing can be specific or widespread in its action depending on the length of the run of nucleotides available within the double-stranded section of RNA. Within longer sections, for example in repetitive elements such as *Alu* elements, A-to-I editing is promiscuous and widespread, with up to 60% of adenosines edited (Carmi *et al*., 2011, Bazak *et al*., 2014b, Nigita *et al*., 2015). In smaller and/or unstable regions of dsRNA, A-to-I editing is more specific, with at least 10% of adenosines undergoing deamination (Wahlstedt and Ohman, 2011, Nigita *et al*., 2015). This suggests that repetitive sequences are preferential for A-to-I RNA editing to take place and that this primarily occurs within non-coding regions of genes (Nishikura, 2010).

The large amount of genetic variation created by A-to-I RNA editing exists due to the expression of both the edited and unedited forms of the encoded protein in the same cell at any one time (Athanasiadis et al., 2004). The ratio of expression can be varied in a cell type-specific, tissue-specific, or time-dependent manner. The consequence of this can vary from little impact, to lethal phenotypes depending on the extent of RNA editing.

2.3 ADAR Enzymes

RNA A-to-I editing is catalysed by ADAR (adenosine deaminase acting on RNA) enzymes (Slotkin and Nishikura, 2013), which usually reside in the nucleus and are responsible for the implementation of many biological pathways (Bass, 2002). ADARs were first discovered in *Xenopus laevis* whilst investigating the effect on antisense RNA on inhibiting mRNA translation (Rebagliati and Melton, 1987). In humans, three members of the ADAR enzyme family exist, ADAR1 (*ADAR*), ADAR2 (*ADARB1*) and ADAR3 (*ADARB2*). ADAR1 and ADAR2 expression occur widely across all mammalian tissues, whilst ADAR3 expression is restricted to the brain (Athanasiadis et al., 2004). Despite the extensive expression of ADARs, almost all ADAR targets identified to date reside in the brain (Bass, 2002, Maas et al., 2003, Athanasiadis et al., 2004).

All ADAR proteins have a common structure which consists of two or three N-terminus double-stranded RNA binding domains (dsRBDs) and a C-terminal deaminase domain (Mannion *et al*., 2015) (Figure 1). The

aforementioned dsRBD, which was originally identified in *Drosophila Staufen* and *Xenopus Xlrbpa* proteins, consists of a highly conserved 65-68 amino acid sequence which forms a secondary structure of three beta pleated sheets with an alpha helix either side; α-helix 1, β-strand 1, β-strand 2, β strand 3, and α-helix 2 (αββββα) (Mannion *et al*., 2015, Bycroft *et al*., 1995, Kharrat *et al*., 1995, Nanduri *et al*., 1998).

ADAR1 is present in two different length isoforms; ADAR1p150 and ADAR1p110, created by the alternate splicing of exon 1A (Patterson and Samuel, 1995, Mannion *et al*., 2015) (Figure 1). Compared to other members of the enzyme family, ADAR1 also comprises of an extended N terminus domain containing Z-DNA binding motifs, Zα; exclusive to ADAR1p150 and Zβ; which resides in both isoforms (Mannion *et al*., 2015). The functions of these domains are currently unknown, although they are thought to be involved in the localisation of the isoforms to the cytoplasm (ADAR1p150) or the nucleus (ADAR1p110) (Herbert and Rich, 2001, Mannion *et al*., 2015).

The A-to-I editing of dsRNA secondary structure substrates is brought about by ADAR1 and ADAR2, whilst ADAR3 lacks a catalytic domain and its primary role is still undetermined (Chen *et al*., 2000, Khermesh *et al*., 2016). However, RNA editing assays have proven that despite having no catalytic activity ADAR3 still acts as a dsRNA binding protein via two dsRNA-binding motifs (DRBM) and also binds to single-stranded RNA (ssRNA) via an arginine and lysine-rich domain (R-domain) located at the N-terminus region (Chen *et al*., 2000). For the catalytic A-to-I editing activity of ADAR1 and ADAR2 to be triggered, homodimerisation is required (Nishikura, 2010, Cho *et al*., 2003). Interestingly ADAR3 does not show homodimerisation in vitro which could provide an answer to its inactivity (Nishikura, 2010).

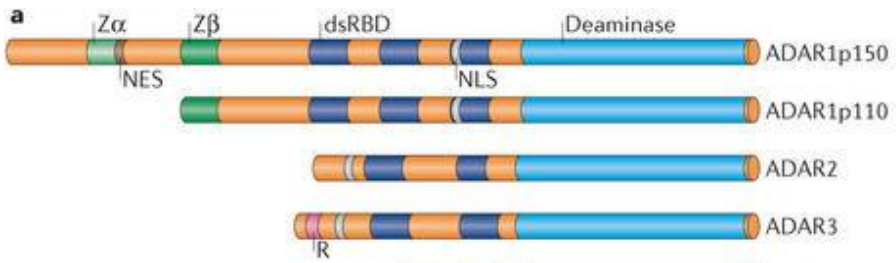


Figure 1: The structure of the human ADAR family, including both isoforms of ADAR1 (ADAR), ADAR2 (ADARB1) and ADAR3 (ADARB2). The shared domains between family members are visible, including the deaminase domain and the double-stranded RNA binding domains. The individual features such as the Z-DNA binding domains and single-stranded RNA R binding domain are also labelled. (NES, nuclear export signal; NLS, nuclear localization signal). Adapted from (Nishikura, 2016).

The extent to which ADARs edit the adenosine residues in a region depends not on the specific sequence of nucleotides but on the secondary structure surrounding the potential editing site (Bass, 1997). Although *in vitro* studies indicate that ADAR1 has a preference to deaminate adenosines with a 5' neighbour of A, U, or C, over G in both selective and non-selective editing, this is not sufficient to account for the highly specific selection of substrates (Lehmann and Bass, 1999). Completely double-stranded RNA substrates have been shown to be edited non-selectively whilst RNAs containing mismatches, bulges and loops within their structure are deaminated more selectively (Lehmann and Bass, 1999). It follows that short duplexes are deaminated more selectively than long duplexes (Nishikura et al., 1991, Polson and Bass, 1994).

The disruption or inactivation of *ADAR* genes results in dramatic phenotypes such as seizure episodes, extreme incoordination, and neurodegeneration in a variety of species (Savva et al., 2012). In *Drosophila*, a homozygous deletion in the *dADAR* gene causes effects such as temperature-sensitive paralysis, irregular movement, and age-dependent neurodegeneration thought to be a result of an absence of A-to-I editing in important ion channels (Palladino *et al*., 2000, Nishikura, 2010). Homozygous ADAR2 null mice display repeated episodes of epileptic seizures due to the surplus influx of calcium ions brought about by a lack of editing of the GluR-B glutamine/arginine site, resulting in premature death (Higuchi *et al*., 2000, Nishikura, 2010). The inactivation of ADAR1 creates an embryonic lethal phenotype due to widespread apoptosis, suggesting that ADAR1 function is imperative for life in mammals (Hartner *et al*., 2004, Wang *et al*., 2000, Wang *et al*., 2004, Nishikura, 2010). By contrast, homozygous ADAR3 null mice appear to be unaffected (Nishikura, 2010).

The conversion of adenosine to inosine occurs through hydrolytic deamination at C6 of the adenine base resulting in the substitution of an amine group for a carbonyl group (Figure 2). By studying the structure of ADAR2 using x-ray crystallography, the deaminase activity of the C-terminal domain has been revealed (Macbeth *et al*., 2005). The globular structure of the domain consists of two α-helices (α2 and α5) and four β-strands (β1, β2, β5, and β8), and also contains a zinc ion and one molecule of inositol hexakisphosphate (IP_6) at its core (Mannion *et al*., 2015). It is the zinc ion which reacts with a water molecule to form zinc hydroxide, which in turn donates a proton to a glutamate residue in ADAR2. This creates a highly-

reactive negatively-charged species which interacts with the C6 of the adenosine residue present in the active site, with ammonia being removed as a by-product (Mannion *et al*., 2015). This disrupts the normal A:U pairing and instead causes binding to cytosine (C). The resulting inosine nucleotide of A-to-I editing exhibits properties most similar to guanosine and is read by the transcriptional machinery as so.

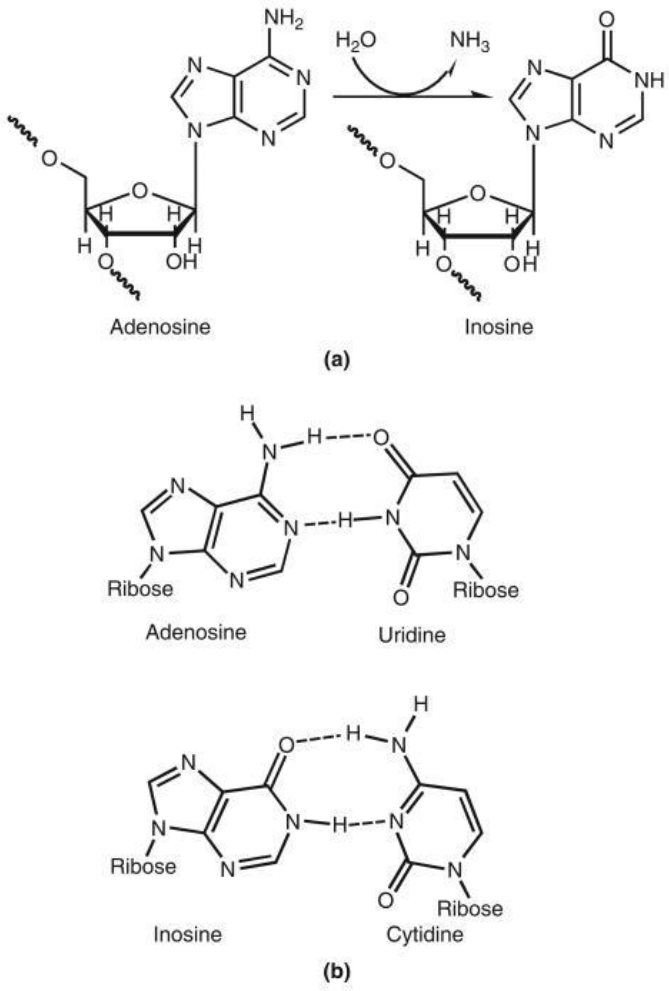


Figure 2: The conversion of Adenosine to Inosine. (a) The conversion of adenosine to inosine through hydrolytic deamination. (b) The A:U pairing of adenosine in comparison to the I:C pairing created by the newly converted inosine base. Taken from (Zinshteyn and Nishikura, 2009).

2.4 RNA editing in Humans, Gene Expression, and Disease

Like many sources of genetic variation, A-to-I editing can alter the amino acid sequence, binding specificity, secondary structure and the creation and elimination of splice sites (Slotkin and Nishikura, 2013). Despite the majority of A-to-I editing occurring in non-coding regions, such as the repetitive elements within intronic or intergenic RNAs and in the 5'UTR and 3'UTR of transcripts (Scadden, 2007), studies tend to focus on those events that take place in coding regions due to the possible functional disruption to the resulting proteins (Sommer *et al*., 1991, Niswender *et al*., 1999, Berg *et*

al., 2001, Hoopengardner *et al*., 2003, Iwamoto and Kato, 2003, Bhalla *et al*., 2004, Cenci *et al*., 2008, Schellekens *et al*., 2012, Khermesh *et al*., 2016). The production of faulty proteins or the deregulation of protein production, in turn, can lead to the increased risk of developing a disease. Specifically, in transcripts corresponding to the brain, the changes in amino acid sequence or expression levels can lead to the manifestation of neurological and neurodegenerative disorders such as depression, epilepsy, schizophrenia, amyotrophic lateral sclerosis, and Alzheimer's disease (Feldmeyer *et al*., 1999, Kortenbruck *et al*., 2001, Vollmar *et al*., 2004, Iwamoto *et al*., 2005, Iwamoto *et al*., 2009, Khermesh *et al*., 2016).

The 5-hydroxytryptamine serotonin receptor 5-HT_{2C}R is a G protein-coupled receptor that undergoes A-to-I editing. Five different residues (A-E) within an intracellular loop are targeted that change the amino acids isoleucine, asparagine, and isoleucine at positions 156, 158, and 160, respectively to several possible amino acids (valine, methionine, aspartic acid, serine and glycine) depending on the number of sites edited (Maas *et al*., 2006). This action, along with two other mechanisms not discussed here, reduces the affinity for G protein-coupling, can change the splice site selection and cell surface expression of the edited HT_{2C}R isoform. The increase in A-to-I editing at locations A and E; two adenosine residues within Ile at position 156 and Asn at position 158 respectively, have been linked to major depression and suicide victims (Niswender *et al*., 2001, Gurevich *et al*., 2002, Iwamoto and Kato, 2003, Maas *et al*., 2006).

Due to the high frequency of A-to-I RNA editing sites in the brain, it follows that such editing is also key in regulating the properties of neurotransmitters and has a heavy involvement in the central nervous system (Athanasiadis et al., 2004, Seeburg et al., 1998, Schmauss and Howe, 2002). Abnormal occurrences or an absence of A-to-I editing in the brain can lead to dangerous and often lethal phenotypes, such as the pathogenesis of amyotrophic lateral sclerosis (ALS) (Casey et al., 2006, Lemmens et al., 2010, Tang et al., 2012). The action of ADAR2 on the pre-mRNA of GluR2 changes the codon at position 586 from glutamine (Q) to arginine (R), leading to the production of a differing protein (Tang et al., 2012, Heale et al., 2009). The incorporation of this protein into the AMPA-type receptor reduces the calcium permeability of the ion channel due to the presence of the positively charged arginine residue (Tang et al., 2012). The absence of RNA A-to-I editing at this position allows the glutamine residue at position 586 in the pre-mRNA to be

translated, leading to an increased permeability and an influx of calcium ions which leaves the motor neuron vulnerable to overload and consequent death (Heale et al., 2009, Tang et al., 2012).

In a survey of RNA editing in the human brain by Blow *et al*., A-to-I editing was found as frequently as 1:2000 nucleotides in >3 Mb investigated (Blow et al., 2004). The vast majority of edits took place in the intergenic and intronic regions of genes, with none occurring in the translated exons and very few in the untranslated exons (Blow et al., 2004). Khermesh *et al.,* found that overall RNA A-to-I editing levels of a selected target set of editing sites, in the brain tissues of Alzheimer's disease patients, were decreased when compared to age and sex-matched non-demented controls (Khermesh *et al*., 2016). This decrease in editing was mainly observed in the hippocampus but was also seen in the temporal and frontal lobes to a smaller degree. Changes in editing rates are associated with diseases such as atherosclerosis, myopathy, epilepsy, major depression disorder, schizophrenia and other mental disorders as well as in cancer and brain tumours (Meier *et al*., 2016). This implies that abhorrent RNA A-to-I editing/disruption of editing levels can lead to disease ontology, making RNA editing events an interesting target for further investigation.

2.5 Identification of Potential Editing Sites using Sequencing technology

Using next-generation sequencing data it is possible to look at the transcriptome and genome at single nucleotide resolution, the comparison of which will allow the identification of any potential editing sites. This is due to observed differences in the RNA-Seq data and the exome sequencing data. As RNA-Seq data and exome sequencing data are available for all three samples, across five different brain regions, comparisons can be made both within the samples for brain region specificity and between samples for an inter-person difference in RNA editing.

RNA-Seq is advantageous over microarrays as it makes use of next-generation sequencing, and allows the detection of transcripts in detail across the entire genome. It is superior in detecting novel transcripts, allele-specific expression and splice junctions in comparison to microarrays (Zhao *et al*., 2014). Despite the advent of many bioinformatics methods to interpret the data produced from RNA-Seq it still poses problems with alignment of reads and quantification of genes (Zhao *et al*., 2014).

During the development of sequencing technologies, identification of RNA A-to-I editing sites was slow and infrequent. The discovery of A-to-I editing sites has been dependent on the sequencing technologies available at the time (Ramaswami and Li, 2014). Alongside the progression of sequencing technologies, the occurrence and expansion of nucleotide databases have dramatically aided the identification of additional editing sites (Ramaswami and Li, 2014). Sakurai *et al.* (Sakurai *et al*., 2010), also developed a biochemical assay to identify inosine in RNA using inosine chemical erasing (ICE). This method uses cyanoethylation combined with reverse transcription (Sakurai *et al*., 2010). However, the majority of RNA A-to-I editing sites have been identified using high-throughput RNA-Seq technologies. Computational pipelines were created to compare genomic DNA and cDNA sequences for one individual, whilst reducing false positive call rates caused either through sequencing or read mapping errors (Bahn *et al*., 2012). Following this, Li *et al.*, developed a method using targeted capture with padlock probes in combination with high-throughput sequencing (Li *et al*., 2009).

Computational frameworks have since been developed to identify RNA A-to-I editing sites by comparing RNA-Seq data with matched genomic DNA sequencing from one individual (Bahn *et al*., 2012, Peng *et al*., 2012, Ramaswami *et al*., 2012, Ramaswami *et al*., 2013). The pipelines used have been executed from the command line using freely available tools from software development platforms. For example, Bahn *et al.*, uses a combination of Bowtie, BLAT and TopHat tools to map RNA-Seq reads, whereas Peng *et al.,* uses SOAP2 (Bahn *et al*., 2012, Peng *et al*., 2012). Ramaswami makes use of Picard and GATK tools such as MarkDuplicates, IndelRealigner and Unified Genotyper to carry out later stages of the pipeline (Ramaswami *et al*., 2012).

Currently, information on previously discovered RNA A-to-I editing sites is collectively stored in databases such as DARNED and RADAR. DARNED (a DAtabase of RNa EDiting in humans) is a centralised database of published RNA editing sites in humans (http://darned.ucc.ie). The database allows the user to search particular coordinates or a gene for potential editing sites and returns information such as strandedness, type of editing and published dataset for each site (Kiran and Baranov, 2010). RADAR (a rigorously annotated database of A-to-I RNA editing) is a database compiling a myriad of A-to-I editing sites in humans (Homo sapiens), mice (Mus musculus) and flies (Drosophila melanogaster) alongside annotations for each editing site using

the hg19 reference genome (available at http://RNAedit.com). In contrast to DARNED, it also contains a listing of tissue-specific editing levels for each editing site (Ramaswami and Li, 2014). Both databases allow records to be viewed using the UCSC browser, with DARNED also allowing visualisation in Ensembl.

2.6 Aims and Objectives

The aim of this research project is to implement a bioinformatics pipeline to identify potential A-to-I editing sites in three samples with both exome sequencing data and RNA sequence data for five different regions of the brain available (Frontal, Temporal, Occipital, Cerebellum, and Parietal). This will enable the characterisation of A-to-I RNA editing in the human brain across different regions. Verification of a number of these potential editing sites will be carried out using Sanger sequencing from these samples.

The main biological question asked of my data is how many novel adenosine-to-inosine editing sites (A-to-I) are found in coding regions of RNA for each brain region tested? Further questions to be asked include; how many editing sites are there in the transcriptome in the brain and how many of these editing sites reside in each region of the gene? Does the number of editing sites vary per chromosome and are there any that show a particularly large amount of RNA editing? Is this editing widespread across the chromosome or clustered together? The overall success of the pipeline and experimental methods used will be evaluated with a conclusion on the effectiveness of RNA-Seq as a tool for the prediction of the location of A-to-I editing sites.

3 Materials and Methods

Brain tissue was obtained from the Brains for Dementia Research (BDR) repository. Tissue from five regions of the brain were provided for each of the three control samples (Frontal, Temporal, Occipital, Cerebellum and Parietal). RNA was extracted from these regions for both RNA sequencing and for subsequent verification experiments detailed in this project. All samples were phenotypically and clinically normal healthy controls, two of the samples were female, with one male sample, the average age of death was 71 years (±14 years), average PMI = 50.3 hours.

3.1 Bioinformatics pipeline

RNA-Seq data was trimmed, removing reads with poor quality scores and with adapter sequences using Sythe (https://github.com/vsbuffalo/scythe) and Sickle (https://github.com/najoshi/sickle). Reads were mapped to the human reference genome (UCSC hg19) using TopHat and annotated with Ensembl VEP. Further filtering according to MAPQ was then applied.

Using the protocol by Altmann (Altmann *et al*., 2012), a specific variant-calling pipeline was generated utilising the Galaxy cloud environment to identify the presence of RNA A-to-I editing sites (Goecks *et al*., 2010). Analysis files were imported, including variant call format files from the 1000 Genomes project (1000G_phase1.indels.hg19.sites.vcf), dbSNP (dbsnp_138.hg19.vcf), HapMap (hapmap_3.3.hg19.sites.vcf) and the Mills set (Mills_and_1000G_gold_standard.indels.hg19.sites.vcf), obtained through the Broad Institute FTP server (ftp.broadinstitute.org/bundle). Alongside these VCF files, the hg19 genome assembly in FASTA format (UCSC.hg19.fasta), and the corresponding FASTA index (UCSC.hg19.fasta.fai) and dictionary file (UCSC.hg19.dict) were uploaded (ftp.broadinstitute.org/bundle). RNA-Seq data files were imported in FASTA format.

To verify DeepSeq QC, the "FASTQC" tool (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used on RNA-Seq data files. As all reads surpassed a Phred score of 20, no trimming was required. "FASTQ Groomer" (Blankenberg *et al*., 2010) was used to convert file format from FASTQ to FASTQSANGER as a prerequisite to mapping. "Map with BWA for Illumina" (Li and Durbin, 2009) was used to align the reads to the hg19 reference genome using the Burrows Wheeler Transform (Burrows and Wheeler, 1994).

Reads were sorted by chromosomal coordinate using SamTools (http://samtools.sourceforge.net/) and alignment summary metrics gathered using Picard (http://broadinstitute.github.io/picard/). Of the reads that passed Illumina's chastity filter, 12,700,297 of the 13,599,735 were aligned to the reference genome, totalling 93.39% of all reads. PCR duplicates were removed using "RmDup" (http://samtools.sourceforge.net/), and reads filtered for those that were definitely mapped and in a proper pair (https://github.com/pezmaster31/bamtools) (Barnett *et al*., 2011). Read groups were allocated and data reordered to match the contig ordering in the reference file (http://broadinstitute.github.io/picard/). Realignment around indels and recalibration of quality scores was carried out using the Mills set,

HapMap and dbSNP VCF files and the Genome Analysis Toolkit (DePristo *et al*., 2011). SNP calling was carried out using dbSNP138 only and the Unified Genotyper tool (GATK).

Filtering was carried out for mapping quality (MQ>30), approximate read depth (DP>10), quality per depth (QD>5.0) and the length of the homopolymer run surrounding the editing site (HRun<6) (https://github.com/vcflib/vcflib). The data was then annotated using ANNOVAR (Wang *et al*., 2010), creating a tabular output of 4,576 SNPS. Editing sites were then filtered for A to G variants only, those not matching intronic or intergenic, and those without rs nomenclature.

3.2 Selection of Potential Editing Sites for experimental validation

Potential Editing Sites (PES) were chosen for verification based on the region they resided in and the type of editing pattern shown based on the pipeline results, then ranked according to the number of supporting reads in the raw bam files. Each PES was also checked against RNA Editing databases RADAR/DARNED to identify if it had already been discovered (Ramaswami and Li, 2014, Kiran and Baranov, 2010). A number of PES were selected from the data in order to verify the effectiveness of the bioinformatics pipeline according to the aforementioned criteria; region resided in (exonic or 3'UTR), sequencing pattern shown (sequencing error, sporadic editing or sample specific editing) and then ranked according to number of supporting reads (See Tables 9A-9C). Several false positive sites were selected to confirm, as a refinement for the bioinformatics pipeline, and one likely true positive PES was selected to confirm the presence of RNA editing detection.

3.3 Designing primers

For each PES, primers were designed using the hg19 reference sequence genome from the Broad Institute server in IGV. Boundaries between exons and introns were identified manually using IGV. Approximately 700bp of transcribed/genomic sequence encompassing the PES was obtained for primer design, via PrimerBLAST using default settings (https://www.ncbi.nlm.nih.gov/tools/primer-blast/). Primers were designed to sequence the genomic DNA (gDNA) content of the site to compare to the cDNA generated from RNA for each brain region.

Although it would be preferable to design gDNA primers to bind in non-transcribed regions of the sequence, e.g. introns, many of the PES were located in the 3'UTR which are transcribed.

Increasing the product size to incorporate non-transcribed genomic regions to allow for substrate specificity would lead to very large amplicons that would be inefficient to PCR amplify and sequence. Therefore, in these cases the same primer pairs were used for both gDNA and cDNA. To ensure no cross-contamination between substrates, a DNA nuclease step was included into the RNA extraction process, and then the extracted RNA was checked for gDNA contamination. Table 1 shows the primers pairs designed for follow-up in the lab.

PES	**Location**	**Primer Sequence**	**Annealing Temperature/°C**	**Product Size/bp**
PES_*KIF1B*_3'UTR (Chr1:*10440678)*	*KIF1B*, 3'UTR	Fwd: TTCTGCATTTCCCTCGTGCT Rev: GAATCACTGGCCCTGGTGAA	Fwd: 57.3 Rev: 59.4	528
PES_*HMP19*_3'UTR (Chr5:*173535468)*	*HMP19,* 3'UTR	Fwd: GCCACAAGCGATAAGCACAG Rev: GCGAATGCCTGATACGGTTC	Fwd: 59.4 Rev: 59.4	328
PES_*NEBL*_3'UTR (Chr10:*21072164)*	*NEBL,* 3'UTR	Fwd: TCCTTGTGCTTCTGTCAGGC Rev: TGCATGCTCTTGAGGCCATA	Fwd: 59.4 Rev: 57.3	348
PES_*MEG3*_ncRNA_exonic (Chr14:*101312647)*	*MEG3,* ncRNA_exonic	Fwd: GGCTTCAAGCTCCAATGACC Rev: GGGCTCGCGTCCTAAACATA	Fwd: 59.4 Rev: 59.4	429

Table 1: Primers designed to target the chosen PES for validation in the laboratory.

3.4 Cryoprepping of brain samples

The brain samples to be used for RNA extraction were CryoPrepped prior to use according to the "Tissue dry Pulverisation using the CryoPrep™ CP02 (Covaris)" protocol, to increase surface area for tissue lysis. Frozen tissue was placed into Tissue Tubes and briefly placed in liquid nitrogen before pulverising to a powder using the CryoPrep™ CP02 (Covaris). Pulverised tissue was stored in 2ml millitubes at -80°C until further use.

3.5 RNA Extraction

RNA extraction followed an in-house method developed from the QIAGEN RNase easy mini kit. For each sample 100mg of cryoprepped tissue was placed in an Eppendorf tube with 1ml of Trizol (Ambion), and centrifuged at 4°C for 10 minutes at 12,000rcf.

Supernatant was transferred to phase gel tubes without disturbing the pellet. Incubation at room temperature was carried out for 5 minutes, before the addition of 200µl of chloroform and samples were shaken vigorously. The samples were then incubated at room temperature for 3 minutes. Samples were centrifuged for 5 minutes at 12,000rcf at room temperature.

The aqueous layer from the phase gel tubes was removed and transferred into new 1.5ml Eppendorf tubes. One volume of 70% Ethanol was then added to the samples (~500µl). Up to 700µl of solution was then added to the QIAGEN RNA easy spin column and the columns centrifuged for 30s at room temperature at 10,000rcf. The flow through was discarded and another 700µl of sample added to the RNA easy spin column and spun according to the previous step.

Flow-through was discarded and 700µl of RW1 wash buffer (QIAGEN) was added to the columns and spun as previously mentioned. Flow-through was again discarded. 80µl of DNaseI mix (10µl DNaseI, 70µl RDD Buffer (QIAGEN)) was added to the spin column and incubated at room temperature for 15 minutes. Following this, 350µl of RWI buffer (QIAGEN) was added to the columns and centrifuged at room temperature for 30s at 10,000rcf.

500µl of RPE buffer (QIAGEN) was added to the spin column and centrifuged at room temperature for 30s at 10,000rcf. The flow through was discarded. Another 500µl of RPE buffer was added to the spin column, and spun for 2 minutes at room temperature at 10,000rcf. Again, the flow through was discarded. The samples were then allowed to air dry by centrifuging them for 1 minute at room temperature at 10,000rcf. The columns were then placed into new RNase-free 1.5ml tubes and 50µl of RNase-free water (Agilent) added. The columns were then centrifuged at room temperature for 1 minute at 10,000rcf. The flow through was retained and stored at -80°C prior to further use.

3.6 Checking for RNA Quality

RNA was tested for quality using the Agilent 2200 TapeStation and for the presence of gDNA by standard PCR using genomic DNA specific primers.

Samples were run on the Agilent 2200 TapeStation following the manufacturer's instructions. Briefly 1µl of RNA Ladder or RNA sample was added to 5µl of RNA sample Buffer. Samples were vortexed using the IKA MS3 Vortexer for 1 minute at 2000rpm, and heated to 72°C for 3 mins, then

put on ice for 2 mins. Samples were centrifuged prior to running on the TapeStation.

3.7 PCR to check for gDNA contamination

To check the extracted RNA for genomic contamination, a Master Mix was made up as stated below (Table 2) using primers designed to specifically amplify non-transcribed genomic sequence.

Reagent	Volume (µl)
H_2O (distilled and deionised)	13.2
10x PCR Buffer + $MgCl_2$ (Roche)	2.5
dNTPs (2mM)	1.0
Primer Forward	0.4
Primer Reverse	0.4
Taq Polymerase (Roche)	0.5
RNA (NOT IN MASTER MIX)	4.0
Total	22.0

Table 2: Reaction volumes for testing RNA for genomic contamination. RNA added individually to each tube, not included in Master Mix.

The reactions were placed on a thermal cycler and the following conditions run.

Temperature (°C)	Time	Number of cycles
94	2 minutes	1
94	30s	35
50-65	30s	
72	45s	
72	7 minutes	1
10	∞	1

Table 3: PCR conditions used for gDNA contamination test. The primers used had an annealing temperature of 58°C.

3.8 cDNA synthesis

cDNA was synthesised from the extracted RNA, using the Stratagene AffinityScript Multiple Temperature cDNA Synthesis Kit. 1.5µg of RNA was used for each sample reaction, made up to 15.7µl with RNase-free water, and 1.0µl of oligo(dT) primer (0.5µg/µl, Agilent) added. Two RT(-) were set up as negative controls.

The synthesis reaction was incubated at 65°C for 5 minutes then cooled to room temperature for 10 minutes. The following components were

added in order; 2.0µl of 10x Affinity Script RT Buffer, 0.8µl of dNTP Mix (25mM each dNTP), 0.5µl of RNase Block Ribonuclease Inhibitor (40U/µl) and 1.0µl of Affinity Script Multiple Temperature Reverse Transcriptase. For the RT(-), 2.0µl of 10x Affinity Script RT Buffer, 0.8µl of dNTP mix and 1.5µl of RNase-free water was added instead. The following programme was run on the thermal cycler.

Temperature (°C)	Time (mins)
25	10
42	60
70	15
10	∞

Table 4: PCR conditions for first-strand cDNA synthesis reaction.

3.9 Checking for successful cDNA synthesis

After cDNA synthesis, PCR was used to check for the presence of cDNA using GAPDH primers, with the reaction volumes listed below (Table 5).

Reagent	Volume (µl) per reaction
H_2O (distilled and deionised)	22.2
10x PCR Buffer + $MgCl_2$ (Roche)	3.0
dNTPs (2mM)	3.0
Primer Forward	0.3
Primer Reverse	0.3
Taq Polymerase (Roche)	0.2
cDNA (NOT IN MASTER MIX)	1.0
Total	30.0

Table 5: Reaction volumes for PCR amplification of cDNA.

Thermal cycler conditions were as stated previously (Table 3) with the annealing temperature set at 58°C.

3.10 Polymerase Chain Reaction amplification of PES

The reaction volumes were calculated for the experiment based on the number of samples and controls being tested (See Table 5).

A Master Mix of all reagents excluding the cDNA was made up in 1.5ml Eppendorf tube and used to prepare several PCR reactions. The buffer, dNTPs and primers were all vortexed before use. 29.0µl of Master Mix was added to each 0.1ml PCR tube before adding 1.0µl of cDNA to all samples bar

controls. After mixing by pipetting, the samples were placed on the thermal cycler and the following cycles run.

Temperature (°C)	Time	Number of cycles
94	2 minutes	1
94	30s	35
50-65	30s	
72	45s	
72	7 minutes	1
10	∞	1

Table 6: PCR conditions for amplification of DNA. The temperature of the third step is changed relative to the specific oligonucleotide primers being used.

Primer pairs were optimised in stock cDNA and gDNA using a veriflex step to set up a temperature gradient. Each primer pair was tested on all six gDNA samples and fifteen cDNA samples using PCR. Amplification was checked on a 1% agarose gel using 10µl of the PCR product.

3.11 Sanger Sequencing

5µl of PCR product was treated with 2µl ExoSAP-IT and heated to 37°C for 15 minutes, followed by 80°C for 15 minutes. 4µl of cleaned PCR sample was added to 1µl of 5pmol/µl sequencing primer, 3µl of BigDye (v3.1, Applied Biosystems) and 2µl of ABI 5X Sequencing Buffer (Big Dye Terminator v1.1 & v3.1, Applied Biosystems) and mixed. The sequencing reaction mix was then placed in the thermal cycler and the following programme run:

Temperature (°C)	Time	Number of cycles
96	30s	25
50	15s	
60	4 mins	
10	∞	1

Table 7: PCR conditions for Sanger Sequencing.

Using Edge Biosystems Performa DTR Gel Filtration Cartridges, column purification was carried out. The columns were spun at 5000rpm for 3 mins, and then placed in new 1.5ml collection tubes. The sequencing reaction products were then added to the columns and spun at 5000rpm for 3 mins. The eluted products were transferred to new 0.5ml ABI tubes and stood in the

Thermoblock at 90°C until all liquid had evaporated. The remaining pellets were then stored at -20°C until sent for analysis by capillary gel electrophoresis.

4 Results

4.1 Bioinformatics Pipeline

4.1.1 Quality control graphs

The following graphs were produced as part of the output from the MultiQC html module. This module combines multiple FastQC outputs in one report and flags any potential problems before proceeding with analysis. Using the tool “MultiQC”, the overall quality of the RNA-Seq data can be assessed by combining the 30 FastQC outputs (2 files per sample). Although possible, the reads were not merged as they were required to be separate for later analysis steps. The html output produced showed basic statistics for the data examined containing percentage of duplicate reads, average GC content, average sequence length and percentage of FastQC modules failed. The average GC content of the data was calculated to be 51%.

Also included were graphs depicting sequence quality, per sequence quality score, per base n content and adapter content, all of which were passed. Despite sequence length distribution, sequence duplication levels and overrepresented sequences modules raising warnings during quality checking, the graphs were deemed satisfactory and as having no impact on the overall data quality. Although the modules representing per base sequence content, and per sequence GC content failed, upon examination these qualities commonly raise errors with RNA-Seq data, and were not thought to impact the following pipeline. As the majority of checks made by the FastQC module were passed, the RNA-Seq data was classified as good and suitable for further analysis.

Figure 3 displays the results for the QC analysis, Graph A represents per base sequence quality. The Y-axis shows quality score for the calling of bases and it hence follows the higher the value, the better the base call. The red, amber, and green sections of the graph background represent poor, average and good quality base calls respectively. Graph A shows that the values for all positions in the read are above a phred score of 30 and reside within the green portion of the graph, indicating they are all good quality. Each sample is plotted as a line on the graph. It is common to see bases tail off at

the end of read due to accumulation of errors as cycle number increases. No trimming was required before proceeding with the pipeline as all bases have quality scores of above phred score 30, meaning there is a less than 1 in 1000 chance of the base being called incorrectly (99.9% accurate).

Graph B shows per sequence quality scores (phred scores) for all sequences within the data. As the lines are skewed heavily to the right of the graph, all sequences have a high phred score. The majority of sequences have a phred score of above 30. The higher the phred score the greater the accuracy of the calls.

Graph C displays sequence length distribution. The vast majority of sequences were of the same length (76bp), indicated by the sharp peak at the right hand side of the graph.

Graph D shows per sequence GC content. It shows the percentage GC content across all sequences with a line per sample. The normalised distribution is not shown in this graph. The red peaks represent the actual GC content of the sequences. This module failed according to the summary statistics, due to the fact the sum of the deviations from the normal distribution curve represents more than 30% of the reads. Sharp peaks indicate specific contaminant for example the presence of overrepresented sequences. However, the overall GC content according to the basic statistics module is 51% which is acceptable.

The final graph in Figure 3, Graph E, shows the sequence duplication levels for all sequences within the data. As RNA-Seq data contains sequences from different transcripts, they will be present in highly varying levels. It is hence common in RNA Seq libraries to see raised peaks in the right-hand side of the graph due to the over sequencing of highly expressed transcripts in order to observe those lowly expressed. Therefore, it is not deemed crucial that the data be edited to de-duplicate reads.

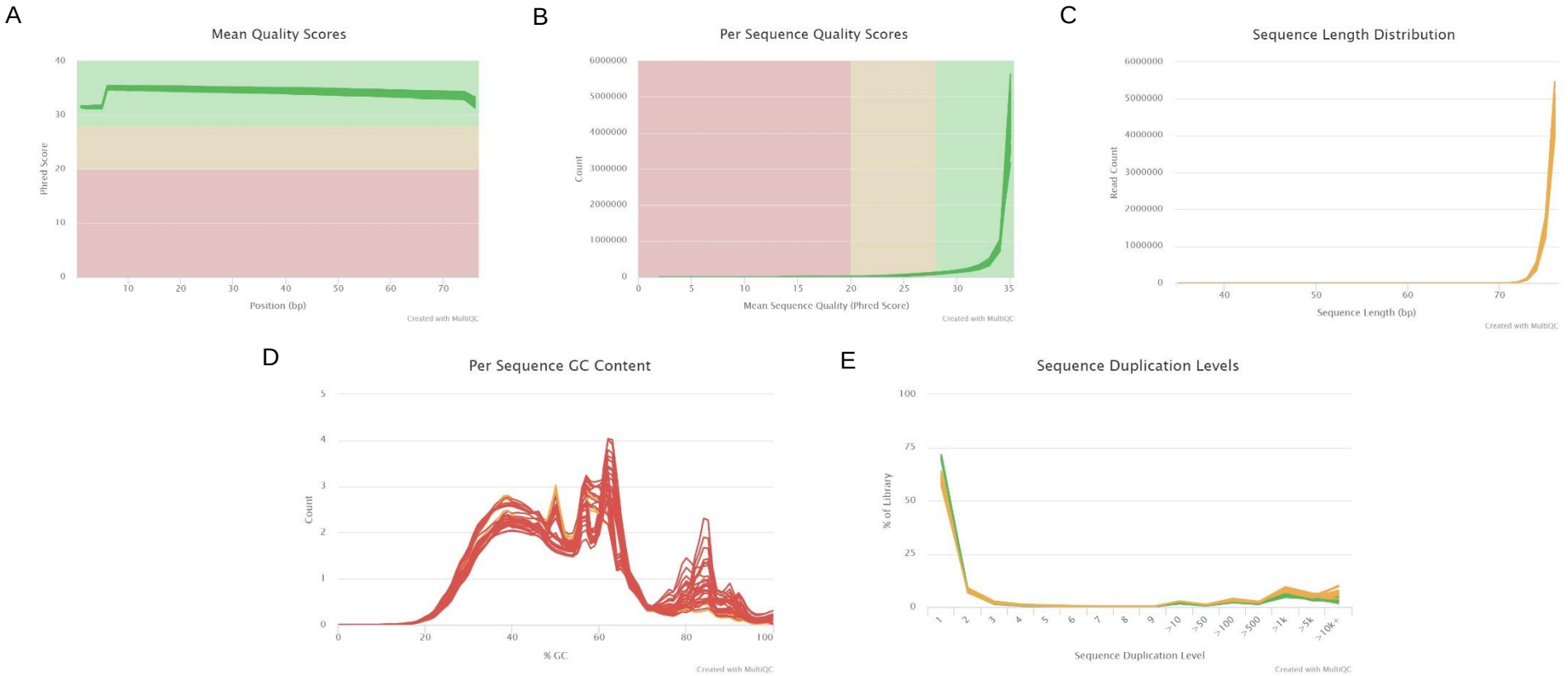


Figure 3: A. Per base sequence quality graph. The Y-axis shows quality score for the calling of bases. Green, amber, and red sections of the graph background represent bases of good, reasonable, and poor quality respectively. B. Per sequence quality score graph showing the number of sequences at each quality score. The higher the phred score the greater the accuracy with which the sequence was called. Again, the background of the graph is coloured according to phred score quality. C. Sequence Length Distribution. This graph shows the distribution of sequences according to length. The vast majority of sequences are of the same length represented by the sharp peak at the right-hand side of the graph. D Per sequence GC content, shown by the red trace. Although this module failed the overall GC content is acceptable. E. Sequence duplication levels, as shown by the amber and green lines. It is common to see this shape graph for RNA-Seq data.

Further to the above graphs there was also a module representing the per base sequence content of the reads, showing the mean percentage of each base at each position. The combined graph for all 30 samples; 2 files (forward and reverse reads) for each of the 15 samples, consisted of a heat map (graph not shown), which could be separated to view data for each individual sample. Theoretically the lines for each base should be parallel as you would expect no bias in a random sample. RNA-Seq also tends to underrepresent GC-rich and GC-poor regions, leading to a more flattened profile of base % than expected. There is variation in the lines and hence the module technically failed as indicated by the tool overview. However, this module is frequently failed when RNA-Seq data is ran through it due to use of random hexamer primers which unintentionally bind with different specificities, creating a biased library. This could only be solved by trimming the first 12 or so nucleotides from the reads; however this may cause mapping problems later on (Hansen *et al*., 2010). Peaks at the end of a read may be due to overrepresented sequences.

Also included in the MultiQC report was a graph representing the per base N content of the RNA-Seq data. The per base N content (graph not shown) showed that no N calls were made for any of the reads in all samples. An N base is called when the sequencing platform cannot determine the base with sufficient confidence.

Also included in the Fast QC module, but not shown in the MultiQC report, were graphs representing the per tile sequence quality, adapter content and Kmer content. The per tile sequence quality graph represents the average quality of sequencing per tile of the flowcell. It shows the deviation of each base call from the average using a cold to hot scale, with the cold colours representing equal to or greater than average and the hotter colours representing worse qualities. This graph (not shown) was solid blue demonstrating a higher than average quality score for all tiles in the flowcell. The graph for adapter content shows the percentage of the library that has seen adapter sequences at each position. A flat line was produced for this module (not shown) meaning no adapter sequences were seen and no trimming was necessary. The overrepresented sequences module raises a warning when a sequence represents over 0.1% of the total data. This module was flagged during quality checking but was deemed trivial because RNA-Seq libraries commonly produce repeated sequences as they are not subjected to random fragmentation. Finally, the

Kmer content graph was produced (graph not shown). This highlights any positions in the reads for which there is an overrepresentation of small fragments/k-mers, again indicating whether the data requires trimming. This module was also failed according to the Fast QC report as at least one k-mer was imbalanced with a binomial p-value < 10^{-5}.; however, this is common for RNA-Seq data and is deemed unimportant as no adapters were detected.

As the quality checks deemed most important to the integrity of the data were passed (Figure 3), and those which raised a warning or technically failed were not regarded as having a negative impact, the RNA-Seq data was considered to be of high quality and can be used to proceed with the downstream pipeline without any adjustments being made.

4.1.2 Number of paired reads vs. number of mapped reads for all samples

The number of paired reads detected from the "Collect Alignment Summary Metrics" file, produced during the bioinformatics pipeline, was plotted against number of paired reads aligned to the UCSC.hg19.fasta reference genome for each sample tested. This was done to evaluate the mapping efficiency of the "Map with BWA for Illumina" tool. The percentage of total reads that are mapped varies by approx. 2.84% with an average of 93.39% of the reads being mapped to the reference genome. This indicates that the mapping step is efficient; however, it could perhaps be improved by the use of a splice aware aligner to allow alignment around introns. "Map with BWA for Illumina" is currently not splice aware.

4.1.3 Comparison of mapping tool efficiency

As part of the bioinformatics pipeline constructed, it was decided to briefly compare the efficiency of three mapping tools available under the "NGS: Mapping" tool-shed, Map with Bowtie, Map with BWA and Map with BWA for Illumina. Map with BWA-MEM was also available; however, this is an alternative for Map with BWA for longer reads and hence was not suitable for the dataset used. The comparison of the aforementioned tools was done to question which is most suitable for use with RNA-Seq data. Map with BWA for Illumina was the most successful mapping tool of the three as it maps the highest amount of reads to the reference genome (93.03%). This was hence the tool chosen for use at the mapping stage of the bioinformatics pipeline.

4.2 Identification of Potential Editing Sites

Figure 4 below shows a flowchart of the main steps in the bioinformatics pipeline). Before commencing with any analysis, the data was quality checked and found to contain 13,599,735 reads. Of these 13,599,735 reads, 12,700,297 were then mapped to the human reference genome hg19 (93.39%). The software RmDup was used to remove the data of PCR duplicates narrowing the number of reads down to 9,337,873. Reads were then filtered dependant on whether they were mapped and in a proper pair, leaving 8,597,698 reads. Realignment around indels, Count covariates and Table Recalibration steps were then carried out with no effect on read number. Using Unified Genotyper to call 68,410 variants from the RNA-Seq data, variants were then filtered on mapping quality and quality depth down to 4,576 individual sites, and then for A-to-G polymorphisms only, leaving 865 variants. Further filtering to discount intronic or intergenic editing sites brought the count down to 477 variants, and removing all those with rs numbers gave a final number of 83 unique A-to-G potential editing sites. All values included in the flowchart are averages gathered across all 15 samples.

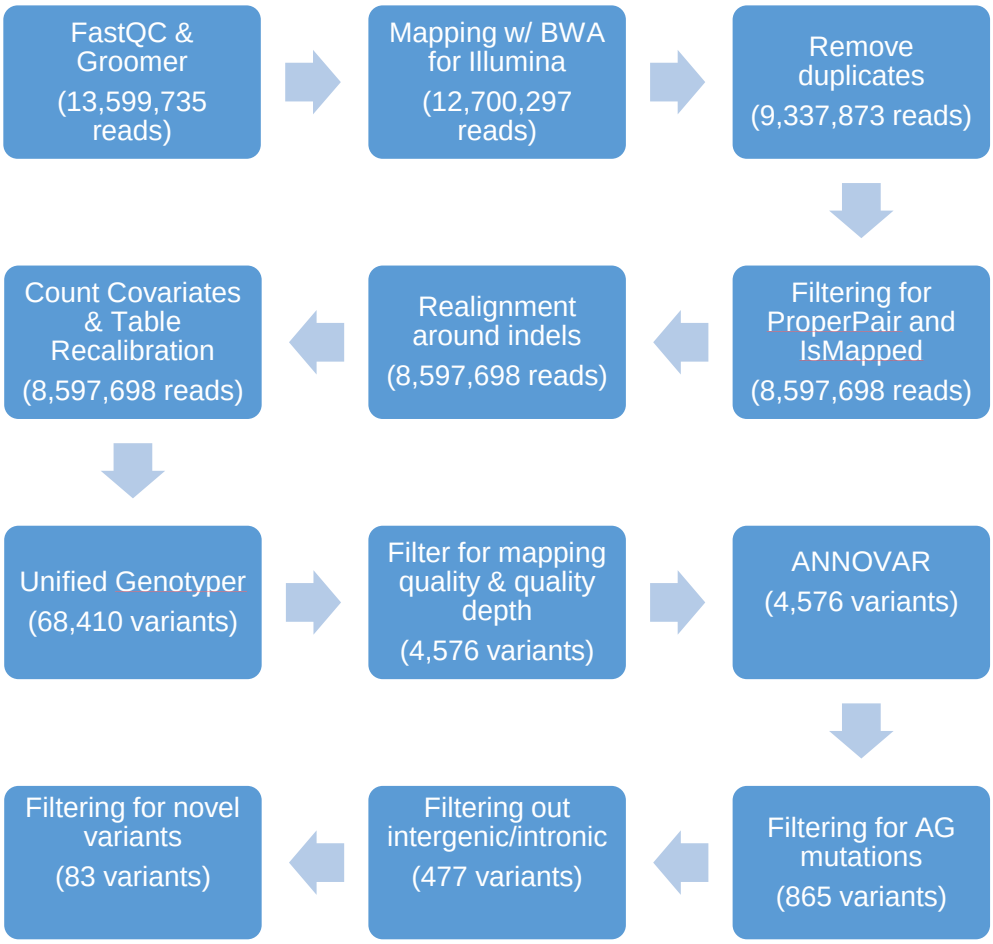


Figure 4: A flowchart representing the main stages of my bioinformatics pipeline in which the number of reads or variants could be called from the data. Values are an average across all 15 samples.

4.2.1 Filtering for AG polymorphisms

Prior to filtering out for intergenic/intronic variants in the bioinformatics pipeline, demographics were collected on the variants called for all samples. These were taken from the "AG mutations" file created as an output from the first of three final filtering steps. At this step in the pipeline, variants with rs numbers would still be included in the output so these are filtered out alongside synonymous and non-synonymous single-nucleotide variants (SNVs).

A total number of 1,486 PES were observed across all the samples, an average of 343 were seen per sample. The PES were located in various places across the genome (Table 8), the majority were located in the intergenic region, with the fewest observed in 3'UTR non-coding RNA. The annotations for gene region were taken from the ANNOVAR tool. A

downstream editing site is classified as residing within the 1kb section downstream of the transcription end site. An intergenic variant is listed as being in the non-coding regions between genes, with intronic variants residing in the introns of genes. PES were also identified in non-coding RNA transcripts both in the exons, introns and 3'UTR. Splicing variants are characterised as located within 2bp of a splicing junction, in the intron. The upstream annotation denotes that a variant is located within 1kb of the transcription start site. 3'UTR and 5'UTR indicate the variant is located in the 3' untranslated or 5' untranslated region respectively.

.

	Number of PES per sample															
	1 Frontal	1 Temporal	1 Occipital	1 Cerebellum	1 Parietal	2 Frontal	2 Temporal	2 Occipital	2 Cerebellum	2 Parietal	3 Frontal	3 Temporal	3 Occipital	3 Cerebellum	3 Parietal	Number of PES across all regions
Downstream	7	11	5	6	12	10	13	7	4	11	8	4	6	2	9	38
Intergenic	238	249	169	108	198	206	189	94	167	147	155	100	97	117	113	593
Intronic	124	135	90	129	101	106	89	67	131	76	83	54	59	149	62	517
ncRNA_Exonic	55	59	40	22	46	48	44	31	35	40	30	30	23	18	23	93
ncRNA_Intronic	19	18	12	7	13	15	15	7	11	10	9	8	5	5	8	29
ncRNA_3'UTR	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	2
Splicing	12	11	7	3	8	6	10	0	1	7	5	1	4	2	1	22
Upstream	1	1	1	0	1	2	0	0	0	1	0	0	0	2	1	5
3'UTR	33	41	28	24	39	24	35	18	37	38	25	15	20	27	24	183
5'UTR	0	2	1	0	0	0	0	0	1	0	1	0	0	1	0	4
Total	489	527	353	299	418	418	395	224	387	330	317	212	214	323	241	

Table 8: This table displays the number of novel A/G polymorphisms and the regions in which they reside, for each RNA-Seq sample. The total number of novel editing sites in each sample is listed in the bottom row. The right-hand most column contains the number of novel editing sites found across all regions. The total number of novel polymorphisms residing across all regions is 1486.

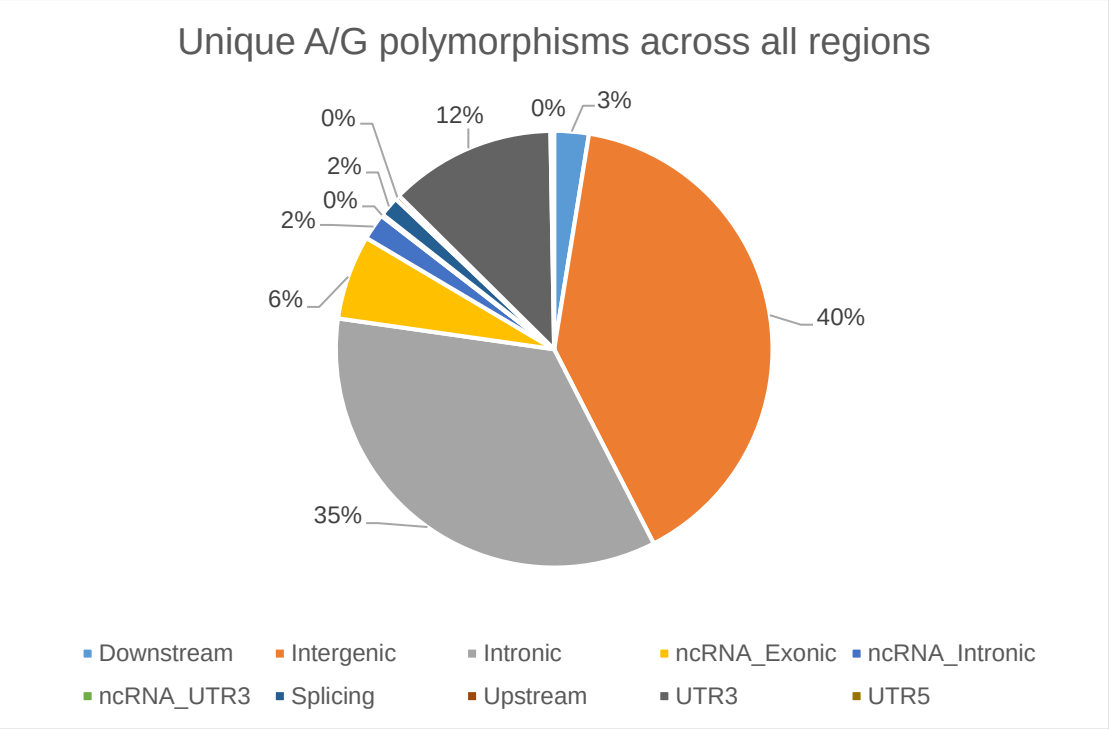


Figure 5: A pie chart displaying the above data listed in Table 10. The number of novel PES detected in each region is displayed across all samples as a percentage of the total.

. The graph below (Figure 6) shows the average number of PES per chromosome, with the standard deviation represented by the error bars. The chromosome with the most polymorphisms is chromosome 1, with an average of 37 editing sites (21-53), despite chromosome X having a higher standard deviation. The fewest PES were present on the mitochondrial chromosome, with an average of 1 PES (0-4). This correlates with the overall length of the chromosomes, with chromosome 1 being the largest and the mitochondrial chromosome the smallest. Chromosomes 21, Y, and those with an unknown allocation did not show any PES on average and were hence disregarded from the graph.

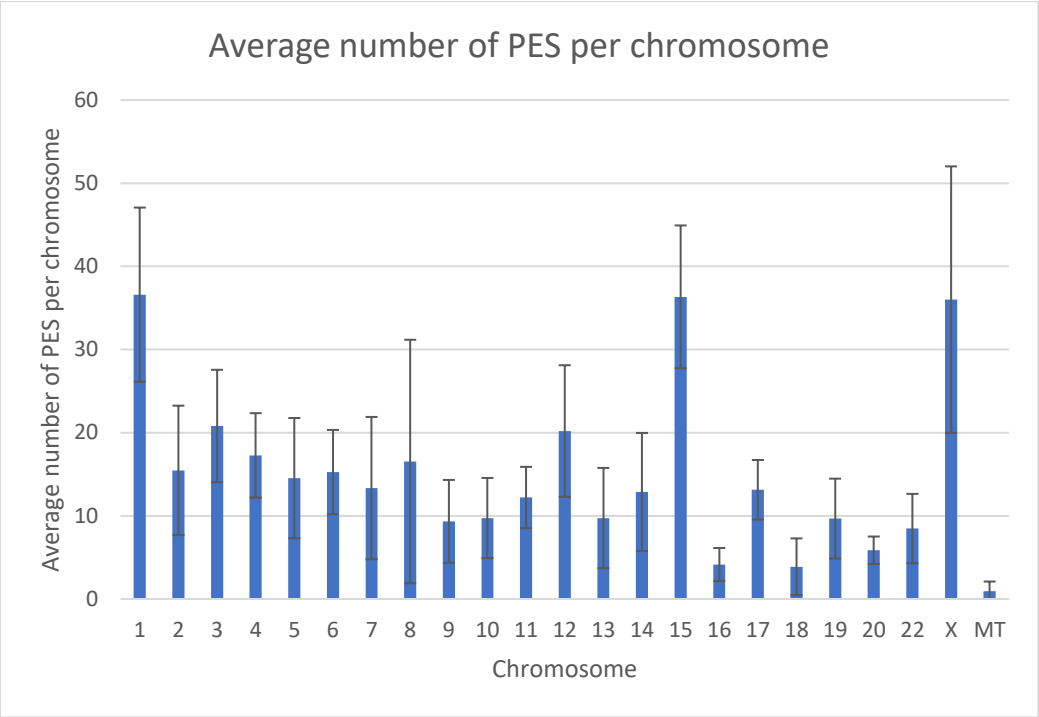


Figure 6: A graph displaying the distribution of PES across all chromosomes. PES with rs numbers excluded. Data compiled from the "AG mutations" file.

Figure 7 shows the average number of PES called by the bioinformatics pipeline per chromosome as calculated for Figure 6, against the length of each chromosome. Figure 7 shows a positive correlation between the two properties.

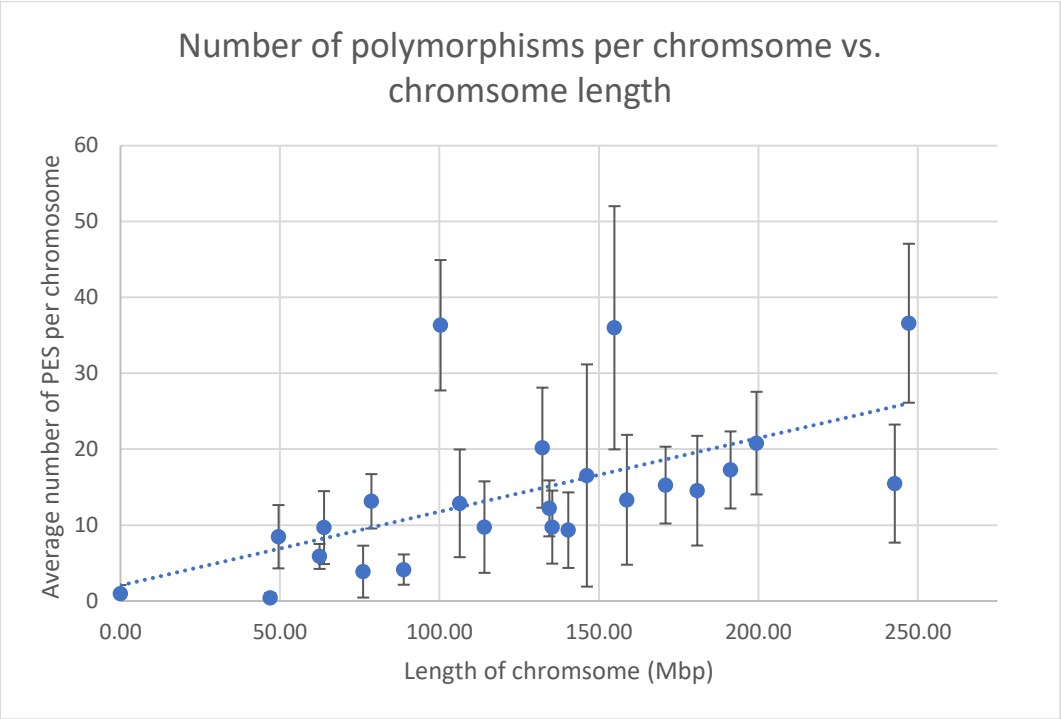


Figure 7: A graph showing the average number of PES found by the bioinformatics pipeline per chromosome against the length of each chromosome. Data taken from the "Editing site demographics" file.

4.2.2 Filtering for novel PES using the "Extracted PES" file

The 83 variants listed in the final output stage of the above flowchart are an average across all samples. By combining the final output file for each sample and filtering for duplicates, it was possible to construct a table containing all novel PES called by the pipeline. Disregarding intronic, intergenic and ncRNA_intronic residing PES, it was found that 339 PES were found in at least one brain region. In hindsight those residing in splicing regions should have been disregarded during the original pipeline as the sites are located 2bp into an intron at a splice junction. Of the total 339 novel PES called, 19 were SNVs and hence removed as they differ from RNA A-to-I editing sites. The remaining 320 sites, consist of 30 downstream PES, 85 PES residing in the ncRNA_exonic regions, 2 ncRNA_3'UTR, 21 PES are splicing variants, 2 are upstream PES, 177 exist in the 3'UTR and 3 in the 5'UTR. Of these PES, 189 were found to reside in 2 or more samples. None of the PES resided in coding exons.

When studying the locations of the PES in all chromosomes using the above mentioned novel PES table, it was noted that some editing sites were present in clusters. To be classified as being present in a cluster the editing site must be within 50bp of another PES (Zhu *et al*., 2013).

PES were ranked for validation in the laboratory via Sanger sequencing based on the editing pattern shown and the average number of raw reads supporting each site. Firstly, a selection of PES showing different editing patterns were chosen. Then data was collected on the number of reads supporting each PES in all fifteen RNA-Seq samples and the average value calculated. This average value was used to rank the editing sites in order of priority for follow up in the laboratory. The sites validated in the laboratory are shown in yellow. PES_*MEG3*,MIR770_ncRNA_exonic was also chosen for validation however it was tested unsuccessfully in the laboratory. Despite being ideal candidates for further testing in the laboratory; due to sufficiently supported gDNA genotypes, the likes of PES_PIN1P1_ncRNA_exonic, PES_RAB11F1P4_3'UTR and PES_RBBP4_3'UTR_2 were not included as they were all supported by less than an average of 100 reads in the raw RNA-Seq data.

Table 9 is split into three parts, Part A containing PES with the exonic DNA taken into consideration, Part B where PES are identified but the exonic DNA cannot be determined and Part C where the exonic DNA shows a heterozygote.

Each PES was also checked against RNA Editing databases RADAR/DARNED to identify if it had already been discovered (Kiran and Baranov, 2010, Ramaswami and Li, 2014). Those discovered in RADAR/DARNED are shown in bold within Tables 9 A-C.

			Sample Number & Brain region																		
PES ID	Chr	Position	1 DNA	1 Frontal	1 Temporal	1 Occipital	1 Cerebellum	1 Parietal	2 DNA	2 Frontal	2 Temporal	2 Occipital	2 Cerebellum	2 Parietal	3 DNA	3 Frontal	3 Temporal	3 Occipital	3 Cerebellum	3 Parietal	Reads capturing the PES
PES_*MALAT1*_ncRNA_exonic	11	65266712	A	A	A	A	A	A	A/G	A/G	A/G	A/G	A/G	A/G	A	A	A	A	A	A	3573
PES_*MEG3*,*MIR770*_ncRNA_exonic	14	101318811	A	A/G	A/G	A/G	A/G*	A/G	A	A/G	A/G	A/G	A/G	A/G	A	A/G	A/G*	A/G	A*	A/G*	217
PES_*HMP19*_3'UTR	5	173535468	A*	A/G	A/G	A/G	A/G*	A/G	A*	A	A	A	A	A	A*	A	A*	A*	A*	A*	175
PES_*RAB11FIP4*_3'UTR	17	29862183	A	A/G*	A	A/G*	A*	A/G*	A	A/G*	A/G	A*	A*	A/G*	A	A/G	G*	A/G	A*	A/G	62
PES_*RBBP4*_3'UTR_2	1	33149502	A*	G*	G	G*	G*	G*	A*	G*	G*	G*	G	G*	A	G*	G*	G*	G*	G*	36
PES_*RBBP4*_3'UTR_5	1	33149414	A*	A/G	A/G*	G*	A/G*	A/G*	A*	G*	G*	G*	A/G*	A*	A*	A/G*	A/G*	A*	A/G*	A/G*	24
PES_*HSP90B2P*_ncRNA_exonic	15	99798257	A	G*	G	G*	A/G*	A/G*	A	G	A/G	G*	G*	A/G*	A	G*	G*	A/G*	A*	G*	17
PES_*RPS14P3*_ncRNA_exonic	1	19935043	A	G	G	G	G	G	A	G	G	G	G	G	A*	G	G	G	A/G	G	14
PES_*PIN1P1*_ncRNA_exonic	1	70385332	A	G	G	G*	G*	G	A	G*	A/G*	G*	G*	G*	A	G*	G*	A/G*	G*	G*	7
PES_*GABARAPL3*_ncRNA_exonic	15	90892244	A	G	G	G	G*	G	A	G	G	G	G	G	A	G	G*	G*	G*	G	2
PES_*PA2G4P4*_ncRNA_exonic	3	156527380	A*	G*	G*	G*	G*	G*	A*	G*	G*	G*	G*	G*	A*	G	G*	G	G*	G*	1

Table 9A: This table shows the PES identified from the pipeline that had corresponding exome sequence reads from corresponding DNA samples, ordered by the number of average supporting reads from the RNA sequence. Exome-seq data from corresponding genomic DNA samples suggests that these may be true positive PES; however, one sample (highlighted in yellow) only shows RNA-editing in one of the samples, and this is supported by less than 10 exome reads. Therefore, this PES was tested in the laboratory to corroborate whether this was a true RNA editing site or a false positive finding in the pipeline. One PES was highlighted in green as this shows that false positives are produced by the pipeline, only highlighted when corresponding genomic DNA is used. Genotypes (taken from the Table Recalibration step) supported by less than 10 reads are marked with an asterisk, and PES found to already exist in DARNED are in bold. PES are named according to the RefGene they reside in followed by the gene region. Those PES which would have the same name using this system are followed by a number to distinguish them. Universal editing is shown across all regions in all samples, Sample-specific editing occurs in five brain regions in one sample only and Sporadic editing appears to be random. For all PES, it was ensured that the RefSeq base at each position was an A.

			Patient Number & Brain region																		
	Chr	Position	1 DNA	1 Frontal	1 Temporal	1 Occipital	1 Cerebellum	1 Parietal	2 DNA	2 Frontal	2 Temporal	2 Occipital	2 Cerebellum	2 Parietal	3 DNA	3 Frontal	3 Temporal	3 Occipital	3 Cerebellum	3 Parietal	Reads capturing the PES
PES_*KIF1B*_3'UTR	**1**	**10440678**	**-**	**A**	**A**	**A**	**A**	**A**	**A***	**A**	**A**	**A**	**A**	**A**	**-**	**A**	**A**	**A**	**A/G**	**A**	**300**
PES_*BTBD3*_3'UTR	20	11906859	A*	A	A	A	A	A	-	A/G	-	A/G	A/G	A/G	-	A	A	A	A	A	235
PES_*NEBL*_3'UTR	10	21072164	-	A/G*	A/G	A/G	A/G	A/G	-	A	A	A*	A	A	A*	A*	A*	A	A	A*	172
PES_*PWARSN*_ncRNA_exonic_1	15	25227768	A*	A/G	A/G	A/G	A/G	A/G	A*	G	A/G	A/G	A/G	A/G	-	A/G	A/G	G	A/G	A/G	133
PES_*CRTC1*_3'UTR	19	18892734	A*	A	A/G*	A/G	A*	A/G	-	A*	A/G*	A*	A*	A/G	-	A/G*	A/G*	A*	A*	A/G	117
PES_*MPRIP*_3'UTR	**17**	**17089926**	**A***	**A/G***	**A/G***	**A**	**A/G***	**A/G***	**-**	**A**	**A/G***	**A/G***	**A/G**	**A/G**	**-**	**A/G***	**A***	**A/G**	**A**	**A***	**92**
PES_*PWARSN*_ncRNA_exonic_2	15	25227838	-	A/G	A/G	A/G	A/G	A/G	A*	A/G	A/G	A/G	A/G	A/G	-	A/G	A/G	A/G	A/G	A/G	75
PES_*PWARSN*_ncRNA_exonic_5	15	25228793	A*	G	A/G	A/G	A/G	G	-	G	G	G	A/G	G	-	G	G	A/G	A/G	G	69
PES_*PWARSN*_ncRNA_exonic_6	15	25228814	A*	A/G	A/G	A/G	A/G	A/G	-	G	A/G	G	A/G	G	-	G	G	A/G	G	A/G	62
PES_*PWARSN*_ncRNA_exonic_7	15	25228815	A*	G	G	G	G	G	-	G	G	G	G	G	-	G	G	A/G	G	G	62
PES_*INADL*_3'UTR	1	62628522	-	-	A*	A/G*	A/G	A*	-	G*	G*	A*	A/G	A/G*	A*	A*	-	-	A/G	A*	34
PES_*RBBP4*_3'UTR_3	1	33149373	A*	A/G	A/G*	A/G*	A/G*	A/G*	-	A/G*	G*	G*	A/G*	A/G*	-	A/G*	A/G*	A/G*	A/G*	A*	20
PES_*RBBP4*_3'UTR_4	1	33149376	A*	G	G*	G*	G*	A/G*	-	A/G*	G*	G*	A/G*	G*	-	G*	G*	G*	A/G*	G*	19

Table 9B: This table shows 13 PES for further investigation, with at least one of genomic DNA sample supported by an A genotype. These PES were chosen primarily on the editing pattern shown across the tissues. However, when checking the corresponding exonic DNA, some PES were not supported by any reads and are hence marked with a hyphen. Genotypes (taken from the Table Recalibration step) supported by less than 10 reads are marked with an asterisk, and PES found to already exist in DARNED are in bold. PES are ranked by the average number of supporting reads in the raw bam file. PES are named according to the RefGene they reside in followed by the gene region. Those PES which would have the same name using this system are followed by a number to distinguish them. Sample-specific editing occurs in five brain regions in one sample only. Those PES suspected of being a sequencing error only reside in one brain region in one sample. Those PES highlighted in yellow were verified in the laboratory. For all PES, it was ensured that the RefSeq base at each position was an A.

			Patient Number & Brain region																		
	Chr	Position	1 DNA	1 Frontal	1 Temporal	1 Occipital	1 Cerebellum	1 Parietal	2 DNA	2 Frontal	2 Temporal	2 Occipital	2 Cerebellum	2 Parietal	3 DNA	3 Frontal	3 Temporal	3 Occipital	3 Cerebellum	3 Parietal	Reads capturing the PES
PES_*PWARSN*_ncRNA_exonic_3	15	25227939	-	G	A/G	A/G	A/G	G	-	A/G	A/G	A/G	A/G	A/G	-	A/G	A/G	A/G	G	A/G	257
PES_*PWARSN*_ncRNA_exonic_4	15	25228419	-	A/G	A/G	A/G	A/G	G	-	G	A/G	G	G	G	-	A/G	G	G	A/G	A/G	246
PES_*MEG3*_ncRNA_exonic	**14**	**101312647**	**-**	**A/G**	**A/G**	**A/G**	**A**	**A/G**	**-**	**A/G**	**A/G**	**A/G**	**A/G**	**A/G**	**-**	**A/G**	**A/G***	**A/G**	**A/G**	**A/G***	**243**
PES_*OIP5-AS1*_ncRNA_exonic	15	41591342	-	A/G	A/G	A/G	A/G	A/G	-	A/G	A/G	A/G	A/G	A/G	-	A/G	A/G	A/G	A/G	A/G	175
PES_*ZYG11B*_3'UTR	1	53291420	-	A/G	G	A/G	A/G*	A/G	-	G	G	G	G	G	-	A/G	A/G	G	A/G	A/G	131
PES_*IFNAR1*_3'UTR	21	34730096	-	A/G*	A	A/G	A/G*	A/G	-	A/G*	-	A*	A/G*	A/G*	-	A/G	A/G	A	A	A/G	106
PES_*CAPZA1*_3'UTR	1	113213682	-	G	G	A*	A/G*	G	-	A	A	A*	A*	A*	-	A*	A*	A*	A*	A*	105
PES_*KIAA1456*_3'UTR_2	8	12881826	-	A/G*	A/G*	G*	A/G	A/G*	-	A/G*	A/G*	G*	A/G	A/G*	-	A/G*	A/G*	T	A/G	A/G*	69
PES_*MRTO4*_3'UTR	**1**	**19585553**	**-**	**G***	**G**	**A/G***	**A/G***	**G***	**-**	**A***	**G***	**A/G***	**G***	**A/G***	**-**	**A/G***	**G***	**A/G***	**-**	**A/G***	**44**
PES_*RBBP4*_3'UTR_1	1	33148789	-	G*	G*	G*	A/G*	A/G*	-	A/G*	G*	G*	A/G	A/G	-	G*	A/G*	G*	A/G*	A/G*	38
PES_*KIAA1456*_3'UTR_1	**8**	**12881759**	**-**	**G***	**A/G***	**G***	**G**	**G***	**-**	**G***	**G***	**-**	**G**	**A/G***	**-**	**G***	**G***	**-**	**G**	**G***	**37**
PES_*KIAA1456*_3'UTR_3	**8**	**12882906**	**-**	**-**	**A***	**G***	**A/G**	**G***	**-**	**-**	**-**	**G***	**A/G***	**G***	**-**	**-**	**-**	**-**	**A/G**	**G***	**8**

Table 9C: Table showing 12 PES for further investigation, however no corresponding genomic DNA was available for these positions. PES were chosen primarily on the editing pattern shown across the tissues. However, when checking the corresponding exonic DNA, the PES were not supported by any reads and are hence marked with a hyphen. Genotypes (taken from the Table Recalibration step) supported by less than 10 reads are marked with an asterisk, and PES found to already exist in DARNED are in bold. PES are ranked by the average number of supporting reads in the RNA-Seq data file. PES are named according to the RefGene they reside in followed by the gene region. Those PES which would have the same name using this system are followed by a number to distinguish them. The PES highlighted in yellow was verified in the laboratory. For all PES, it was ensured that the RefSeq base at each position was an A.

4.3 Validating PES in the laboratory

4.3.1 RNA Quantity and Quality

Tapestation results for the RNA extracted from each region of the brain for all three samples demonstrates that good quality RNA was achieved (Table 10). The average RIN was 5.5 (3.5-6.5) with DV200 percentages all above the threshold of 70% (Illumina, Technical Note RNA Sequencing "Evaluating RNA Quality from FFPE Samples") with an average of 87.61%.

Sample	RIN	Conc. [ng/µl]	DV200
1 Frontal	6.5	106	90.09
1 Temporal	6.5	282	91.09
1 Occipital	6.2	255	89.32
1 Cerebellum	5.7	201	87.00
1 Parietal	6.3	362	90.37
2 Frontal	6	158	89.14
2 Temporal	6.3	305	90.17
2 Occipital	4.6	370	86.19
2 Cerebellum	5.9	174	89.84
2 Parietal	4.3	526	81.52
3 Frontal	5.3	125	86.92
3 Temporal	4.8	688	88.40
3 Occipital	3.5	260	78.45
3 Cerebellum	4.5	372	81.94
3 Parietal	4.3	165	73.79

Table 10: A table showing the fifteen RNA samples ran on the Agilent 2200 Tapestation to check RNA quality. The RIN is shown against the sample concentration and the DV200 value.

Genomic DNA contamination was tested for by PCR using primers designed to capture intergenic genome sequence, and in all samples no gDNA was observed when gel electrophoresis of the PCR products was undertaken (Data not shown).

4.3.2 Sanger Sequencing

The Sanger sequencing of gDNA from each of the three samples was carried out for 4 PES, alongside the sequencing of cDNA; synthesised from RNA extraction, for each of the five regions across all three samples. Of these four PES tested, one was found to be a sequencing error, two were not validated and one was verified as an RNA A-to-I editing site.

PES_*KIF1B*_3'UTR: The exome sequencing data suggested that this site was an A in sample 2, supported by less than 10 reads, and the

remaining two gDNA samples were not supported by any reads at all. Contrary to this, Sanger sequencing validated that the nucleotide at this position was an A in all three samples. The RNA-Seq data suggested that there might be a PES in the Cerebellum region of sample 3. Due to the lack of editing pattern shown this site was thought to be a sequencing error. Sanger sequencing revealed that there was an A base in the Cerebellum, see chromatograph of PES (Figure 8), with gDNA on the left panel and cerebellum on the right. The PES in question is highlighted in all cases. All samples and regions showed an A at this position (see Table 11). This was verified by the alignment of fasta sequence to the reference using Clustal Omega (data not shown). This PES was hence thought to be an error made by the sequencing platform, despite literature, and it was verified as such.

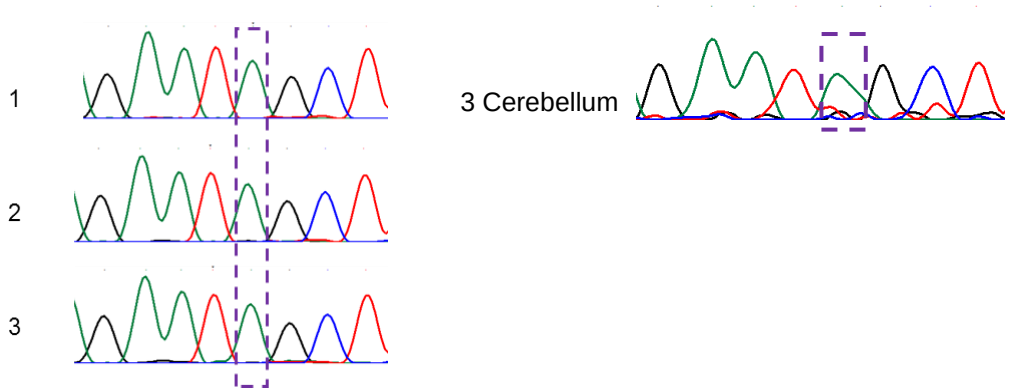


Figure 8: Chromatogram sections for gDNA samples 1-3 on the left, and the cDNA Sample 3 Cerebellum on the right, for PES_KIF1B_3'UTR, with the PES highlighted by dashed lines. All samples show an adenine base at that position.

PES_*HMP19*_3'UTR: The exome sequencing data for this site suggested the presence of an A nucleotide at position 173535468 in all three gDNA samples which is confirmed in two of three samples by Sanger sequencing. However, in sample 1, the sequencing quality is not particularly good, so it is difficult to speculate on whether there is a heterozygote present or not (see chromatogram in Figure 9). The RNA-Seq data predicted from the bioinformatics pipeline suggested that the PES may be present in all five brain regions in Sample 1 only. This is validated by Sanger sequencing (see Figure 9). Sample 1 Parietal is called by the sequencing platform as a guanine base according to the alignment; however, when viewing the chromatogram, it is visible that this is an A/G heterozygote. The two other samples show an adenine base at the PES in all regions. As this PES was thought to be a heterozygote in both sample 1 gDNA and in all sample 1 cDNA brain regions, this suggests there is no

PES present, as it would be expected to see both alleles reflected in the RNA.

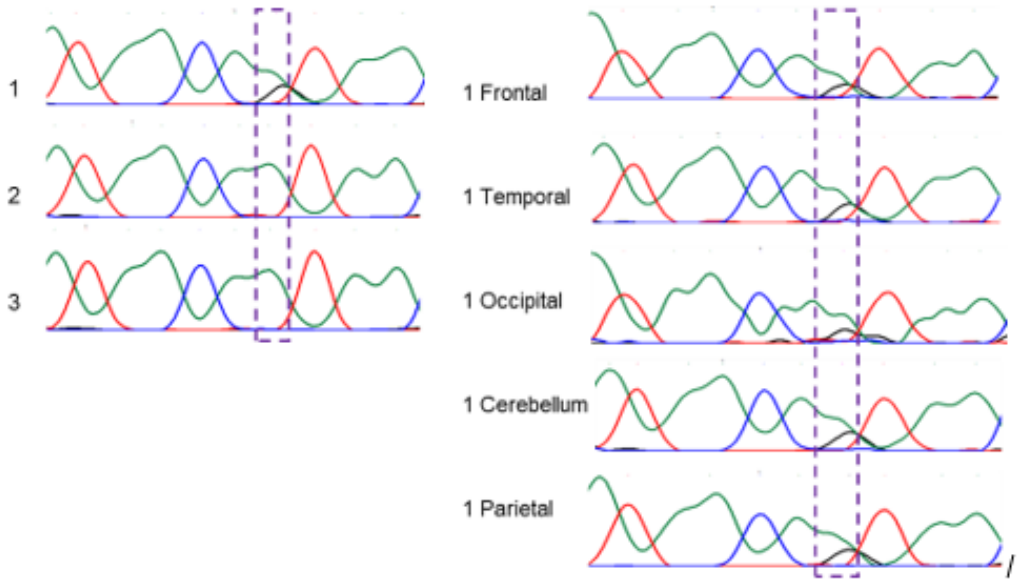


Figure 9: Chromatogram sections showing gDNA samples 1-3 on the left and cDNA sample 1, all brain regions on the right for PES_HMP19_3'UTR. gDNA sample 1 shows an A/G heterozygote whereas samples 2 and 3 show an adenine base at the PES. All brain regions in cDNA sample 1 show an A/G heterozygote.

PES_*NEBL*_3'UTR: The exome sequencing data for this site in samples 1 and 2 cannot predict a genotype as it is not supported by any reads. However, in sample 3, an A nucleotide is predicted at this position despite being supported by less than 10 reads. The Sanger sequencing in gDNA, shows the presence of an A nucleotide at the PES for samples 2 and 3 (see Figure 10). For sample 1, two peaks are observed, suggesting an A/G heterozygote. As the gDNA is heterozygous the site could be a mutation or unknown polymorphism and it was hence tested. From the RNA-Seq data, the PES is predicted in all brain regions within sample 1 but not in any regions for samples 2 and 3. Sanger sequencing confirms the presence of the PES, indicated by A/G heterozygote in all sample 1 brain regions. At the PES in samples 2 and 3 an A nucleotide is observed in all brain regions. As the gDNA for Sample 1 shows the presence of an A/G heterozygote, and all brain regions for Sample 1 in cDNA also show an A/G heterozygote, there is no RNA A-to-I editing taking place at this PES as it would be expected to see both alleles reflected in the RNA. Sample 1 Occipital was not confirmed by Sanger sequencing due to the poor quality of the sequencing chromatogram.

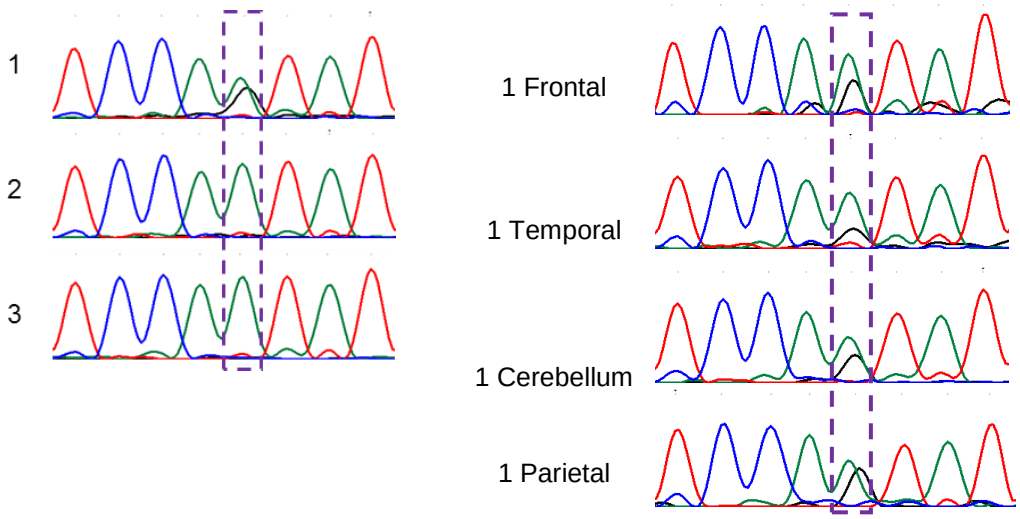


Figure 10: Chromatogram sections for all three gDNA samples on the left and cDNA sample 1, frontal, temporal, cerebellum, and parietal regions respectively on the right. gDNA sample 1 shows an A/G heterozygote at PES_NEBL_3'UTR, whilst samples 2 and 3 show an adenine base. For all four brain regions in cDNA sample 1, an A/G heterozygote is present, suggesting no editing has taken place.

PES_*MEG3*_ncRNA_exonic: Exome sequencing data was unable to predict the genotype of the PES in gDNA as the site was not supported by any reads. However, Sanger sequencing confirmed that for all three samples an A base was present (see Figure 11). The RNA-Seq data predicts that the PES will be present in all three samples, in all brain regions with the exception of sample 1 cerebellum. For all samples which were predicted to undergo A-to-I editing, Sanger sequencing validated the presence of the PES (see Figure 12). Although sample 1 Cerebellum shows both a peak for adenosine and one for guanosine, the guanosine peak is significantly smaller than in the other brain regions, suggesting it undergoes editing at a lower level, if at all. Sample 3 Occipital and Parietal were not sequenced due to poor quality.

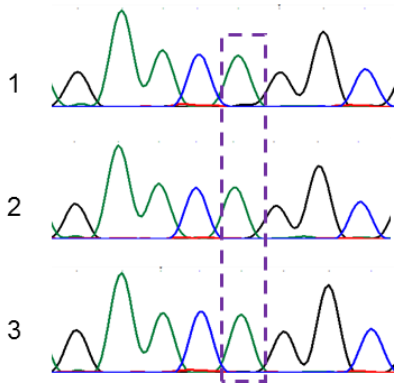


Figure 11: Chromatogram sections showing gDNA samples 1-3 for PES_MEG3_ncRNA_exonic. All samples show an adenine base at the PES, highlighted by dashed lines.

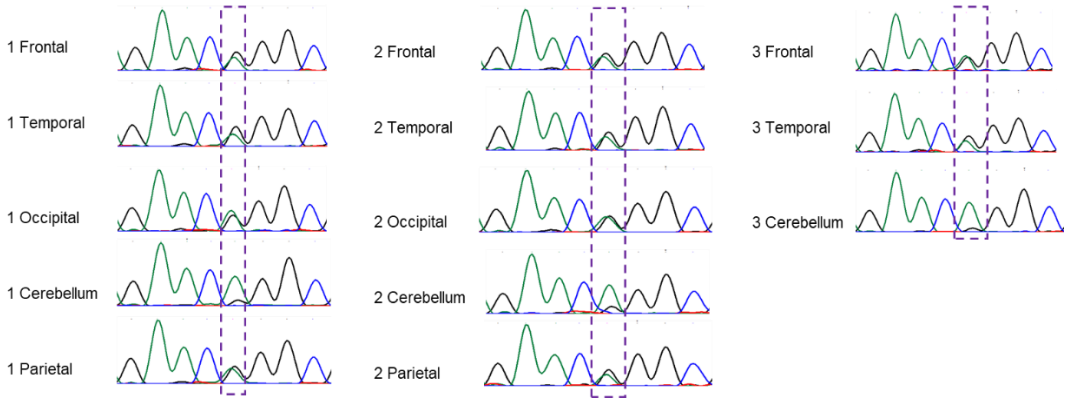


Figure 12: Chromatogram sections for all cDNA samples sequenced for PES_MEG3_ncRNA_exonic, excluding sample 3 occipital and parietal which were not sequenced due to poor quality. All cDNA samples show the presence of an A/G heterozygote at the PES indicative of editing, except for sample 1 Cerebellum which displays a noticeably smaller guanosine peak. In all three samples, in the cerebellar region, the peak representing guanosine is distinctly lower than at the PES for the other brain regions.

4.3.3 Summary of Validation Results

PES ID	Chr	Position	1 DNA	1 Frontal	1 Temporal	1 Occipital	1 Cerebellum	1 Parietal	2 DNA	2 Frontal	2 Temporal	2 Occipital	2 Cerebellum	2 Parietal	3DNA	3 Frontal	3 Temporal	3 Occipital	3 Cerebellum	3 Parietal
PES_*KIF1B*_3'UTR	1	10440678	-	A	A	A	A	A	A*	A	A	A	A	A	-	A	A	A	A/G	A
			A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
PES_*HMP19*_3'UTR	5	173535468	A*	A/G	A/G	A/G	A/G*	A/G	A*	A	A	A	A	A	A*	A	A*	A*	A*	A*
			A/G	A/G	A/G	A/G	A/G	G	A	A	A	A	A	A	A	A	A	A	A	A
PES_*NEBL*_3'UTR	10	21072164	-	A/G*	A/G	A/G	A/G	A/G	-	A	A	A*	A	A	A*	A*	A*	A	A	A*
			A/G	A/G	A/G	-	A/G	A/G	A	A	A	A	A	A	A	A	A	A	A	A
PES_*MEG3*_ncRNA_exonic	14	101312647	-	A/G	A/G	A/G	A	A/G	-	A/G	A/G	A/G	A/G	A/G	-	A/G	A/G*	A/G	A/G	A/G*
			A	A/G	A/G	A/G	A/G	A/G	A	A/G	A/G	A/G	A/G	A/G	A	A/G	A/G	-	A/G	-

Table 11: Table showing the predicted editing patterns for each PES investigated in the laboratory (highlighted grey), with Sanger sequencing verification shown in the second line. Those genotypes marked with an asterisk are supported by less than 10 reads in the genomic DNA or RNA-Seq data. Those marked with a hyphen were not present, either through a lack of supporting reads in the predicted pattern or no available sequencing data in the Sanger verification.

For each PES investigated, the above table shows the bioinformatics prediction taken from the Table Recalibration step in the first row, followed by the sequencing results in the second. It is possible to see that for PES_*KIF1B*_3'UTR, the prediction was not verified in the tissue. For PES_*HMP19*_3'UTR and PES_*NEBL*_3'UTR, the editing pattern predicted from the RNA-Seq data has been verified but the gDNA genotype is different from that predicted in sample 1. This shows that no editing has taken place as an A/G heterozygote would be expected in the cDNA samples. Finally, for PES_*MEG3*_ncRNA_exonic, the editing pattern predicted is verified by Sanger sequencing.

5 Discussion

This project has used RNA sequencing data from five brain regions in three control samples to identify potential RNA-editing sites in the transcriptome of the brain. A bioinformatics pipeline has been developed to identify potential RNA editing sites from RNA sequence, comparing sites to a reference sequence and screening out known polymorphisms. In order to evaluate the pipelines ability to discern true RNA editing sites, PES were compared with corresponding exome sequencing from each sample and a number of PES were verified using independent Sanger sequencing methodology to confirm suspected false positive and true positive findings. A single PES in the *MEG3* gene was verified and suggested RNA editing occurring across the different brain regions tested with the exception of the Cerebellum.

5.1 Identification of PES

There were 1,486 PES identified by the bioinformatics pipeline across all brain tissues sampled. The majority resided in intergenic (39.9%) and intronic regions (34.8%). However, these regions were deliberately excluded from the final output due to a filtering stage in the pipeline as the initial brief was to study those PES in the coding regions. Of the remaining PES, 12.3% were located in the 3' untranslated region of a gene. The fewest PES were located in the ncRNA_3'UTR (0.13%). None were located in the exonic coding region of genes and few were observed in exonic regions of non-coding RNA transcripts (6.3%). This may be due to the fact that RNA A-to-I editing is found primarily in inverted *Alu* repeats (Sakurai *et al*., 2014), which most commonly reside in non-coding regions such as the intergenic and introns rather than coding sequence (Zinshteyn and Nishikura, 2009). These figures indicate that A-to-I editing is most prevalent in the intergenic and intronic regions for the tissues tested, followed by the 3'UTR. These findings agree with literature outlining the locations of RNA A-to-I editing sites (Blow et al., 2004, Nishikura, 2010, Park et al., 2012). However, the figures outlined in the paper by Blow *et al* vary widely from those gathered above, with 70.30% of editing sites residing in intronic regions and 19.28% in intergenic regions. Nishikura (2010) and Park *et al*., (2012) do not include numbers within their studies. The difference in proportions of editing sites may be due to the fact that the study undertaken

by Blow *et al* did not include the cerebellum, however when re-evaluating the figures collected from our data without the presence of the cerebellar regions, the percentages were not vastly affected (intronic – 44.08%, intergenic – 27.17%). It is therefore likely that the difference in percentages is due to a difference in technique used to call the editing sites.

Blow *et al*. carried out a study of RNA editing in the human brain by comparing sequences from a single sample human brain cDNA library; created from cerebral cortex tissue, to the reference human genome sequence (hg16) and to genomic DNA for the same individual. The issue with using reference libraries to align sequencing reads is that despite constantly advancing, there are relatively large regions where the exact sequence is not known, particularly in repetitive regions, leading to mapping errors (Miga et al., 2015). As well as identifying the intronic region as the most highly edited (70.30%), this study found that the majority of edits are located in high-copy-number repeats such as *Alu* elements and the likelihood of editing is increased by the occurrence of two inverted copies of a sequence within the same intron, the distance between the two inverted sequences and by the density of inverted copies in the region. To determine whether the potential editing sites called by the bioinformatics pipeline are located within or near to *Alu* elements, the database *Alu*Gene (Dagan *et al*., 2004) could be searched for the coordinates of the PES. This was not initially done due to time constraints.

Alu regions are prime targets for RNA A-to-I editing; due to their ability to form intramolecular inverted dsRNA structures, a prerequisite for editing to take place via the action of ADARs. With 75% of all known genes containing *Alu* elements within their introns and/or UTRs (Kim et al., 2004), it suggests that a large amount of editing could take place in these regions. It is thought that *Alu* elements tend to reside in these areas because they are gene-rich and contain high levels of G+C (Deininger, 2011). These editing events could be involved in the elimination or creation of splice sites as well as many aspects of RNA metabolism such as trafficking and microRNA binding and regulation (Li and Church, 2013). The deletion of existing splice sites and the creation of new ones can lead to the inclusion or exclusion of *Alu* exons and introns, generating a new variant protein with a novel function (Nishikura, 2010). Currently, very little is known about the full extent of RNA A-to-I editing, due to the technical challenges of deciphering the functions of an individual RNA editing event (Li and

Church, 2013). However, it is speculated that despite not directly altering the protein function, high levels of RNA A-to-I editing occurring in the intronic/UTR regions of the genes can impact gene regulation through the action of transcriptional enhancers/silencers and further regulatory RNAs that affect translation and mRNA stability of the encoded gene. Particular interest in the effect of RNA A-to-I editing on miRNA biogenesis and the recognition site interactions of mature miRNA is noted (Picardi *et al*., 2015).

The number of editing sites varies from chromosome to chromosome. Chromosome 1 has the highest average number of editing sites including those residing in intronic/intergenic regions, with the mitochondrial chromosome having the fewest. This correlates with the overall length of the chromosomes, with chromosome 1 being the largest and the mitochondrial chromosome the smallest. However, chromosome X shows the highest standard deviation in the number of editing sites present when compared to the other chromosomes. This suggests that there is the greatest amount of variability i.e. the highest range in the number of editing sites across samples. This implies that there may be greater tissue differences/sample differences within chromosome X.

Editing sites were also observed to cluster. The clustering of editing sites occurs due to the promiscuous editing of dsRNA, predominantly in 3'UTR and introns (Nishikura et al., 1991, Polson and Bass, 1994). These clusters are abundant in repetitive genomic regions and are rarely seen in coding regions (Porath *et al*., 2017). An editing site is classified as showing clustering; or being in an "editing box", if it resides within 50bp of another editing site (Zhu *et al*., 2013). For example, in chromosome 15, 26 of the 37 potential editing sites present are clustered within 50bp of another editing site. Overall, 51 clusters were observed across all chromosomes, mainly located within 3'UTR, ncRNA_exonic and downstream regions. For the 21 chromosomes containing potential editing sites (none were observed in chromosome 18, MT or Y due to initially low numbers being filtered out), 8 chromosomes (57%) showed clustering. The number of clusters per chromosome varied from 1 (chr2, 3, 6, 7, 10, 19, 21) to 8 (chr12). There appears to be no correlation between chromosome length and overall clustering. The chromosomes which do not show clustering do not appear to have any properties which would distinguish them from those which are clustered.

The clustering of multiple editing sites is caused by the non-selective editing of several adenosines by ADARs simultaneously, made possible by the presence of stretches of dsRNA >100bp in length (Nishikura, 2016). Non-selective or promiscuous editing causes the deamination of 50–60 % of the adenosines in a section of dsRNA (Lehmann and Bass, 1999). The purpose of clustering with relation to RNA A-to-I editing is not speculated often in literature.

5.2 Evaluation of PES called from pipeline

In the bioinformatics analysis of the RNA data, the hg19 genome build was included as a genomic DNA comparison to reduce the number of false positive PES findings, alongside excluding known polymorphisms sites via the 'rs' identifiers. When comparing PES with corresponding DNA sequencing obtained from exome sequencing for the same samples, only one PES was shown to be false positive before laboratory validation (Table 9A). As the RNA bases observed at that point matched that of the corresponding DNA sample, suggesting that unknown DNA A/G polymorphisms or mutations were occurring at that site, and that the perceived PES was just a reflection of both alleles being transcribed from the DNA. As a result, a number of PES with no or limited DNA sequence information were tested for verification of PES status. Three out of the four PES were found to be false positives, indicating the importance of obtaining corresponding DNA information for the RNA samples.

The four PES tested in the laboratory were PES_*KIF1B*_3'UTR, PES_*HMP19*_3'UTR, PES_*MEG3*_ncRNA_exonic and PES_*NEBL*_3'UTR. PES_*KIF1B*_3'UTR is located at position 10440678 (amino acid 3116) of chromosome 1 in exon 47, at the 3' untranslated region of *KIF1B* and is a known editing site in DARNED. PES_*HMP19*_3'UTR resides at position 173535468 (amino acid 492) of chromosome 5 in exon 5, within the 3'UTR of gene *HMP19.* Gene *HMP19* is also known as *NSG2*.
PES_*MEG3*_ncRNA_exonic is also a known RNA A-to-I editing site, listed in DARNED. PES_*MEG3*_ncRNA_exonic is located at position 101312647 of chromosome 14 in the exonic region of the non-coding gene *MEG3.* There are 16 transcripts of *MEG3,* with the PES residing in the untranslated transcript 16 which is a truncated RNA. The final PES tested in the laboratory was PES_*NEBL*_3'UTR, an unknown PES located at position 21072164 of chromosome 10. This PES resides in the 3'UTR of *NEBL*.

Depending on the transcript, PES_*NEBL*_3'UTR resides at amino acid 1108/1109 within exon 7, or at amino acid 1853 in exon 28.

Of the PES entered in the ranking table, six were found to be previously known A-to-I editing sites in the RNA editing database DARNED. The first of which is PES_KIF1B_3'UTR, located at position 10440678 of chr1 as mentioned above, in the 3'UTR of *KIF1B.* It was discovered in the Substantia nigra of the midbrain. PES_*MRTO4*_3'UTR is located within exon 8, amino acid 317, at position 19585553 of chromosome 1. The editing site resides in the 3'UTR of gene *MRTO4*, and was discovered in a lymphoblastoid cell line. PES_MEG3_ncRNA_exonic, as described above, is located at position 101312647 of chromosome 14 in the exonic region of the non-coding gene *MEG3.* The RNA A-to-I editing site was discovered in the Cerebellum. PES_MPRIP_3'UTR resides at position 17089926 of chr17, in the 3'UTR of *MPRIP.* Depending on the transcript*,* the PES is located in either exon 23 at amino acid position 1612, or in exon 24. The PES was first found in a lymphoblastoid cell line. PES_KIAA1456_3'UTR_1 was first discovered in several tissues including the liver, prostate, uterus, kidney, colon, normal and tumourous lung, glioma brain tumour, cerebellum and frontal lobe. The PES resides at position 12881759 of chromosome 8 within the 3'UTR of gene *KIAA1456.* Again, relative to the transcript produced, the PES is located in either exon 5, at amino acid 1191, or exon 2, at amino acid position at 1065. PES_KIAA1456_3'UTR_3 is also located on chromosome 8 at position 12882906, within the 3'UTR of *KIAA1456.* Depending on the transcript produced the PES can reside in exon 5 at amino acid 1573 or in exon 2 at amino acid position 1447. The PES was initially discovered in numerous tissues such as the liver, prostate, uterus, kidney, colon, normal and tumourous lung, glioma brain tumour, cerebellum, and frontal lobe.

When evaluating the PES, in the absence of corresponding DNA information, it was noted that two of the PES selected for validation in the laboratory namely PES_*MALAT1*_ncRNA_exonic and PES_BTBD3_3'UTR showed potential editing in one sample but not the other two, making them interesting candidates for further study. However, when including the sequencing information from the corresponding DNA samples, it showed that these two samples were heterozygous for the samples displaying potential RNA A-to-I editing, therefore, these PES were disregarded.

The three PES found to be false positives using PCR and Sanger Sequencing all resided in the 3'UTR region of their corresponding genes, and were supported by less than 10 reads in the genomic DNA. This suggests that sites must be supported by more than 10 reads to be accurately called. Therefore, those PES with poorly supported sequencing data from the exome sequencing should be viewed with caution and require additional verification.

5.3 *MEG3* non-coding RNA Editing site

. Sanger sequence verification of the DNA and RNA from all samples at the *MEG3* non-coding exonic PES suggested that all regions of the brain are edited with the exception of the cerebellum. All samples were found to be A/A homozygous at the PES site in the DNA, but RNA sequencing indicated the presence of G alleles as well as A alleles, suggesting RNA editing was occurring.

In the cerebellum, minimal observation of the G allele was seen, suggesting that either no editing was occurring or that it was occurring at a much lower rate than in the other regions. However, Sanger sequencing is not quantitative. Contrary to these findings, PES_*MEG3*_ncRNA_exonic is a known RNA A-to-I editing site, identified in DARNED and was discovered in the cerebellum by Li *et al* (Li *et al*., 2009).

PES_*MEG3*_ncRNA_exonic resides within a non-repetitive element in *MEG3,* a maternally expressed imprinted gene which acts as a tumour suppressor. MEG3 also interacts with the tumour suppressor p53, and regulates p53 target gene expression (O'Leary *et al*., 2016). The downregulation of *MEG3* gene expression promotes tumour cell proliferation and angiogenesis in vitro (Peng *et al*., 2015). The downregulation of *MEG3* expression is caused by multiple mechanisms such as gene deletion, promoter hypermethylation, and hypermethylation of the inte rgenic differentially methylated region. MEG3 produces 16 alternatively spliced transcripts, all of which are long non-coding RNAs. PES_*MEG3*-_ncRNA_exonic is only located in transcript 16, which contains an alternate and large 3' exon, producing a truncated lncRNA. In the other 15 transcripts, the PES would be present within an intron. This implies that transcript 16 has a differing function to the others.

Long non-coding RNAs (lncRNAs) can act as enhancers, influencing gene regulation over several Mb with their expression being correlated to that of associated genes (Zhuang *et al*., 2015). LncRNA *MEG3* is intergenic and found to inhibit cell proliferation, migration and invasion (Peng *et al*., 2015).

RNA editing has the potential to affect the nucleotide sequence of the lncRNA, changing the amino acid composition and impacting the secondary structure of the lncRNA. Modifying the dsRNA substrate by altering base pairing can lead to a change in structural stability. The stabilisation of the dsRNA substrate should not negatively impact the gene expression and the functionality of the gene remains unaltered. However, the destabilisation of the dsRNA, can lead to many changes in the expression levels or result in the creation of new variants, with potentially devastating consequences. The disruption or ablation of function of *MEG3* can lead to the uncontrolled proliferation of tumour cells, as well as increased angiogenesis (He *et al*., 2017).

5.4 Evaluation of Bioinformatics Pipeline

The bioinformatics pipeline consisted of 21 steps to identify potential RNA A-to-I editing sites within the RNA-Seq datasets and followed that of the previously mentioned pipeline by Altmann et al (2012). The protocol was used as a guide, with the new pipeline including additional steps as prerequisites to tools mentioned by Altmann et al., or for data formatting purposes. Although available, the pipeline created did not utilise exonic DNA sequencing, instead it compared RNA-Seq data to the human reference genome hg19 and relied on the removal of known polymorphisms by rs identifiers as well as realignment around indels using dbSNP build 138.

FastQC, used to check the quality of the data, was recommended for use as a prerequisite to analysis by Altmann *et al*., whose paper provides the outline for this pipeline (Altmann *et al*., 2012). The FastQC tool was successful at assessing the data quality and determined that it did not need to undergo trimming. Further tools which worked well in the pipeline include Map with BWA for Illumina, RmDup and Unified Genotyper.

The Map with BWA for Illumina tool aligns short sequences to a large sequence database such as the human reference genome hg19, using the Burrows-Wheeler transform. This tool was suggested in the paper by

Altmann *et al*., alongside Bowtie, MAQ and Stampy. If possible a hashing algorithm based aligner such as Stampy would have been used due to a higher sensitivity (superior indel detection and mapping quality calibration (Lunter and Goodson, 2011)), however, these are not available within Galaxy. Of the remaining suggested tools for this alignment step, BWA for Illumina seemed the most suitable as it works with Illumina encoded reads, accounts for small indels as it is a gapped aligner and has a relatively high processing speed. The Collect Alignment Summary Metrics tool showed that 93.39% reads were mapped to the reference genome.

To prevent the introduction of any artefacts, PCR duplicates were removed using the “SamTools” RmDup tool. Using PCR for amplification and the addition of adapters will introduce reads with the same insert length and position, which can cause problems in downstream analysis. It is not considered detrimental to remove duplicates in SNP/PES calling pipelines, whereas it would be for differential expression analysis. The RmDup tool was successful, decreasing the number of reads within the dataset from an average of 12,700,297 to 9,337,873. The necessity of removing PCR duplicates is debated, and it may be more useful to merely mark them.

The Unified Genotyper tool (GATK toolkit) was chosen to call SNPs and indels using dbSNP138 reference ordered data as indicated by the Altmann protocol. From the 8,597,698 reads present after filtering for properly mapped reads, 68,410 variants were called.

One potential issue of the bioinformatics identification of the PES is that a reference DNA sequence is used, which although comprehensive could be inconsistent with RNA-sequencing reads. This is due to differences in polymorphisms. However as this is a common practice in previous studies, it was deemed appropriate for use.

When compared to an RNA-Seq variant calling pipeline by Ramaswami *et al*., (Ramaswami *et al*., 2013), the same mapping tool (BWA) was used to align reads to the reference genome despite debate over its suitability for this purpose. Contrary to removing duplicates Ramaswami *et al*, simply marked them using Picard Tools, which in hindsight may reduce the possibility of removing information on differential gene expression. The pipeline outlined by Ramaswami *et al*., also makes use of GATK tools IndelRealigner and TableRecalibration to perform local realignment around indels and to recalibrate base quality scores. As per the pipeline outlined above, UnifiedGenotyper was used by Ramaswami *et al*., to call variants.

The pipeline outlined by Ramaswami *et al*. (2013), does not make use of matched genomic DNA sequences for the same individual as the RNA-Seq data, to call RNA A-to-I editing sites. This suggests it is possible to call RNA A-to-I editing sites accurately without accompanying gDNA sequence. This protocol instead uses a variety of filters to dispose of false positives such as filtering all intronic sites that are within 4bp of splice junctions, not considering variation if it occurs only at read ends, discarding non-*Alu* sites in repetitive regions and only reporting sites in regions for which no other highly similar genomic location exists.

Given that the pipeline created identified known RNA A-to-I editing sites present in the databases RADAR and DARNED, and lead to the verification of PES_*MEG3*_ncRNA_exonic in the laboratory, the pipeline described above is deemed successful. Although, there is scope for improvement and it would benefit from the editing of basic parameters to those more specific to the data used, to hopefully reduce the number of false positives called.

5.5 Ranking of PES for validation

After a selection of PES were chosen from the final output by editing pattern shown, the RNA-Seq data was compared to exome sequencing for these sites. This allowed the removal of a single PES immediately as it showed heterozygote gDNA. PES supported by exome sequencing data in all three samples were separated into one table (Table 9A). Sites that were partially or completely missing exome sequencing data were split into two further tables, with those supported by a low number of reads covering the DNA at the PES (<10) marked with an asterisk (Table 9B and 9C respectively). Each table was then ranked by the average number of supporting reads in the raw bam file calculated across all 15 samples. Four PES that were supported with more than 100 reads suggesting a high level of gene expression, and therefore more likely to be picked up by PCR and Sanger sequencing were selected to verify.

Two of the PES chosen for follow up in the laboratory, were supported by less than 10 reads in the gDNA, and it was discovered via Sanger sequencing that these calls were in fact incorrect and the gDNA was heterozygous. This meant no RNA A-to-I editing had taken place as a heterozygote would be expected in the cDNA and implies that 10 reads is insufficient to reliably call the genotype of a site. An alternative to this would have been to construct several ranking tables, one for each type of editing

pattern and then order the suspected PES by the number of supporting reads.

When ranking PES by the number of supporting reads in the raw RNA-Seq file, it is possible that as this value has not undergone mapping to the reference sequence or recalibration that it gives a false representation of expression at each site. Therefore, when ranking the PES for validation in the laboratory, it is advised to check the RPKM for each gene to ensure they are highly expressed. It may also be useful to check the position of the PES in the gene, as any transcript specific editing sites may be difficult to target via PCR. This would hopefully ensure the PES chosen for verification are more likely to be detected via PCR and Sanger sequencing.

Region-specific editing shows editing particular to one brain region across all three samples, whereas sample-specific editing shows editing reserved to one sample only. Those editing sites marked as a sequencing error only occurred in one brain region in one sample, hence making it highly unlikely that they are real editing sites. Yet this type of editing was necessary for testing as a control and would be required in future studies. Sporadic editing is classed as editing that occurs across brain regions in multiple samples with no clear editing pattern whilst universal editing occurs across all brain regions in all samples. Of the 339 PES called in the combined output of the pipeline, only 11 were called across all 15 samples. In comparison to this, 150 were only called once, and hence likely to be a sequencing error. Therefore, it may be wise to consider how many tissues the PES is expressed across before testing it further. The ranking table revealed that whilst the editing sites chosen for validation in the laboratory were suitable for further testing, the number of supporting reads in the genotype should also be taken into account to reduce the number of false positive discoveries.

5.6 Future work

With more time, I would verify further PES from the data generated here, in order to estimate the number of false positive PES from using a reference sequence in the data. Secondly in this project the PES verified in the laboratory were selected due to the read coverage for the PES having over 100 reads to support the finding. Therefore, another follow-up option would be to investigate how low the threshold for this cut-off can go, by testing PES with much lower coverage and still being able to get good PCR

and Sanger sequencing data to confirm or refute the presence of an editing site.

To enable maximum detection and verification of potential editing sites, changes need to be made to the bioinformatics pipeline. Given further time to work on the bioinformatics pipeline, the parameters at each step would be tailored specifically to each process rather than using basic settings. This would allow the accuracy of the pipeline to be improved and reduce the number of false positives called. It is not necessary to import the 1000G vcf, HapMap 3.3, Mills and 1000G gold standard datasets if they are not used further down the pipeline as reference ordered data. DbSNP is the most complete database of genetic variants and is hence sufficient for use as ROD on its own. The addition of a contaminant list would enable the removal of any known primers or repeated sequences, however this information would need to be acquired from DeepSeq.

In the aforementioned pipeline Map with BWA for Illumina was used due to its suitability for Illumina reads, and superiority in the number of reads mapped when compared to other tools under the NGS: Mapping toolkit. However, when reviewing the literature to call editing sites from RNA-Seq data, it was noted that TopHat was more commonly used over BWA as it is a splice-aware aligner whereas BWA is not. This ensures that transcripts are mapped over exon-intron boundaries, so in hindsight this would be a more suitable mapper for RNA-Seq data. Also, a “CleanSam” step should be implemented before the “SortSam” step in the pipeline to allocate “MAPQ=0” to those reads that are unmapped.

The combined final output of the bioinformatics pipeline would be revisited and searched for further PES for validation in the laboratory. RNA-Seq data could be compared to the exome sequence data as opposed to the hg19 reference genome. This would allow the discovery of RNA-DNA differences, and the exclusion of sites with heterozygous DNA. PES would also be checked for the number of supporting reads in the raw bam file, ensuring that the value is over 100 reads. This would hopefully decrease the likelihood of the editing site being a false positive. To screen out false positives without the use of exome sequencing data, specific filtering steps can be used to remove errors at the 5′ read ends due to random-hexamer priming, those sites within 4bp of splice junctions and those within simple repeats to name but a few (Ramaswami *et al*., 2012).

Further to this, the combined final output could be explored for known RNA A-to-I editing sites as listed in DARNED. These known RNA A-to-I editing sites would be tested in the laboratory as positive controls. If few known editing sites were discovered in the final table, this could be due to a difference in the type of tissue tested, or if the tissue used to verify the known site was a control or diseased.

To supplement the findings on *MEG3*, the functionality of the different RNA transcripts produced could be investigated, along with the action of the PES on lncRNA function. The difference in results obtained by Li et al. (2009) and this study; namely the finding of MEG3 in the cerebellum by Li et al, could be explored in further detail. This could be done by examining the methodology used for Li's study and properties such as tissue source.

The functional changes brought about by RNA A-to-I editing are known to be contributing factors to the development of a wide range of diseases, particularly those residing in the nervous system and brain (Maas et al., 2006). Further to identifying novel RNA A-to-I editing sites, the PES could be investigated to see if they reside near candidate genes for related disorders such as amyotrophic lateral sclerosis, schizophrenia, and epilepsy. It would also be interesting to investigate any overlap between genes implicated in other psychiatric and brain disorders. In the final table of A-to-I editing sites called from the RNA-Seq data, all sites were explored to see if they resided in any of the susceptibility loci for Alzheimer's disease (AD) listed in (Lambert *et al*., 2013). AD was focused on in this instance as it a well characterised polygenic disease with known GWAS hits, previously proven to show RNA A-to-I editing (Khermesh *et al*., 2016).

Only two potential editing sites were found to be located within the AD candidate gene, *SORL1*, at positions 121358739 and 121415950 in the splicing region. However, as these sites resided in an intron close to a splice junction of *SORL1* they were excluded from the final laboratory validation process. These two sites would not be investigated in the laboratory if further work were to be undertaken as PES located close to splicing junctions are more prone to mismapping artefacts and are hence usually discarded as false positives (Park *et al*., 2012). However, it may be useful to explore intronic or intergenic sites that were removed from the final analysis to see whether they should in fact be included.

Given the opportunity to investigate further editing sites, a wider range of editing patterns would be chosen in addition to those mentioned in the ranking of PES for validation section. This would allow the verification of different types of RNA A-to-I editing. It may also be worth testing a greater number of ncRNA_exonic PES to identify whether these are easier to target than 3'UTR PES.

6 Acknowledgments

I would like to thank my supervisors Professor Kevin Morgan and Dr Keeley Brookes for all their guidance and support throughout my Masters year, as well as Dr James Turton and Dr Tamar Guetta-Baranes for all their hard work. I would like to thank my family and friends for their support and patience. Finally, I would like to thank and dedicate my work to the late Dr Sally Chappell, for believing in me and this project; I hope I have done her proud.

7 References

Altmann, A., Weber, P., Bader, D., Preuss, M., Binder, E. B. and Müller-Myhsok, B. (2012) 'A beginners guide to SNP calling from high-throughput DNA-sequencing data', *Human Genetics,* 131(10), pp. 1541-54.

Athanasiadis, A., Rich, A. and Maas, S. (2004) 'Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome', *PLoS Biology,* 2(12), pp. e391.

Bahn, J. H., Lee, J. H., Li, G., Greer, C., Peng, G. and Xiao, X. (2012) 'Accurate identification of A-to-I RNA editing in human by transcriptome sequencing', *Genome Research,* 22(1), pp. 142-50.

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P. and Marth, G. T. (2011) 'BamTools: a C++ API and toolkit for analyzing and managing BAM files', *Bioinformatics,* 27(12), pp. 1691-1692.

Bass, B. L. (1997) 'RNA editing and hypermutation by adenosine deamination', *Trends in Biochemical Sciences,* 22(5), pp. 157-162.

Bass, B. L. (2002) 'RNA editing by adenosine deaminases that act on RNA', *Annual Review of Biochemistry,* 71, pp. 817-846.

Bazak, L., Haviv, A., Barak, M., Jacob-Hirsch, J., Deng, P. and Zhang, R. (2014a) 'A-to-I RNA editing occurs at over a hundred million

genomic sites, located in a majority of human genes', *Genome Research.,* 24.

Bazak, L., Levanon, E. Y. and Eisenberg, E. (2014b) 'Genome-wide analysis of Alu editability', *Nucleic Acids Research.,* 42(11), pp. 6876-84.

Benne, R., Van Den Burg, J., Brakenhoff, J. P. J., Sloof, P., Van Boom, J. H. and Tromp, M. C. (1986) 'Major transcript of the frameshifted coxll gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA', *Cell,* 46(6), pp. 819-826.

Berg, K. A., Cropper, J. D., Niswender, C. M., Sanders-Bush, E., Emeson, R. B. and Clarke, W. P. (2001) 'RNA-editing of the 5-HT(2C) receptor alters agonist-receptor-effector coupling specificity', *British Journal of Pharmacology,* 134(2), pp. 386-92.

Bhalla, T., Rosenthal, J. J., Holmgren, M. and Reenan, R. (2004) 'Control of human potassium channel inactivation by editing of a small mRNA hairpin', *Nature Structural and Molecular Biology,* 11(10), pp. 950-6.

Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J. and Nekrutenko, A. (2010) 'Manipulation of FASTQ data with Galaxy', *Bioinformatics,* 26(14), pp. 1783-5.

Blow, M., Futreal, P. A., Wooster, R. and Stratton, M. R. (2004) 'A survey of RNA editing in human brain', *Genome Research,* 14(12), pp. 2379-87.

Burrows, M. and Wheeler, D. J. (1994) *A block-sorting lossless data compression algorithm.* (SRC-RR-124).

Bycroft, M., Grünert, S., Murzin, A. G., Proctor, M. and St Johnston, D. (1995) 'NMR solution structure of a dsRNA binding domain from *Drosophila* Staufen protein reveals homology to the N-terminal domain of ribosomal protein S5', *The EMBO Journal,* 14(14), pp. 3563-71.

Carmi, S., Borukhov, I. and Levanon, E. Y. (2011) 'Identification of widespread ultra-edited human RNAs', *PLoS Genetics,* 7(10), pp. e1002317.

Casey, J. L., Tennant, B. C. and Gerin, J. L. (2006) 'Genetic changes in hepatitis delta virus from acutely and chronically infected woodchucks', *Journal of Virology,* 80(13), pp. 6469-77.

Cenci, C., Barzotti, R., Galeano, F., Corbelli, S., Rota, R., Massimi, L., Di Rocco, C., O'Connell, M. A. and Gallo, A. (2008) 'Down-regulation

of RNA editing in pediatric astrocytomas: ADAR2 editing activity inhibits cell migration and proliferation', *The Journal of Biological Chemistry,* 283(11), pp. 7251-60.

Chen, C. X., Cho, D. S., Wang, Q., Lai, F., Carter, K. C. and Nishikura, K. (2000) 'A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains', *RNA,* 6(5), pp. 755-67.

Chen, L. L., DeCerbo, J. N. and Carmichael, G. G. (2008) 'Alu element-mediated gene silencing', *The EMBO Journal,* 27(12), pp. 1694.

Cho, D. S., Yang, W., Lee, J. T., Shiekhattar, R., Murray, J. M. and Nishikura, K. (2003) 'Requirement of dimerization for RNA editing activity of adenosine deaminases acting on RNA', *The Journal of Biological Chemistry,* 278(19), pp. 17093-102.

Dagan, T., Sorek, R., Sharon, E., Ast, G. and Graur, D. (2004) 'AluGene: a database of Alu elements incorporated within protein-coding genes', *Nucleic Acids Research,* 32, pp. D489-D492.

'Database Resources of the National Center for Biotechnology Information', (2017) *Nucleic Acids Research,* 45(D1), pp. D12-D17.

Deininger, P. (2011) 'Alu elements: know the SINEs', *Genome Biology,* 12(12), pp. 236.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D. and Daly, M. J. (2011) 'A framework for variation discovery and genotyping using next-generation DNA sequencing data', *Nature Genetics,* 43(5), pp. 491-8.

Farajollahi, S. and Maas, S. (2010) 'Molecular diversity through RNA editing: a balancing act', *Trends in Genetics,* 26(5), pp. 221-30.

Feldmeyer, D., Kask, K., Brusa, R., Kornau, H. C., Kolhekar, R., Rozov, A., Burnashev, N., Jensen, V., Hvalby, O., Sprengel, R. and Seeburg, P. H. (1999) 'Neurological dysfunctions in mice expressing different levels of the Q/R site-unedited AMPAR subunit GluR-B', *Nature Neuroscience,* 2(1), pp. 57-64.

Goecks, J., Nekrutenko, A. and Taylor, J. (2010) 'Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent

computational research in the life sciences', *Genome Biology,* 11(8), pp. R86.

Gott, J. M. and Emeson, R. B. (2000) 'Functions and mechanisms of RNA editing', *Annual Review of Genetics,* 34, pp. 499-531.

Gray, M. W. (2012) 'Evolutionary origin of RNA editing', *Biochemistry,* 51(26), pp. 5235-42.

Gurevich, I., Tamir, H., Arango, V., Dwork, A. J., Mann, J. J. and Schmauss, C. (2002) 'Altered editing of serotonin 2C receptor pre-mRNA in the prefrontal cortex of depressed suicide victims', *Neuron,* 34(3), pp. 349-56.

Hansen, K. D., Brenner, S. E. and Dudoit, S. (2010) 'Biases in Illumina transcriptome sequencing caused by random hexamer priming', *Nucleic Acids Research,* 38(12), pp. e131.

Hartner, J. C., Schmittwolf, C., Kispert, A., Muller, A. M., Higuchi, M. and Seeburg, P. H. (2004) 'Liver disintegration in the mouse embryo caused by deficiency in the RNA-editing enzyme ADAR1', *The Journal of Biological Chemistry,* 279(6), pp. 4894-902.

He, C., Yang, W., Yang, J., Ding, J., Li, S., Wu, H., Zhou, F., Jiang, Y. and Teng, L. (2017) 'Long noncoding RNA MEG3 negatively regulates proliferation and angiogenesis in vascular endothelial cells', *DNA and Cell Biology,* 36(6), pp. 475-481.

Heale, B. S. E., Keegan, L. P., McGurk, L., Michlewski, G., Brindle, J., Stanton, C. M., Caceres, J. F. and Connell, M. A. (2009) 'Editing independent effects of ADARs on the miRNA/siRNA pathways', *The EMBO Journal,* 28(20), pp. 3145.

Herbert, A. and Rich, A. (2001) 'The role of binding domains for dsRNA and Z-DNA in the in vivo editing of minimal substrates by ADAR1', *Proceedings of the National Academy of Sciences of the United States of America,* 98(21), pp. 12132-7.

Higuchi, M., Maas, S., Single, F. N., Hartner, J., Rozov, A., Burnashev, N., Feldmeyer, D., Sprengel, R. and Seeburg, P. H. (2000) 'Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2', *Nature,* 406(6791), pp. 78-81.

Holley, R. W., Everett, G. A., Madison, J. T. and Zamir, A. (1965) 'Nucleotide sequences in the yeast alanine transfer ribonucleic acid.', *The Journal of Biological Chemistry,* 240, pp. 2122-8.

Hoopengardner, B., Bhalla, T., Staber, C. and Reenan, R. (2003) 'Nervous system targets of RNA editing identified by comparative genomics', *Science,* 301(5634), pp. 832-6.

Huang, W. H., Tseng, C. N., Tang, J. Y., Yang, C. H., Liang, S. S. and Chang, H. W. (2013) 'RNA editing and drug discovery for cancer therapy', *Scientific World Journal,* 2013, pp. 804505.

Iwamoto, K., Bundo, M. and Kato, T. (2009) 'Serotonin receptor 2C and mental disorders: genetic, expression and RNA editing studies', *RNA Biology,* 6(3), pp. 248-53.

Iwamoto, K., Bundo, M., Yamada, K., Takao, H., Iwayama-Shigeno, Y., Yoshikawa, T. and Kato, T. (2005) 'DNA Methylation Status of SOX10 correlates with its downregulation and oligodendrocyte dysfunction in Schizophrenia', *The Journal of Neuroscience,* 25(22), pp. 5376-5381.

Iwamoto, K. and Kato, T. (2003) 'RNA editing of serotonin 2C receptor in human postmortem brains of major mental disorders', *Neuroscience Letters,* 346(3), pp. 169-72.

Kharrat, A., Macias, M. J., Gibson, T. J., Nilges, M. and Pastore, A. (1995) 'Structure of the dsRNA binding domain of *E. coli* RNase III', *The EMBO Journal,* 14(14), pp. 3572-84.

Khermesh, K., D'Erchia, A. M., Barak, M., Annese, A., Wachtel, C., Levanon, E. Y., Picardi, E. and Eisenberg, E. (2016) 'Reduced levels of protein recoding by A-to-I RNA editing in Alzheimer's disease', *RNA,* 22(2), pp. 290-302.

Kim, D. D., Kim, T. T., Walsh, T., Kobayashi, Y., Matise, T. C., Buyske, S. and Gabriel, A. (2004) 'Widespread RNA editing of embedded Alu elements in the human transcriptome', *Genome Research,* 14(9), pp. 1719-25.

Kiran, A. and Baranov, P. V. (2010) 'DARNED: a DAtabase of RNa EDiting in humans', *Bioinformatics,* 26(14), pp. 1772-6.

Kortenbruck, G., Berger, E., Speckmann, E. J. and Musshoff, U. (2001) 'RNA editing at the Q/R site for the glutamate receptor subunits GLUR2, GLUR5, and GLUR6 in hippocampus and temporal cortex from epileptic patients', *Neurobiology of Disease,* 8(3), pp. 459-68.

Kwak, S. and Kawahara, Y. (2005) 'Deficient RNA editing of GluR2 and neuronal death in amyotropic lateral sclerosis', *Journal of Molecular Medicine (Berlin),* 83(2), pp. 110-20.

Lambert, J.-C. and Ibrahim-Verbaas, C. A. and Harold, D. and Naj, A. C. and Sims, R. and Bellenguez, C. and Jun, G. and DeStefano, A. L. and Bis, J. C. and Beecham, G. W. and Grenier-Boley, B. and Russo, G. and Thornton-Wells, T. A. and Jones, N. and Smith, A. V. and Chouraki, V. and Thomas, C. and Ikram, M. A. and Zelenika, D. and Vardarajan, B. N. and Kamatani, Y. and Lin, C.-F. and Gerrish, A. and Schmidt, H. and Kunkle, B. and Dunstan, M. L. and Ruiz, A. and Bihoreau, M.-T. and Choi, S.-H. and Reitz, C. and Pasquier, F. and Hollingworth, P. and Ramirez, A. and Hanon, O. and Fitzpatrick, A. L. and Buxbaum, J. D. and Campion, D. and Crane, P. K. and Baldwin, C. and Becker, T. and Gudnason, V. and Cruchaga, C. and Craig, D. and Amin, N. and Berr, C. and Lopez, O. L. and De Jager, P. L. and Deramecourt, V. and Johnston, J. A. and Evans, D. and Lovestone, S. and Letenneur, L. and Morón, F. J. and Rubinsztein, D. C. and Eiriksdottir, G. and Sleegers, K. and Goate, A. M. and Fiévet, N. and Huentelman, M. J. and Gill, M. and Brown, K. and Kamboh, M. I. and Keller, L. and Barberger-Gateau, P. and McGuinness, B. and Larson, E. B. and Green, R. and Myers, A. J. and Dufouil, C. and Todd, S. and Wallon, D. and Love, S. and Rogaeva, E. and Gallacher, J. and St George-Hyslop, P. and Clarimon, J. and Lleo, A. and Bayer, A. and Tsuang, D. W. and Yu, L. and Tsolaki, M. and Bossù, P. and Spalletta, G. and Proitsi, P. and Collinge, J. and Sorbi, S. and Sanchez-Garcia, F. and Fox, N. C. and Hardy, J. and Deniz Naranjo, M. C. and Bosco, P. and Clarke, R. and Brayne, C. and Galimberti, D. and Mancuso, M. and Matthews, F. and European Alzheimer's disease, I. and Genetic and Environmental Risk in Alzheimer's, D. and Alzheimer's Disease Genetic, C. and Cohorts for Heart and Aging Research in Genomic, E. and Moebus, S. and Mecocci, P. and Zompo, M. D. and Maier, W. and Hampel, H. and Pilotto, A. and Bullido, M. and Panza, F. and Caffarra, P. and Nacmias, B. and Gilbert, J. R. and Mayhaus, M. and Lannfelt, L. and Hakonarson, H. and Pichler, S. and Carrasquillo, M. M. and Ingelsson, M. and Beekly, D. and Alvarez, V. and Zou, F. and Valladares, O. and Younkin, S. G. and Coto, E. and Hamilton-Nelson, K. L. and Gu, W. and Razquin, C. and Pastor, P. and Mateo, I. and Owen, M. J. and Faber, K. M. and Jonsson, P. V. and Combarros, O. and O'Donovan, M. C. and Cantwell, L. B.

and Soininen, H. and Blacker, D. and Mead, S. and Mosley, T. H. and Bennett, D. A. and Harris, T. B. and Fratiglioni, L. and Holmes, C. and de Bruijn, R. F. A. G. and Passmore, P. and Montine, T. J. and Bettens, K. and Rotter, J. I. and Brice, A. and Morgan, K. and Foroud, T. M. and Kukull, W. A. and Hannequin, D. and Powell, J. F. and Nalls, M. A. and Ritchie, K. and Lunetta, K. L. and Kauwe, J. S. K. and Boerwinkle, E. and Riemenschneider, M. and Boada, M. and Hiltunen, M. and Martin, E. R. and Schmidt, R. and Rujescu, D. and Wang, L.-s. and Dartigues, J.-F. and Mayeux, R. and Tzourio, C. and Hofman, A. and Nöthen, M. M. and Graff, C. and Psaty, B. M. and Jones, L. and Haines, J. L. and Holmans, P. A. and Lathrop, M. and Pericak-Vance, M. A. and Launer, L. J. and Farrer, L. A. and van Duijn, C. M. and Van Broeckhoven, C. and Moskvina, V. and Seshadri, S. and Williams, J. and Schellenberg, G. D. and Amouyel, P. (2013) 'Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease', *Nature Genetics,* 45(12), pp. 1452-1458.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M. and Fitz Hugh, W. e. a. (2001) 'Initial sequencing and analysis of the human genome', *Nature,* 409(6822), pp. 860-921.

Lehmann, K. A. and Bass, B. L. (1999) 'The importance of internal loops within RNA substrates of ADAR111', *Journal of Molecular Biology,* 291(1), pp. 1-13.

Lemmens, R., Moore, M. J., Al-Chalabi, A., Brown, R. H. and Robberecht, W. (2010) 'RNA metabolism and the pathogenesis of motor neuron diseases', *Trends in Neurosciences,* 33(5), pp. 249-258.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics,* 25(14), pp. 1754-60.

Li, J. B. and Church, G. M. (2013) 'Deciphering the functions and regulation of brain-enriched A-to-I RNA editing', *Nature Neuroscience,* 16(11), pp. 1518-1522.

Li, J. B., Levanon, E. Y., Yoon, J. K., Aach, J., Xie, B., Leproust, E., Zhang, K., Gao, Y. and Church, G. M. (2009) 'Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing', *Science,* 324(5931), pp. 1210-3.

Lunter, G. and Goodson, M. (2011) 'Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads', *Genome Research,* 21(6), pp. 936-9.

Maas, S., Kawahara, Y., Tamburro, K. M. and Nishikura, K. (2006) 'A-to-I RNA editing and human disease', *RNA Biology,* 3(1), pp. 1-9.

Maas, S., Rich, A. and Nishikura, K. (2003) 'A-to-I RNA editing: recent news and residual mysteries', *The Journal of Biological Chemistry,* 278(3), pp. 1391-4.

Macbeth, M. R., Schubert, H. L., Vandemark, A. P., Lingam, A. T., Hill, C. P. and Bass, B. L. (2005) 'Inositol hexakisphosphate is bound in the ADAR2 core and required for RNA editing', *Science,* 309(5740), pp. 1534-9.

Mannion, N., Arieti, F., Gallo, A., Keegan, L. P. and O'Connell, M. A. (2015) 'New insights into the biological role of mammalian ADARs; the RNA editing proteins', in Gerber, A.P. (ed.) *Biomolecules: Vol. 4*, pp. 2338-62.

Meier, J. C., Kankowski, S., Krestel, H. and Hetsch, F. (2016) 'RNA editing — systemic relevance and clue to disease mechanisms?', *Frontiers in Molecular Neuroscience,* 9, pp. 124.

Miga, K. H., Eisenhart, C. and Kent, W. J. (2015) 'Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments', *Nucleic Acids Research,* 43(20), pp. e133-e133.

Moreira, S., Valach, M., Aoulad-Aissa, M., Otto, C. and Burger, G. (2016) 'Novel modes of RNA editing in mitochondria', *Nucleic Acids Research,* 44(10), pp. 4907-4919.

Nanduri, S., Carpick, B. W., Yang, Y., Williams, B. R. and Qin, J. (1998) 'Structure of the double-stranded RNA-binding domain of the protein kinase PKR reveals the molecular basis of its dsRNA-mediated activation', *The EMBO Journal,* 17(18), pp. 5458-65.

Nigita, G., Alaimo, S., Ferro, A., Giugno, R. and Pulvirenti, A. (2015) 'Knowledge in the investigation of A-to-I RNA editing signals', *Frontiers in Bioengineering and Biotechnology,* 3.

Nishikura, K. (2010) 'Functions and regulation of RNA editing by ADAR deaminases', in Kornberg, R.D., Raetz, C.R.H., Rothman, J.E. & Thorner, J.W. (eds.) *Annual Review of Biochemistry, Vol 79 Annual Review of Biochemistry*. Palo Alto: Annual Reviews, pp. 321-349.

Nishikura, K. (2016) 'A-to-I editing of coding and non-coding RNAs by ADARs', *Nature Reviews Molecular Cell Biology,* 17(2), pp. 83-96.

Nishikura, K., Yoo, C., Kim, U., Murray, J. M., Estes, P. A., Cash, F. E. and Liebhaber, S. A. (1991) 'Substrate specificity of the dsRNA unwinding/modifying activity', *The EMBO Journal,* 10(11), pp. 3523-32.

Niswender, C. M., Copeland, S. C., Herrick-Davis, K., Emeson, R. B. and Sanders-Bush, E. (1999) 'RNA editing of the human serotonin 5-hydroxytryptamine 2C receptor silences constitutive activity', *Journal of Biological Chemistry,* 274(14), pp. 9472-9478.

Niswender, C. M., Herrick-Davis, K., Dilley, G. E., Meltzer, H. Y., Overholser, J. C., Stockmeier, C. A., Emeson, R. B. and Sanders-Bush, E. (2001) 'RNA editing of the human serotonin 5-HT2C receptor, alterations in suicide and implications for serotonergic pharmacotherapy', *Neuropsychopharmacology,* 24(5), pp. 478-91.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B. and Ako-Adjei, D. (2016) 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic Acids Research,* 44(D1), pp. D733-45.

Palladino, M. J., Keegan, L. P., O'Connell, M. A. and Reenan, R. A. (2000) 'A-to-I pre-mRNA editing in *Drosophila* is primarily involved in adult nervous system function and integrity', *Cell,* 102(4), pp. 437-49.

Park, E., Williams, B., Wold, B. J. and Mortazavi, A. (2012) 'RNA editing in the human ENCODE RNA-seq data', *Genome Research,* 22(9), pp. 1626-1633.

Patterson, J. B. and Samuel, C. E. (1995) 'Expression and regulation by interferon of a double-stranded-RNA-specific adenosine deaminase from human cells: evidence for two forms of the deaminase', *Molecular and Cellular Biology,* 15(10), pp. 5376-88.

Peng, W., Si, S., Zhang, Q., Li, C., Zhao, F., Wang, F., Yu, J. and Ma, R. (2015) 'Long non-coding RNA MEG3 functions as a competing endogenous RNA to regulate gastric cancer progression', *Journal of Experimental & Clinical Cancer Research : CR,* 34(1), pp. 79.

Peng, Z., Cheng, Y., Tan, B. C., Kang, L., Tian, Z., Zhu, Y., Zhang, W., Liang, Y., Hu, X., Tan, X., Guo, J., Dong, Z., Bao, L. and Wang, J. (2012) 'Comprehensive analysis of RNA-Seq data reveals extensive

RNA editing in a human transcriptome', *Nature Biotechnology,* 30(3), pp. 253-60.

Picardi, E., Manzari, C., Mastropasqua, F., Aiello, I., D'Erchia, A. M. and Pesole, G. (2015) 'Profiling RNA editing in human tissues: towards the inosinome atlas', *Scientific Reports,* 5, pp. 14941.

Polson, A. G. and Bass, B. L. (1994) 'Preferential selection of adenosines for modification by double-stranded RNA adenosine deaminase', *The EMBO Journal,* 13(23), pp. 5701-11.

Porath, H. T., Knisbacher, B. A., Eisenberg, E. and Levanon, E. Y. (2017) 'Massive A-to-I RNA editing is common across the metazoa and correlates with dsRNA abundance', *Genome Biology,* 18(1), pp. 185.

Ramaswami, G. and Li, J. B. (2014) 'RADAR: a rigorously annotated database of A-to-I RNA editing', *Nucleic Acids Research,* 42, pp. D109-D113.

Ramaswami, G., Lin, W., Piskol, R., Tan, M. H., Davis, C. and Li, J. B. (2012) 'Accurate identification of human Alu and non-Alu RNA editing sites', *Nature Methods,* 9(6), pp. 579-581.

Ramaswami, G., Zhang, R., Piskol, R., Keegan, L. P., Deng, P., O'Connell, M. A. and Li, J. B. (2013) 'Identifying RNA editing sites using RNA sequencing data alone', *Nature Methods,* 10(2), pp. 128-132.

Rebagliati, M. R. and Melton, D. A. (1987) 'Antisense RNA injections in fertilized frog eggs reveal an RNA duplex unwinding activity', *Cell,* 48(4), pp. 599-605.

Sakurai, M., Ueda, H., Yano, T., Okada, S., Terajima, H., Mitsuyama, T., Toyoda, A., Fujiyama, A., Kawabata, H. and Suzuki, T. (2014) 'A biochemical landscape of A-to-I RNA editing in the human brain transcriptome', *Genome Research,* 24(3), pp. 522-534.

Sakurai, M., Yano, T., Kawabata, H., Ueda, H. and Suzuki, T. (2010) 'Inosine cyanoethylation identifies A-to-I RNA editing sites in the human transcriptome', *Nature Chemical Biology,* 6(10), pp. 733-40.

Savva, Y. A., Rieder, L. E. and Reenan, R. A. (2012) 'The ADAR protein family', *Genome Biology,* 13(12), pp. 252.

Scadden, A. (2007) 'Inosine-containing dsRNA binds a stress-granule-like complex and downregulates gene expression in trans', *Molecular Cell: Vol. 3*, pp. 491-500.

Schellekens, H., Clarke, G., Jeffery, I. B., Dinan, T. G. and Cryan, J. F. (2012) 'Dynamic 5-HT2C receptor editing in a mouse model of obesity', *PLoS One,* 7(3), pp. e32266.

Schmauss, C. and Howe, J. R. (2002) 'RNA editing of neurotransmitter receptors in the mammalian brain', *Science's STKE,* 2002(133), pp. pe26.

Seeburg, P. H., Higuchi, M. and Sprengel, R. (1998) 'RNA editing of brain glutamate receptor channels: mechanism and physiology', *Brain Research/Brain Research Reviews,* 26(2-3), pp. 217-29.

Slotkin, W. and Nishikura, K. (2013) 'Adenosine-to-inosine RNA editing and human disease', *Genome Medicine,* 5(11), pp. 105.

Sommer, B., Kohler, M., Sprengel, R. and Seeburg, P. H. (1991) 'RNA editing in brain controls a determinant of ion flow in glutamate-gated channels', *Cell,* 67(1), pp. 11-9.

Tang, W., Fei, Y. and Page, M. (2012) 'Biological significance of RNA editing in cells', *Molecular Biotechnology,* 52(1), pp. 91-100.

Vollmar, W., Gloger, J., Berger, E., Kortenbruck, G., Kohling, R., Speckmann, E. J. and Musshoff, U. (2004) 'RNA editing (R/G site) and flip-flop splicing of the AMPA receptor subunit GluR2 in nervous tissue of epilepsy patients', *Neurobiology of Disease,* 15(2), pp. 371-9.

Wahlstedt, H. and Ohman, M. (2011) 'Site-selective versus promiscuous A-to-I editing', *Wiley Interdisciplinary Reviews RNA,* 2(6), pp. 761-71.

Wang, K., Li, M. and Hakonarson, H. (2010) 'ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data', *Nucleic Acids Research,* 38(16), pp. e164-e164.

Wang, Q., Khillan, J., Gadue, P. and Nishikura, K. (2000) 'Requirement of the RNA editing deaminase ADAR1 gene for embryonic erythropoiesis', *Science,* 290(5497), pp. 1765-8.

Wang, Q., Miyakoda, M., Yang, W., Khillan, J., Stachura, D. L., Weiss, M. J. and Nishikura, K. (2004) 'Stress-induced apoptosis associated with null mutation of ADAR1 RNA editing deaminase gene', *The Journal of Biological Chemistry,* 279(6), pp. 4952-61.

Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K. and Liu, X. (2014) 'Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells', in Zhang, S.D. (ed.) *PLoS One: Vol. 1*. San Francisco, USA.

Zhu, S., Xiang, J.-F., Chen, T., Chen, L.-L. and Yang, L. (2013) 'Prediction of constitutive A-to-I editing sites from human transcriptomes in the absence of genomic sequences', *BMC Genomics,* 14, pp. 206-206.

Zhuang, W., Ge, X., Yang, S., Huang, M., Zhuang, W., Chen, P., Zhang, X., Fu, J., Qu, J. and Li, B. (2015) 'Upregulation of lncRNA MEG3 promotes osteogenic differentiation of mesenchymal stem cells from multiple myeloma patients by targeting BMP4 transcription', *Stem Cells,* 33(6), pp. 1985-1997.

Zinshteyn, B. and Nishikura, K. (2009) 'Adenosine-to-inosine RNA editing', *Wiley Interdisciplinary Reviews Systems Biology and Medicine,* 1(2), pp. 202-9.