From THE DEPARTMENT OF NEUROSCIENCE
Karolinska Institutet, Stockholm, Sweden

Exploring the transcriptome of the brain

Allissa Dillman

Karolinska Institutet

Stockholm 2014

Published by Karolinska Institutet. Printed by US-AB.

ISBN 978-91-7549-600-9

Karolinska Institutet

Institutionen för Neurovetenskap

Exploring the transcriptome of the brain

THESIS FOR DOCTORAL DEGREE (Ph.D.)

AKADEMISK AVHANDLING
som för avläggande av medicine doktorsexamen vid Karolinska Institutet offentligen försvaras i föreläsningssal Lennart Nilsson, Nobelsväg 15A.

Måndag den 2 Juni, kl 09.00

av

Allissa Dillman
BS

Principal Supervisor:
Senior Investigator Mark Cookson
National Institutes of Health
National Institute on Aging
Laboratory of Neurogenetics
Cell Biology and Gene Expression Unit

Co-supervisor(s):
Docent Dagmar Galter
Karolinska Institutet
Department of Neuroscience

Opponent:
Professor Chris Ponting
University of Oxford
Department of Physiology, Anatomy, and Genetics

Examination Board:
Docent Rickard Sandberg
Karolinska Institutet
Department of Cell and Molecular Biology

Professor Brun Ulfhake
Karolinska Institutet
Department of Neuroscience

Assistant Professor
Uppsala Universitet
Department of Genetics and Pathology

Stockholm 2014

ABSTRACT

Our knowledge of the transcriptome has become much more complex since the days of the central dogma of molecular biology. We now know that splicing takes place to create potentially thousands of isoforms from a single gene, and we know that RNA does not always faithfully recapitulate DNA if RNA editing occurs. Collectively, these observations show that the transcriptome is amazingly rich with intricate regulatory mechanisms for overall gene expression, splicing, and RNA editing.

Genetic variability can play a role in controlling gene expression, which can be identified by examining expression quantitative trait loci (eQTLs). eQTLs are genomic regions where genetic variants, including single nucleotide polymorphisms (SNPs) show a statistical association with expression of mRNA transcripts. In humans, many SNPs are also associated with disease, and have been identified using genome wide association studies (GWAS) but the biological effects of those SNPs are usually not known. If SNPs found in GWAS are also found in eQTLs, then one could hypothesize that expression levels may contribute to disease risk. Performing eQTL analysis with GWAS SNPs in both blood and brain, specifically the frontal cortex and the cerebellum, we found both shared and tissue unique eQTLS. The identification of tissue-unique eQTLs supports the argument that choice of tissue type is important in eQTL studies (Paper I).

Aging is a complex process with the mechanisms underlying aging still being poorly defined. There is evidence that the transcriptome changes with age, and hence we used the brain dataset from our first paper as a discovery set, with an additional replication dataset, to investigate any aging-gene expression associations. We found evidence that many genes were associated with aging. We further found that there were more statically significant expression changes in the frontal cortex versus the cerebellum, indicating that brain regions may age at different rates. As the brain is a heterogeneous tissue including both neurons and non-neuronal cells, we used LCM to capture Purkinje cells as a representative neuronal type and repeated the age analysis. Looking at the discovery, replication and Purkinje cell datasets we found five genes with strong, replicated evidence of age-expression associations (Paper II).

Being able to capture and quantify the depth of the transcriptome has been a lengthy process starting with methods that could only measure a single gene to genome-wide techniques such as microarray. A recently developed technology, RNA-Seq, shows promise in its ability to capture expression, splicing, and editing and with its broad dynamic range quantification is accurate and reliable. RNA-Seq is, however, data intensive and a great deal of computational expertise is required to fully utilize the strengths of this method. We aimed to create a small, well-controlled, experiment in order to test the performance of this relatively new technology in the brain. We chose embryonic versus adult cerebral cortex, as mice are genetically homogenous and there are many known differences in gene expression related to brain development that we could use as benchmarks for analysis testing. We found a large number of differences in total gene expression between embryonic and adult brain. Rigorous technical and biological validation illustrated the accuracy and dynamic range of RNA-Seq. We

were also able to interrogate differences in exon usage in the same dataset. Finally we were able to identify and quantify both well-known and novel A-to-I edit sites. Overall this project helped us develop the tools needed to build usable pipelines for RNA-Seq data processing (Paper III).

Our studies in the developing brain (Paper III) illustrated that RNA-Seq was a useful unbiased method for investigating RNA editing. To extend this further, we utilized a genetically modified mouse model to study the transcriptomic role of the RNA editing enzyme ADAR2. We found that ADAR2 was important for editing of the coding region of mRNA as a large proportion of RNA editing sites in coding regions had a statistically significant decrease in editing percentages in $Adar2^{-/-}$ $Gria2^{R/R}$ mice versus controls. However, despite indications in the literature that ADAR2 may also be involved in splicing and expression regulatory machinery we found no changes in gene expression or exon utilization in $Adar2^{-/-}$ $Gria2^{R/R}$ mice as compared to their littermate controls (Paper IV).

In our final study, based on the methods developed in Papers III and IV, we revisited the idea of age related gene expression associations from Paper II. We used a subset of human frontal cortices for RNA sequencing. Interestingly we found more gene expression changes with aging compared to the previous data using microarrays in Paper II. When the significant gene lists were analysed for gene ontology enrichment, we found that there was a large number of downregulated genes involved in synaptic function while those that were upregulated had enrichment in immune function. This dataset illustrates that the aging brain may be predisposed to the processes found in neurodegenerative diseases (Paper V).

LIST OF PUBLICATIONS

I. Hernandez DG, Nalls MA, Moore M, Chong S, Dillman A, Trabzuni D, Gibbs JR, Ryten M, Arepalli S, Weale ME, Zonderman AB, Troncoso J, O'Brien R, Walker R, Smith C, Bandinelli S, Traynor BJ, Hardy J, Singleton AB, Cookson MR. (2012): Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiol Dis.* 47(1):20-8

II. Kumar A, Gibbs JR, Beilina A, Dillman A, Kumaran R, Trabzuni D, Ryten M, Walker R, Smith C, Traynor BJ, Hardy J, Singleton AB, Cookson MR. (2013). Age-associated changes in gene expression in human brain and isolated neurons. *Neurobiol Aging*. Apr;34(4):1199-209.

III. Dillman AA, Hauser DN, Gibbs JR, Nalls MA, McCoy MK, Rudenko IN, Galter D, Cookson MR. (2013). mRNA expression, splicing and editing in the embryonic and adult mouse cerebral cortex. *Nat Neurosci.* Apr;16(4):499-506.

IV. Dillman AA, Galter D, Cookson M ADAR2 affects mRNA coding sequence edits but not expression or splicing *in vivo. Manuscript*

V. Dillman AA, Majounie E, Gibbs JR, Hernandez D, Arepalli S, Traynor BJ, Singleton AB, Galter D, Cookson MR. Transcript profiling of the human brain reveals diminished synaptic function and accumulated inflammatory signaling are associated with chronological aging. *Manuscript*

TABLE OF CONTENTS

1 Introduction....1
1.1 Measuring RNA....1
1.2 RNA sequencing....3
1.3 Regulation of the transcriptome....5
1.3.1 Gene expression....5
1.3.2 Splicing....6
1.3.3 A-to-I RNA editing....8
1.4 Genetic variability....9
1.5 Cellularity....11
1.6 The developing brain....12
1.7 Aging in the brain....13
2 Aims....15
3 Materials and methods....16
3.1 RNA extraction....16
3.2 Microarray....16
3.3 qRT-pcr....16
3.4 Genotyping....17
3.5 Expression quantitative trait loci....17
3.6 Calculating assocaitions bewteen aging and expression....18
3.7 WGCNA analysis....18
3.8 RNA sequencing....18
3.9 Anaylsis of gene expression in RNA-Seq data....19
3.10 Statistical analysis of alternative exon utilization....19
3.11 Editing analysis....20
3.12 RNA editing validation by RT-PCR and sequencing....20
4 Results and Discussion....21
4.1 Paper I....21
4.2 Paper II....23
4.3 Paper III....25
4.4 Paper IV....27
4.5 Paper V....28
5 Conclusions....30
6 Acknowledgements....33
7 References....34

LIST OF ABBREVIATIONS

AD	Alzheimer's Disease
ADAR	adenosine deaminase
BAM	binary sequence alignment map format
Cdh22	cadherin 22
CFH	complement factor H
Cyfip2	Cytoplasmic FMR1 interacting protein 2
DAVID	database for annotation, visualization, and integrated
Dscam	down sydrome cell adhesion molecule
eQTL	Expression quantitative trait loci
Flnb	filamin B, beta
FTLD	frontotemporal lobar degeneration
FDR	false discovery rate
Gabra3	gamma-aminobutyric acid (GABA) A receptor, subunit alpha 3
Gria2	glutamate receptor, ionotropic, AMPA2 (alpha 2)
GRIK1	glutamate receptor, ionotropic, kainate 1
GO	gene ontology
GTF	gene transfer format
GWAS	Genome-wide association studies
HWE	Hardy-Weinberg equilibrium
KCNMA1	potassium large conductance calcium-activated channel, subfamily M, alpha member 1
LCM	laser capture microdissection
MAF	minor allele frequency
Mapt	microtubule-associated protein tau
PMI	post-mortem interval
PPID	Peptidylprolyl isomerase D (cyclophilin D)
qRT-PCR	quantitative reverse transcription polymerase chain reaction
RIN	RNA integrity number
RNA-Seq	RNA sequencing

Robo3	roundabout homolog 3
snRNPs	Small nuclear ribonucleic particles
SNP	single nucleotide polymorphism
STREX	stress axis-regulated exon
TSS	transcriptional start site
Ubc	ubiquitin C
UTR	untranslated region
WGCNA	weighted gene correlation network analysis

1 INTRODUCTION

The "central dogma" of molecular biology states that biological information is transferred from nucleic acids to proteins as DNA is transcribed into RNA and RNA is translated to protein, wherein a single gene made a single protein[1]. Although DNA holds the genetic information to code for an entire multicellular organism, the select creation of a subset of RNA species is needed to specify identity in any given cell or tissue at any given time. In recent years, we have come to realize that the transcriptome is much more complex than the central dogma might imply. Messenger RNA molecules are heavily regulated through methods such as RNA splicing, where single exons are included or excluded from the mRNA. RNA is also editing by several enzymes, leading to differences in sequence from the encoded genome. These processes lead to diversification in the encoded proteins. Thus, the transcriptome is central to the dynamic transfer of information from DNA to protein.

1.1 MEASURING RNA

As mRNA is the key intermediate between DNA and protein, it is critical to be able to describe the mRNA content in a cell or tissue qualitatively and quantitatively. Historically, many methods have been developed to look at RNA, which will be outlined briefly here.

In northern blotting, RNA samples are separated based on size using electrophoresis, then transferred via capillary to a membrane that can be blotted using a labeled probe complementary to the gene of interest [2,3]. As alternate splicing can lead to differences in sizes of the RNA, Northern blots can be used to estimate splicing events in some cases. Furthermore, Northern blots can be used for quantification of major species.

Accurate quantification of mRNA can also be achieved using quantitative reverse transcription polymerase chain reaction (qRT-PCR). In this method, the first step is to reverse transcribe RNA into cDNA. The cDNA is then amplified using polymerase chain reaction with quantification occurring in real time after each cycle. Detection can be done in two ways, using either a non-specific fluorescent reporter such as SYBR green or a sequence specific probe such as those found in TaqMan assays. SYBR green intercalates with double stranded DNA and as amplification using gene specific primers

proceeds, more DNA product is made and fluorescence increases. TaqMan uses a sequence specific probe that contains a fluorophore at its 5' end and a 3' quencher, which anneals with its complementary DNA region. As the Taq polymerase synthesizes a new copy of the DNA, it degrades the probe through its exonuclease activity. This degradation releases the fluorophore from the quencher, allowing for detection. Both SYBR Green and Taqman assays can be highly accurate with a large dynamic range when probes are designed correctly and reference genes are chosen appropriately[4–8].

Another method, *in situ hybridization*, uses labeled probes to show the localization of individual RNA species within tissues[9]. In this approach, sections of tissues are incubated with probes of interest therefore allowing for identification of the cells in which a given sequence is expressed. This technology has been particularly useful in complex tissues such as the brain that contain many cell types and is amenable to large-scale examination of many genes such as in the Allen Brain Atlas project (http://mouse.brain-map.org/). Furthermore, quantification of *in situ* signals can be achieved by counting silver grains after development of radiolabeled sections with photographic emulsion[10] or with fluorescent probes[11].

All of these methods are generally limited to quantifying a single gene or a small handful of genes per experiment and are therefore most effectively applied when there is a prior knowledge of which genes to investigate. However, unbiased genome-wide approaches would be useful when trying to find new molecular events associated with a given biological events. Microarray chips contain tens to hundreds of thousands of probes providing an ability to assay multiple genes in a single experiment. Generally, the first step in this method is to use reverse transcriptase to convert RNA to cDNA and then tag the cDNA with biotin. Biotinylated cDNA is then hybridized to the microarray with cDNA binding to probes that represent the complementary sequences and are attached to a support (usually a glass slide). Next, a series of washes is applied to remove unbound cDNA then a fluorescent dye containing streptavidin is added to bind to the biotinylated cDNA. The overall fluorescence at each prove is quantified using an array scanner giving relative expression levels of each gene interrogated by the array[12–14].

There are several known limitations in microarrays, including problems with probe

design[15,16], a compressed range of sensitivity[17], and statistical concerns[18] due to the large number of probes measured. For example, non-specific hybridization to microarray probes leads to a limited dynamic range[19,20]. Because of these limitations, validation is generally required for any proposed differences in gene expression between groups, with qRT-PCR being the favoured method. Interestingly there is little in the literature that deals with which method is "correct" when these two techniques are in not in agreement, either by the differences magnitude of expression differences or even in the direction of effect reported. Etienne at al found that correlation between RT-PCR and microarray was highest with moderately expressed genes[21], and another study found that there was less correlation between the two methods for down-regulated genes versus those that were up-regulated[22]. Both studies indicate that microarray is weak at accurately measuring genes that are expressed at low levels. Because of these problems with the accuracy of microarray approaches, other methods of quantifying gene expression on a genome-wide level are worth exploring.

1.2 RNA SEQUENCING

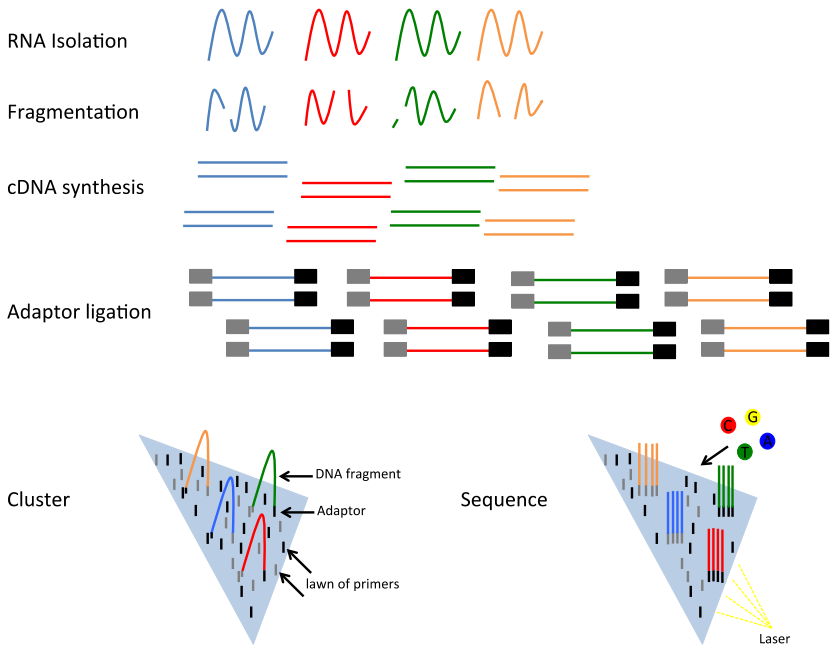

Figure 1. This figure was adapted from van der Brug *et al.* and Mardis *et al.* to illustrate the steps in library preparation and sequencing for the RNA-Seq method[23,24].

Interrogating the complexity of the entire transcriptome in an unbiased manner has become easier in recent years due to the development of RNA-sequencing (RNA-Seq). RNA-Seq is a high throughput method of sequencing cDNA libraries and the general workflow is illustrated in Figure 1. RNA input for library creation can either be selected for mature, polyadenylated RNA using oligo-dT magnetic beads, or total RNA can be used as input with rRNA removal beads to clear ribosomal RNA. The next steps, cDNA synthesis and fragmentation can be interchanged depending on the aim of the study. Fragmenting the RNA before cDNA synthesis gives a more uniform sequence coverage over the gene body but fewer reads at the ends of transcripts[25], while fragmenting after shows a strong 3' bias[26]. Strand specific information can also be obtained if uracil is incorporated during the 2nd strand synthesis portion of the cDNA protocol and then later degraded using uracil-DNA glycosylase. It is also possible, however, to skip cDNA synthesis entirely and sequence the RNA directly such as in the method described by Ozsolak *et al.*[27], but this method was not utilized in the thesis work described herein. Adapters are then ligated so that each fragment has the same 3' and 5' sequence for PCR amplification. Libraries are attached to a flowcell containing a dense lawn of primers. For downstream signal detection, amplification of each cDNA fragment is needed. Colonies of each sequence are generated by bridge PCR. Sequencing with illumina technology uses fluorescently labeled dNTPs with a reversible terminator so only a single nucleotide can be added each cycle. Once the microscope captures the fluorescent signal the terminator is cleaved so the next base can be added. This process can be repeated up to 150 times with little error and then the process can be repeated using the opposite ended primer (paired end) for a potential 300 base pairs of sequence information. Each sequenced fragment is then stitched back together using a reference genome where overall counts per gene can be used for gene expression analysis.

RNA-Seq has advantages over microarray in gene expression quantification, as it does not rely on probe design and, as such, both known and novel genes can be interrogated. RNA-Seq has much greater dynamic range than microarrays [28,29], spanning over five orders of magnitude[25]. Correlation between protein and gene quantification is stronger when using RNA-Seq over microarray[30,31], suggesting more accurate capture of biological information. Information pertaining to splicing can also be extracted using coverage per exon and reads that span more than one exon, which are also known as

junction reads[32]. As RNA-Seq contains sequence-specific information, sequence mismatches between RNA and DNA, such as those resulting from RNA editing, can also be identified[33]. With the proper depth of coverage[34], RNA-Seq is capable of giving us a robust and complex view of the transcriptome.

Collectively, these different methodological approaches have been used to generate a view of the complement of mRNA that includes data at the level of whole gene, exon and single base pair level. Regulation of mRNA expression at each of these levels includes distinct biological mechanisms and is therefore worth discussing separately.

1.3 REGULATION OF THE TRANSCRIPTOME

1.3.1 Gene expression

Regulating the overall amount of RNA molecules produced from a given gene can be accomplished in multiple ways. First, RNA polymerase must be able to bind the DNA in order for the DNA to be transcribed. If a section of DNA is tightly bound to histones, RNA polymerase cannot gain access and no gene product can be made. Modifications to histone proteins[35] and DNA methylation[36] dictate the strength of their interaction and therefore how accessible any particular gene is for transcription. If the nucleoprotein complex containing histones and DNA is in an open conformation, transcription factors can bind and either recruit or impede RNA polymerase[37]. Subsequently, transcribed RNA is converted into mature RNA via post-transcriptional modifications each having a potential impact on overall mRNA quantity, which will be discussed further below.

Modulation of gene expression is critical to many biological processes including morphogenesis and cell differentiation. In multi-tissue gene expression studies using microarrays, only 1-7.5% of the genes measured were ubiquitously expressed, suggesting that gene regulation contributes to the transcriptome complexity needed to derive different cell and tissue types[38–40] . In these studies, only a small fraction of genes were utilized in any single tissue type, indicating that gene expression is not regulated as a simple on/off mechanism. Ramskold *et al.* used RNA-Seq to measure gene expression in multiple tissues and found that approximately 75% of all genes are ubiquitously expressed[41]. In the same study, it was found that of all the tissues sequenced, brain was one of the most complex. With such a large number of genes

sharing expression across tissue types, mechanisms such as splicing and editing may also play a role in tissue specificity.

1.3.2 Splicing

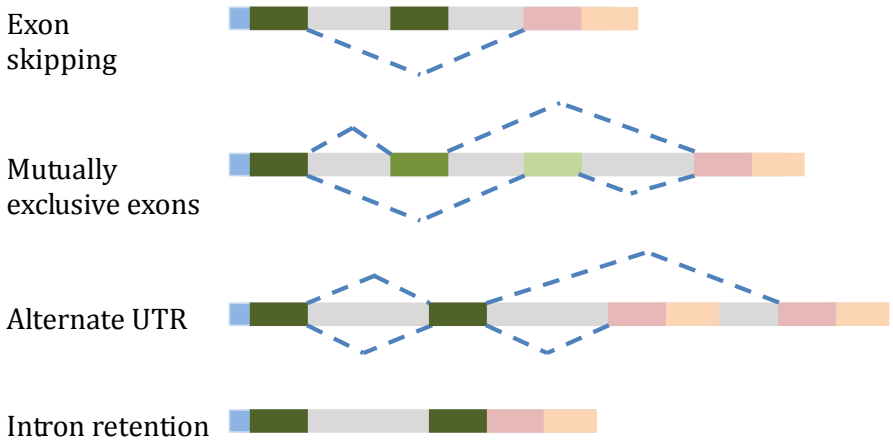

Figure 2. This figure was adapted from Licatalosi *et al.* demonstrates the different types of alternative splicing that can occur as indicated by the dotted lines[42]. Exons are in green, introns in grey, the 5'UTR is in blue and the 3'UTR is in pink.

Splicing, the removal of non-coding intronic sequences and the joining of exons, is required in order to form mature RNA. This process occurs in the nucleus by the spliceosome, a large complex of proteins and small nuclear ribonucleic particles (snRNPs)[43,44]. Introns contain a specific 5' donor site, a 3' acceptor site, and a branch site near the 3' end for recognition and removal by the spliceosome. Non-splicsosomal proteins can enhance or repress the use of splice sites by binding to exonic sequences, allowing for these exons to be included or excluded. This mechanism of alternative splicing leads to the creation of multiple isoforms from the same gene and thus proteomic diversity (Figure 2). Studies utilizing exon junction microarrays have found that as many as 74% of multi-exon genes have alternative splicing events[45–47] and RNA-Seq estimates that an even higher proportion of genes may be spliced [48,49]. Splicing plays an important role in neurodevelopment, synaptic plasticity and strength[50], and the mammalian brain has been found to have the one of the highest number of alternative transcripts compared to other tissue types[48,51]. For instance,

isoforms of the drosophila homologue of *dscam* mediate axon guidance[52], while neurexin splice variants in the rat brain have critical roles in synaptogenesis[53].

There are also cases in which intronic sequences are retained. This form of alternative splicing is thought to be less common than other forms of alternate splicing but there have been relatively few studies investigating intron retention on a transcriptome wide level to ascertain their true frequency. In two RNA-Seq studies, the human brain was shown to have higher levels of intron retention compared to liver[54,55]. Intron retention may play a vital role in the brain, for example by generating calcium channel diversity[56]. Intron retention in KCNMA1 RNA promotes the inclusion of a stress axis-regulated exon (STREX) and reduction of either the intron or the STREX exon reduced burst firing in hippocampal neurons[56].

Alternative splicing can also influence the regulation of transcript expression by changing mRNA stability, affecting the efficiency of translation, altering the number of miRNA sites or switching localization signals[42,57]. Many regulatory elements that are involved in these processes are found in the untranslated regions (UTRs) of genes. For instance, the subcellular localization of WT1, a tumor suppressor gene, depends on the alternative splicing of a gene region including its zinc finger[58]. Isoforms of the divalent metal transporter known as nramp2 can either include or exclude an iron response element found in its 3'UTR[59], where is it then expressed primarily in the duodenum[60] or in erythroid cell precursors respectively[61]. Interestingly, the 3'UTR regions in the mammalian brain are much longer than in other tissues[41,62].

1.3.3 A-to-I RNA editing

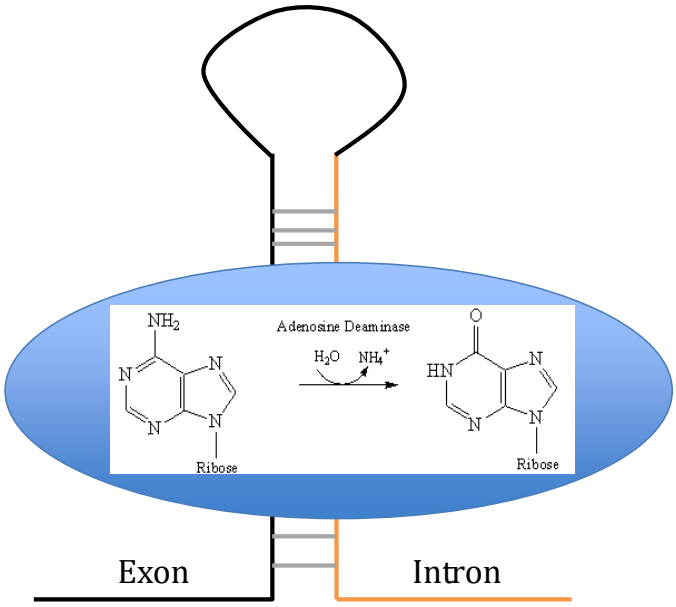

Figure 3. This figure was adapted from Hogg *et al.* and demonstrates ADAR binding to double stranded RNA from both intronic and exonic sequences and then the deamination that occurs[63].

An additional source of transcriptome diversity is generated at the single base level via RNA editing illustrated in Figure 3. Adenosine to inosine substitutions in mammalian RNA are the most widely understood RNA editing events and are carried out by adenosine deaminases (ADARs), of which there are three isoforms; ADAR, ADAR2, and ADAR3. ADAR1 and ADAR2 are ubiquitously expressed, with expression levels are highest in the brain[64,65] while ADAR3 is exclusively expressed in the brain[66,67]. ADARs act on double stranded RNA specific and may require dimerization to be enzymatically active[68,69], although ADAR3 does not seem dimerize[68] and is also enzymatically inactive to known substrates and synthetic dsRNA[67]. Though other double stranded binding proteins have been found to have structural rather than sequence specificity[70], Lehmann *et al.* found ADAR1 and ADAR2 share a 5' sequence preference while only ADAR2 has a 3' preference[71]. ADARs are localized primarily in the nucleolus and are bound to ribosomal RNA[72] but can translocate to the nucleus

upon expression of ADAR substrates such as *Gria2*[73].

Inosine is recognized as guanosine in translation and as such editing in the coding region of a gene can result in a change in the amino acid sequence[74]. Editing may be particularly important in the brain as there are multiple isoforms of neurotransmitter receptors that are targeted by ADARs[75]. In mice, editing of *Gria2* leads to the impermeability of this glutamate channel to calcium ions. If only the unedited isoform is present mice die approximately two weeks after birth due to seizures[76]. Looking across mouse strains the majority of A-to-I editing sites are highly conserved[77] supporting the idea that editing is biologically important. Editing is also found in non-mammalian species, in octopuses RNA editing plays a role in the temperature adaptability of potassium channels[78], again showing that RNA editing may influence neuronal excitability.

RNA editing in the noncoding region of a gene can result in altered splicing, nuclear retention, or stability of the transcript[79–82]. For example, ADAR2 may to regulate its own expression via the creation of an alternate splice site by auto-editing[80]. Hyperedited RNAs are bound and anchored to the nuclear matrix via the p54nrb complex to prevent them from being exported from the nucleus[79]. Several studies have reported that viral RNA undergoes a high level of editing[83,84] and it possible that nuclear retention of these promiscuously edited RNAs is therefore an anti-viral mechanism[85].

1.4 GENETIC VARIABILITY

Variation in the genome can play role in controlling gene expression, thereby adding another layer of complexity to the transcriptome. Single nucleotide polymorphisms (SNPs) a common form of variation in mammalian genomes. With the development of high throughput SNP arrays, capturing this variation on a genome-wide scale is relatively straightforward and cost-effective. Associations between genetic variants and gene expression, where gene expression is treated as a quantitative trait, identify expression quantitative trait loci (eQTLs). The majority of eQTLs are found in *cis* rather than in *trans*. The strongest signals between correlated SNPs and gene expression are within 1 Mb on either side of the transcriptional start site (TSS) of the gene of origin[86–88]. SNPs in such regions may affect the strength of a transcription

factor binding to the promoter site, thereby modulating the amount of mRNA transcribed.

Many previous studies examining eQTLs have used microarrays to estimate gene expression[86–89] potentially missing additional transcriptomic events such as splicing. A study by Kwan *et al.* using exon arrays found that out of all transcript-SNP pairs, 55% were isoform changes but only 39% were whole gene expression level changes[90], which stresses the importance of measuring the expression level of each transcript. RNA-Seq may therefore be suitable for assessing transcriptomic genetic associations. Pickrell *et al.* using RNA-Seq in lymphoblastoid cell lines and reported measurable associations between genetic variation and splicing[91].

One might typically expect equal expression of each allele of a heterozygous haplotype, but this is not always the case. For example, X-inactivation is the random silencing of one of the two copies of the X chromosome in each cell in females, thereby balancing the amount of expression to be equivalent to males. However, about 15% of X-linked genes in human and 3% in mice escape this inactivation[92]. Another form of monoallelic gene expression is genomic imprinting where autosomal genes are expressed in a parent-of-origin fashion, via inherited epigenetic marks such as DNA methylation and histone modification[93]. Additionally, expression of some genes is subject to allelic imbalance, where in a heterozygous haplotype one allele is expressed more than the other. RNA-Seq may be able to find such allele-specific expression information in a transcriptome wide manner. In the same study mentioned, Pickrell *et al.* could also quantify the amount each allele was expressed. Of the 222 putative cis-eQTLS with quantifiable exonic SNPs 88% were driven by more reads from the higher expressing haplotype[91].

Genetic variation can be a risk factor for disease[94]. SNP chip data along with phenotypic information can be used to investigate associations between genetic variation and disease as in genome-wide association studies (GWAS). GWAS can be used to find SNPs that mark chromosomal regions and disease but they don't indicate any biological relevance for that association. By definition variants found in GWAS are present in the control population, so it is possible to overlay these with eQTLs. Any SNPs that are associated with risk of disease and that are also in eQTL would suggest that expression levels of that specific gene are associated with disease. Supporting this

idea, previous studies show that GWAS hits paired with eQTL data can improve the ability to find true associations[95,96]. However, these associations are more likely to be found when the gene is expressed in the tissue type observed[97].

1.5 CELLULARITY

The brain is a complex tissue with a heterogeneous cellular composition. For example, the cerebral cortex is organized into six layers each having differing densities and distributions of neuronal cell populations[98]. These neurons also project to other regions in the brain depending on their function. For instance in the primary visual cortex, layer 4 is thicker than other layers as it receives a large number of projections from the thalamus[99,100], whereas in the primary motor cortex the thickest layer is 5 as it contains a large number of neurons projecting to the spinal cord[101]. Another type of organization found in subcortical structures are tight clusters of neurons in nuclei, an example being the substantia nigra which contains a large number of dopaminergic projection neurons. With all of this heterogeneity, homogenization of brain tissue may complicate measures of gene expression if certain cell types age more quickly or are more affected by disease.

Cell specificity is of particular importance in neurodegenerative diseases such as Alzheimer's and Parkinson's disease. In many of these diseases there is evidence of preferential neuronal vulnerability, such as the prominent but not exclusive loss of dopaminergic neurons of the *substantia nigra pars compacta* in Parkinson's disease. Being able to separate out specific cell types for gene expression quantification may therefore be useful in understanding the molecular basis of neurodegenerative diseases. Laser capture microdissection (LCM) is one technique that can be utilized to isolate subpopulations of cells with the *caveat* that any method used for identification of the cells of interest cannot degrade the RNA needed for downstream analysis. If the cell type has a relatively unique morphology, such is the case for Purkinje cells, cresyl violet staining alone may suffice for identification and capture. In these types of experiments population each neuron must be individually captured and the amount of RNA acquired is small. For example, approximately 300 dopaminergic cells contain around 10-20 ng of RNA[102]. With two rounds of amplification microarray analysis is possible but RNA-Seq has required substantially more RNA input. Picelli *et al.* created a method called Smart-seq2 for creating cDNA libraries from single captured cells with

pictogram amounts of RNA[103], and Deng *et al.* used this method with single cells from mouse embryos demonstrating that minimal amounts of laser captured cells can be used for robust RNA-Seq analysis[104]. In their study analyzing gene expression in hundreds of singular mouse cells they found a number of genes show expression for only one allele in a completely stochastic manner illustrating the heterogeneity in individual cells.

1.6 THE DEVELOPING BRAIN

The cellular and molecular complexity of the mature adult brain discussed is influenced by processes in development and by the formation of neuronal circuits that are experience dependent[105,106]. Brain development requires the coordinated expression of many genes in a spatially and temporally appropriate context[107,108]. A number of studies have employed microarray technology in an attempt to study the molecular changes occurring during brain development on a genome wide scale[109–112].

There are many examples of regulated alternative splicing in neuronal development[50,113]. In mice, fetal *Mapt* has only minor incorporation of exon 10 but by postnatal day 24 all tau contains this exon[114]. The glutamate receptor gene *Gria2* has a pair of mutually exclusive exons leading to two splice isoforms, flip and flop[115], and in rats the isoform containing the flip exon is expressed at stable levels throughout development while the isoform incorporating flop is low until postnatal day 8[116]. In general, retention of introns is also high in the brain and is developmentally regulated as levels of retention are higher in the fetal brain as compared to the adult[54]. One example of intron retention during development is in the role Robo3 plays in axon guidance[117]. The Robo3 isoform captaining an intronic sequence (Robo3.2) is expressed but translationally repressed allowing for attraction to the spinal cord midline. Once the axon crosses the spinal cord midline, it receives signals from the floor plate to translate Robo3.2 allowing nonsense-mediated decay to occur, causing repellence away from midline. The brain also has been found to have longer 3'UTR regions than [41,62] and this lengthening of UTRs occurs during development[118].

A-to-I editing also seems to be important in development. ADAR2 knock out mice demonstrate perinatal lethality, dying of seizures a few weeks after birth[76], while knocking out ADAR1 is embryonic lethal, although this seems to be due to lack of

erythropoiesis in the liver[119]. Looking at 28 specific editing sites across development in the mouse brain Wahlstedt *et al.* found that the majority showed a relatively low editing efficiency that increased over time[33].

1.7 AGING IN THE BRAIN

One of the strongest risk factors for neurodegenerative diseases is age[120,121]. Aging of the brain is a complex set of events that include both aspects under genetic control and stochastic events[122]. While the molecular mechanisms underlying aging are poorly understood, several studies have shown correlations in gene expression levels and aging in the human brain[123–126]. However, comparing these datasets reveals that only a few age-related expression changes have been consistently found. Kang *et al.* observed that while abundant and drastic expression changes were seen during prenatal development, the changes in adult aging were fewer and tended to be more modest[127]. To date, no reports using RNA-Seq to study aging in the brain have been published and so it remains possible that there are additional associations with aging that are yet to be identified.

Singular examples of aberrant splicing have been shown to play a role in neurodegenerative diseases. For example, an unbalanced ratio of splice isoforms of *MAPT* has been found in FTLD (frontotemporal lobar degeneration)[128,129]. There is some evidence that changes in splicing occur during normal aging. For example in, Harries *et al*. have reported that deregulation of mRNA processing occurs with aging in human blood[130]. In a study done in human brain, splicing changes associated with normal aging were also found in FTLD and AD (Alzheimer's Disease) cases irrespective of age[131], illustrating how aging may predispose the brain to certain neurodegenerative diseases.

Finally there are potential links between RNA editing and aging. ADAR2 and ADAR3 are reported to contain SNPs that are associated with extreme old age in the New England Centenarian study, and these findings were replicated in three independent centenarian datasets[132]. Knocking out either of the ADAR orthologues decreased lifespan in the model organism *Caenorhabditis elegans*. ADAR2 knockout mice die shortly after birth, but if these mice also carry a mutation for the edited base in Gria2 they have normal life spans[76] indicating that ADAR2 may have a stronger role in development versus aging. Focusing on ADAR output, Holmes *et al.* quantified the

amount of editing per site in a selected number of sites in rat but did not find any differences associated with aging[62]. However, a study in human brain found an age related decrease in editing of *Cyfip2* but not in *Gabra3*[134]. The differences between studies may be explained by the fact that mice and rats lack a long enough life span to reach these age related transcriptomic changes, or that an association between aging and ADAR activity may be unique to humans. One possible way to interrogate associations of editing with aging in an unbiased manner would be to use RNA-Seq on a human brain cohort with a suitably large age range.

2 AIMS

The goal of my thesis project is to gain a better understanding of the complexity of the brain transcriptome at several different levels, using both human brain samples and model organisms. I initially focused on the interplay between genetic variation and gene expression in the human brain, generating a data set that can be used to elucidate potential associations between normal genetic variation and disease. As well as genetic variants, a major risk factor for neurological diseases is aging. I therefore also focused on normal aging in human post mortem brain to find age related gene expression changes that may prime the brain for neurodegenerative conditions.

A major aim of my thesis work was to shift the gene expression quantification technique from microarray to RNA-Seq. As there are many well-documented transcriptomic changes during development, I generated a preliminary dataset of embryonic versus adult mouse cerebral cortex. The discovery that RNA-Seq could be used to find and quantify RNA-editing sites lead to further characterization of A-to-I editing in a knock out mouse model. Finally, I then applied the RNA-Seq methods to human brain samples with a view to revisiting the relationship between age and gene expression.

Aim 1: Quantify mRNA abundance in human brain and blood samples using microarrays and query the relationship between GWAS risk variants and eQTLs.
Aim 2: Quantify mRNA abundance in human brains and laser-captured Purkinje cells across age ranges using microarray to find potential associations between gene expression and aging.
Aim 3: Quantify expression, splicing, and editing in E17 and adult mouse cerebral cortex using RNA-Seq to discover differential transcriptomic events.
Aim 4: Quantify expression, splicing, and editing in Adar2$^{-/-}$ Gria2$^{R/R}$ and Adar2$^{+/+}$ Gria2$^{R/R}$ littermate controls using RNA-Seq to find ADAR2 specific substrates and examine the potential role of ADAR2 in gene expression and splicing regulation.
Aim 5: Quantify expression, splicing, and editing in the human brain series using RNA-Seq to find associations between transcriptomic changes and aging.

3 MATERIALS AND METHODS

3.1 RNA EXTRACTION

I extracted total RNA from fresh frozen brain using a tissue homogenizer and trizol (papers I-V). Peripheral blood was collected using PAXgene tubes and RNA was extracted using the PAXgene Blood mRNA kit according to the manufactures protocols (paper I). For Purkinje cell isolation (paper II) tissue was immersed in Shandon M-1 embedding matrix and stored at −80°C until use. Cryostat sections (7–8 μm thick) were cut and stained with Cresyl Violet. Laser-capture microdissection was performed with ArcturusXT microdissection system. Between 70 and 150 excised Purkinje cells were selected from the slide surface and captured on LCM Macro Caps. Cellular RNA was recovered from the collected cells using PicoPureTM RNA isolation kit and treated with RNase-free DNase. RNA quality was measured using either the 260 to 280 ratio using a nanodrop (paper I) or the Agilent 2100 Bioanalyzer RNA Nano Chip (papers II-V).

3.2 MICROARRAY

Samples were biotinylated and amplified using the Illumina® TotalPrep-96 RNA Amplification Kit (papers I-II). Two rounds of amplification was necessary for Purkinje cell RNA and was carried out with the Ambion MessageAmp II aRNA kit (paper II). Amplified RNA was hybridized onto HumanHT-12_v3 Expression BeadChips (Illumina). These arrays contain 48,804 probes estimating expression of ~25,000 annotated genes from the RefSeq (Build 36.2, release 22) and Unigene (Build 199). Raw intensity values for each probe were normalized using cubic spline in BeadStudio (Illumina) then $\log_2$ transformed. Individual probes were included in analysis if they were detected ($P < 0.01$) in more than 95% of samples in the series.

3.3 QRT-PCR

We synthesized cDNA from Trizol-extracted RNA using the SuperScript III First-Strand Synthesis System. We measured cDNA abundance using Sybr Green on an Applied BioSystems HT-7900 qRT-PCR system. We performed serial dilutions of cDNA to find primer pairs with 100% efficiency and a single product on the disassociation curve. We used *Ppid* and *Ubc* (paper III) or β-actin (paper II) as normalization genes. In paper III regression was used to compare estimates of fold

differences found in RNA-Seq as compared to qRT-PCR. In paper II relative expression levels were then plotted against age after correction for other known covariates.

3.4 GENOTYPING

In paper I genomic DNA was extracted using phenol–chloroform and genotyping was performed using the Illumina Infinium HumanHap550 v3, Human610-Quad v1 or Human660W-Quad v1 Infinium Beadchip and common SNPs across all platforms were identified for each sample. SNPs were excluded if they showed < 95% genotyping success rate per SNP, minor allele frequency (MAF) < 0.01 or Hardy–Weinberg equilibrium (HWE) p-value < 1×10^{-7}. Markov Chain based haplotyper was used to impute non-assayed genotypes for blood and brain datasets independently using the June 2010 release of the 1000 Genomes Project build-36 reference panel, using default settings for MACH. Imputed SNPs were excluded from the analysis if their minor allele frequency (MAF) was < 0.01 and if their r^2 was < 0.3. Trait and disease associated SNPs were extracted from the NHGRI catalog of published GWAS at http://www.genome.gov/gwastudies/ on July 30th 2011. Analyses were restricted to the following criteria: discovery p-value < 1×10^{-8}, initial sample size > 1000 (or 1000 cases in binomial analyses), replication sample size > 500 (or 500 cases in binomial analyses), number of SNPs > 100,000, samples of European ancestry and risk allele frequency of SNP(s) greater or equal to 0.01.

3.5 EXPRESSION QUANTITATIVE TRAIT LOCI

In each brain region, mRNA probes within 500 kb of the chromosomal location of each SNP were incorporated into linear regression modeling using MACH2QTLv1.08 in paper I. Estimates of the association between the allelic dose of each SNP as a predictor of proximal gene expression levels were generated. These linear regression models were adjusted for both biological and methodological covariates which are known to affect expression profiles[135]. SNPs with fewer than 3 minor homozygotes detected (based on either genotyped SNPs or maximum likelihood genotypes from imputation) were excluded from analyses. A consensus set of results was extracted from the frontal cortex, cerebellum and blood eQTL datasets with identical overlapping combinations of GWAS SNPs and proximal *cis* mRNA probes.

Significant associations were determined within each tissue type using a 5% FDR adjustment for multiple testing. Proportions of tested associations were calculated per tissue based on this subset of the eQTL results, and were compared using chi-squared tests.

3.6 CALCULATING ASSOCAITIONS BEWTEEN AGING AND EXPRESSION

In paper II we used genotype data to perform a principal components analysis for identity by state of genotype and took the first two principal components, PC1 and PC2, from this analysis to estimate overall genetic distance within the sample series. Cubic spline normalization was applied to raw output from array scans, then expression values were corrected for known covariates of gender, post-mortem interval (PMI), principal components PC1 and PC2 from genotyping and hybridization batch using multivariate regression the same as in eQTL analysis above except that age was not specified in the model. The residuals for expression after covariate correction were then tested against age using linear regression. *P*-values were adjusted for multiple testing using a false discovery rate (FDR) correction set at 0.05. Probes were included if detected in >95% of samples in each series for frontal cortex and cerebellum, while in the Purkinje cells dataset only probes that had shown prior evidence of association in the above datasets were tested.

3.7 WGCNA ANALYSIS

In paper II the R-based package WGCNA (Weighted gene correlation network analysis) was used for a weighted network level analysis[136]. Twenty-seven consensus modules for genes with similar expression patterns were identified using the discovery and replication dataset in the frontal cortex and cerebellum as well as the Purkinje cell dataset. Covariates were tested for correlation with each probe and then tested for module membership. Enrichment for Gene Ontology (GO) terms within modules was performed using DAVID[137,138].

3.8 RNA SEQUENCING

We used either purified poly(A)$^+$ RNA (papers III, V) or total RNA (paper IV) for cDNA library synthesis. mRNA libraries were created using the mRNA-Seq prep kit

with oligo(dT) priming as per the manufacturer's protocol (papers III, V). Total RNA was processed using ribo-Zero gold rRNA removal beads and then libraries were synthesized using the Tru-Seq stranded total RNA sample preparation kit as per the manufacturer's protocol (paper IV). In paper III we hybridized a single library to each lane of a flow cell and 80-bp single end sequences were generated using an Illumina GA-II sequencer. In paper IV samples were multiplexed and six libraries were hybridized to a lane with 100-bp paired end sequences being generated using the Hi-Seq. In paper V a single library was hybridized to each lane with 100-bp paired end sequences being generated using the Hi-Seq 2000.

3.9 ANAYLSIS OF GENE EXPRESSION IN RNA-SEQ DATA

Fastq files were aligned to the mouse reference genome (mm9) using Tophat and Bowtie (paper III and IV) applying the ensembl gtf option with the Mus_musculus.NCBIM37.61 gtf (a gene transfer format file) to build bowtie indexes (paper IV)**.** For paper V fastq files were aligned to the human reference genome (hg19) using Tophat and Bowtie applying the ensembl gtf option with the GRCh37.59 gtf to build bowtie indexes. Reads were annotated and quantified to a given gene using the Python module HT-SEQ. For gene counts the same ensembl gtf mentioned above was used to provide reference boundaries. We used either the R/Bioconductor package DESeq (paper III) or DESeq2 (paper V) for library size normalize and to perform a variance-stabilizing transformation[139]. Poisson distributions of normalized counts for each transcript were then compared across samples either using a negative binomial test (paper III and IV) or a generalized linear model (paper V). Multiple testing was corrected for using the Benjamini-Hochberg procedure.

3.10 STATISTICAL ANALYSIS OF ALTERNATIVE EXON UTILIZATION

In paper III an in house gtf was built with exon start and stop locations to quantify reads that fell within exon boundaries, and junction reads were analyzed separately with the start position and end position being treated as independent measures. We then used the ratio of gene count to exon count as input for DE-SEQ, specifically using gene/exon rather than exon/gene because the latter would be fractions and not suitable for DE-SEQ, which requires integers. In paper IV the NCBIM37.61 gtf was flattened with the python script found in HT-SEQ to create the appropriate counting bins needed for downstream exon analysis. We then used R/Bioconductor package DEXSeq[140] for

library size normalize, variance-stabilizing transformation, and differential usage.

3.11 EDITING ANALYSIS

For papers III and IV BAMS (binary sequence alignment map format files) were first filtered for only singly mapped reads. Variants were then called for each sample using SAMtools pileup with mm9 as a reference. We created a merged table to include only variants shared by all samples within a given group. We extracted variants within a gene boundary and with A-to-I changes. Finally, we submitted sequences containing the candidate edit sites to UCSC BLAT to distinguish reads with single mismatches to the genomic sequence due to RNA editing from those due to inappropriate mapping to another part of the genome. We accepted candidate edits if the best scored alignment of the read included the site of the edited base from the original alignment; if there were multiple genomic alignments of equal or higher score, then the candidate edit site was discarded. For paper V variants for each of the 43 conserved sites found by Pino *et al.*[141] were called as a listed input for each sample using SAMtools pileup with hg19 as a reference. We calculated percentages using the total number of called bases at each site and dividing them by the number of edited sites.

3.12 RNA EDITING VALIDATION BY RT-PCR AND SEQUENCING

In paper III cDNA was made from RNA extracted with trizol using the Supercript III First-Strand Synthesis System from Invitrogen. We designed primer pairs that were approximately 150 base pairs away from the editing site and spanned an exon junction to avoid genomic DNA amplification. We also designed a genomic DNA primer to span the same site that amplified a product of similar size when used with one of the cDNA primers. Amplification of cDNA and genomic DNA was then performed with Fast Start Master Mix. Products were separated on a 1.5% agarose gel, cut out and extracted DNA using a QIAquick gel extraction kit. We cloned DNA into the PCR8-TA cloning vector (Invitrogen) in One Shot Mach1-T1 Chemically competent cells and sequenced 24 colonies from three biological replicates for each edit site and 12 colonies from genomic DNA using a 3730 capillary sequencer.

4 RESULTS AND DISCUSSION

4.1 PAPER I

Genome-wide association studies (GWAS) have played a useful role in finding many single nucleotide polymorphisms (SNPs) associated with common traits, including diseases. These studies generally do not, however, demonstrate direct biological relevance of disease associated SNPs. Expression quantitative trait loci (eQTL) studies look at the association between SNPs and gene expression levels. Therefore, in paper I we extracted all known SNPs associated with human traits based on GWAS and used these SNPs for eQTL discovery. eQTLs discovered in this manner would suggesting altered gene expression levels may underlie the GWAS association. Choosing the tissue type that is relevant to the disease or phenotype observed is an important consideration when measuring gene expression levels in eQTL experiments. While it may be easier to procure a blood sample than brain, one must consider that there are likely to be unique expression patterns in each tissue type not shared with the other. We therefore compared eQTLs in both blood[142] and brain[86] to gain a better understanding of how informative blood can be for brain related phenotypes. Like previous eQTL studies involving more than one tissue or cell type, we found a small set of shared eQTLs across both tissues. eQTLs that were shared across tissues tended to have a larger effect sizes than tissue specific eQTLs[143–147]. Looking at tissue specific eQTLs there was a subset in which the gene was not expressed in measurable levels in other tissues. An example of this is the blood specific eQTL found for LRAP/ERAP2 and rs2549794 a SNP associated with Crohn's disease[148], which was not expressed in brain at detectable levels. However, when restricting the analysis to genes expressed in both tissue there were still examples of genes with similar expression levels but where there was a difference in how the genetic variants associated with transcript levels (Figure 4). This study underlines the importance of tissue type in expression studies, and demonstrates GWAS data paired with eQTL data can yield biologically informative results.

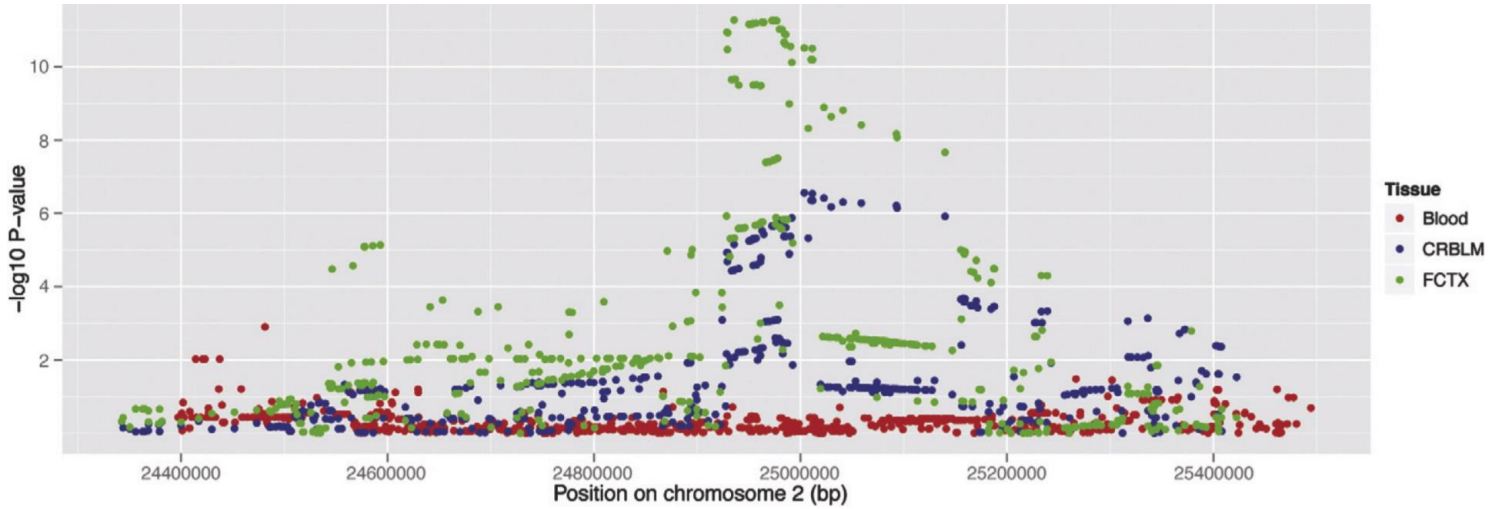

Figure 4. This figure shows an example of a frontal cortex specific eQTL.

Understanding the interplay between genetic variation and gene expression in the human brain can be used to help elucidate associations between normal genetic variation and disease. We conclude that eQTL data from an appropriate tissue can be compared to GWAS nominated SNPs, giving a potential explanation of expression level changes as the cause of disease associations.

4.2 PAPER II

In paper II, we measured gene expression from bulk extracted frontal cortex and cerebellum from human brains across a wide age range, a replication dataset, and in isolated Purkinje cells collected via LCM. We found that several associations between age and gene expression carried over from the discovery dataset to the replication dataset, with the caveat that the replication set had a smaller sample size. Interestingly, more genes with age-associated changes were found in frontal cortex versus cerebellum. This was consistent with data published by Fraser *et al.*[149], and may support the theory that differing brain regions age at differing rates. In the laser captured Purkinje microarray expression dataset we noted increases in cell specific makers such as PCP2[150] indicating effectiveness of the technique. Looking only at significant expression age associations from the frontal cortex or cerebellum to interrogate potential correlations in LCM there were five genes with the greatest support for significance including some that were not significant in bulk cerebellum (Figure 5). Comparing these hits to another brain dataset from Colantunoni *et al*. we found four out of these five to be significant illustrating the robustness of our findings[151]. This work is of particular importance as it demonstrates that LCM samples can be used as input to measure gene expression using microarray technologies.

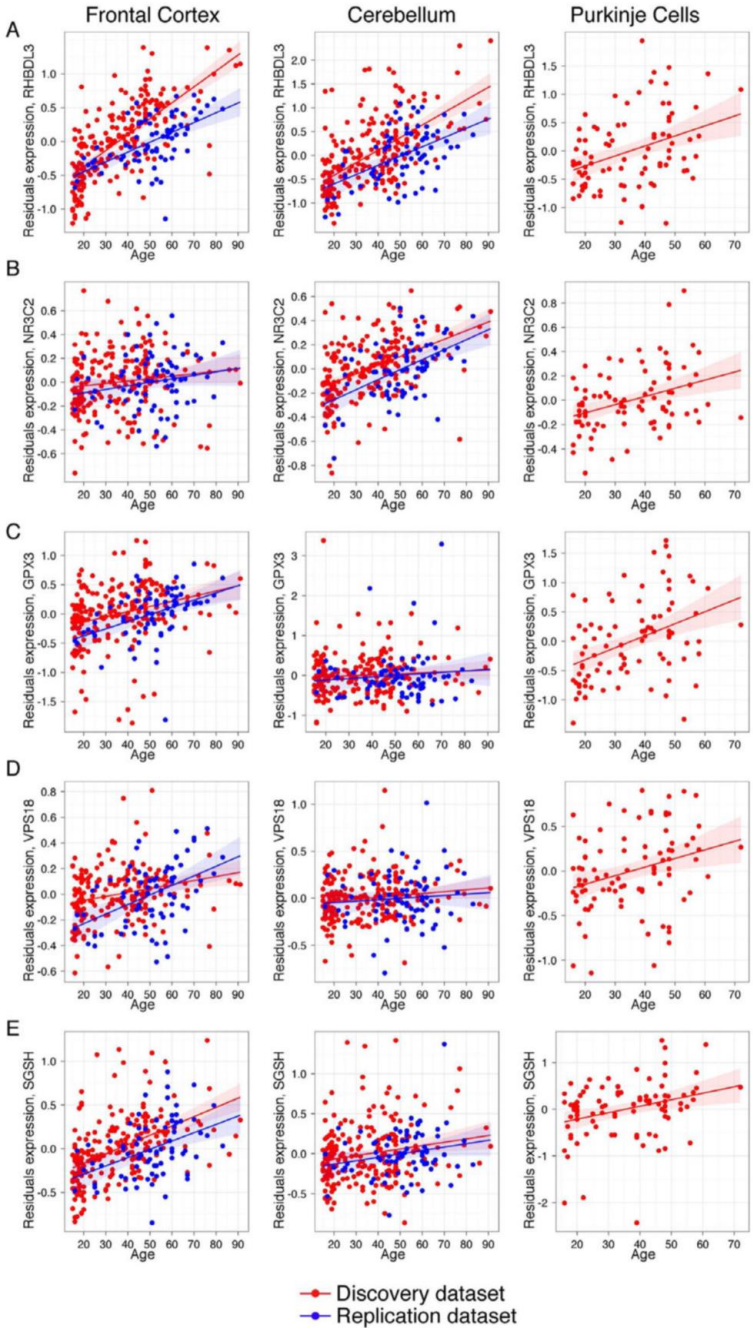

Figure 5. This figure illustrates the replicated age associated, with each plot showing associations between age and residual expression. Frontal cortex is on the left hand panels, cerebellum is the central panels while Purkinje cells are on the right.

4.3 PAPER III

Although measurements of gene expression have been performed using microarrays, this approach generally considers each 'gene' as a single unit. However, many tissues including the brain, in particular, show a large number of splicing and editing events. With this depth of complexity in mind, we chose to use newly developed RNA-Seq techniques to capture and quantify transcriptomic events in a global and unbiased manner. RNA-Seq is, however, still in its nascent stages with little consensus on how to handle data processing and quantification. We therefore generated a preliminary dataset in paper III using RNA sequencing on samples from the cerebral cortex of embryonic day 17 and adult mouse brains. We decided to use mouse as it is a commonly used laboratory model organism, and there is a wealth of published information on differential expression, splicing, and RNA editing in pre- and post-natal tissue. These published differences were therefore used to estimate the accuracy of the results found by deep sequencing methods. Our analysis revealed that there are numerous differences in RNA expression, splicing, and editing between embryonic and adult tissue. There were over 4,000 genes that were differentially expressed (Figure 6).

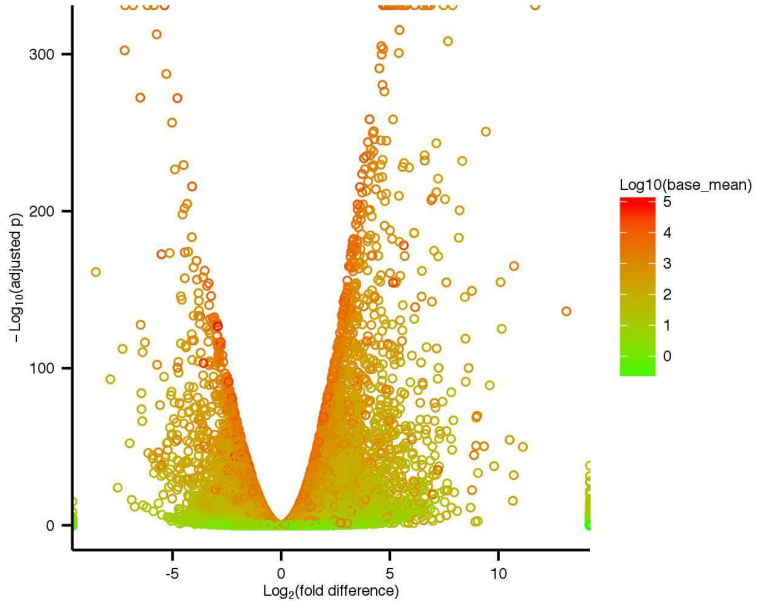

Figure 6. Differential expression, with log_2 fold differences in normalized expression in adult and embryonic cerebral cortex plotted on the x-axis and the – log_{10} adjusted p-value on the y-axis. Genes are colored based on the log_{10} base mean expression

Gene ontology analysis illustrated that genes more highly expressed in the adult brain had more enrichment in neurotransmission or ion homeostasis while the majority of genes with higher expression in the embryonic brain were involved with cell division, a finding that is recapitulated in other studies[109–111,152]. We found 387 exons that were differentially expressed with a fairly equal division of significance between adult and embryonic brains (Figure 7).

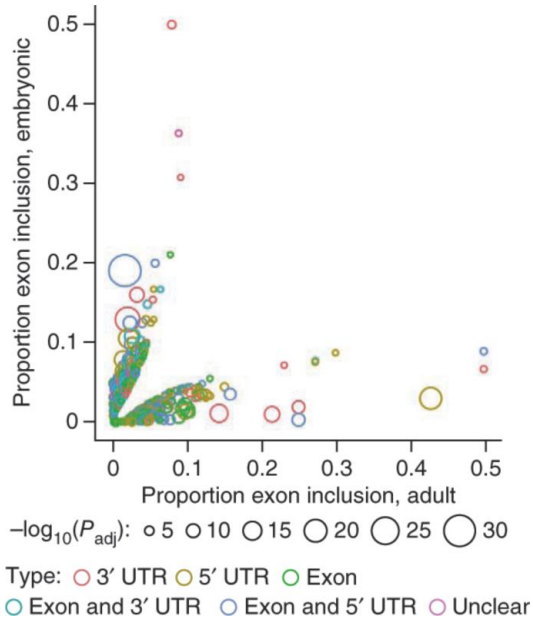

Figure 7. This is the alternative exon utilization with proportion of inclusion of each event in the adult on the x-axis and the proportion of embryonic on the y-axis.

In our analysis of A-to-I editing we discovered 176 sites with a majority of these residing in the 3'UTR repeat regions of genes (Figure 8), a finding that has been reported in previous studies[153–156].

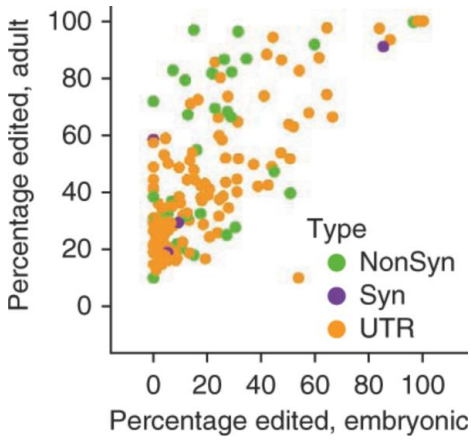

Figure 8. The is the average percentage edited for each A-to-I site in the adult cortex on the y-axis and the embryonic cortex on the x-axis.

Interestingly there is an increase in editing sites, mostly in the coding region, over development. These results were biologically robust as they validated in a second animal cohort, and accurate as they were confirmed with more traditional quantification methods. Our RNA-Seq dataset has also been made publically available for additional data mining.

4.4 PAPER IV

We performed RNA-Seq on the frontal cortices of ADAR2$^{-/-}$ GRIA2$^{R/R}$ mice and their ADAR2$^{+/+}$ GRIA2$^{R/R}$ littermates, quantifying overall gene expression splicing and most importantly editing transcriptome wide. We found 71 editing sites that had at least partial ADAR2 specificity with a majority of these being in the coding region of genes (figure 9)

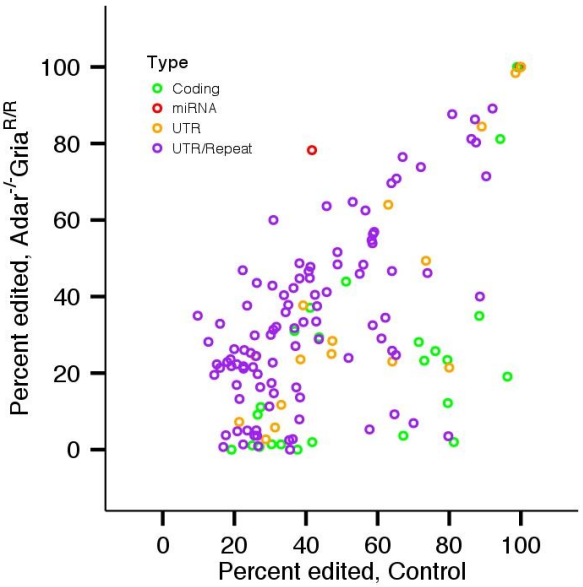

Figure 9. A-to-I editing percentages with control on the x-axis and ADAR2$^{-/-}$ GRIA2$^{R/R}$ on the y-axis.

It has been found in previous studies that ADAR2 shows a base preference immediately before and after the edit site[71], but aligning all differential edits in our dataset did not yield any sequence specificity. Interestingly, besides ADAR2 expression, there was only a single statistically significant differentially expressed gene, *Flnb*. Also, despite the finding that ADAR2 may associate with spliceosomal proteins[157], there were only three exons that showed statistically significant usage with none of these signals reaching two fold. This work illustrates the importance of ADAR2 in site-specific changes of protein coding sequences but also suggests that ADAR2 does not have a major role in splicing or gene expression in vivo.

4.5 PAPER V

Previously we have used microarray technologies to show correlations in gene expression levels and aging in the human brain in an attempt to better understand the molecular mechanisms of aging. In paper II we only found five age-related expression changes that were consistently significant in multiple datasets. To further explore the association between aging and gene expression we decided to employ RNA-Seq on 60 neurologically normal human brains of varying ages. We found a large number of

genes that are associated with aging including several that had not previously been identified by microarrays. These gene expression changes also had consistent biological relevance as those that decreased with age were more involved in synaptic transmission while those that increased had more involvement in immune response (examples in Figure 10).

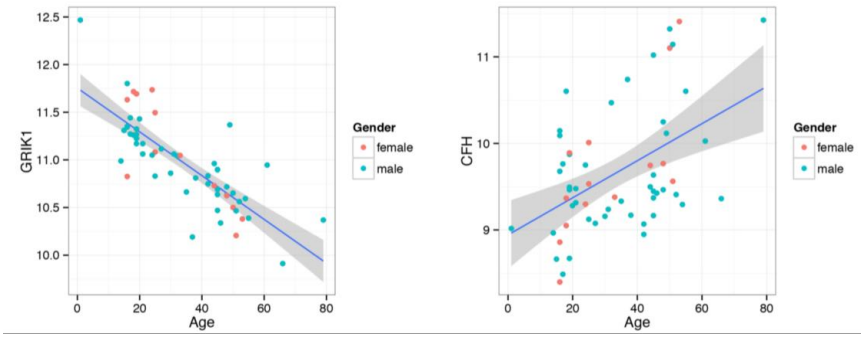

Figure 10. This is a plot illustrating examples of age associated gene expression changes. On the left panel is *GRIK1,* a gene involved in synaptic transmission, while the right panel is *CFH,* a gene involved in the immune response.

Although we only found sex chromosomal genes were differential expressed when comparing male versus female frontal cortices, a study using exon arrays found splicing differences in autosomal genes in frontal cortex when comparing across gender[158]. This finding illustrates the need for an in depth splicing analysis of out dataset for both potential gender and age differences. Finally we found a single editing site, in the gene *cdh22*, to be associated with aging. These results show that RNA-Seq can be used to identify novel associations between aging and gene expression, including editing changes.

5 CONCLUSIONS

Paper I:

- Tissue specific eQTLs were found
- Some unique eQTLs were due to the gene only being expressed in that particular tissue
- This study illustrates the importance of using the correct tissue in eQTL mapping
- It also demonstrated that SNPs found in GWAS paired with expression can yield biologically informative results

Paper II:

- Many genes show some evidence of age-association in the human brain but they are often modest
- Only a small subset of associations carried over from the discovery set to the replication set
- More genes with age-associated gene changes were found in the frontal cortex as compared to the cerebellum supporting the theory that brain regions age at different rates
- Comparing LCM to both datasets there were five genes with the greatest support for significance

Paper III:

- There was a large set of statistically significant differentially expressed genes
- Differential exon usage was found in both adult and embryonic
- Both novel and previously cited editing sites were found and quantified
- Some edit sites increased with age, most commonly those found within the coding region of genes

Paper IV:

- $Adar2^{-/-}/Gria2^{R/R}$ mice have decreases in multiple editing sites with the majority being found in the coding regions of genes as compared to their littermate controls

- Adar2$^{-/-}$/Gria2$^{R/R}$ mice do not have major differences in total gene expression or splicing as compared to their littermate controls, arguing against ADAR2 playing an important regulatory role in these pathways.

Paper V:

- Batch, PMI, and RIN have an affect on gene expression as measured by RNA-Seq
- Only genes found on sex chromosomes were differentially expressed when comparing male versus female samples but the strength of these genes were large
- A large number of genes had a more modest association with age
- Those genes with a decrease in expression across age had a gene ontology enrichment in synaptic transmission. Genes that were more expressed with age showed the strongest association with GO categories in immune response

6 ACKNOWLEDGEMENTS

Mark Cookson
Thank you for taking me on as a summer student all those years ago, I bet you didn't know what you signed up for back then! I appreciate all the time you spent giving me advice on how to present and write in the scientific world. You have helped mould me into what I hope is a great scientist.

Dagmar Galter
Thank you for all your positivity and kind words. You have been wonderful helping to guide me through the Swedish process, thank you for being so tolerant of my many mistakes.

The Olson Group
Thank you for your support and for welcoming me into your lab. I enjoyed sharing all the Swedish traditions and science with you.

The Cookson Group
Thank you for the scientific and emotional support throughout the years, and thanks for all the moves you helped me make throughout DC and to Sweden and back again!

Lars Olson
I have been amazed at your breadth and depth of knowledge and your ability to embrace new things.

Brian Traynor
Thanks for letting me be a part of the ALS world through participating in bench work with your group and for taking me to meetings.

Andy Singleton
Thank you for putting up with my random questions and for with our cubicles noise levels!

My Family
Thank you for all the love and support. Thanks mom and dad for dealing with an exceptionally precocious child!

7 REFERENCES

1. Crick, F. Central dogma of molecular biology. *Nature* **227,** 561–563 (1970).
2. Thomas, P. S. Hybridization of denatured RNA transferred or dotted nitrocellulose paper. *Methods Enzymol.* **100,** 255–266 (1983).
3. Khandjian, E. W. & Méric, C. A procedure for Northern blot analysis of native RNA. *Anal. Biochem.* **159,** 227–232 (1986).
4. Siebert, P. D. Quantitative rt-PCR. *Methods Mol. Med.* **26,** 61–85 (1999).
5. Udvardi, M. K., Czechowski, T. & Scheible, W.-R. Eleven golden rules of quantitative RT-PCR. *Plant Cell* **20,** 1736–1737 (2008).
6. Vandesompele, J. *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3,** RESEARCH0034 (2002).
7. Joyce, C. Quantitative RT-PCR. A review of current methodologies. *Methods Mol. Biol. Clifton NJ* **193,** 83–92 (2002).
8. Freeman, W. M., Walker, S. J. & Vrana, K. E. Quantitative RT-PCR: pitfalls and potential. *BioTechniques* **26,** 112–122, 124–125 (1999).
9. Hoefler, H. *et al.* In situ hybridization methods for the detection of somatostatin mRNA in tissue sections using antisense RNA probes. *Histochem. J.* **18,** 597–604 (1986).
10. Lewis, M. E., Krause, R. G., 2nd & Roberts-Lewis, J. M. Recent developments in the use of synthetic oligonucleotides for in situ hybridization histochemistry. *Synap. N. Y. N* **2,** 308–316 (1988).
11. Kwon, S. Single-molecule fluorescence in situ hybridization: quantitative imaging of single RNA molecules. *BMB Rep.* **46,** 65–72 (2013).
12. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–470 (1995).
13. McLachlan, G. J., Do & Ambroise, C. *Analyzing microarray gene expression data*. (Wiley-Interscience, 2004). at <http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=128122>
14. Epstein, C. B. & Butow, R. A. Microarray technology - enhanced versatility, persistent challenge. *Curr. Opin. Biotechnol.* **11,** 36–41 (2000).
15. Harbig, J., Sprinkle, R. & Enkemann, S. A. A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. *Nucleic Acids Res.* **33,** e31 (2005).
16. Mecham, B. H. *et al.* Increased measurement accuracy for sequence-verified microarray probes. *Physiol. Genomics* **18,** 308–315 (2004).
17. Sharov, V. *et al.* The limits of log-ratios. *BMC Biotechnol.* **4,** 3 (2004).
18. Nadon, R. & Shoemaker, J. Statistical issues with microarrays: processing and analysis. *Trends Genet. TIG* **18,** 265–271 (2002).
19. Okoniewski, M. J. & Miller, C. J. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics* **7,** 276 (2006).
20. Binder, H. & Preibisch, S. Specific and nonspecific hybridization of oligonucleotide probes on microarrays. *Biophys. J.* **89,** 337–352 (2005).
21. Etienne, W., Meyer, M. H., Peppers, J. & Meyer, R. A., Jr. Comparison of mRNA gene expression by RT-PCR and DNA microarray. *BioTechniques* **36,** 618–620, 622, 624–626 (2004).

22. Morey, J. S., Ryan, J. C. & Van Dolah, F. M. Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. *Biol. Proced. Online* **8,** 175–193 (2006).
23. Van der Brug, M., Nalls, M. A. & Cookson, M. R. Deep sequencing of coding and non-coding RNA in the CNS. *Brain Res.* **1338,** 146–154 (2010).
24. Mardis, E. R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9,** 387–402 (2008).
25. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5,** 621–628 (2008).
26. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320,** 1344–1349 (2008).
27. Ozsolak, F. *et al.* Direct RNA sequencing. *Nature* **461,** 814–818 (2009).
28. 't Hoen, P. A. C. *et al.* Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res.* **36,** e141 (2008).
29. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5,** 613–619 (2008).
30. Fu, X. *et al.* Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* **10,** 161 (2009).
31. Ning, K., Fermin, D. & Nesvizhskii, A. I. Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. *J. Proteome Res.* **11,** 2261–2271 (2012).
32. Bryant, D. W., Jr, Priest, H. D. & Mockler, T. C. Detection and quantification of alternative splicing variants using RNA-seq. *Methods Mol. Biol. Clifton NJ* **883,** 97–110 (2012).
33. Wahlstedt, H., Daniel, C., Ensterö, M. & Ohman, M. Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. *Genome Res.* **19,** 978–986 (2009).
34. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* **15,** 121–132 (2014).
35. Kouzarides, T. Chromatin modifications and their function. *Cell* **128,** 693–705 (2007).
36. Keshet, I., Lieman-Hurwitz, J. & Cedar, H. DNA methylation affects the formation of active chromatin. *Cell* **44,** 535–543 (1986).
37. Lemon, B. & Tjian, R. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.* **14,** 2551–2569 (2000).
38. Su, A. I. *et al.* Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 4465–4470 (2002).
39. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 6062–6067 (2004).
40. Zhang, W. *et al.* The functional landscape of mouse gene expression. *J. Biol.* **3,** 21 (2004).
41. Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* **5,** e1000598 (2009).
42. Licatalosi, D. D. & Darnell, R. B. RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.* **11,** 75–87 (2010).
43. Jurica, M. S. & Moore, M. J. Pre-mRNA splicing: awash in a sea of proteins. *Mol. Cell* **12,** 5–14 (2003).

44. Rino, J. & Carmo-Fonseca, M. The spliceosome: a self-organized macromolecular machine in the nucleus? *Trends Cell Biol.* **19,** 375–384 (2009).
45. Johnson, J. M. *et al.* Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302,** 2141–2144 (2003).
46. Pan, Q. *et al.* Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol. Cell* **16,** 929–941 (2004).
47. Le, K. *et al.* Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic Acids Res.* **32,** e180 (2004).
48. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40,** 1413–1415 (2008).
49. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476 (2008).
50. Li, Q., Lee, J.-A. & Black, D. L. Neuronal regulation of alternative pre-mRNA splicing. *Nat. Rev. Neurosci.* **8,** 819–831 (2007).
51. Yeo, G., Holste, D., Kreiman, G. & Burge, C. B. Variation in alternative splicing across human tissues. *Genome Biol.* **5,** R74 (2004).
52. Wojtowicz, W. M., Flanagan, J. J., Millard, S. S., Zipursky, S. L. & Clemens, J. C. Alternative splicing of Drosophila Dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. *Cell* **118,** 619–633 (2004).
53. Ullrich, B., Ushkaryov, Y. A. & Südhof, T. C. Cartography of neurexins: more than 1000 isoforms generated by alternative splicing and expressed in distinct subsets of neurons. *Neuron* **14,** 497–507 (1995).
54. Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.* **18,** 1435–1440 (2011).
55. Kapranov, P. *et al.* The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. *BMC Biol.* **8,** 149 (2010).
56. Bell, T. J. *et al.* Intron retention facilitates splice variant diversity in calcium-activated big potassium channel populations. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 21152–21157 (2010).
57. Kalsotra, A. & Cooper, T. A. Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.* **12,** 715–729 (2011).
58. Larsson, S. H. *et al.* Subnuclear localization of WT1 in splicing or transcription factor domains is regulated by alternative splicing. *Cell* **81,** 391–401 (1995).
59. Lee, P. L., Gelbart, T., West, C., Halloran, C. & Beutler, E. The human Nramp2 gene: characterization of the gene structure, alternative splicing, promoter region and polymorphisms. *Blood Cells. Mol. Dis.* **24,** 199–215 (1998).
60. Canonne-Hergaux, F., Gruenheid, S., Ponka, P. & Gros, P. Cellular and subcellular localization of the Nramp2 iron transporter in the intestinal brush border and regulation by dietary iron. *Blood* **93,** 4406–4417 (1999).
61. Canonne-Hergaux, F., Zhang, A. S., Ponka, P. & Gros, P. Characterization of the iron transporter DMT1 (NRAMP2/DCT1) in red blood cells of normal and anemic mk/mk mice. *Blood* **98,** 3823–3830 (2001).
62. Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J. O. & Lai, E. C. Widespread and extensive lengthening of 3' UTRs in the mammalian brain. *Genome Res.* **23,** 812–825 (2013).
63. Hogg, M., Paro, S., Keegan, L. P. & O'Connell, M. A. RNA editing by mammalian ADARs. *Adv. Genet.* **73,** 87–120 (2011).
64. Melcher, T. *et al.* A mammalian RNA editing enzyme. *Nature* **379,** 460–464 (1996).

65. O'Connell, M. A. *et al.* Cloning of cDNAs encoding mammalian double-stranded RNA-specific adenosine deaminase. *Mol. Cell. Biol.* **15,** 1389–1397 (1995).
66. Chen, C. X. *et al.* A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains. *RNA N. Y. N* **6,** 755–767 (2000).
67. Melcher, T. *et al.* RED2, a brain-specific member of the RNA-specific adenosine deaminase family. *J. Biol. Chem.* **271,** 31795–31798 (1996).
68. Cho, D.-S. C. *et al.* Requirement of dimerization for RNA editing activity of adenosine deaminases acting on RNA. *J. Biol. Chem.* **278,** 17093–17102 (2003).
69. Gallo, A., Keegan, L. P., Ring, G. M. & O'Connell, M. A. An ADAR that edits transcripts encoding ion channel subunits functions as a dimer. *EMBO J.* **22,** 3421–3430 (2003).
70. Ryter, J. M. & Schultz, S. C. Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. *EMBO J.* **17,** 7505–7513 (1998).
71. Lehmann, K. A. & Bass, B. L. Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. *Biochemistry (Mosc.)* **39,** 12875–12884 (2000).
72. Sansam, C. L., Wells, K. S. & Emeson, R. B. Modulation of RNA editing by functional nucleolar sequestration of ADAR2. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 14018–14023 (2003).
73. Desterro, J. M. P. *et al.* Dynamic association of RNA-editing enzymes with the nucleolus. *J. Cell Sci.* **116,** 1805–1818 (2003).
74. Sommer, B., Köhler, M., Sprengel, R. & Seeburg, P. H. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell* **67,** 11–19 (1991).
75. Seeburg, P. H. RNA helicase participates in the editing game. *Neuron* **25,** 261–263 (2000).
76. Higuchi, M. *et al.* Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature* **406,** 78–81 (2000).
77. Danecek, P. *et al.* High levels of RNA-editing site conservation amongst 15 laboratory mouse strains. *Genome Biol.* **13,** 26 (2012).
78. Garrett, S. & Rosenthal, J. J. C. RNA editing underlies temperature adaptation in K+ channels from polar octopuses. *Science* **335,** 848–851 (2012).
79. Zhang, Z. & Carmichael, G. G. The fate of dsRNA in the nucleus: a p54(nrb)-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. *Cell* **106,** 465–475 (2001).
80. Rueter, S. M., Dawson, T. R. & Emeson, R. B. Regulation of alternative splicing by RNA editing. *Nature* **399,** 75–80 (1999).
81. Serra, M. J., Smolter, P. E. & Westhof, E. Pronounced instability of tandem IU base pairs in RNA. *Nucleic Acids Res.* **32,** 1824–1828 (2004).
82. Nishikura, K. Functions and regulation of RNA editing by ADAR deaminases. *Annu. Rev. Biochem.* **79,** 321–349 (2010).
83. Bass, B. L., Weintraub, H., Cattaneo, R. & Billeter, M. A. Biased hypermutation of viral RNA genomes could be due to unwinding/modification of double-stranded RNA. *Cell* **56,** 331 (1989).
84. Hajjar, A. M. & Linial, M. L. Modification of retroviral RNA by double-stranded RNA adenosine deaminase. *J. Virol.* **69,** 5878–5882 (1995).
85. Kumar, M. & Carmichael, G. G. Nuclear antisense RNA induces extensive adenosine modifications and nuclear retention of target transcripts. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 3542–3547 (1997).

86. Gibbs, J. R. *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* **6,** e1000952 (2010).
87. Veyrieras, J.-B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* **4,** e1000214 (2008).
88. Myers, A. J. *et al.* A survey of genetic human cortical gene expression. *Nat. Genet.* **39,** 1494–1499 (2007).
89. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477,** 289–294 (2011).
90. Kwan, T. *et al.* Genome-wide analysis of transcript isoform variation in humans. *Nat. Genet.* **40,** 225–231 (2008).
91. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).
92. Berletch, J. B., Yang, F., Xu, J., Carrel, L. & Disteche, C. M. Genes that escape from X inactivation. *Hum. Genet.* **130,** 237–245 (2011).
93. Reik, W. & Walter, J. Genomic imprinting: parental influence on the genome. *Nat. Rev. Genet.* **2,** 21–32 (2001).
94. Tsuji, S. Genetics of neurodegenerative diseases: insights from high-throughput resequencing. *Hum. Mol. Genet.* **19,** R65–70 (2010).
95. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452,** 423–428 (2008).
96. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6,** e1000888 (2010).
97. Grundberg, E. *et al.* Population genomics in a disease targeted primary cell model. *Genome Res.* **19,** 1942–1952 (2009).
98. Economo, C. *Cellular structure of the human cerebral cortex*. (Karger, 2009).
99. Hirsch, J. A. & Martinez, L. M. Laminar processing in the visual cortical column. *Curr. Opin. Neurobiol.* **16,** 377–384 (2006).
100. Montagnini, A. & Treves, A. The evolution of mammalian cortex, from lamination to arealization. *Brain Res. Bull.* **60,** 387–393 (2003).
101. Rivara, C.-B., Sherwood, C. C., Bouras, C. & Hof, P. R. Stereologic characterization and spatial distribution patterns of Betz cells in the human primary motor cortex. *Anat. Rec. A. Discov. Mol. Cell. Evol. Biol.* **270,** 137–151 (2003).
102. Pietersen, C. Y., Lim, M. P., Macey, L., Woo, T.-U. W. & Sonntag, K. C. Neuronal type-specific gene expression profiling and laser-capture microdissection. *Methods Mol. Biol. Clifton NJ* **755,** 327–343 (2011).
103. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9,** 171–181 (2014).
104. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343,** 193–196 (2014).
105. Innocenti, G. M. & Price, D. J. Exuberance in the development of cortical networks. *Nat. Rev. Neurosci.* **6,** 955–965 (2005).
106. Sur, M. & Rubenstein, J. L. R. Patterning and plasticity of the cerebral cortex. *Science* **310,** 805–810 (2005).
107. Miller, J. A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508,** 199–206 (2014).
108. Liscovitch, N. & Chechik, G. Specialization of gene expression during mouse brain development. *PLoS Comput. Biol.* **9,** e1003185 (2013).
109. Mody, M. *et al.* Genome-wide gene expression profiles of the developing mouse hippocampus. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 8862–8867 (2001).

110. Semeralul, M. O. *et al.* Microarray analysis of the developing cortex. *J. Neurobiol.* **66,** 1646–1658 (2006).
111. Matsuki, T., Hori, G. & Furuichi, T. Gene expression profiling during the embryonic development of mouse brain using an oligonucleotide-based microarray system. *Brain Res. Mol. Brain Res.* **136,** 231–254 (2005).
112. Díaz, E. *et al.* Molecular analysis of gene expression in the developing pontocerebellar projection system. *Neuron* **36,** 417–434 (2002).
113. Lee, C. J. & Irizarry, K. Alternative splicing in the nervous system: an emerging source of diversity and regulation. *Biol. Psychiatry* **54,** 771–776 (2003).
114. McMillan, P. *et al.* Tau isoform regulation is region- and cell-specific in mouse brain. *J. Comp. Neurol.* **511,** 788–803 (2008).
115. Sommer, B. *et al.* Flip and flop: a cell-specific functional switch in glutamate-operated channels of the CNS. *Science* **249,** 1580–1585 (1990).
116. Monyer, H., Seeburg, P. H. & Wisden, W. Glutamate-operated channels: developmentally early and mature forms arise by alternative splicing. *Neuron* **6,** 799–810 (1991).
117. Colak, D., Ji, S.-J., Porse, B. T. & Jaffrey, S. R. Regulation of axon guidance by compartmentalized nonsense-mediated mRNA decay. *Cell* **153,** 1252–1265 (2013).
118. Ji, Z., Lee, J. Y., Pan, Z., Jiang, B. & Tian, B. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 7028–7033 (2009).
119. Wang, Q., Khillan, J., Gadue, P. & Nishikura, K. Requirement of the RNA editing deaminase ADAR1 gene for embryonic erythropoiesis. *Science* **290,** 1765–1768 (2000).
120. Fjell, A. M. *et al.* What is normal in normal aging? Effects of aging, amyloid and Alzheimer's disease on the cerebral cortex and the hippocampus. *Prog. Neurobiol.* (2014). doi:10.1016/j.pneurobio.2014.02.004
121. Tan, C.-C. *et al.* Autophagy in aging and neurodegenerative diseases: implications for pathogenesis and therapy. *Neurobiol. Aging* **35,** 941–957 (2014).
122. Martin, G. M. The biology of aging: 1985-2010 and beyond. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **25,** 3756–3762 (2011).
123. Lu, T. *et al.* Gene regulation and DNA damage in the ageing human brain. *Nature* **429,** 883–891 (2004).
124. Berchtold, N. C. *et al.* Gene expression changes in the course of normal brain aging are sexually dimorphic. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 15605–15610 (2008).
125. Erraji-Benchekroun, L. *et al.* Molecular aging in human prefrontal cortex is selective and continuous throughout adult life. *Biol. Psychiatry* **57,** 549–558 (2005).
126. Kumar, A. *et al.* Age-associated changes in gene expression in human brain and isolated neurons. *Neurobiol. Aging* **34,** 1199–1209 (2013).
127. Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478,** 483–489 (2011).
128. Hong, M. *et al.* Mutation-specific functional impairments in distinct tau isoforms of hereditary FTDP-17. *Science* **282,** 1914–1917 (1998).
129. Jiang, Z., Cote, J., Kwon, J. M., Goate, A. M. & Wu, J. Y. Aberrant splicing of tau pre-mRNA caused by intronic mutations associated with the inherited dementia frontotemporal dementia with parkinsonism linked to chromosome 17. *Mol. Cell. Biol.* **20,** 4036–4048 (2000).

130. Harries, L. W. *et al.* Human aging is characterized by focused changes in gene expression and deregulation of alternative splicing. *Aging Cell* **10,** 868–878 (2011).
131. Tollervey, J. R. *et al.* Analysis of alternative splicing associated with aging and neurodegeneration in the human brain. *Genome Res.* **21,** 1572–1582 (2011).
132. Sebastiani, P. *et al.* RNA editing genes associated with extreme old age in humans and with lifespan in C. elegans. *PloS One* **4,** e8210 (2009).
133. Holmes, A. P., Wood, S. H., Merry, B. J. & de Magalhães, J. P. A-to-I RNA editing does not change with age in the healthy male rat brain. *Biogerontology* **14,** 395–400 (2013).
134. Nicholas, A. *et al.* Age-related gene-specific changes of A-to-I mRNA editing in the human brain. *Mech. Ageing Dev.* **131,** 445–447 (2010).
135. Trabzuni, D. *et al.* Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies. *J. Neurochem.* **119,** 275–282 (2011).
136. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).
137. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37,** 1–13 (2009).
138. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2009).
139. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
140. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22,** 2008–2017 (2012).
141. Pinto, Y., Cohen, H. Y. & Levanon, E. Y. Mammalian conserved ADAR targets comprise only a small fragment of the human editosome. *Genome Biol.* **15,** R5 (2014).
142. Wood, A. R. *et al.* Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. *Hum. Mol. Genet.* **20,** 4082–4092 (2011).
143. Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325,** 1246–1250 (2009).
144. Greenawalt, D. M. *et al.* A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome Res.* **21,** 1008–1016 (2011).
145. Nica, A. C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* **7,** e1002003 (2011).
146. Bullaughey, K., Chavarria, C. I., Coop, G. & Gilad, Y. Expression quantitative trait loci detected in cell lines are often present in primary tissues. *Hum. Mol. Genet.* **18,** 4296–4303 (2009).
147. Ding, J. *et al.* Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. *Am. J. Hum. Genet.* **87,** 779–789 (2010).
148. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42,** 1118–1125 (2010).
149. Fraser, H. B., Khaitovich, P., Plotkin, J. B., Pääbo, S. & Eisen, M. B. Aging and gene expression in the primate brain. *PLoS Biol.* **3,** e274 (2005).

150. Oberdick, J., Levinthal, F. & Levinthal, C. A purkinje cell differentiation marker shows a partial dna sequence homology to the cellular sis/pdgf2 gene. *Neuron* **3,** 386 (1989).
151. Colantuoni, C. *et al.* Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature* **478,** 519–523 (2011).
152. Kagami, Y. & Furuichi, T. Investigation of differentially expressed genes during the development of mouse cerebellum. *Brain Res. Gene Expr. Patterns* **1,** 39–59 (2001).
153. Athanasiadis, A., Rich, A. & Maas, S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol.* **2,** e391 (2004).
154. Neeman, Y., Levanon, E. Y., Jantsch, M. F. & Eisenberg, E. RNA editing level in the mouse is determined by the genomic repeat repertoire. *RNA N. Y. N* **12,** 1802–1809 (2006).
155. DeCerbo, J. & Carmichael, G. G. SINEs point to abundant editing in the human genome. *Genome Biol.* **6,** 216 (2005).
156. Blow, M., Futreal, P. A., Wooster, R. & Stratton, M. R. A survey of RNA editing in human brain. *Genome Res.* **14,** 2379–2387 (2004).
157. Raitskin, O., Cho, D. S., Sperling, J., Nishikura, K. & Sperling, R. RNA editing activity is associated with splicing factors in lnRNP particles: The nuclear pre-mRNA processing machinery. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 6571–6576 (2001).
158. Trabzuni, D. *et al.* Widespread sex differences in gene expression and splicing in the adult human brain. *Nat. Commun.* **4,** 2771 (2013).