RNA-seq analysis of hippocampi in *mei2.5* mice reveals two unannotated splice isoforms of *Akap9* in the mm10 build of the Ensembl database

Will Henriques

Dr. Jeremy O. Ward
Thesis Advisor

Senior Thesis
Submitted in partial fulfillment of the requirements for
High Honors
in
the Independent Scholar Program
Computational, Genetic, and Behavioral Approaches to Neurobiology

Middlebury College
Middlebury, Vermont
Friday, 16 December 2016

Committee in Charge:

Jeremy O. Ward, Ph.D.

Clarissa C. Parker, Ph.D.

Michael D. Linderman, Ph.D.

Abstract

In the *mei2.5* mouse model, a mutation in a conserved splice site at the end of exon 13 in the 48 exon gene *Akap9*, leads to aberrant spermatogenesis and infertility in male mice. *Akap9*, an A kinase anchoring protein, is involved in numerous signal transduction pathways and undergoes extensive alternative splicing. Transcripts originating from the *Akap9* genomic locus have been implicated in signal transduction events connected to NMDA receptor function, cardiac potassium channel function, Golgi ribbon organization, centrosomal organization, and microtubule modulation. Here, we conducted a study of *mei2.5* hippocampal gene expression using RNA-seq, piloting a cloud-like bioinformatics work-station in the process. We identified two novel splice isoforms of AKAP9 from the hippocampus which putatively contain a centrosomal binding domain, and identified genes connected to both Alzheimer's disease and Long QT syndrome that may be modulated by signal transduction pathways coordinated by isoforms of AKAP9.

Acknowledgements

This project represents, in many ways, the culmination of a long and twisting intellectual journey at Middlebury, and it could not have been possible without the support of many, but in particular, Professor Jeremy Ward. Since "Cell Hell" in the fall of 2012, Jeremy has been a mentor, role-model, supporter, and friend. Jeremy advised my team in the first year of the STEM Pilot Program in the summer of 2013. He took a risk with me when he agreed to be my primary advisor as I developed an Independent Scholar program to explore the intersection of neuroscience, molecular biology, mathematics, and computer science. He supported my decision to study abroad in Switzerland despite its disruption to my research progress in his lab. Finally, he supported me as I endeavored on this thesis upon my return, and has been instrumental in guiding me towards the light in moments I have despaired about the plausibility of the project.

I would like to thank the members of my thesis committee, Professor Clarissa Parker and Professor Michael Linderman, for their gentle guidance when I have come careening into their office seeking advice. I would also like to give a special thanks to Dr. Grace Spatafora, who reignited my passion for research at a time when I was filled with doubt about my interest and motivation. Nancy Graham was a sage mentor in my early days in the lab.

This project would not have been possible without the staff of Bicentennial Hall. Joanna Shipley's patience and calm is unparalleled, except perhaps by the staff members of Middlebury's Animal Facilities, especially Vicki Major, Alexis Paquette, and Brittany Young or by that of Cathy Ekstrom. To them all I am deeply grateful. I want to thank Tim Allen for his good-natured handling of my perpetual and persistent question: "Do you know where I could find...?", and Tony Desautels and Chris Goodrich for their technical support, Carrie Donohue for her help with orders, and the stock room for providing emergency supplies. Of course, none of this work would have been possible without Tim Wickland, the BiHall manager.

I would also like to thank the current and former Deans of Curriculum, Dean Suzanne Gurland and Professor Bob Cluss, for helping me to develop my independent scholar curriculum (Professor Cluss) and for working with me when I decided to redirect in the middle of my college career (Dean Gurland). I would also like to thank Janis Audet and the Curriculum Committee for its support. Alongside Professor Ward, Professor Noah Graham and Professor Frank Swenton pushed me to envision an education beyond the boundaries of a department through the STEM Pilot Program, so they too deserve acknowledgement in this thesis.

Finally, I would like to thank the National Science Foundation, the Biology Department, the Senior Project Research Supplement, and the Elizabeth Miller Palen '40 Fund for providing the funding necessary to support this project.

Introduction

The dynamic interplay between genes and the environment determine organismal phenotypes. On that, most biologists can agree, but the question of how much the environment contributes and how much the genome contributes to that phenotype remains open. For *Homo sapiens*, we perpetually gravitate towards this question to understand how and why variation occurs in our population. Some of this variation leads to disease, some to difference coded as "disease", and some variation is coded as "normal." Mouse models provide a relevant proxy organism in which to probe this question for those interested in *H. sapiens* due to their close relation to *H. sapiens* on an evolutionary scale, and their history as an extensively researched model organism in the last century. In that time, a number of inbred lines have been bred and characterized, and – second only to the *H. sapiens* genome – the mouse genome is the best-characterized mammalian genome. To understand how the mouse genome gives rise to phenotype, the mouse genetics community endeavored to undertake a coordinated, large-scale mutagenesis project in 1997 [26].

The project began with a forward genetics, "phenotype-driven" approach, in which N-ethyl-N-nitrosourea (ENU) mutagens were used to generate arbitrary nucleotide changes in the genome, and the resulting mice were screened for dominant or recessive phenotypes that differed compared to their un-mutagenized counterparts. Once the sequencing of the mouse genome was completed in 2002, a second phase of the project was proposed, in which investigators would create knockout(KO) lines by systematically knocking out all annotated genes, and conducting several tiers of phenotyping and transcriptome analysis. In the United States, this effort became the Knockout Mouse Project (KOMP) and in Europe, the European Conditional Mouse Mutagenesis (EuCOMM) project. With the advent of next-generation sequencing technologies, high-throughput transcriptional analysis has become a

reality, but the time required to thoroughly phenotype remains a bottleneck [26]. The KO and ENU approaches dovetail nicely in the field of functional genomics: KO models provide a blunt-force method to understand a gene's function broadly, while ENU mutagenesis-derived models facilitate a more subtle examination of how slight variation at highly conserved loci may give rise to diverse phenotypes, some of which may be associated with disease.

High-throughput transcriptome analysis: RNA-Seq

Different tissues within an organism are made up of unique cell types. The central "dogma" of biology explains the basic process by which DNA determines these phenotypes: the enzyme RNA polymerase transcribes deoxyribonucleic acid (DNA) - long, helical, paired strands of the nucleotides adenine (A), thymine (T), guanine (G), and cytosine (C) - into single stranded molecules of ribonucleic acid (RNA). These RNA transcripts fill diverse roles in the cell: long non-coding RNA (lncRNA) and micro-RNAs (miRNA) can regulate the expression of other genes, ribosomal RNA (rRNA) is a component of the ribosome (the organelle responsible for protein synthesis), tRNA guides amino acids to the ribosome where they are incorporated into the growing amino acid chain in protein synthesis, messenger RNA (mRNA) is marked for protein synthesis. mRNA species are tagged are polyadenylated post-transcription and undergo alternative splicing. They are then transported to the ribosome where they are translated into polypeptide chains that fold and develop tertiary and quaternary structure to become functional proteins in the cell. Diverse suites of functional proteins determine cell type and functionality. The "genome" refers to the sequence of nucleotides packed into each chromosome, which contains the organismal blueprint [51]. The "transcriptome" of the cell refers to the entire collection of RNA transcripts present in a cell or tissue at a given moment in time [73].

In 1977, Dr. Frederick Sanger's lab developed the first DNA-sequencing technology by replacing 3' hydroxyl groups with hydrogen atoms. These "dideoxynucleotides" terminate strand elongation by DNA polymerase, creating variable-length DNA fragments. Sanger *et al.* combined dideoxynucleotides and ^{32}P incorporation, and resolved the products on a slab gel, which was subsequently exposed to X-ray film from which the original DNA sequence could be read. Two important next steps in DNA sequencing technology were

the development of base-specific fluorescently-labelled nucleotides in the late 1980's, and the invention of the polymerase chain reaction (PCR), a method for amplifying small amounts of a template sequence with DNA polymerase and a thermal cycler. The introduction of capillary sequencing instruments in 1999 expedited sequencing by removing the need to pour and run slab gels. Since 2005, diverse sequencing technologies – referred to collectively as "Next Generation Sequencing" (NGS) methods – have emerged, which, coupled with advances in computational techniques, have enabled the analysis of whole genomes in a matter of days at relatively low cost [44].

Next-generation methods have revolutionized the field of transcriptomics through the development of RNA-sequencing (RNA-seq) methods. RNA-Seq provides a window into the transcriptional activity of a cell or tissue by enabling the deep-sequencing of all or a subset of the RNA species present in a cell. Through RNA-Seq, the sequence of RNA molecules can be converted into short digitial sequences, and subsequent computational analysis allows these reads to be mapped back to their location of origin on the genome, allowing the quantification of gene expression across the entire genome. RNA-seq experiments provide snapshots in time and space of transcript species structure and expression levels [73]. RNA-seq experiments can be divided into three distinct components: RNA extraction from the biological system of interest (be it tissue or cell type), library preparation and sequencing, and computational analysis.

RNA extraction

The first step of the RNA-Seq pipeline is extraction of total RNA from biological (either tissue or individual cells) or environmental samples. To examine the transcripts destined for protein synthesis, mRNA is purified from the total RNA by passing the total RNA over a column containing magnetic beads with oligo-dT nucleotide chains . The dT chains bind the complementary poly-adenylated tails of the mRNA and capture them while other RNA species pass through the column [49].

Library preparation and sequencing

After polyA selection, the mRNA is eluted from the column and one of two courses can be followed. First, the mRNA may be reverse transcribed into cDNA using an oligo-dT adapter [77] or a combination of oligo-dT adapters and random-hexamer priming [49]. Second, the mRNA can be hydrolized into smaller fragments [48] and primed with random-hexamer primers to synthesize a single cDNA strand by reverse transcription. After synthesis of the second strand of cDNA, a barcode (adapter sequence) is ligated onto the cDNA strands. Barcodes are unique to each sample, allowing libraries to be combined in the downstream analysis without losing sample-specific information. After adapter ligation, a PCR amplification of 12-16 cycles enriches the library. At this point, the cDNA library is considered to be representative of the original mRNA content of the sample, and is ready for sequencing [35].

The development of sequencing technology has focused on developing efficient, high-throughput, high-quality sequences of short fragments of single DNA molecules. There are two general methods for short-read sequencing: sequencing by ligation, and sequencing by synthesis. The biotech company Illumina dominates the short-read sequencing industry with its sequencing by synthesis methodology [27], which was the cannonical platform for early RNA-Seq (see [48], [49], [77]) that continues to be popular, so the remainder of this section will focus on providing a general conceptual overview of sequencing by synthesis, and specifically Illumina's cyclic reversible terminator chemistry (as opposed to the single-nucleotide addition technology used in Ion Torrent technology).

The sequencing reaction (visualized in Figure 1 and Figure 2) takes place in a flow cell, an eight-channel sealed glass microfabricated device. In each of the eight lanes, short adapter sequences are covalently bound to the glass surface in a tightly regulated spatial pattern that enables the generation of clusters. The adapter-ligated cDNA fragments of the library are pumped over the lane at a concentration such that only one fragment binds to the adapters at each cluster site. Cycles of bridge amplification generate approximately a million copies of each fragment at each cluster (Figure 1). The amplification reagents are washed away and sequencing begins. In the incorporation step, all four nucleotides

(labelled with base-specific fluorophores and a 3' block to ensure that every incorporation is a unique event) and DNA polymerase are added simultaneously to the flow cell lanes (Figure 2) . Complementary nucleotides incorporate into the growing second strand, while others are washed away. Each flow cell lane is imaged in tile segments, with a specific cluster density per tile (for example, a flow cell may be imaged in three 100-tile segments with approximately 30,000 clusters per tile), allowing the simultaneous resolution of massive numbers of unique sequences. At each cluster, the incorporated nucleotides are imaged and a base-calling software assigns a base identity [6], [43], [27].

Computational analysis

The final step in an RNA-seq analysis involves the analysis of the sequencing data, a non-trivial task given that a single whole transcriptome can be several gigabytes of data, depending on the model organism. A general conceptual scaffold of the analysis pathway follows: In the pre-processing steps, read quality should be verified and low-quality reads should be filtered out of the analysis. Then, in the case where no reference genome or transcriptome exists, reads can be used to create a *de novo* transcriptome assembly. The downstream analysis largely revolves around annotation of the transcriptome, to build a reference for future work [35]. In the case where a well-annotated reference genome or transcriptome exists – as in this study [2] – reads are aligned to the reference genome or transcriptome. Next, aligned reads are assembled into transcripts, based on the reference annotation, and finally gene and transcript expression is analyzed.

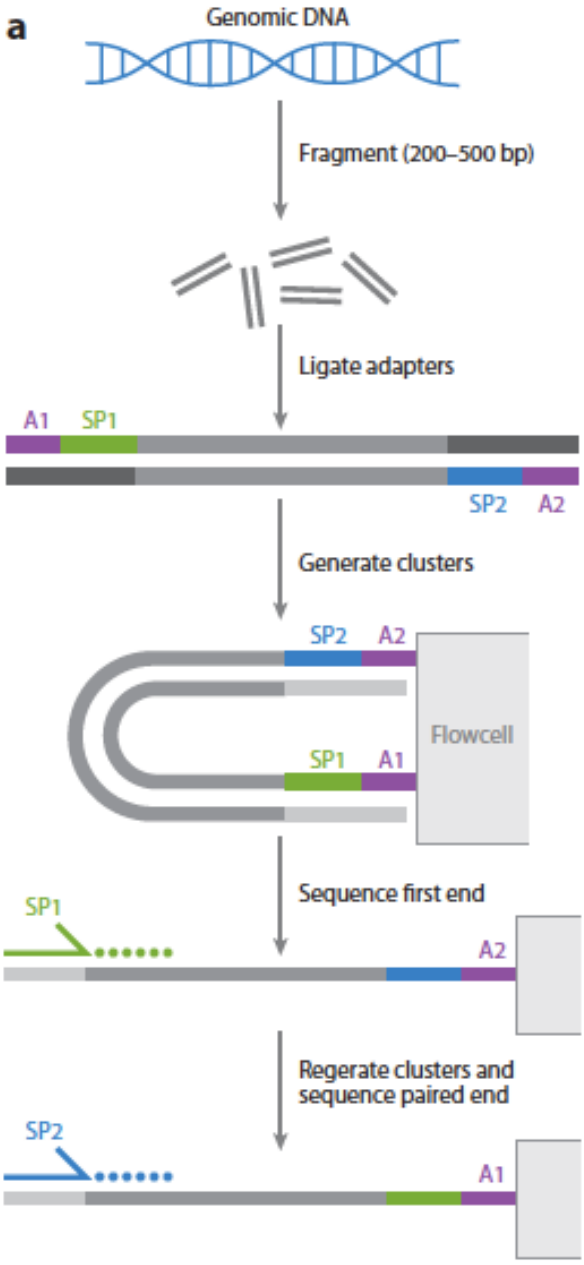


Figure 1: The Illumina sequencing-by-synthesis approach uses bridge amplification to generate clusters of identical sequence which can be sequenced from both ends using adapter-specific primers. Figure courtesy of Mardis, 2013 [44].

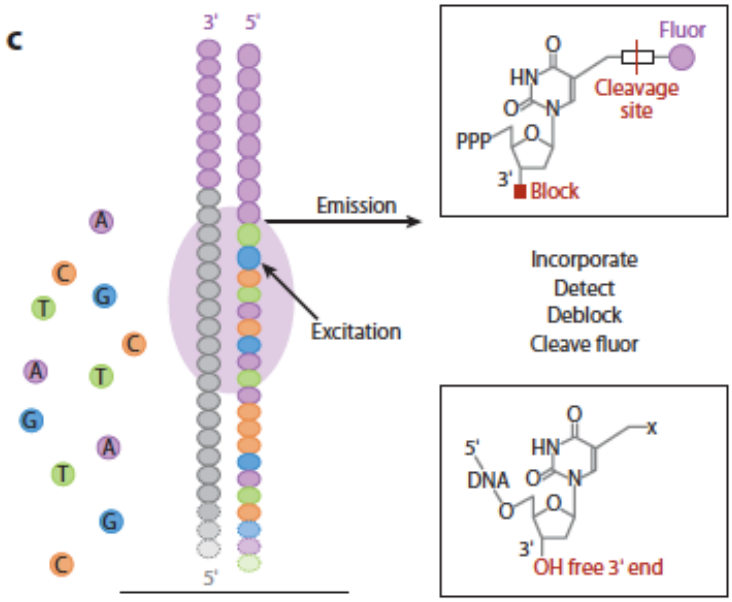


Figure 2: The Illumina sequencing-by-synthesis approach uses a reversible dye terminator to ensure unique nucleotide incorporation events. Following imaging, both fluorophore and terminator are cleaved, enabling the next round of incorporation. Figure courtesy of Mardis, 2013 [44].

AKAP Proteins: Facilitators of efficient signal transduction

Complex, differentiated multicellular life arises, in part, out of the ability of individual cells to coordinate their development, growth, and death. The adenosine 3′,5′-cyclic monophosphate(cAMP)-activated protein kinase A (PKA) signal transduction pathway was the first signal cascade to be elucidated, and since then, the vital role signal transduction cascades play in coordinating intra- and intercellular communication has become increasingly apparent [13]. Since its initial description in 1968, the mechanism by which PKA activation occurs has been exhaustively documented and is considered to be the most extensively characterized signal transduction cascade [17]. As such, a thorough description of the PKA signal transduction cascade here will provide a productive foundation upon which to base further discussion of signal transduction more generally in the context of research in the Ward lab.

An important component of signal transduction is protein phosphorylation (or dephosphorylation). The ready availability of adenosine triphosphate facilitates the use of this simple system ubiquitously across cell types for diverse signals. The phosphorylation state of a protein can increase or decrease the biological activity of an enzyme, facilitate or inhibit movement between subcellular compartments, stabilize or mark for destruction a protein product, or initiate or disrupt a protein-protein interaction [13]. Most eukaryotic processes are regulated at some step by phosphorylation of protein substrates by kinases. In humans, there are at least 518 protein kinases, divided into typical (478) and atypical (40) kinases. Typical kinases are classified into two main groups, based on the amino acid residues they phosphorylate: serine/threonine kinases (388) and tyrosine kinases (90) [19].

PKA is a holoenzyme composed of two regulatory subunits (RI and RII) and two catalytic (C) subunits. The C subunit is a broad-spectrum serine/threonine kinase [79]. The active site of the catalytic domain of PKA (conserved across typical kinases) consists of an N-terminal lobe composed of a β-sheet and a single α-helix opposite a C-terminal α-helical lobe. Sandwiched together, these two lobes create the ATP binding pocket. In the inactive state of a typical kinase, a phenylalanine residue blocks the active site. In its active state, an aspartate residue chelates Mg^{2+} to orient the ATP substrate in the ATP binding pocket with the triphosphate group pointing out of the pocket to facilitate the transfer of the γ-

phosphate to the peptide substrate. In all typical kinases, the target hydroxyl group on the substrate peptide (on serine, threonine, or tyrosine) is oriented towards a catalytic aspartate, in a position that facilitates the transition of the γ-phosphate to the peptide substrate [19].

However, PKA does not phosphorylate in a vacuum. cAMP must first bind two sites on the R subunit, inducing extensive conformational changes that result in the dissociation of the C subunit and the R subunit, and the subsequent activation of the C subunit [34]. In the archetypal PKA signal transduction cascade, G-protein-coupled receptors at the plasma membrane bind to a ligand – often an intercellular signalling molecule such as adrenocorticotropin, glucagon, or adrenaline. G_s-protein activated adenyl cyclases synthesize cAMP from ATP while phosphodiesterases (PDE) hyrolyze cAMP to AMP, attenuating the cAMP signal [79], [5].

In the 1980's, the PKA research community began to understand that cAMP does not diffuse passively through the cytosol to bind to its target molecules (eg PKA). Rather, cAMP production and its range of action localize to specific subcellular compartments [13]. Unsuprisingly, cAMP-activated protein kinases and phosphatases – two families of proteins that work in tandem to tightly regulate the phosphorylation state of target proteins – colocalize to those same subcellular compartments as well.

Further investigation into the localization of cAMP signal transduction cascades led to the discovery of A-kinase anchoring proteins (AKAP), a family of at least 50 structurally diverse proteins classified by their ability to copurify with PKA catalytic activity [79]. The proteins of the AKAP family share three common characteristics. First they contain a PKA anchoring domain that binds to N-terminal dimerization domain of the R subunit of PKA with an amphipathic helix of 14-18 residues. Second, AKAPs possess a unique targeting domain for specific subcellular compartments. Several AKAPs may be targeted to the same subcellular compartment, or conversely, different isoforms of the same gene for an AKAP may be uniquely targeted. Third – and perhaps most significant from a biological standpoint – AKAPs colocalize simultaneously with multiple enzymes capable of signal transduction (e.g. kinases, phosphatases, and phosphodiesterase) [5].

These properties taken together suggest a dizzying, combinatorial array of possibilities for coordinating diverse signal transduction cascades. Research up to this point suggests

that a single AKAP may play tissue – and perhaps even subcellular compartment-specific – roles in coordinating cellular communication. Localization of AKAP complexes can be modulated by competitive binding, modification of the targeting domain, or phosphorylation of the AKAP, which in turn provides a mechanism for the dynamic recruitment of binding partners and subsequent regulation of enzyme activity. Additionally, though the cAMP signal transduction pathway has been used as an illustrative example in this introduction, it is important to keep in mind that the integrative roles AKAPs play in signal transduction is not limited to cAMP-driven cascades [5].

AKAP9 in focus: a literature review

In 1998, Lin *et al.* screened human brain cDNA clones for interactions with the NMDA receptor subunit NR1. Specifically, they examined the C-terminal tail, which contains three exon cassettes. They discovered a novel 272 amino acid protein - named "yotiao" by the investigators - that interacted with the NR1 subunit in a C1-exon-cassette-dependent manner. 5' RACE identified a 5.1 kb open reading frame, and subsequent Northern blot analysis revealed an 11 kb yotiao transcript strongly expressed in skeletal muscles muscle and pancreas, and weakly expressed in brain, placenta, and heart tissues [39]. A separate study confirmed expression of this 11 kb transcript in the brain and heart, and showed expression of the transcript in the thyroid and testes [22]. The fractionation profile of yotiao from rat brain extract resembles a cytoskeleton-associated protein. Anti-yotiao antibodies revealed yotiao expression in the rat cerebellum, cortex, and CA1 pyramidal cell layer. Staining patterns in the rat cerebral cortex and the CA1 region of the hippocampus were consistent with synaptic localization [39], and a NR1-PKA-yotiao complex coimmunoprecipitates from synaptosomal membranes [22]. In cortical pyramidal neurons, yotiao colocalized with the NR1 subunit. At rat neuromuscular junctions, yotiao showed concentrated colocalization with acetylcholine receptors [39].

A follow up study by Westphal *et al.*(1999) confirmed that yotiao binds the NR1A subunit of the NMDA receptor. Yotiao contains a 17-residue sequence with the essential amino acid residues for RII interaction; expression of full-length yotiao fused to GFP confirmed that a 210-kD protein bound the RII subunit of PKA. Additionally, protein phos-

phatase 1 (PP1) co-precipitated with yotiao from rat brain extracts. Yotiao binding to PKA and PP1 does not inhibit their activity, suggesting a scaffolding function. PP1 activation reduced current through the NMDA receptor and PKA activation increased current through the NMDA receptor, suggesting that yotiao functions to localize PP1 and PKA in proximity to NMDA receptors for rapid modulation of NMDA receptor current flow [75]. Indeed, yotiao may be the limiting quantity determining PKA-NR1 interaction: increasing yotiao concentration artificially increases PKA activation [22].

In addition to PKA and PP1 interaction domains, yotiao contains several leucine-isoleucine zipper motifs [58], one of which mediates the interaction between yotiao and the Type 1 Inositol 1,4,5-triphosphate receptor ($InsP_3R$), an intracellular calcium (Ca^{2+}) release channel that plays an important role in intracellular Ca^{2+} signalling [69]. Yotiao has also been found to associate with the outward potassium channel of the slow delayed rectifier current $(I_{Ks})^2$ KCNQ1-E1. Co-immunoprecipitation studies suggest that yotiao complexes with KCNQ1-E1 and AC2 and AC9 in cell culture [38].

Approximately 50% of yotiao's coding region overlaps with the exonic regions of AKAP350, suggesting that yotiao is a 5' splice variant of AKAP350. Additionally, AKAP350 appears to undergo tissue-specific alternative splicing; northern blot analysis revealed an 11 kb transcript and a 9.5 kb transcript. Both transcripts were found in human kidney and skeletal muscle, while only the 9.5 kb transcript was detected in the liver, and only the 11 kb transcript was detected in the heart and brain. In rabbit tissues, the larger transcript was found in gastric wall smooth muscle, spleen, heart, forebrain, hindbrain, and cerebellum. In MDCK cells (a well established polarized kidney cell line), AKAP350 co-localized with the RII subunit of PKA, and with γ-tubulin at the centrosome. AKAP350 staining in the testes revealed AKAP350 localized to the basal membrane and the adluminal compartment [58].

In non-polarized HCT116 cells, AKAP350 localizes to distinct sub-cellular locations during the cell cycle. During interphase, AKAP350 localizes with γ-tubulin, while during metaphase, AKAP350 localizes to the centrosomal poles. During anaphase, AKAP350 localized to centrosomal poles, but also exhibits diffuse cytosolic immunostaining and localizes to the forming cleavage furrow. Localization to the poles continues and localization to the cleavage furrow strengthens during telophase. These results paint a picture of a protein with

dynamic, cell-cycle-dependent localization, with a strong affinity for the centrosome [58].

A separate study identified a 3,899 amino acid, 451 kDa protein - named CG-NAP - present in neuroblastoma, HeLA cells, and the human hippocampus that interacts with PKN, coimmunoprecipitates with PKA, and contains the conserved RII binding motif from hAKAP350 and rAKAP120. Additionally, CG-NAP interacts with the catalytic subunit of PP1 and with PP2A through its regulatory subunit PR120 [65].

The N-terminus of hAKAP450 targets to the Golgi apparatus, and association of hAKAP450 and the Golgi protein GM130 is essential for microtubule nucleation at the Golgi apparatus. In immortalized human pigment epithelial and immortalized nonmalignant human breast epithelial MMCF10A cells, AKAP450 binding to the Golgi apparatus interferes with directional migration and the formation of primary cilium [31]. The long AKAP9 isoform (AKAP450/AKAP350/CG-NAP) is the predominantly expressed isoform in human umbilical endothelial cells, and reduced Akap9 expression in these cells inhibits microtubule growth. Additionally, this isoform is important for Epac-promoted adhesion [60]. AKAP350 recruits PKA to the Golgi apparatus and subapical centrosomes in HepG2 cells [45]. Terrin *et al.* suggest that PDE4D3 recruitment by AKAP450 to the centrosome lowers basal cAMP concentrations at the centrosome compared to the cytosol, facilitating AKAP450-bound autophosphorylation by PKA, which in turn increases the cAMP sensitivity of PKA [66]. AKAP350 facilitates the initiation of DNA synthesis by scaffolding Cdk2 to the centrosome at the G_1/S transition [46].

In summary, the isoform yotiao localizes important components of intracellular signalling to various receptors in the cell membrane, while longer isoforms of yotiao are associated with microtubule spindle formation, the Golgi apparatus, and the centrosome. The length of the gene and the large number of exons suggest extensive alternative splicing, and the presence of a similarly sized 11 kb transcript suggests that the full-length gene may be transcribed in many different cell and tissue types before it undergoes alternative splicing.

Disruption of gene splicing of *Akap9* in the *mei2.5* mouse strain

The *mei2.5* mouse model was generated by Ward *et al.* (2003) via ethylmethansulfonate mutagenesis while conducting a forward genetic screen for infertility phenotypes that would

elucidate the underlying genetics of mammalian reproduction. The induced mutation labeled as mei2.5 was characterized by an infertility phenotype in homozygous recessive (-/-) males. Histological analysis revealed disorganized structure in the seminiferous tubules, mislocalized spermatocytes, and an absence of round spermatids and luminal spermatozoa, all suggesting a disruption of gametogenesis [74].

Genetic mapping followed by sequencing identified the gene *Akap9* on Chromosome 5 as the gene containing the causative mutation, a G to A transition at a conserved splice site at the exon 13 and intron 13-14 boundary. As a result of the splice site disruption, a stop codon in the intronic region is believed to truncate protein isoforms containing exon 13 [57].

Further immunohistochemical analysis revealed a normal progression of spermatocyte development up to prophase I, but aberrantly localization of spermatocytes to the adluminal compartment. IHC with the Sertoli cell marker WT-1 revealed a significant increase in the number of Sertoli cells in the seminiferous tubules of -/- mice. Additional analysis with the antibodies raised against the negative cell cycle regulator $p27^{Kip1}$ (a marker of cell maturity) revealed an absence of $p27^{Kip1}$ expression in the seminiferous tubules of -/- mice. Coupled with an RT-PCR analysis that reveals increased expression of markers for immature Sertoli cells in -/- males compared to wild-type males, these results suggest that Sertoli maturation is disrupted in mei2.5 male mice. IHC staining for the junctional complexes ZO-1 and Cx43 reveals consistent mislocalization in -/- seminiferous tubules, suggesting a possible mechanism for the alteration of Sertoli cell maturation pathways. Schimenti *et al.* put forward a proposed model that suggests ZO-1 binds Cx43 and AKAP9, targeting Cx43 to the plasma membrane. Disruption of AKAP9 function by the mei2.5 could lead to disrupted gap junction formation and function in Sertoli cells, affecting their maturation, which in turn affects their capabilities as support cells in spermatogenesis [57]

Feuer suggests that the gap junction communication between Sertoli cells is disrupted, and immunohistochemistry conducted by Bovet on mouse embryonic fibroblasts supports this hypothesis [23]. Isoforms of AKAP9 colocalize with Cx43 around Golgi and centrosomal structures, and a separate isoform appears to colocalize at gap junctions in cells. [8]. Though preliminary RT-PCR results suggest that AKAP9 is disrupted by the mei2.5

mutation in a teste-specific and developmentally regulated manner, the somatic nature of the disruption poses the question: does the mei2.5 mutation disrupt AKAP9 expression in any other tissues, resulting in as of yet undetected phenotypes?

Investigating hippocampal gene expression in the *mei2.5* model

Disruption of alternative splicing can cause disease directly, modify the severity of the disease, or be linked to disease susceptibility. Up to 50% of disease-causing mutations have been shown to affect splicing [70]. Given that the *mei2.5* mutation disrupts a conserved splice site in a gene that displays tissue-specific alternative splicing, a more complete characterization of the *mei2.5* model requires expanding inquiry beyond the testes. Both the yotiao and CG-NAP/AKAP350/AKAP450 isoforms of AKAP9 have been detected in the brain, suggesting it as a tissue beyond the testes worth investigating. The enormous complexity of cell and tissue types suggest ample opportunities for alternative splicing, and the *mei2.5* model could provide insight into the mechanisms by which alternative splicing functions in the brain.

AKAP9 and Cx43 have not been linked directly by co-immunoprecipitation. However, Feuer and Bovet provided evidence for an association of an expressed isoform of Akap9 and Cx43 at gap junctions. Interestingly, undifferentiated neural progenitor cells isolated from the mouse striatal germinal zone and grown *in vitro* differentiate into three main cell types: astrocytes, neurons, and oligodendrocytes. Undifferentiated cells within the neurosphere express unphosphorylated Cx43 and display coupling, while cells with astrocytic morphology in the neurosphere's outgrowth regions express phosphorylated Cx43. Cells exhibiting neuronal and oligodendritic morphology are uncoupled, and inhibiting gap junctions reduces neural progenitor cell viability and alters morphology of differentiated cells [18]. Additionally, Cx43 was the most abundant connexin transcript in radial glial (RG)-like cells, and Cx30/Cx43 double knockout mice exhibited reduced neurogenesis in the dentate gyrus [36]. Taken together, these findings implicate Cx43 in hippocampal neurogenesis.

Lin *et al.* initially characterized the yotiao isoform of Akap9 as an NR-associated protein expressed in the brain, and immunohistochemistry with α-yotiao antibodies in coronal sections of the rat brain reveal labelling of the CA1 pyramidal cell layer in the hippocampus [39]. Activation of NMDA receptors induces long-term potentiation and long-term

depression at synapses between CA1 and CA3 pyramidal neurons [42].

Taken together, these findings suggest that AKAP9 may play a role in either long-term potentiation or neurogenesis in the hippocampus, making the hippocampus and attractive region of the brain to explore. In this study, we conducted RNA-seq on six mice from the *mei2.5* line to further elucidate the role of AKAP9 in the brain.

Materials and Methods

Mouse Colony Maintenance

The *mei2.5* strain is on a C57BL/6J background. Standard animal husbandry practices as established by Middlebury's IACUC and described in IACUC protocol #220-15 were followed. Mice were housed in One Cage ventilated 8" x 7" x 12" (W x H x L) cages in a RAIR HD One Cage High Density Ventilated Rack (Lab Products, Inc, Seaford, DE) on 7099 Tek-Fresh laboratory animal bedding (Envigo, Indianapolis, IN). Cages contained enrichment in the form of nestlets (Ancare, Bellmore, NY), Enviro-dri crinkled paper, shepherd shacks (Shepherd Specialty Papers, Milford, NJ), and Kimtech Science Kimwipe Delicate Task Wipers (Kimberly-Clark, Irving, TX). Mice were fed 2020X Teklad global soy protein-free extruded rodent diet (Envigo, Indianapolis, IN) and water *ad libitum*.

Genotyping

Earclips taken at weaning on post-natal day 21 (p21) were taken for genotyping using a 2 mm ear punch (Fine Science Tools, Foster City, CA) rinsed with 70% EtOH. DNA was extracted from the earclips using the HotShot method [68]: 75 μL of HotShot buffer (25 mM NaOH, 0.2 mM EDTA) were added to a 1.5 ml Eppendorf tube containing the earclip. The tubes were incubated for 45 min at 95°C on a heating block, and disrupted by flicking every 15 minutes. The solution was then frozen for at least 1 hour (or until the solution had frozen solid) at -20°C, before thawing and neutralization with 75 μL of neutralization buffer (40 mM Tris-HCl). Samples were either used directly as a template in subsequent PCR reactions or stored at -20°C.

Genotype was established with an allele-specific PCR reaction previously developed

Primer name	Primer sequence
mei2.5_allele_specific_F	5'-GGATAGAGTAATCTTGACTCA-3'
mei2.5_allele_specific_WT_R	5'-CGGCAGCATGTGCATACC-3'
mei2.5_allele_specific_Mut_R	5'-CGGCAGCATGTGCATACT-3'
mei2.5_sequencing_F2	5'-GTGATCCTAAGCTATGAGG-3'
mei2.5_sequencing_R2	5'-AGGCACAGAATACACG-3'

Table 1: Primers used for allele-specific genotyping PCR reactions and sequencing

in this lab [23]. 10X Standard *Taq* Reaction Buffer, 10 mM dNTPs (New England Biolabs, Ipswich, MA), 10 μM allele-specific forward and reverse primers (Table 1)(Eurofins MWG Operon Genomics, Louisville, KY), nuclease free water (Qiagen, Germantown, MD), and 5000 units/ml *Taq* DNA Polymerase (New England Biolabs, Ipswich, MA) were combined with 1 μL of HotShot extraction product in a 25 μL final volume PCR reaction (Final concentrations: 1X Standard *Taq* Reaction Buffer, 200 μM dNTPs, 0.4 μM forward and reverse primers, 0.04 units/μL *Taq*). PCR reactions were run in an S1000 Thermal Cycler (Bio-Rad, Hercules, CA) using the following protocol: initial denaturation for 3:00m at 94°C, followed by 30 cycles of a 0:30s denaturation step at 94°C, a 0:30s annealing step at 55°C, and a 1:00m extension step at 72.00°C. A final extension step for 7:00m at 72°C is followed by a 4.0°C hold.

Genotypes were confirmed by sequencing. The region of interest in *Akap9* was amplified using 10 μM sequencing primers (Table 1), and DNA from the HotShot extraction protocol. 10X High Fidelity PCR Buffer, 50 mM $MgSO_4$ (Thermo Fisher Scientific, Waltham, MA), nuclease free water, 10 mM dNTP mix, 1 μL of sample DNA, and 5 units/μL of Platinum® *Taq* DNA High Fidelity Polymerase in a 50 μL reaction (Final concentrations: 1X High Fidelity PCR Buffer, 0.2 mM each dNTP, 2 mM $MgSO_4$, 1 units/μL). PCR reactions were purified using the QIAquick® PCR purification (Qiagen, Germantown, MD, Cat. No. 28104) according to the kit instructions, and sent for sequencing with Eurofins Genomics Sequencing Department (Louisville, KY).

Tissue Collection

Mice at p180 were euthanized by cervical dislocation followed by decapitation, in accordance with a Protocol 220-12, approved by Middlebury's IACUC. Whole brains were frozen in liquid RNA*later*® (Sigma-Aldrich, St. Louis, MO) chilled with dry ice until they achieved the consistency of stiff gelatin, then the left hippocampus was removed and stored in 1.3 ml of RNA*later*® on ice for the duration of the collection session. Dissection tools: Straight Sharp/Sharp scissors (12 cm), delicate pattern straight sharp/sharp scissors (9 cm), Straight Sharp/Blunt Scissors (18.5 cm), Serrated Graefe Forceps, tip width 0.8 mm (Fine Science Tools, Foster City, CA). All tools and surfaces were sterilized with 70% EtOH and RNase*Zap*®. Before storage at -80°C, the supernatant was drawn off (Protocol from Professor Clarissa Parker, Personal communication).

RNA Extraction

Tissue samples were sent to the University of Vermont's DNA Analysis Facility on dry ice, where 1 ml TRIzol® was added to each sample before tissue homogenization on a FastPrep® (MP Biomedicals, Solon, OH). Following homogenization, 100 μL of pure 1-Bromo-3-chloropropane (Sigma-Aldrich, St. Louis, MO) was added to each sample (Personal communication, UVM DNA Analysis Facility) [12], before proceeding with the standard TRIzol® Reagent protocol (Invitrogen, Carlsbad, CA, Cat. No. 15596-018). RNA concentration and quality were measured (See Supplemental table) on 2100 Expert Bioanalyzer according to standard protocols for the Eukaryote Total RNA Nano (Agilent, Santa Clara, CA).

RNA Sequencing

A total of 3 μg of total RNA for each sample was diluted in diethyl dicarbonate (DEPC) to a concentration of 30 ng/μL, and concentrations were verified using a Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA) and the quality verified on a Nanodrop. Samples were shipped to the New York Genome Center for RNA sequencing. An mRNA TruSeq Stranded, 175 bp library preparation protocol was used to construct a sequencing library [55] [82], and 50 bp paired-end reads were sequenced on on a HiSeq2500 to a depth of 30 million

reads. The New York Genome Center conducted a basic bioinformatic analysis using Star, FeatureCounts, and DESeq2 [1].

Virtual Machine and Pipeline Assembly

A cloud-based Red Hat Enterprise Linux 7 (RHEL 7), 64-bit virtual machine with 98 GB of storage and 16 GB of RAM with 8 CPUs, hereafter referred to as "Middgenpilot," was instantiated on a Middlebury College server through Middlebury's Information Technology Department. An additional 1 TB Alpaca storage volume was associated with this virtual machine. All packages were installed on the virtual machine and scripts were run from the "executing directory" of the virtual machine (See Appendix D for all scripts used in the analysis). All working files were stored in the Alpaca directory (Figure 3).

Data files were checked for corruption using md5sum. FastQC is a software package developed to provide a rapid summary of raw read quality [3]. Following quality verification by FastQC, the pre-processing program Trimmomatic was chosen to trim low-quality reads [7] [21]. The package RSeQC was chosen to perform annotation-based quality control [71]. Tophat 2 and Bowtie 2 were chosen for alignment to the Ensembl reference genome [2], Cufflinks was used for transcript assembly. Cuffmerge and Cuffdiff were used for for differential expression analysis. The successful installation of the "Tuxedo Suite" was verified using an *in silico* experiment contrived for pipeline validation. The experimental sequences were download from the Gene Expression Omnibus at accession GSE32038. A reference annotation of the *Drosophila melanogaster* genome was downloaded from Illumina's iGenome support site (Build BDGP5.25, `http://support.illumina.com/sequencing/sequencing_software/igenome.html`) [67]. The R packages CummeRbund and Gviz were used to visualize differential expression data [25] [28].

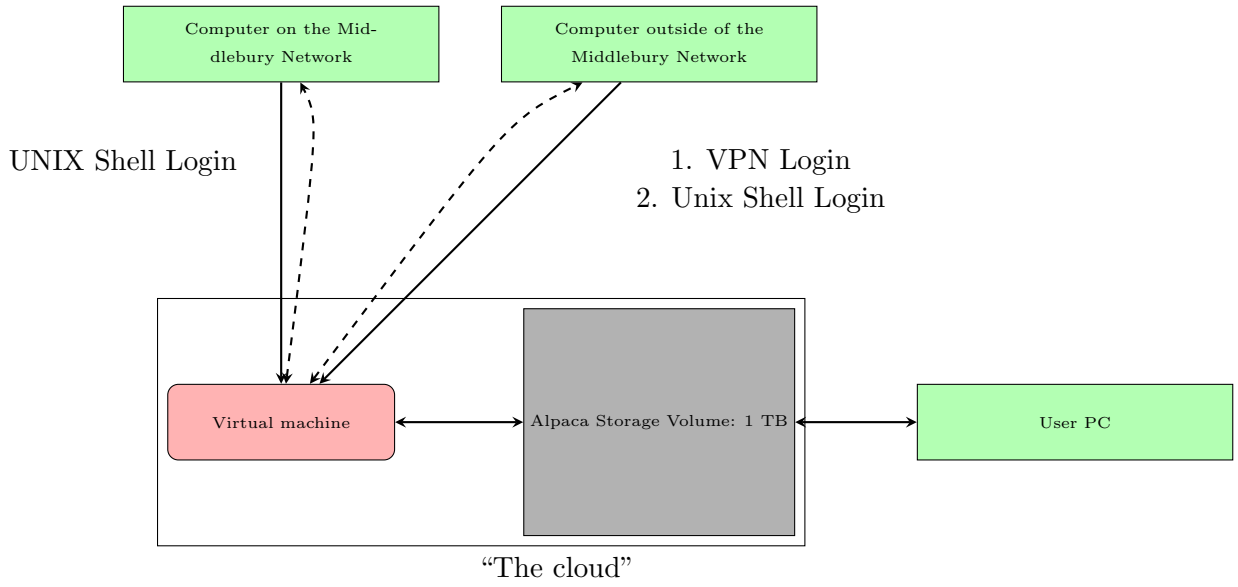
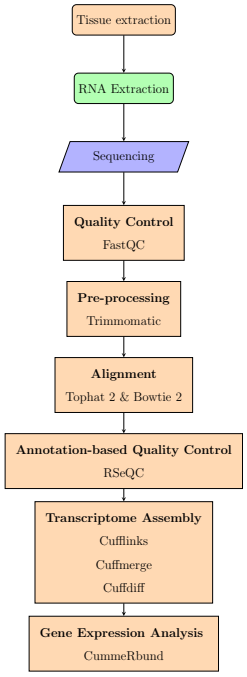


Figure 3: A pilot cloud-based bioinformatic workstation at Middlebury College. From a UNIX shell, a user accesses the virtual machine housed on Middlebury's servers (solid unidirectional arrows). Graphical information can be passed back to the user's personal computer (indicated by dashed lines). Users not comfortable with the command line may access the Alpaca storage volume from their personal desktop to add, view, or remove files on the Alpaca storage volume (horizontal bidirectional arrow), but do not have access to the virtual machine. When the VM is accessed from the command line, the Alpaca volume integrates seamlessly

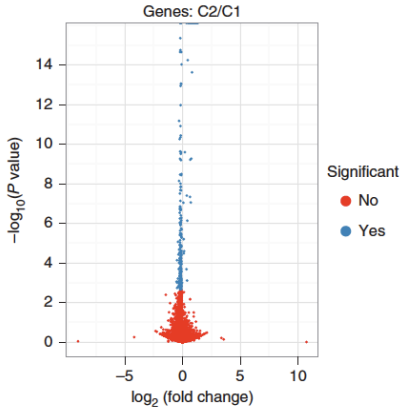
Results

Middgenpilot: A cloud-like bioinformatics workstation

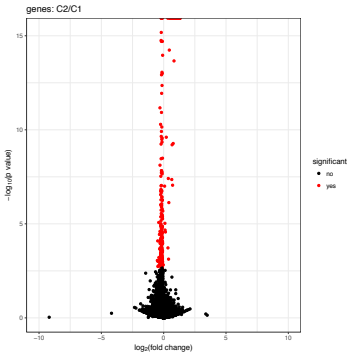
As a proof of concept of a cloud-based genomic and transcriptomic workstation at Middlebury College, an RNA-seq data analysis pipeline was installed on the virtual machine Middgenpilot. Successful processing and analysis of a test data set from a contrived *Drosophila melanogaster* experiment (as described by Trapnell *et al.* in [67]) verified the successful installation of all necessary packages, establishing the workstations capabilities for RNA-seq data analysis. A volcano plot plots the p-values of differentially expressed genes against the change in gene expression between conditions, giving a visual overview of differential expression between samples. Figure 4c displays a volcano plot constructed from *D. melanogaster* data processed in the pipeline, which matches the volcano plot provided by Trapnell *et al.* (Figure 4b). This visual effectively verifies the successful installation of the pipeline because every preceding piece must provide functional output to generate this figure. See Appendix D for bash scripts used to process the data in this study.



(a) RNA-seq workflow



(b) Trapnell *et al.* 2011 volcano plot



(c) Middgenpilot volcano plot

Figure 4: The Middgenpilot workstation effectively processes RNA-seq data. (a) The hippocampus was extracted at Middlebury College (yellow), then RNA was extracted at the University of Vermont (green) (rounded corners: biological samples) before sequencing at the New York Genome Center (blue) (trapezoid: biological to digital samples). Sequence analysis (yellow) was conducted on the Middgenpilot workstation. (b) The volcano plot published by Trapnell *et al.* displaying differentially expressed genes in a contrived *in silico* experiment matches (c) the volcano plot generated by the Middgenpilot workstation.

Experimental design

Hippocampi were extracted from six male mice aged to p180. For analysis, homozygous *mei2.5/mei2.5* (-/-) mice were binned in a "non-functional AKAP9" group, and heterozygotes (*mei2.5*/+ or +/-) and wild-type (+/+) were binned together in a "functional AKAP9" group (Table 2).

RNA quality

After extraction, a cDNA library was constructed and sequenced from RNA samples.

Sample ID	2197-4	2197-3	2197-1	2199-1	2199-7	2199-6
Genotype	mei2.5 -/-	mei2.5 -/-	mei2.5 -/-	mei2.5 +/-	mei2.5 +/+	mei2.5 +/-
rRNA Ratio	1.3	1.3	1.3	1.3	1.3	1.3
RIN	8.6	8.8	8.4	8.6	8.6	8.8
Concentration (ng/μL)	32.6	27.5	24.6	23	26.6	23.5
Nanodrop 260/280	2.05	2.02	2.06	1.94	2	1.99
Total RNA (ng)	2934	2475	2214	2070	2394	2115

Table 2: RNA quality data for sample submissions to the New York Genome Center

Data Quality

Two of the samples did not reach the expected number of reads in the first sequencing run, so they were resequenced by the NYGC to generate a second set of fastq files. These files were treated separately in FastQC and Trimmomatic and aligned separately to the reference genome, then the BAM files were merged for further downstream processing.

Quality control and pre-processing of reads

FastQC is an open source quality control program for verifying the quality of high-throughput sequencing data [3]. Each column in Figure 5 represents the FastQC test output for raw FastQ sequencing files returned by the New York Genome Center for each sample. Each test measures a different component of read quality.

The Per Tile Sequence Quality module takes advantage of Illumina's data encoding, which include the flow cell tile from which a read originates, enabling a tile-level view of

sequence quality. Illumina also returns Phred scores – quality scores based on the probability of an error for a given base call [20] – which are used by FastQC for quality contorl. The Per Tile Sequence Quality module will issue warnings if a tile shows a mean Phred score between 2 and 5 Phred units lower than the mean for that base across all tiles, and a failure if the mean Phred score for a specific tile falls more than 5 Phred units below the mean tile value across all tiles. FastQC issued eight warnings and two failures for our samples.

The Per Base Sequence Quality test displays a box whisker plot of the base quality score (Phred) at each sequence position for every sequence in the sample. A Phred score represents the certainty of any given base call; the higher the quality score, the more certain base calling program was in establishing the identity of that base. FastQC will issue a warning if the lower quartile for any base is less than 10 or if the median for any base is less than 25, and a failure if the lower quartile for any base is less than 5 or the median score is less than 20. None of the samples were issued a base quality warning,

The Per Sequence Quality Score module takes a slightly different approach to looking at sequence quality. The module examines the distribution of the mean Phred score across all reads for the sample. If the most frequently observed mean is below a 27, the module issues a warning; If the most frequently observed mean is below 20, the module issues a failure. Warning and failures of this module would indicate overall low sequence quality; however, none of the samples were issued a warning or a failure.

The Per Base Sequence Content measures the relative ratio of bases at each position in the read. The module assumes a random, unbiased library in which each base is represented roughly equally, and issues a warning if A and T or G and C frequencies differ by more than 10%. The module issues a failure if the difference between A and T or G and C frequencies differ by more than 20%.

The Per Sequence GC content module measures the GC content of each read and compares it to a normal distribution of GC content modeled from the sequence data. The module gives a warning if the sum of the deviations from the normal distribution represents more than 15% of the reads, and a failure if the sum of the deviations from the normal distribution represents more than 30% of the reads. All samples received warnings from this module.

When a sequencer is unable to make a base call, it fills in that position in the sequence with an N. A high percentage of Ns in sequence files reflects low-quality samples. The Per Base N Content module issues a warning if any position has an N content greater than 5%, and a failure if any postion has an N content greater than 20%. All samples passed this module, suggesting that our sample quality was not compromised.

The Sequence Length Distribution module returns a distribution of sequence lengths, and it raises a warning if not all the sequences are the same length or any sequence have zero length. We expected 50 bp paired end reads, and all of the reads returned were 50 bp long; there were no warnings or failures for this module.

The Duplicate Sequences module tracks the first 100,000 sequences in a FastQ file, and scans the rest of the file for duplicates of those sequences. It reports the number of duplicates for each sequence as a percent of the total number of sequences. If non-unique sequences make up more than 20% of the sequence file, this module issues a warning; if non-unique sequences make up more than 50% of the sequence file, this module issues a failure.

The Overrepresented Sequences module lists all of the sequences that make up more than 0.1% of the total number of sequences; it lists a warning if any sequence makes up more than 0.1% of the total, and a failure if any sequence makes up more than 1% of the total number of sequences. None of our sequences contained over-represented sequences.

The Adapter Content module searches sequences for known Illumina, Nextera, and SOLID adapter sequences, and issues a warning if any sequence is present in more than 5% of reads, and a failure if any sequence is present in more than 10% of reads. FastQC failed to find any adapter sequences in our sample sequences, and every sequence passed this module.

The Kmer Content module looks for positional bias of k-mers. The module issues a warning if any kmer is imbalanced with a binomial p value < 0.01. The module issues a failure if any kmer is imbalanced with a binomial p value $< 10^{-5}$. Fifteen of our samples triggered a failure on this module, with one exception, which triggered a warning.

Pre-processing reads to eliminate low quality reads increased the number of reads that uniquely align to the reference genome [7]. To maximize the number of uniquely aligned reads, raw reads were processed using the ILLUMINACLIP and MAXINFO trimmers of the

Java-based filtering program Trimmomatic. The ILLUMINACLIP command trims standard Illumina adapter sequences from reads. The MAXINFO command trims low-quality reads with increasing stringency over the length of the read, with the goal of maximizing the number of reads that may be used during alignment[7]. The specific parameters used with each parameter can be found in Appendix A. Less than 0.01% of reads were trimmed for low quality or adapter contamination, and the number of reads to be aligned exceeded 30 million for all samples. However, not all samples had the same number of reads: Samples 2197-1, 2197-3, and 2197-4 (all members of the non-functional group) had more reads than samples 2199-1, 2199-6, and 2199-7 (from the functional group) (Figure 6).

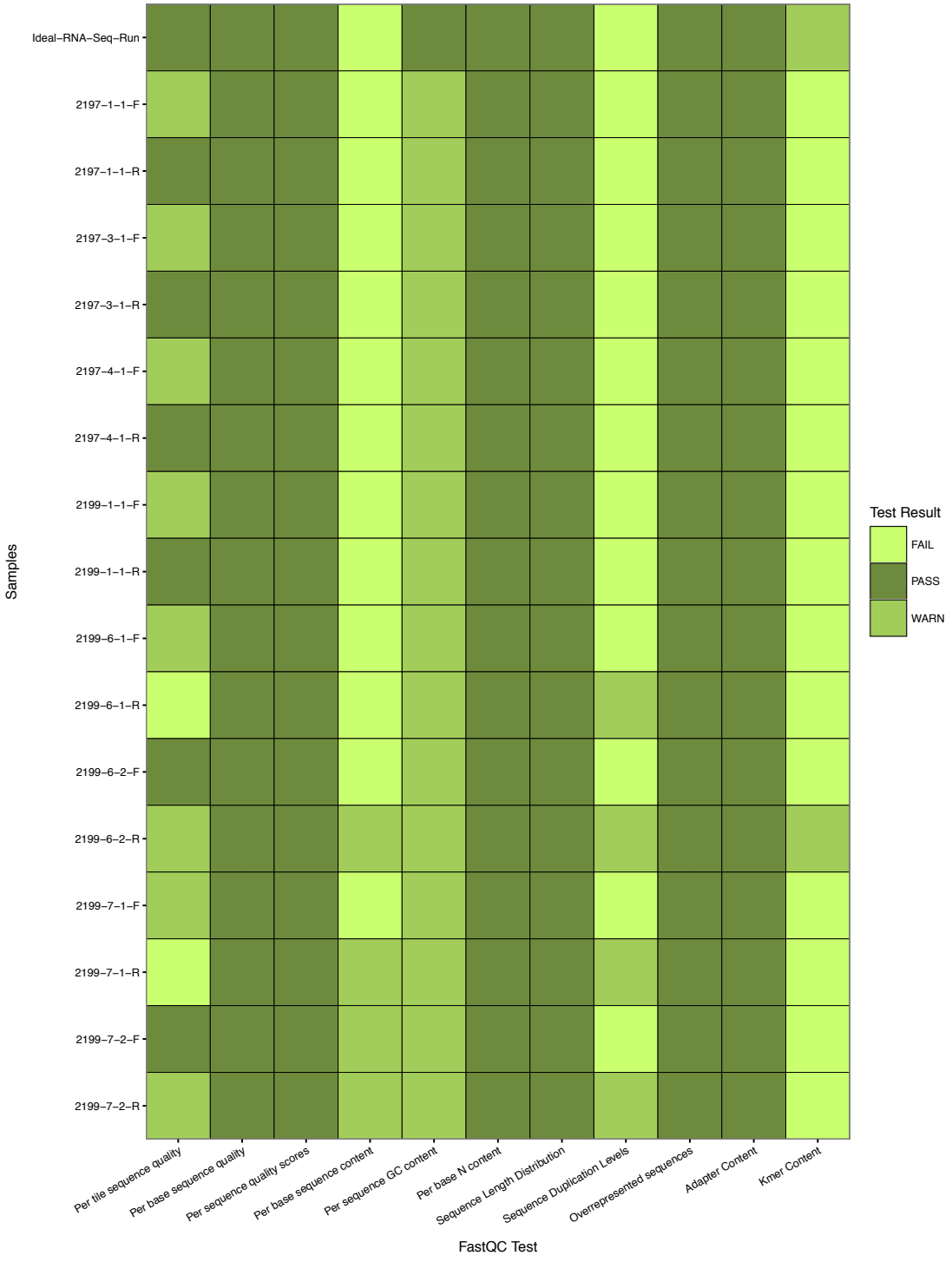


Figure 5: FastQC conducts twelve tests on measures of basic sequence quality. The row "Ideal-RNA-Seq-Run" represents the expected FastQC output for an RNA-Seq experiment with high-quality sequence reads

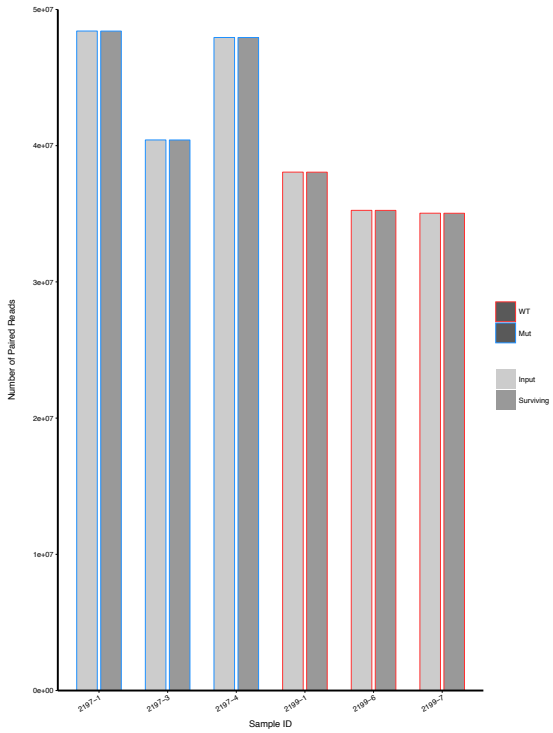


Figure 6: The majority of reads were kept for TopHat alignment after quality filtering with Trimmomatic. Input: all raw paired-end reads received from the NYGC. Surviving: All of the read pairs that were kept after adapter and base-score quality trimming

Annotation-Based Quality Control

Following alignment with Tophat 2, sample files underwent a second round of quality control, based on annotation features from the reference genome. Using the Samtools "idxstats" tool, the number of reads that map to each chromosome was counted (Figure 7) The non-functional group had consistently more reads mapped to each chromosome, which is to be expected given that the non-functional group had more reads per sample overall. However, the pattern of variation between chromosomes was consistent.

The RSeQC package consists of a variety of Python programs that can be used to assess data quality in RNA-sequencing experiments [35], [71]. The package read_distribution.py counts the number of tags per kilobase covering important genomic features. "Tags" are an indicator of read alignment. A uniquely mapped read will have a single tag associated with it; a split read will have two tags associated with it, as each mapping location will have a tag associated with it. Figure 8 demonstrates that more reads mapped to exons and the 3' untranslated regions (UTRs) than the 5' UTRs. Introns, the genomic regions upstream of transcription start sites (TSS), those regions downstream of transcription end sites (TES) should not have many tags in an mRNA sequencing experiment. The data show that most reads map to either exons or UTRs, with very little intronic or intergenic mapping.

The package geneBody_coverage.py scales assembled transcripts to 100 nucleotides and calculates the theoretical number of reads that cover each theoretical nucleotide, converting that number to a coverage ratio. Figure 8 reveals consistent 3' bias across all samples.

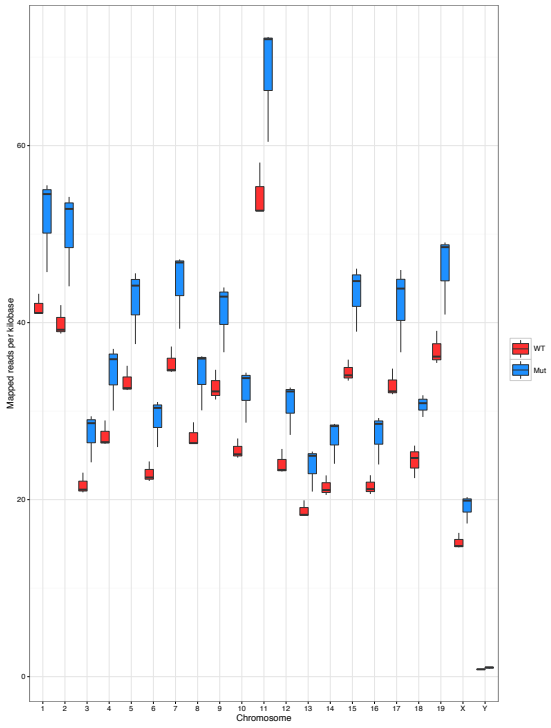
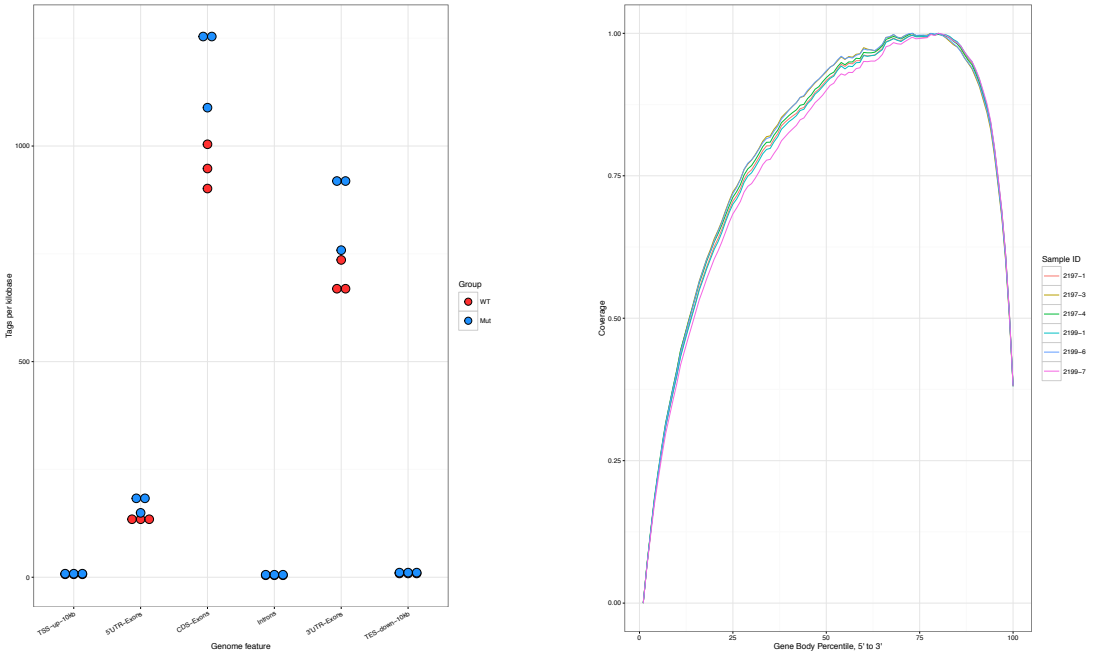


Figure 7: Read mapping distribution (measured by reads mapped per kilobase) between chromosomes varies, the Mut group, which had more reads, contains more reads mapped per kilobase. Read mapping was counted using the Samtools package "idxstats."



(a) **(b)**

Figure 8: Mapped reads display consistent 3' bias across samples. (a) A plot of tags per kilobase for genomic features by experimental grouping reveals bias towards the 3' UTR over the 5' UTR. TSS-up-10kb: The 10 kb region upstream of the transcription start site. 5'UTR-Exons: The 5' untranslated region of exons. CDS-Exons: Reference-annotated exons. 3'UTR-Exons: The 3' untranslated region of exons. TES-down-10kb: The 10 kb region downstream of the transcription end site. (b) A plot of the read coverage over the scaled length of the gene body reveals 3' bias across all samples

Hippocampal Gene Expression

Overall Gene Expression

The software package Cuffdiff creates an output directory that contains all of the necessary files to conduct differential expression analysis between experimental groups using the R package cummeRbund [25]. Gene expression is expressed in fragments per kilobase of exon per million of reads mapped (FPKM), a within-sample normalization method [48] [76]. The squared coefficient of variation (CV^2) plotted along a log scale of FPKM values is a measure of cross-replicate variability for the samples; differences in variability between samples may result in fewer than expected differentially expressed genes (due to wider confidence intervals). Figure 9 shows that cross-replicate variability is similar between samples, and tends to drop for both genes and isoforms as FPKM values increase. Figure 9c displays the density of gene expression over ten FPKM orders of magnitude. The majority of genes are expressed between 0.001 FPKM and 1000 FPKM, with a bimodal density distribution. The group with non-functional AKAP9 ("Mut"), has a slightly higher peak around 0.1 FPKM, suggesting that more genes are expressed at this lower level. At the lower peak around 10 FPKM, the group with functional AKAP9 has a slightly larger peak, suggesting that more genes are expressed at this level in this group. A large portion of genes have levels of expression (FPKM) that fall below 1 FPKM, suggesting that many genes are expressed at very low levels in the hippocampal extracts.

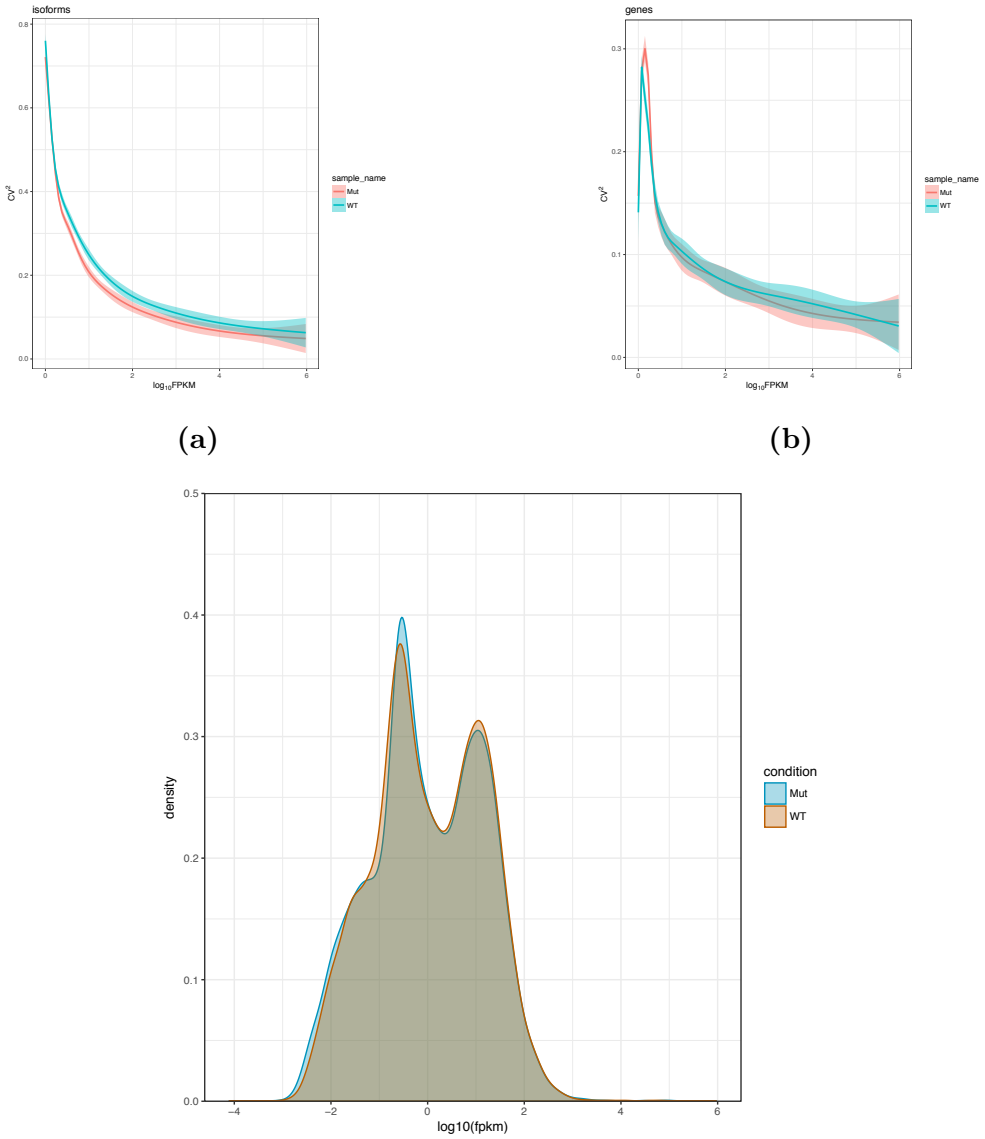


Figure 9: FPKM estimates are similar across experimental groups. The squared coefficient of variation, a measure of cross-replicate variability for (a) genes and (b) isoforms. (c) A density plot of FPKM values for each group.

Akap9 Expression

Visual inspection of the BAM files generated by TopHat suggest very low coverage of the *mei2.5* mutation site (Appendix A). However, from the sequencing data, Cuffdiff was still able to generate a robust assembly. Cuffdiff's transcriptome assembly was guided by the Ensembl reference genome which contains 15 predicted transcripts; however, Cuffdiff predicted two novel isoforms of Akap9 in the 3' end of the gene, TCONS_00099633 and TCONS_00099642 (Figure 10). It is interesting to note that these isoforms both demonstrate alternative splicing patterns in the last four exons, a region of the gene that overlaps with exon 44 in the full-length gene and that contains the PACT interaction domain. ENSMUST00000200591 and ENSMUST00000198500 are two transcripts within the Akap9 region according to Ensembl, but they represent Wrd46, a retrotransposed pseudogene, and GM43031, a predicted gene, neither of which is associated with Akap9 at this juncture.

Figure 11 depicts Cuffdiff-predicted protein products for five of the expressed transcripts. TCONS_00099638 appears to correspond to P31955, and TCONS_00099640 appears to correspond to P31958. Interestingly, Ensembl has annotated TCONS_00099640 as a nonsense-mediated decay transcript [2]. The confidence intervals overlap for all FPKM values; therefore we cannot conclude from this study that the isoform expression of Akap9 transcripts differs between the the two experimental groups.

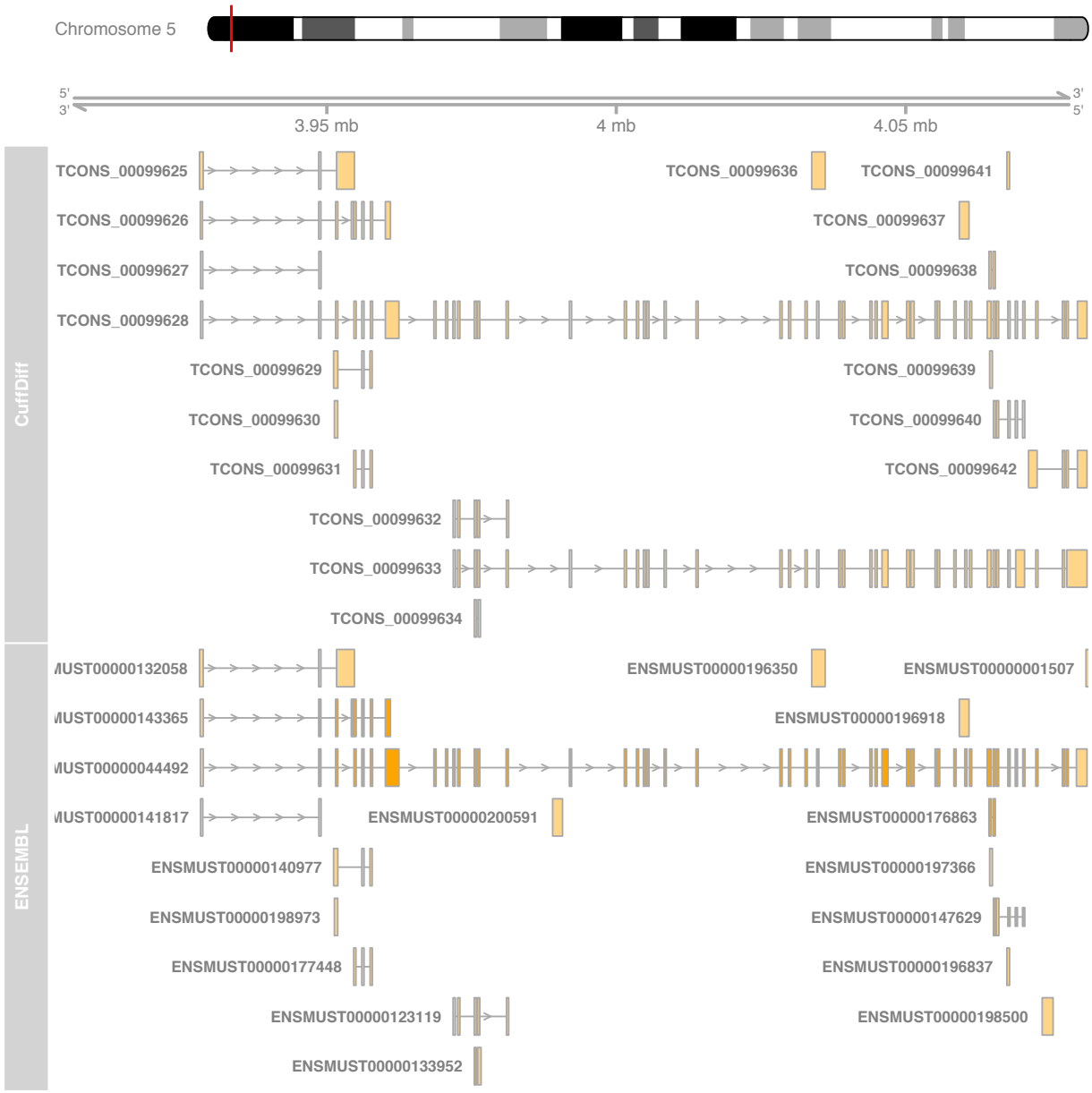
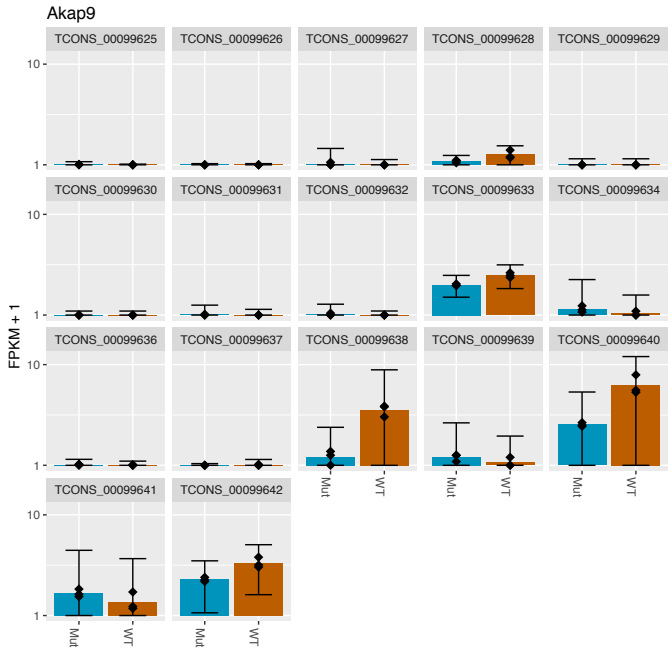
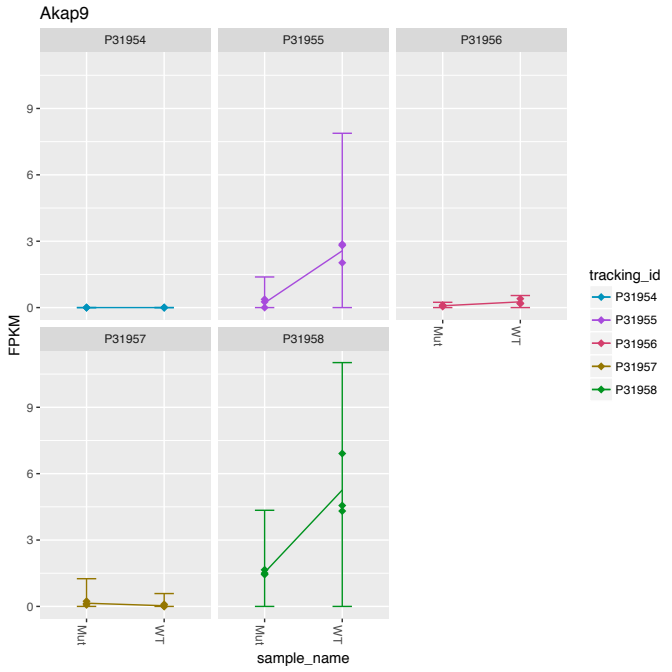


Figure 10: Cuffdiff predicts seventeen Akap9 isoforms in the hippocampus, two of which are not present in the mm10 Ensembl build.



(a)



(b)

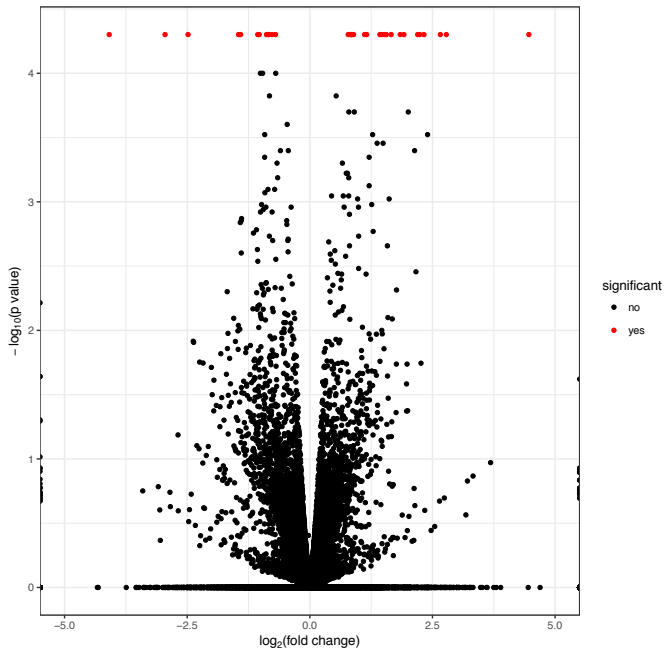
Figure 11: Several transcript variants of AKAP9 are expressed at low levels in the hippocampus, five of which Cuffdiff predicts as protein products. (a) expression barplots of seventeen Akap9 isoforms in the hippocampus, on a $\log_10$ scale. (b) Expression levels of transcripts predicted by Cuffdiff to have protein products. Error bars represent 95% confidence intervals for the FPKM value for the transcript.

Differential expression

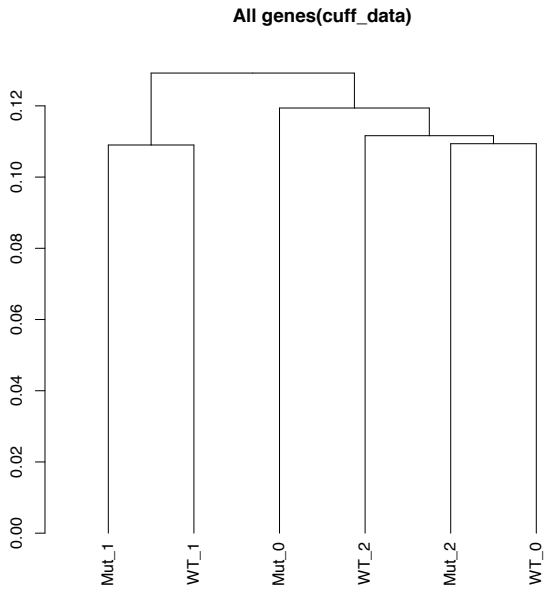
Figure 12 depicts a plot of the fold-change in gene expression between the two conditions plotted against the p-value (uncorrected for multiple testing) calculated for that difference. The volcano plot reveals a trend of difference between the two conditions in global gene expression, however, only 32 of the genes have differences in expression that are large enough to be considered significant by Cuffdiff after correction for multiple testing. (See Appendix B for a full table of genes). Interestingly, a dendogram drawn from Jensen-Shannon distances reveals that samples do not cluster by grouping, and appear to cluster randomly.

Figure 13 depicts a heatmap of differentially expressed genes deemed significant by Cuffdiff, clustered by grouping. At the top of the figure is a dendrogram depicting the clustering of the samples by expression levels of only significantly expressed genes (again, clustered by their Jensen-Shannon distances). The samples do not cluster by grouping and many genes have a single outlier that drives differential expression.

Of the differentially expressed genes, several interesting genes stand out, given the association of AKAP9 with microtubule organization and connection to Alzheimer's and long-QT syndrome. The sodium voltage-gated channel beta subunit 4 (Sbcn4) has been associated with long-QT syndrome, and displays upregulated expression in mice with non-functional AKAP9 (Figure 14) [59]. Amyloid Beta Precursor Protein Binding Family A Member 3 (Apba3) is thought to be a protein that interacts with amyloid precursor protein (thought to play an important role in Alzheimer's) [4]. Apba3 exhibited higher levels of expression in the group lacking functional AKAP9 (Figure 14) . However, as the heatmap in Figure 13 demonstrates, and the isoform plots further reinforce, often a single replicate drives the FPKM value up for a specific isoform.

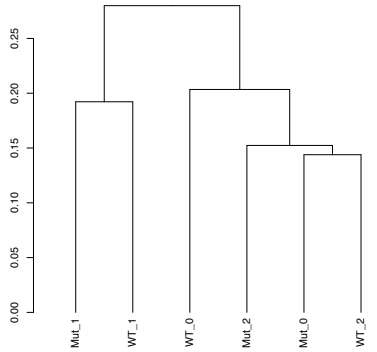


(a)

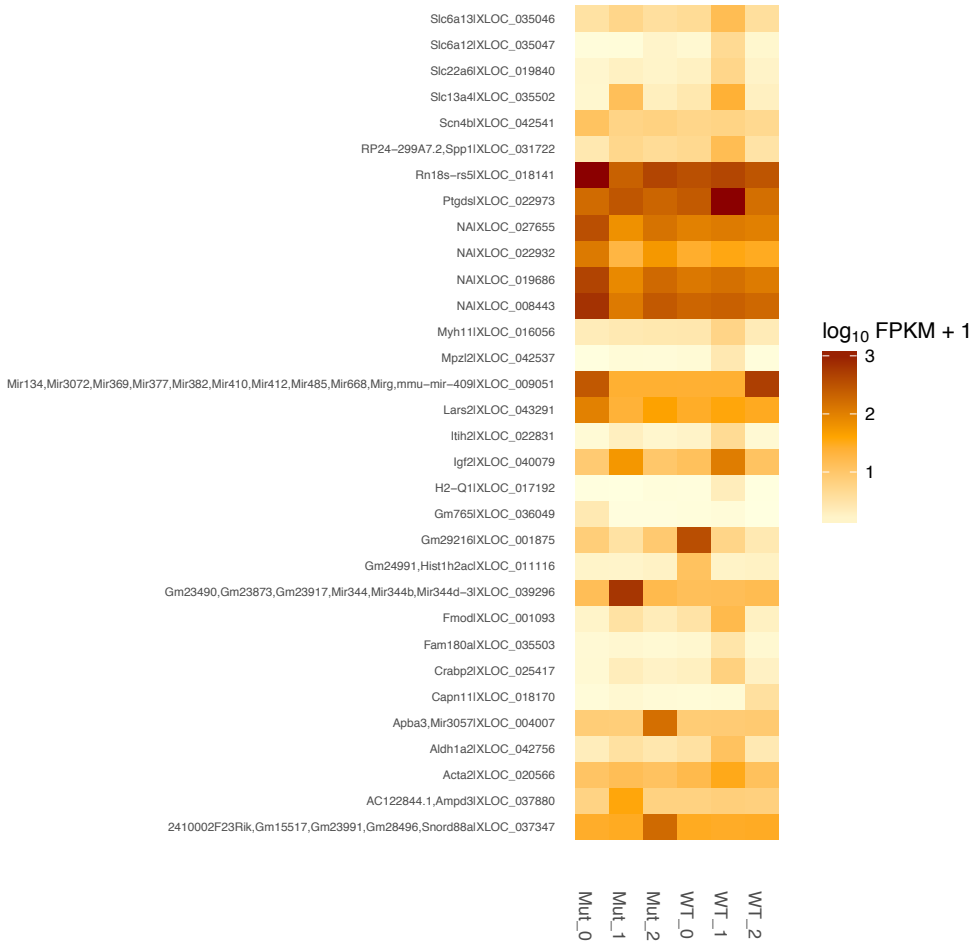


(b)

Figure 12: Global gene expression differs between sample groups, but gene expression for individuals does not cluster by grouping (a)The log_2 of the fold change of expression (FPKM values) for each gene between samples (values denoted as "significant" corrected for multiple testing) (b) dendrogram depicting Jensen-Shannon distances calculated for gene expression between all replicates.

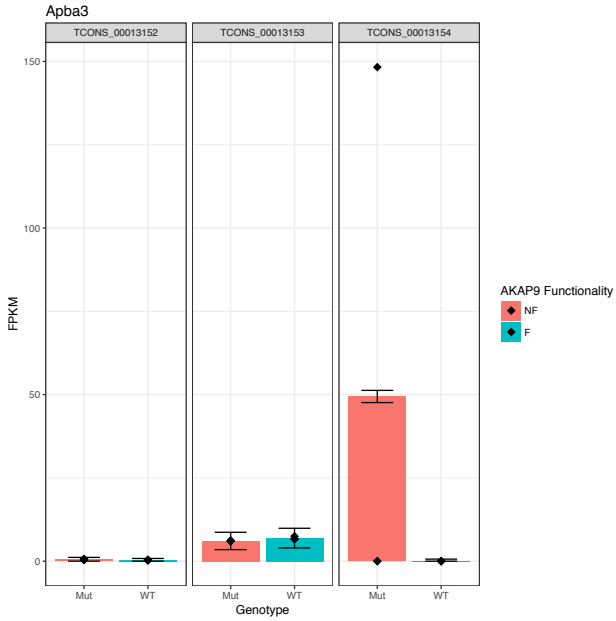


(a)

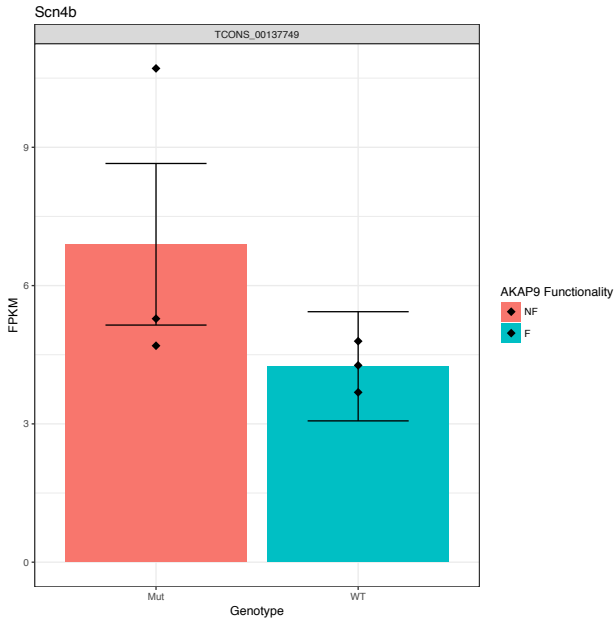


(b)

Figure 13: Differentially expressed genes are driven by outliers and do no cluster by group. (a) A differentially expressed gene clustering by Jensen-Shannon distance. (b) A heatmap of differentially expression between conditions for genes above the significance threshold



(a)



(b)

Figure 14: Apba3 and Scn4b are two genes up-regulated in the group lacking functional AKAP9, while Fmod is downregulated. Barplots of mean FPKM values for individual isoforms of each gene, where error bars represent confidence intervals around the FPKM value

Discussion

The goals of this study were twofold: first, to study the biological effects of the *mei2.5* mutation on the hippocampal transcriptome as a step towards understanding more fully the biology of the *mei2.5* mouse model. Second, to pilot a bioinformatics workstation at Middlebury College to explore how computational biology and bioinformatic education might be implemented at an undergraduate institution.

Middgenpilot: An effective tool for bioinformatics education and research at Middlebury College

The Middgenpilot workstation proved to be an effective tool for facilitating a deeper understanding of the components of an RNA-seq analysis pipeline, though it's potential is certainly not limited to RNA-seq. The environment provided a platform on which to assemble the software components of an analysis pipeline piece-by-piece. It could be expanded to compare the same experiment across analysis platforms, or utilized to conduct several different types of analysis. It could easily be used as a platform for processing microarray data, genomic, or proteomic data. Though the virtual machine was a Linux, command-line environment, its connection to an Alpaca storage volume offered the advantage of access to data and analysis files from a standard desktop computer. This enabled file-sharing beyond the Linux machine. This was an effective method for file sharing between investigators, but it could easily be extended to a bioinformatics classroom environment, with students submitting assignments from the virtual machine to the storage volume, where a professor without access to the individual virtual machine could review assignment output files.

Several comments about the virtual machine: First, the bash scripts used to run

the pipeline are rudimentary (See Appendix C). Conveniently, scripts could be executed in a users home directory that contained all of the necessary pipeline packages, but input could be seamlessly streamed in from and output deposited in the Alpaca volume in a specified working directory. One of the lessons learned from this analysis was the importance of an organized directory structure. To avoid copying files (and possibly corrupting them), each step in the pipeline deposits outputs into a pre-specified directory. Processing re-sequenced reads led to a slightly more complicated directory structure then necessary, and ultimately to confusion. Were this analysis to be run again, the directory structure should be as streamlined as possible. This would enable more sophisticated bash scripts that could make the pipeline itself more streamlined (with one step flowing directly into the next).

Finally, with respects to the virtual machine: the R work environment leaves something to be desired. Due to outdated laptop software, the entire cummeRbund analysis in R was conducted from an Xterm window, which was inconvenient, and far more time consuming than necessary. Running R scripts involved opening the Xterm window from the virtual machine, and then re-importing the source code each time an edit to the code was made. This could be fixed by operating from a desktop or laptop machine with the most up-to date version of R, capable of installing all necessary packages. Cuffdiff output could be funneled back down to the desktop or laptop machine for final analysis, which could then be deposited on the cloud again. The lack of a graphical user interface meant that integrated development environments (IDEs) could not be used for writing R scripts or bash scripts, and so all script writing and editing was conducted in the Linux editor VIM. Though this meant testing scripts was a cumbersome process, it was ultimately encouraged deeper learning about the Linux system.

Where does a tool such as Middgenpilot fit into the current bioinformatics landscape? The cost of sequencing has dropped at an exponential rate in recent years: it drops by half approximately every five months, meaning that the capacity to generate data increases 5-fold every year. However, computing technology continues to improve at the rate predicted by Moore's law, doubling in capacity every 18-24 months [64]. As the fields of bioinformatics and computational biology expand, they will hit the very real barriers of computational and storage capacity. Increasingly, computing is turning to the cloud for solutions. The

cloud refers to large collections of servers (owned an operated by a service provider) that are accessible through the internet, that provide resources that are accessible on-demand, in a pay-as-you-go model. These resources can be divided conceptually into "Data as a service" (DaaS - e.g. publically accessible databases such as GenBank, Ensembl), "Software as a service" (SaaS - cloud-based software services that eliminate the need to download and update bioinformatic software packages), "Platform as a Service" (PaaS - a platform for developing and deploying cloud applications, where computational resources scale automatically, e.g. Galaxy, Eoulsan), and "Infrastructure as a service" (IaaS - delivers virtualized hardware and software, e.g. Amazon Web Services, Cloud BioLinux) [15]. These resources are far more powerful then a small virtual machine on a Middlebury server, and are the tools that are used at the cutting edge of bioinformatics today. Learning how to use these tools will be a vital component of undergraduate biology education moving forward.

However, having a "local" VM installed on Middlebury servers confers several benefits. First, it gives the user a window into the infrastructure needed to support a high-throughput analysis, providing an "under the hood" glimpse on a small scale of what is going on in an environment such as Galaxy. The VM was a barebones installation, and installing the pipeline was a valuable lesson in troubleshooting software dependencies. Second, the VM provides the user with full control over all steps of the analysis, forcing the user to pay attention to small details such as md5sum checks to verify the integrity of transferred data, directory structures, and the full range software open-source software. Third, the VM provides a sunk-cost way to experientially investigate the potentials of cloud-computing, without the risk of incurring exorbitant fees by exceeding the requested computational limit. Middlebury College has an extensive computational infrastructure that is professionally maintained. The Biology and Biochemistry departments could benefit from taking advantage of that infrastructure to develop a bioinformatics curriculum.

The Middgenpilot model facilitates fluency in the computational infrastructure necessary for bioinformatics, allows full user control, and is a cost-effective "sandbox" environment where mistakes need not be expensive. Dovetailed with an overview of cloud-based computing resources being used at the forefront of bioinformatcs, it has the potential to be a powerful tool for undergraduate education in bioinformatics.

Data Quality

Pre-processing: FastQC and Trimmomatic

Visual inspection of the tiles in the per tile sequence section of the FastQC output file reveals several tiles with lower quality than the surrounding tiles. However, the quality reductions are not consistent in a single tile or distributed across the same position for all reads, suggesting that the drop in sequence quality was likely due to a bubble in the flow cell at that specific position.

The Per Base Sequence Content module is expected to issue failures for RNA-Seq data. Random hexamer priming - the library construction method used for our RNA-seq library - leads to the enrichment of certain kmers at the 5' end of reads [29], and this is observed in the samples given warnings and failures by FastQC.

Our samples failed the FastQC per sequence GC content module. However, they have normal-looking GC distributions with a mean centered around 48%, though the distributions are systematically shifted to the left with a GC content distribution compared to FastQC's theoretical GC content prediction. indicating lower GC content than the FastQC model predicts. This could either be the result of systematic bias in the library preparation, or simply characteristic of the distribution of GC content in C57BL/6J background strain of *Mus musculus*, which has been estimated to have a mean GC content of 51% with a standard deviation of 7.8% [56]. The FastQC module does not take into account theoretical GC distributions by species, but rather calculates it from the sample. Given the systematic shift, the relatively normal-looking distribution, and a GC content that falls within a single standard deviation of the estimated species GC content, we decided that the warnings received by this module could be disregarded.

All of our samples failed the FastQC Duplicate Sequence module, which suggests that duplicate sequences make up anywhere from 20% to over 50% of our overall sequence content. This is to be expected in RNA-Seq experiments, because of varying levels of transcript expression. Highly expressed transcripts will have high levels of sequence duplication, whereas low-level transcripts will see less transcript. The failures on this module do not give any cause for concern. ‘

All of our samples failed the Kmer Content module. The module may issue a warning or failure for duplicate sequences that don't trigger the Overrepresented Sequences module. Examination of the plots for each sample reveal that no one kmer is consistently over-represented at a single location. The lack of consistent kmer over-representation between samples and the presence of high levels of duplication suggest that duplicate sequences - an expected RNA-Seq bias caused by high levels of transcript expression - are triggering failures and warnings in the Kmer Content module.

Trimming low-quality reads – determined by base-by-base Phred scores – is generally beneficial in sequencing experiments. In RNA-seq experiments, trimming low-quality reads can increase the percentage of reads aligning to the reference genome, increasing the reliability of the results. In an analysis of nine trimming programs, Trimmomatic performed well for RNA-seq data using a sliding-window quality filtering approach with an intermediate quality threshold [21]. In Trimmomatic's sliding window quality filtering algorithm, a set-length window scans from the 5' to 3' end of a read and removes the 3' end when the average quality drops below a specified quality parameter Q [7]. However the algorithm's performance (as measured by percentage of reads mapping to the reference genome after trimming) dropped as the Q was increased. However, the sliding window approach utilized by Trimmomatic and five other trimming programs doesn't take into account the drop in percentage of aligned reads when quality parameters are too stringent [21]. In 2014, the developers of Trimmomatic developed a maximum information quality filtering algorithm with an increasingly strict trimming process as the length of the read being trimmed increases. This algorithm, when combined with adapter trimming for Illumina adapter sequences, outperformed the sliding window algorithm [7]. Though trimming is recommended practice, Trimmomatic trimmed very few reads in this study (Fig 6).

Annotation-based Quality Control

Following alignment of filtered reads to the reference genome, chromosomal read mapping was inspected (Figure 7). The Samtools command "idxstats" reports the number of fragments that map to each chromosome. More reads map to each chromosome in the non-functional Akap9 group (Mut), which is to be expected, given that there are more reads overall in this

group, and the measure is of fragments mapped, a raw, un-normalized measure of reads.

The annotation-based quality control package RSeQC demonstrates consistent 3' bias across all of our samples (Figure 8). Early RNA-seq experiments used oligo-dT priming to synthesize cDNA strands from oligo-dT purified mRNA samples [49] [77], however, these were found to introduce 3' bias, where 3' ends of transcripts were over-represented compared to 5' ends [73]. A method to correct this bias was developed that involved RNA fragmentation followed by cDNA synthesis using random hexamer primers [48]. Though this method has been shown to introduce nucleotide biases in the first 13 bp of a read [29], it provides more evenly distributed coverage globally along a gene [73]. To avoid 3' bias in the preparation, the mRNA pool in this study was converted into a cDNA sequencing library using the mRNA TruSeq Stranded, 175 bp library preparation protocol [55], [82]. This protocol entails polyA selection followed by RNA fragmentation, cDNA synthesis with random hexamer primers, and size selection for 175 bp cDNAs before sequencing. Despite these measures, 3' bias persisted.

Wang *et al.* suggest that RNA degradation may occur preferentially at the 5' end due to oligo-dT selection [72]. However, transcriptome-wide, spontaneous, global 5' RNA degradation seems an unlikely cause at the temperatures and timescales in the mRNA TruSeq protocol [55]. Another study found that RNA degradation (as measured by RIN over a range of values from 10 to 2) leads to increasing 3' bias associated with decreasing mRNA quality. As RNA quality diminishes from 10, 3' bias increases. Interesting, longer RNAs tend to have higher expression in groups with higher RINs, suggesting that longer genes are preferentially degraded from the 5' end [62]. The genomic region for Akap9 is over 150 kb and contains 48 exons, while the average mouse gene was estimated to have a length of 7902 bp [32]. Given the length of Akap9, and the 3' bias that particularly affects long genes as RNA quality drops, it is unsurprising that we observed such low coverage of the mutation site at exon 13.

An examination of RNA degradation pathways in *Saccharomyces cerevisae* suggests a potential biological mechanism for this degradation. In *S. cerevisae*, RNA degradation plays an important role in transcriptional expression and regulation. Cytoplasmic Xrn1 and nuclear Rat1 are primarily responsible for 5' to 3' degradation of RNA, while the exosome - a conserved complex with an endonuclease cleavage site - is responsible for 3' to 5' exonuclease

acitivty. Cytoplasmic RNA degradation begins with the shortening of the 3' polyA tail by the Pan2/Pan3 complex or the Ccr4/Pop2/NOT complex. After deadenylation, 3' to 5' degradation may occur directly or the RNA may undergo 5' decapping followed by 5' to 3' degradation. Studies suggest that decapping and 5' to 3' degradation occurs preferentially over 3' to 5' degradation, which is also 1.5-6 times slower [53]. In mice, RNA degradation is also temperature-sensitive, and occurs more rapidly at 37^oC than 4^oC [83]. In this study, we conducted sacrifice and tissue extraction at room temperature. Though the interval between death and whole brain freezing in chilled RNA*later* was on the order of minutes, the gene body coverage data (Figure 8) and RIN scores (Figure 2) suggest the interval was long enough for some 5' degradation to occur. Further work on this dataset should implement Cuffdiff's -frag-bias-correct option, which finds and corrects biases in the dataset [67], or a 3' tag counting method described in [62].

The presence of microRNAs among the differentially expressed genes identified by Cuffdiff was unexpected. There is, however, some evidence to suggest that some microRNAs may be transcribed in large blocks and capped and polyadenylated before processing [9]. It is possible that the dT capture step managed to capture some of these unprocessed microRNAs.

An open question that remains to be answered in this analysis is the interpretation of physiological relevance of the expressed genes. Figure 9 displays the distribution of FPKM values. Should one peak of the bimodal distribution be considered physiologically relevant, and the other not? Should there be an FPKM threshold, below which a transcript ins considered physiologically irrelevant? Mortazavi *et al.* demonstrate that with RNA standards combined with information on cellular RNA content, RPKM values can be translated into absolute transcript levels [48]. This approach is unfeasible in the hippocampus, with many different cell types [81], as different cell types will display unique transcriptional profiles. Hart *et al.* combined an ENCODE 2.0 RNA-seq data set from 17 human cell lines with ENCODE ChIP-seq data to determine whether promoters at a given FPKM value were active or repressed. They determined physiological relevance to be the point where the ratio of active to repressed promoters drops below 1. When standardized across samples, this results in a raw FPKM threshold for physiological relevance at $\geq$0.1 [30]. This value should be taken with a grain of salt, as it was derived from human cell lines. However, it serves as the best

Transcript ID	Mut	WT	Cuffdiff Protein ID	Ensembl Annotation (rel 86)
TCONS_00099626	No	No	P31954	Protein coding
TCONS_00099628	No	Yes	P31956	Protein coding
TCONS_00099633	Yes	Yes		None
TCONS_00099634	Yes	No	P31957	Protein coding
TCONS_00099638	Yes	Yes	P31955	Protein coding
TCONS_00099639	Yes	No		Retained intron
TCONS_00099640	Yes	Yes	P31958	Nonsense-mediated decay
TCONS_00099641	Yes	Yes		Retained intron
TCONS_00099642	Yes	Yes		None

Table 3: Table of Akap9 transcripts identified by Cuffdiff that meet the criteria for physiological relevance.

current estimate, given the limited nature of this study, of a threshold to determine which transcripts to consider and which to disregard.

Akap9 expression in the hippocampus

Cuffdiff detected 17 transcripts of Akap9 expressed in the *mei2.5* hippocampus. Of those 17 transcripts, only eight have FPKM values that meet the criteria established above for physiological relevance. However, some of those transcripts only meet the criteria for physiological relevance in one sample group, but not the other (Table 3 (See Table **??** in Appendix B for confidence intervals and FPKM values). Interestingly, full length Akap9 (TCONS_00099628), group meets the criteria for physiological relevance while the other does not. However, the confidence interval for the Mut group (0:0.239531) and the WT group (0:0.545391) overlap, and both contain zero. A separate analysis by the NYGC found the only differentially expressed gene between groups to be Akap9, which was reduced in the Mut group (NYGC, personal communication).

Interestingly, Akap9 was initially described in four different studies, all of which detected an 11.7 kb mRNA transcript (detected as yotiao [39], AKAP350 [58], CG-NAP [65], and AKAP450 [78] expressed in the brain. Perhaps this transcript is a precursor transcript

that is further spliced to form the various isoforms. Reduced expression of this transcript may lead to reduced expression of other isoforms. Some gene classes contain exonic "stop" codons, an their inclusion into a splice isoform marks them for nonse-mediated decay (NMD) [50]. Perhaps the *mei2.5* mutation abolishing the exon 13 conserved splice site resulting in the inclusion of intron 13-14 during splicing leads to nonsense mediated decay of some transcripts, reducing overall pre-mRNA available for splicing.

Two other transcripts identified as physiologically relevant in the Mut group but not in the WT group are TCONS_00099634 and TCONS_00099639 . However, inspection of the confidence intervals for each (TCONS_00099634 – WT CI: [0:0.580874]; Mut CI: [0:1.25176]) (TCONS_00099639 – WT CI:[0:0.951379]; Mut CI: [0:1.64014]) reveals that for both transcripts, both sample CIs contain 0 and overlap extensively; any biological conclusions drawn from these data would be baseless.

Cuffdiff identified TCONS_00099640 as protein coding in the hippocampus, however the Ensembl annotation identifies this transcript as tagged for NMD. This could be an aberrant annotation by Cuffdiff, or it could be that the transcript codes a hippocampus-specific protein product. This transcript had one of the largest discrepancies in confidence interval lengths, though the difference between the Mut and WT groups is not significant. However, the first exon of this transcript overlaps with exon 44 of TCONS_00099628, but extends in the 5' direction. Interestingly, the pericentrin-AKAP450 centrosomal targeting (PACT) domain identified in hAKAP450 by Gillingham and Munro as being recruited to the centrosome also extends in the 5' direction from exon 44 of murine TCONS_00099628 [24].

Another interesting aspect of the Cuffdiff transcript assemblies lies in the organization of the 3' transcripts of TCONS_00099628, TCONS_00099633, TCONS_00099640, and TCONS_00099642. Using TCONS_00099628 as a reference in the Cuffdiff assembly (Figure 10), each of the other three transcripts displays a unique splicing profile. This region overlaps with the exchange protein directly activated by cAMP (Epac1) interaction domain [60] in addition to the previously described PACT domain interaction.

The AKAP450 isoform of AKAP9 localizes the PKA-regulated phosphodiesterase PDE4D3 to the centrosome, reducing cAMP concentrations in a centrosomal pocket and fine-

tuning the sensitivity of PKA [66], implicating hAKAP450 in centrosomal cAMP signalling pathways. AKAP9 regulates microtubule dynamics through its association with Epac1, which coordinates cadherin and integrin-mediated adhesion via the actin cytoskeleton [60]. Targeting of AKAP450 to the centrosome occurs via the PACT, which recruits PKA type II α, an interaction shown to be critical for centrosomal integrity and duplication; disruption of this interaction induces cytokinesis defects and G1 arrest [33]. The presence of three AKAP9 isoforms in the hippocampus suggests that perhaps hippocampal AKAP9 plays a role in modulating microtubule dynamics rather than modulating gap junctions or NMDA receptors as initially hypothesized.

The expression of all Akap9 isoforms falls in the saddle between the two peaks of the density distribution in Figure 9, suggesting relatively low abundance of all of the AKAP9 isoforms in the hippocampus. Staining of the hippocampus for yotiao isoforms localized their expression to the pyramidal neurons of the CA1, CA3, and dentate gyrus areas[39]. These cells make up a small fraction of the hippocampal tissue. A possible reason for low overall expression could be that Akap9 expression is cell-type specific, and is only actively transcribed in pyramidal cells (as an example), while being repressed in other cell types. An alternate hypothesis could be that AKAP9 is constitutively expressed at low levels across the hippocampus.

Global differential expression

A volcano plot of fold change of FPKM values between groups suggests that there may be meaningful changes in the transcriptome between groups. However, the fold-change value taken into account in this plot are average FPKM values, and do not reflect the sample variance (and the sample size is small, with only $n = 3$ in each group). A dendrogram of Jensen-Shannon distances (offered as part of the cummeRbund package [25]) reveals that samples do not cluster by sample grouping. This clustering could be explained by an unintentional mixing of two *mei2.5* congenic lines previously maintained in the colony, one on a C3H background and one on a C57BL/6J background. The samples could be clustering based on background strain.

A heatmap of genes marked by Cuffdiff as differentially expressed at a level of

significance above the multiple-test correction threshold $\alpha < 0.05$ (Figure 13 reveals that often a single outlier with expression an order of magnitude different from the other replicates will throw off the FPKM estimate. However, given the small sample size and the nature of binning (with two heterozygotes binned with a single wild-type mouse), the estimates of FPKM values and fold-change between conditions will be tenuous at best. A differential expression analysis conducted by the NYGC using Star, FeatureCounts, and DESeq2 found that the only gene with significant changes in expression between the two conditions was Akap9 (NYGC, personal communication). This reinforces what has been demonstrated in the literature [37], [63]: different analysis methods produce different results, and the differences are exacerbated with small sample sizes. The statistical methodologies of the "Tuxedo Suite" is beyond the purview of this study, but given the limited sample size of this study, a more careful selection of statistical methodologies would be a benefit in any further analysis.

With those caveats, several of the genes identified by CuffDiff as differentially expressed between groups are worth noting. The sodium voltage-gated channel beta subunit 4 (Scn4b) is expressed in the brain, spinal cord, and some sensory neurons [80], and a SNP in its sequence has been associated with congenital LongQT syndrome [47] and Brugada syndrome [54], both characterized by malignant ventricular arrhythmias. AKAP9 interacts with the cardiac potassium channel KCNQ1 and regulates its phosphorylation state through recruitment of PKA and PP1 [11], and it has been shown to be a genetic modifier of the corrected QT interval, risk, and severity of cardiac events in a founder population carrying a KCNQ1 mutation [16]. The increased expression of Scn4b in this study was driven by a single replicate, however, the result suggests a possible connection worth investigating further. Perhaps the yotiao isoform of Akap9 regulates the Scn4b phosphorylation state in the nervous system (or beyond), regulating activity.

Amyloid beta precursor protein binding family A member 3 (Apba3/Mint3/X11L2) interacts with amyloid precursor protein (APP), regulating APP trafficking between the *trans*-Golgi and the membrane [61] [10]. Cleaving of the amyloid precursor protein produces β-amyloid peptide (Aβ), the accumulation of which is a hallmark of Alzheimer's disease, though it's role in disease etiology remains unclear [52]. Increased expression of Apba3 in the Mut group was driven by a single replicate. However, taken together with a recent study

that discovered a SNP in the PACT of AKAP9 associated with Alzheimer's disease [41], a tenuous connection of AKAP9 to pathways associated with Alzheimer's disease begins to emerge. AKAP9 may coordinate the localization of the centrosome and the Golgi apparatus, but though disruption of that connection has been shown to affect directional cell migration and ciliogenesis, it did not affect global secretion [31]. Nonetheless, disruption of an infrastructural protein such as AKAP9 that coordinates aspects of microtubule dynamics [60] and Golgi organization [31] may affect APP trafficking; this would be an interesting avenue for future inquiry.

Moving forward with the *mei2.5* mouse model

This study represents an initial attempt to characterize the *mei2.5* phenotype beyond the testes. The dataset generated in this study should be more thoroughly examined. A basic first step would be to add Gene Ontology terms to the dataset, to examine if the significantly different genes identified by Cuffdiff share common ontologies. A KEGG pathway analysis may shed more light on the nature of the differential expression between conditions: though individual genes may not meet the threshold for significant expression, perhaps overall pathways expression is affected in a significant way. The sequencing files should be re-aligned and sent through several different differential expression analysis programs, to provide a more comprehensive picture of differential expression (a comprehensive review of programs to try can be found here: [14]). Furthermore, though the sample size is small, future work might take advantage of new statistical methods being developed [40]. Future work can build on this study through qRT-PCR validation of differential expression findings in the hippocampus. Akap9 expression should be quantified via qRT-PCR expression studies in other regions of the brain, and in other tissues. The tissue and cell-type specificity of AKAP9 isoforms remains largely undocumented, and should be explored further. Tenuous connections to both Alzheimer's disease and longQT syndrome suggest that the *mei2.5* model may be an effective model for disease; at the very least, the connections are worth investigating further.

Situated in the larger context of functional genomics and large-scale efforts to understand the relationship between genotype and phenotype, the *mei2.5* mouse model provides a unique opportunity to understand the mechanisms by which single nucleotide base changes in

the genome can, by subtly or not so subtly altering cellular processes, radiate up through cell types and tissues to observable, macro-scale phenomena in vertebrates, ultimately offering a window of insight into mechanisms of both disease and evolution.

Bibliography

[1] Namiko Abe. *Re:RNASeq Pricing*. E-Mail. Mar. 3, 2016.

[2] Bronwen L. Aken et al. "The Ensembl Gene Annotation System". In: *Database (Oxford)* 2016 (2016). ISSN: 1758-0463. DOI: 10.1093/database/baw093. pmid: 27337980.

[3] S Andrews. *FastQC A Quality Control Tool for High Throughput Sequence Data*. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (visited on 11/21/2016).

[4] *APBA3 Gene - GeneCards — APBA3 Protein — APBA3 Antibody*. URL: http://www.genecards.org/cgi-bin/carddisp.pl?gene=APBA3 (visited on 12/02/2016).

[5] Darren L Beene and John D Scott. "A-Kinase Anchoring Proteins Take Shape". In: *Curr Opin Cell Biol* 19.2 (Apr. 2007), pp. 192–198. ISSN: 0955-0674. DOI: 10.1016/j.ceb.2007.02.011. pmid: 17317140. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3521038/ (visited on 10/17/2016).

[6] David R. Bentley et al. "Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry". In: *Nature* 456.7218 (Nov. 6, 2008), pp. 53–59. ISSN: 0028-0836. DOI: 10.1038/nature07517. URL: http://www.nature.com/nature/journal/v456/n7218/abs/nature07517.html (visited on 11/26/2016).

[7] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data". In: *Bioinformatics* (Apr. 1, 2014), btu170. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btu170. pmid: 24695404. URL: http://bioinformatics.oxfordjournals.org/content/early/2014/04/01/bioinformatics.btu170 (visited on 09/19/2016).

[8] Claire A. Bovet. "Akap9 and Alternative Splicing". Middlebury College, Apr. 27, 2012. 84 pp.

[9] XUEZHONG CAI, CURT H. HAGEDORN, and BRYAN R. CULLEN. "Human microRNAs Are Processed from Capped, Polyadenylated Transcripts That Can Also Function as mRNAs". In: *RNA* 10.12 (Dec. 2004), pp. 1957–1966. ISSN: 1355-8382. DOI: 10.1261/rna.7135204. pmid: 15525708. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1370684/ (visited on 12/07/2016).

[10] Amanda H. Caster and Richard A. Kahn. "Recruitment of the Mint3 Adaptor Is Necessary for Export of the Amyloid Precursor Protein (APP) from the Golgi Complex". In: *J. Biol. Chem.* 288.40 (Oct. 4, 2013), pp. 28567–28580. ISSN: 1083-351X. DOI: 10.1074/jbc.M113.481101. pmid: 23965993.

[11] Lei Chen and Robert S. Kass. "Dual Roles of the A Kinase-Anchoring Protein Yotiao in the Modulation of a Cardiac Potassium Channel: A Passive Adaptor versus an Active Regulator". In: *European Journal of Cell Biology*. Anchored cAMP Signaling Pathways1st International Meeting on Anchored cAMP Signaling Pathways 85.7 (July 5, 2006), pp. 623–626. ISSN: 0171-9335. DOI: 10.1016/j.ejcb.2006.03.002. URL: http://www.sciencedirect.com/science/article/pii/S0171933506000550 (visited on 11/14/2016).

[12] P. Chomczynski and K. Mackey. "Substitution of Chloroform by Bromochloropropane in the Single-Step Method of RNA Isolation". In: *Analytical Biochemistry* 225.1 (Feb. 1, 1995), pp. 163–164. ISSN: 0003-2697. DOI: 10.1006/abio.1995.1126. URL: http://www.sciencedirect.com/science/article/pii/S0003269785711268 (visited on 11/22/2016).

[13] Philip Cohen. "The Origins of Protein Phosphorylation". In: *Nat Cell Biol* 4.5 (May 2002), E127–E130. ISSN: 1465-7392. DOI: 10.1038/ncb0502-e127. URL: http://www.nature.com/ncb/journal/v4/n5/full/ncb0502-e127.html (visited on 10/21/2016).

[14] Ana Conesa et al. "A Survey of Best Practices for RNA-Seq Data Analysis". In: *Genome Biology* 17 (2016), p. 13. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0881-8. URL: http://dx.doi.org/10.1186/s13059-016-0881-8 (visited on 05/02/2016).

[15] Lin Dai et al. "Bioinformatics Clouds for Big Data Manipulation". In: *Biol Direct* 7 (Nov. 28, 2012), p. 43. ISSN: 1745-6150. DOI: 10.1186/1745-6150-7-43. pmid: 23190475. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3533974/ (visited on 12/05/2016).

[16] Carin P. de Villiers et al. "AKAP9 Is a Genetic Modifier of Congenital Long-QT Syndrome Type 1". In: *Circ Cardiovasc Genet* 7.5 (Oct. 2014), pp. 599–606. ISSN: 1942-325X. DOI: 10.1161/CIRCGENETICS.113.000580. pmid: 25087618. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4270884/ (visited on 12/07/2016).

[17] D. Diviani and J. D. Scott. "AKAP Signaling Complexes at the Cytoskeleton". In: *Journal of Cell Science* 114.8 (Apr. 15, 2001), pp. 1431–1437. ISSN: 0021-9533, 1477-9137. pmid: 11282019. URL: http://jcs.biologists.org/content/114/8/1431 (visited on 09/23/2016).

[18] Nathalie Duval et al. "Cell Coupling and Cx43 Expression in Embryonic Mouse Neural Progenitor Cells". In: *J. Cell. Sci.* 115 (Pt 16 Aug. 15, 2002), pp. 3241–3251. ISSN: 0021-9533. pmid: 12140256.

[19] Jane A. Endicott, Martin E. M. Noble, and Louise N. Johnson. "The Structural Basis for Control of Eukaryotic Protein Kinases". In: *Annual Review of Biochemistry* 81.1 (2012), pp. 587–613. DOI: 10.1146/annurev-biochem-052410-090317. pmid: 22482904. URL: http://dx.doi.org/10.1146/annurev-biochem-052410-090317 (visited on 10/21/2016).

[20] Brent Ewing and Phil Green. "Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities". In: *Genome Res.* 8.3 (Jan. 3, 1998), pp. 186–194. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.8.3.186. pmid: 9521922. URL: http://genome.cshlp.org/content/8/3/186 (visited on 12/16/2016).

[21] Cristian Del Fabbro et al. "An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis". In: *PLOS ONE* 8.12 (Dec. 23, 2013), e85024. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0085024. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0085024 (visited on 09/19/2016).

[22] Antonio Feliciello et al. "Yotiao Protein, a Ligand for the NMDA Receptor, Binds and Targets cAMP-Dependent Protein Kinase II1". In: *FEBS Letters* 464.3 (Dec. 31, 1999), pp. 174–178. ISSN: 0014-5793. DOI: 10.1016/S0014-5793(99)01585-9. URL: http://www.sciencedirect.com/science/article/pii/S0014579399015859 (visited on 10/26/2016).

[23] Sky K. Feuer. "Akap9 Is Required for Spermatogenesis". Middlebury College, Nov. 13, 2009. 64 pp.

[24] Alison K. Gillingham and Sean Munro. "The PACT Domain, a Conserved Centrosomal Targeting Motif in the Coiled-Coil Proteins AKAP450 and Pericentrin". In: *EMBO Rep* 1.6 (Dec. 15, 2000), pp. 524–529. ISSN: 1469-221X. DOI: 10.1093/embo-reports/kvd105. pmid: 11263498. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1083777/ (visited on 12/04/2016).

[25] Loyal Goff, Cole Trapnell, and D Kelley. *cummeRbund: Analysis, Exploration, Manipulation, and Visualization of Cufflinks High-Throughput Sequencing Data*. Version 2.16.0. 2013.

[26] Yoichi Gondo. "Trends in Large-Scale Mouse Mutagenesis: From Genetics to Functional Genomics". In: *Nat Rev Genet* 9.10 (Oct. 2008), pp. 803–810. ISSN: 1471-0056. DOI: 10.1038/nrg2431. URL: http://www.nature.com/nrg/journal/v9/n10/full/nrg2431.html (visited on 12/02/2016).

[27] Sara Goodwin, John D. McPherson, and W. Richard McCombie. "Coming of Age: Ten Years of next-Generation Sequencing Technologies". In: *Nat Rev Genet* 17.6 (June 2016), pp. 333–351. ISSN: 1471-0056. DOI: 10.1038/nrg.2016.49. URL: http://www.nature.com/nrg/journal/v17/n6/full/nrg.2016.49.html (visited on 11/26/2016).

[28] Florian Hahne and Robert Ivanek. "Visualizing Genomic Data Using Gviz and Bioconductor". In: *Statistical Genomics*. Ed. by Ewy Mathé and Sean Davis. Methods in Molecular Biology 1418. Springer New York, Jan. 1, 2016, pp. 335–351. ISBN: 978-1-4939-3576-5. DOI: 10.1007/978-1-4939-3578-9_16. URL: http://dx.doi.org/10.1007/978-1-4939-3578-9_16 (visited on 12/04/2016).

[29] Kasper D. Hansen, Steven E. Brenner, and Sandrine Dudoit. "Biases in Illumina Transcriptome Sequencing Caused by Random Hexamer Priming". In: *Nucleic Acids Res* 38.12 (July 2010), e131. ISSN: 0305-1048. DOI: 10.1093/nar/gkq224. pmid: 20395217. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2896536/ (visited on 11/20/2016).

[30] Traver Hart et al. "Finding the Active Genes in Deep RNA-Seq Gene Expression Studies". In: *BMC Genomics* 14 (Nov. 11, 2013), p. 778. ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-778. pmid: 24215113. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3870982/ (visited on 12/01/2016).

[31] Lidia Hurtado et al. "Disconnecting the Golgi Ribbon from the Centrosome Prevents Directional Cell Migration and Ciliogenesis". In: *J Cell Biol* 193.5 (May 30, 2011), pp. 917–933. ISSN: 0021-9525. DOI: 10.1083/jcb.201011014. pmid: 21606206. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3105543/ (visited on 11/27/2016).

[32] Niclas Jareborg, Ewan Birney, and Richard Durbin. "Comparative Analysis of Noncoding Regions of 77 Orthologous Mouse and Human Gene Pairs". In: *Genome Res.* 9.9 (Jan. 9, 1999), pp. 815–824. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.9.9.815. pmid: 10508839. URL: http://genome.cshlp.org/content/9/9/815 (visited on 12/05/2016).

[33] Guy Keryer et al. "Dissociating the Centrosomal Matrix Protein AKAP450 from Centrioles Impairs Centriole Duplication and Cell Cycle Progression". In: *Mol Biol Cell* 14.6 (June 2003), pp. 2436–2446. ISSN: 1059-1524. DOI: 10.1091/mbc.E02-09-0614. pmid: 12808041. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC194891/ (visited on 12/05/2016).

[34] Choel Kim et al. "PKA-I Holoenzyme Structure Reveals a Mechanism for cAMP-Dependent Activation". In: *Cell* 130.6 (Sept. 21, 2007), pp. 1032–1043. ISSN: 0092-8674. DOI: 10.1016/j.cell.2007.07.018. URL: http://www.sciencedirect.com/science/article/pii/S0092867407009555 (visited on 10/22/2016).

[35] Eija Korpelainen et al. *RNA-Seq Data Analysis: A Practical Approach*. Google-Books-ID: u5fNBQAAQBAJ. CRC Press, Sept. 19, 2014. 314 pp. ISBN: 978-1-4665-9501-9.

[36] Albrecht Kunze et al. "Connexin Expression by Radial Glia-like Cells Is Required for Neurogenesis in the Adult Dentate Gyrus". In: *Proc Natl Acad Sci U S A* 106.27 (July 7, 2009), pp. 11336–11341. ISSN: 0027-8424. DOI: 10.1073/pnas.0813160106. pmid: 19549869. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2700144/ (visited on 11/07/2016).

[37] Vanessa M. Kvam, Peng Liu, and Yaqing Si. "A Comparison of Statistical Methods for Detecting Differentially Expressed Genes from RNA-Seq Data". In: *Am. J. Bot.* 99.2 (Feb. 1, 2012), pp. 248–256. ISSN: 0002-9122, 1537-2197. DOI: 10.3732/ajb.1100340. pmid: 22268221. URL: http://www.amjbot.org/content/99/2/248 (visited on 12/06/2016).

[38] Yong Li et al. "The A-Kinase Anchoring Protein Yotiao Facilitates Complex Formation between Adenylyl Cyclase Type 9 and the IKs Potassium Channel in Heart". In: *J. Biol. Chem.* 287.35 (Aug. 24, 2012), pp. 29815–29824. ISSN: 0021-9258, 1083-351X. DOI: 10.1074/jbc.M112.380568. pmid: 22778270. URL: http://www.jbc.org/content/287/35/29815 (visited on 11/29/2016).

[39] Jerry W. Lin et al. "Yotiao, a Novel Protein of Neuromuscular Junction and Brain That Interacts with Specific Splice Variants of NMDA Receptor Subunit NR1". In: *J. Neurosci.* 18.6 (Mar. 15, 1998), pp. 2017–2027. ISSN: 0270-6474, 1529-2401. pmid: 9482789. URL: http://jneurosci.org/content/18/6/2017 (visited on 10/23/2016).

[40] Zhixiang Lin et al. "A Markov Random Field-Based Approach for Joint Estimation of Differentially Expressed Genes in Mouse Transcriptome Data". In: *Statistical Applications in Genetics and Molecular Biology* 15.2 (2016), pp. 139–150. ISSN: 2194-6302. DOI: 10.1515/sagmb-2015-0070. URL: http://www.degruyter.com/view/j/sagmb.2016.15.issue-2/sagmb-2015-0070/sagmb-2015-0070.xml (visited on 05/10/2016).

[41] Mark W. Logue et al. "Two Rare AKAP9 Variants Are Associated with Alzheimer Disease in African Americans". In: *Alzheimers Dement* 10.6 (Nov. 2014), 609–618.e11. ISSN: 1552-5260. DOI: 10.1016/j.jalz.2014.06.010. pmid: 25172201. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4253055/ (visited on 11/29/2016).

[42] Christian Lüscher and Robert C. Malenka. "NMDA Receptor-Dependent Long-Term Potentiation and Long-Term Depression (LTP/LTD)". In: *Cold Spring Harb Perspect Biol* 4.6 (Jan. 6, 2012), a005710. ISSN: , 1943-0264. DOI: 10.1101/cshperspect.a005710. pmid: 22510460. URL: http://cshperspectives.cshlp.org/content/4/6/a005710 (visited on 11/06/2016).

[43] Elaine R. Mardis. "Next-Generation DNA Sequencing Methods". In: *Annual Review of Genomics and Human Genetics* 9.1 (2008), pp. 387–402. DOI: 10.1146/annurev.genom.9.081307.164359. pmid: 18576944. URL: http://dx.doi.org/10.1146/annurev.genom.9.081307.164359 (visited on 10/23/2016).

[44] Elaine R. Mardis. "Next-Generation Sequencing Platforms". In: *Annual Review of Analytical Chemistry* 6.1 (2013), pp. 287–303. DOI: 10.1146/annurev-anchem-062012-092628. pmid: 23560931. URL: http://dx.doi.org/10.1146/annurev-anchem-062012-092628 (visited on 11/07/2016).

[45] Stella M. Mattaloni et al. "AKAP350 Is Involved in the Development of Apical Canalicular Structures in Hepatic Cells HepG2". In: *J Cell Physiol* 227.1 (Jan. 2012), pp. 160–171. ISSN: 0021-9541. DOI: 10.1002/jcp.22713. pmid: 21374596. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3899033/ (visited on 11/28/2016).

[46] Stella M Mattaloni et al. "Centrosomal AKAP350 Modulates the G1/S Transition". In: *Cell Logist* 3.1 (Jan. 1, 2013). ISSN: 2159-2780. DOI: 10.4161/cl.26331. pmid: 24475373. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3891632/ (visited on 11/29/2016).

[47] Argelia Medeiros-Domingo et al. "SCN4B-Encoded Sodium Channel 4 Subunit in Congenital Long-QT Syndrome". In: *Circulation* 116.2 (July 10, 2007), pp. 134–142. ISSN: 0009-7322, 1524-4539. DOI: 10.1161/CIRCULATIONAHA.106.659086. pmid: 17592081. URL: http://circ.ahajournals.org/content/116/2/134 (visited on 12/07/2016).

[48] Ali Mortazavi et al. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq". In: *Nat Meth* 5.7 (July 2008), pp. 621–628. ISSN: 1548-7091. DOI: 10.1038/nmeth.1226. URL: http://www.nature.com/nmeth/journal/v5/n7/full/nmeth.1226.html (visited on 11/26/2016).

[49] Ugrappa Nagalakshmi et al. "The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing". In: *Science* 320.5881 (June 6, 2008), pp. 1344–1349. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1158441. pmid: 18451266. URL: http://science.sciencemag.org/content/320/5881/1344 (visited on 11/26/2016).

[50] Julie Z. Ni et al. "Ultraconserved Elements Are Associated with Homeostatic Control of Splicing Regulators by Alternative Splicing and Nonsense-Mediated Decay". In: *Genes Dev.* 21.6 (Mar. 15, 2007), pp. 708–718. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.1525507. pmid: 17369403. URL: http://genesdev.cshlp.org/content/21/6/708 (visited on 12/06/2016).

[51] NIH. *A Brief Guide to Genomics: DNA, Genes, and Genomes*. Aug. 27, 2015. URL: https://www.genome.gov/18016863/A-Brief-Guide-to-Genomics (visited on 11/26/2016).

[52] Richard J. O'Brien and Philip C. Wong. "Amyloid Precursor Protein Processing and Alzheimer's Disease". In: *Annu Rev Neurosci* 34 (2011), pp. 185–204. ISSN: 0147-006X. DOI: 10.1146/annurev-neuro-061010-113613. pmid: 21456963. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3174086/ (visited on 12/07/2016).

[53] Roy Parker. "RNA Degradation in Saccharomyces Cerevisae". In: *Genetics* 191.3 (July 2012), pp. 671–702. ISSN: 0016-6731. DOI: 10.1534/genetics.111.137265. pmid: 22785621. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3389967/ (visited on 12/06/2016).

[54] Uschi Peeters et al. "Contribution of Cardiac Sodium Channel -Subunit Variants to Brugada Syndrome". In: *Circulation Journal* 79.10 (2015), pp. 2118–2129. DOI: 10.1253/circj.CJ-15-0164.

[55] Illumina Proprietary. *TruSeq Stranded mRNA Sample Preparation Guide*. Oct. 2013.

[56] Jonathan Romiguier et al. "Contrasting GC-Content Dynamics across 33 Mammalian Genomes: Relationship with Life-History Traits and Chromosome Sizes". In: *Genome Res* 20.8 (Aug. 2010), pp. 1001–1009. ISSN: 1088-9051. DOI: 10.1101/gr.104372.109. pmid: 20530252. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2909565/ (visited on 11/21/2016).

[57] Kerry J. Schimenti et al. "AKAP9 Is Essential for Spermatogenesis and Sertoli Cell Maturation in Mice". In: *Genetics* 194.2 (June 1, 2013), pp. 447–457. ISSN: 0016-6731, 1943-2631. DOI: 10.1534/genetics.113.150789. pmid: 23608191. URL: http://www.genetics.org/content/194/2/447 (visited on 10/13/2015).

[58] P. Henry Schmidt et al. "AKAP350, a Multiply Spliced Protein Kinase A-Anchoring Protein Associated with Centrosomes". In: *J. Biol. Chem.* 274.5 (Jan. 29, 1999), pp. 3055–3066. ISSN: 0021-9258, 1083-351X. DOI: 10.1074/jbc.274.5.3055. pmid: 9915845. URL: http://www.jbc.org/content/274/5/3055 (visited on 10/23/2016).

[59] *SCN4B Gene - GeneCards — SCN4B Protein — SCN4B Antibody*. URL: http://www.genecards.org/cgi-bin/carddisp.pl?gene=SCN4B (visited on 12/02/2016).

[60] Seema Sehrawat et al. "AKAP9 Regulation of Microtubule Dynamics Promotes Epac1-Induced Endothelial Barrier Properties". In: *Blood* 117.2 (Jan. 13, 2011), pp. 708–718. ISSN: 0006-4971. DOI: 10.1182/blood-2010-02-268870. pmid: 20952690. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3031489/ (visited on 11/28/2016).

[61] Punya Shrivastava-Ranjan et al. "Mint3/X11 Is an ADP-Ribosylation Factor-Dependent Adaptor That Regulates the Traffic of the Alzheimer's Precursor Protein from the Trans-Golgi Network". In: *Mol Biol Cell* 19.1 (Jan. 2008), pp. 51–64. ISSN: 1059-1524. DOI: 10.1091/mbc.E07-05-0465. pmid: 17959829. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2174186/ (visited on 12/07/2016).

[62] Benjamín Sigurgeirsson, Olof Emanuelsson, and Joakim Lundeberg. "Sequencing Degraded RNA Addressed by 3' Tag Counting". In: *PLOS ONE* 9.3 (Mar. 14, 2014), e91851. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0091851. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0091851 (visited on 12/05/2016).

[63] Charlotte Soneson and Mauro Delorenzi. "A Comparison of Methods for Differential Expression Analysis of RNA-Seq Data". In: *BMC Bioinformatics* 14 (2013), p. 91. ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-91. URL: http://dx.doi.org/10.1186/1471-2105-14-91 (visited on 12/06/2016).

[64] Lincoln D Stein. "The Case for Cloud Computing in Genome Informatics". In: *Genome Biol* 11.5 (2010), p. 207. ISSN: 1465-6906. DOI: 10.1186/gb-2010-11-5-207. pmid: 20441614. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2898083/ (visited on 12/05/2016).

[65] Mikiko Takahashi et al. "Characterization of a Novel Giant Scaffolding Protein, CG-NAP, That Anchors Multiple Signaling Enzymes to Centrosome and the Golgi Apparatus". In: *J. Biol. Chem.* 274.24 (Nov. 6, 1999), pp. 17267–17274. ISSN: 0021-9258, 1083-351X. DOI: 10.1074/jbc.274.24.17267. pmid: 10358086. URL: http://www.jbc.org/content/274/24/17267 (visited on 10/26/2016).

[66] Anna Terrin et al. "PKA and PDE4D3 Anchoring to AKAP9 Provides Distinct Regulation of cAMP Signals at the Centrosome". In: *J Cell Biol* 198.4 (Aug. 20, 2012), pp. 607–621. ISSN: 0021-9525. DOI: 10.1083/jcb.201201059. pmid: 22908311. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3514031/ (visited on 11/29/2016).

[67] Cole Trapnell et al. "Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks". In: *Nat. Protocols* 7.3 (Mar. 2012), pp. 562–578. ISSN: 1754-2189. DOI: 10.1038/nprot.2012.016. URL: http://www.nature.com/nprot/journal/v7/n3/full/nprot.2012.016.html (visited on 05/02/2016).

[68] G. E. Truett et al. "Preparation of PCR-Quality Mouse Genomic DNA with Hot Sodium Hydroxide and Tris (HotSHOT)". In: *BioTechniques* 29.1 (July 2000), pp. 52, 54. ISSN: 0736-6205. pmid: 10907076.

[69] Huiping Tu et al. "Association of Type 1 Inositol 1,4,5-Trisphosphate Receptor with AKAP9 (Yotiao) and Protein Kinase A". In: *J. Biol. Chem.* 279.18 (Apr. 30, 2004), pp. 19375–19382. ISSN: 0021-9258, 1083-351X. DOI: 10.1074/jbc.M313476200. pmid: 14982933. URL: http://www.jbc.org/content/279/18/19375 (visited on 10/24/2016).

[70] Guey-Shin Wang and Thomas A. Cooper. "Splicing in Disease: Disruption of the Splicing Code and the Decoding Machinery". In: *Nat Rev Genet* 8.10 (Oct. 2007), pp. 749–761. ISSN: 1471-0056. DOI: 10.1038/nrg2164. URL: http://www.nature.com/nrg/journal/v8/n10/full/nrg2164.html (visited on 12/03/2016).

[71] Liguo Wang, Shengqin Wang, and Wei Li. "RSeQC: Quality Control of RNA-Seq Experiments". In: *Bioinformatics* 28.16 (Aug. 15, 2012), pp. 2184–2185. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts356. URL: https://academic.oup.com/bioinformatics/article/28/16/2184/325191/RSeQC-quality-control-of-RNA-seq-experiments (visited on 11/20/2016).

[72] Lin Wang et al. "A Low-Cost Library Construction Protocol and Data Analysis Pipeline for Illumina-Based Strand-Specific Multiplex RNA-Seq". In: *PLoS One* 6.10 (Oct. 19, 2011). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0026426. pmid: 22039485. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3198403/ (visited on 11/26/2016).

[73] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: A Revolutionary Tool for Transcriptomics". In: *Nat Rev Genet* 10.1 (Jan. 2009), pp. 57–63. ISSN: 1471-0056. DOI: 10.1038/nrg2484. URL: http://www.nature.com/nrg/journal/v10/n1/full/nrg2484.html (visited on 09/26/2016).

[74] Jeremy O. Ward et al. "Toward the Genetics of Mammalian Reproduction: Induction and Mapping of Gametogenesis Mutants in Mice". In: *Biol Reprod* 69.5 (Jan. 11, 2003), pp. 1615–1625. ISSN: 0006-3363, 1529-7268. DOI: 10.1095/biolreprod.103.019877. pmid: 12855593. URL: http://www.biolreprod.org/content/69/5/1615 (visited on 09/23/2016).

[75] Ryan S. Westphal et al. "Regulation of NMDA Receptors by an Associated Phosphatase-Kinase Signaling Complex". In: *Science* 285.5424 (July 2, 1999), pp. 93–96. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.285.5424.93. pmid: 10390370. URL: http://science.sciencemag.org/content/285/5424/93 (visited on 10/23/2016).

[76] *What the FPKM? A Review of RNA-Seq Expression Units*. 2014-05-08T18:55:06+00:00. URL: https://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/ (visited on 12/01/2016).

[77] Brian T. Wilhelm et al. "Dynamic Repertoire of a Eukaryotic Transcriptome Surveyed at Single-Nucleotide Resolution". In: *Nature* 453.7199 (June 26, 2008), pp. 1239–1243. ISSN: 0028-0836. DOI: 10.1038/nature07002. URL: http://www.nature.com/nature/journal/v453/n7199/full/nature07002.html (visited on 11/26/2016).

[78] O Witczak et al. "Cloning and Characterization of a cDNA Encoding an A-Kinase Anchoring Protein Located in the Centrosome, AKAP450." In: *EMBO J* 18.7 (Apr. 1, 1999), pp. 1858–1868. ISSN: 0261-4189. DOI: 10.1093/emboj/18.7.1858. pmid: 10202149. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1171271/ (visited on 12/04/2016).

[79] Wei Wong and John D. Scott. "AKAP Signalling Complexes: Focal Points in Space and Time". In: *Nat Rev Mol Cell Biol* 5.12 (Dec. 2004), pp. 959–970. ISSN: 1471-0072. DOI: 10.1038/nrm1527. URL: http://www.nature.com/nrm/journal/v5/n12/full/nrm1527.html (visited on 10/17/2016).

[80] Frank H. Yu et al. "Sodium Channel beta4, a New Disulfide-Linked Auxiliary Subunit with Similarity to beta2". In: *J. Neurosci.* 23.20 (Aug. 20, 2003), pp. 7577–7585. ISSN: 1529-2401. pmid: 12930796.

[81] Amit Zeisel et al. "Cell Types in the Mouse Cortex and Hippocampus Revealed by Single-Cell RNA-Seq". In: *Science* 347.6226 (Mar. 6, 2015), pp. 1138–1142. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aaa1934. pmid: 25700174. URL: http://science.sciencemag.org/content/347/6226/1138 (visited on 12/06/2016).

[82] Silin Zhong et al. "High-Throughput Illumina Strand-Specific RNA Sequencing Library Preparation". In: *Cold Spring Harb Protoc* 2011.8 (Aug. 1, 2011), pp. 940–949. ISSN: 1559-6095. DOI: 10.1101/pdb.prot5652. pmid: 21807852.

[83] Yi Zhu et al. "[Relationship between RNA degradation and postmortem interval in mice]". In: *Fa Yi Xue Za Zhi* 27.3 (June 2011), pp. 161–163, 177. ISSN: 1004-5619. pmid: 21899002.

Appendix

Appendix A: *mei2.5* mutation site coverage

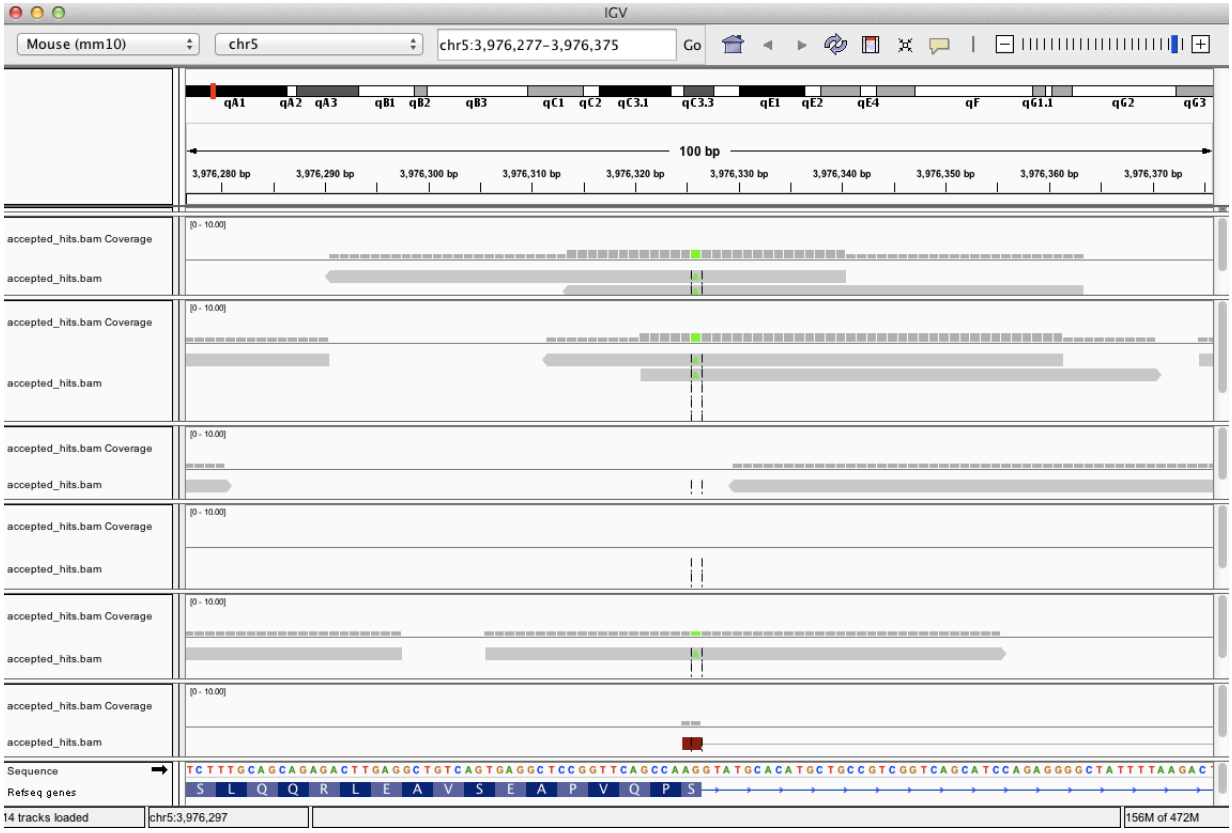


Figure 15: The mutation site received sparse coverage across all samples. Top three rows: Mut. Bottom three rows WT/Het.

Appendix B: Akap9 Isoform Expression Levels

Isoform ID	Sample Group	FPKM	Conf Hi	Conf Lo
TCONS_00099625	Mut	0.00516065	0.0742716	0
TCONS_00099626	Mut	0	0.0273135	0
TCONS_00099627	Mut	0.0198642	0.45424	0
TCONS_00099628	Mut	0.0828765	0.239531	0
TCONS_00099629	Mut	0	0.148343	0
TCONS_00099630	Mut	0	0.0957462	0
TCONS_00099631	Mut	0.00635531	0.254499	0
TCONS_00099632	Mut	0.0148311	0.281703	0
TCONS_00099633	Mut	0.991405	1.47768	0.505134
TCONS_00099634	Mut	0.144354	1.25176	0
TCONS_00099636	Mut	0.0127162	0.146062	0
TCONS_00099637	Mut	0	0.0393095	0
TCONS_00099638	Mut	0.211095	1.38415	0
TCONS_00099639	Mut	0.201526	1.64014	0
TCONS_00099640	Mut	1.53309	4.34404	0
TCONS_00099641	Mut	0.667538	3.44403	0
TCONS_00099642	Mut	1.27881	2.49367	0.0639568
TCONS_00099625	WT	0	0.0156321	0
TCONS_00099626	WT	0	0.0271951	0
TCONS_00099627	WT	0	0.127947	0
TCONS_00099628	WT	0.259437	0.545391	0
TCONS_00099629	WT	0	0.148343	0
TCONS_00099630	WT	0	0.0957462	0
TCONS_00099631	WT	0	0.139613	0
TCONS_00099632	WT	0	0.0972831	0
TCONS_00099633	WT	1.49215	2.15196	0.832343
TCONS_00099634	WT	0.0310203	0.580874	0
TCONS_00099636	WT	0.0055743	0.100196	0
TCONS_00099637	WT	0.00727687	0.142851	0
TCONS_00099638	WT	2.56687	7.87862	0
TCONS_00099639	WT	0.0669399	0.951379	0
TCONS_00099640	WT	5.26037	11.0194	0
TCONS_00099641	WT	0.374021	2.67398	0
TCONS_00099642	WT	2.33266	4.05358	0.611728

Table 4: Table of average FPKM values for AKAP9 isoforms and the upper and lower boundares of 95% confidence intervals for the given FPKM value.

Appendix C: Differentially Expressed Genes

XLOC	Gene Name	Mut FPKM	WT FPKM	$\log_2$(fold change)
XLOC_039296	"Gm23490,Gm23873,Gm23917,Mir344,Mir344b,Mir344d-3"	225.451	13.2377	-4.09009
XLOC_004007	"Apba3,Mir3057"	55.9758	7.23883	-2.95097
XLOC_036049	Gm765	0.536399	0.0960086	-2.48207
XLOC_037880	"AC122844.1,Ampd3"	15.6027	5.69314	-1.45449
XLOC_037347	"2410002F23Ri,Gm15517,Gm23991,Gm28496,Snord88a"	77.8668	29.306	-1.40981
XLOC_022932	NA	63.5446	30.4563	-1.06103
XLOC_018141	Rn18s-rs5	739.126	361.342	-1.03245
XLOC_019686	NA	239.818	130.047	-0.882912
XLOC_008443	NA	366.051	204.452	-0.840284
XLOC_027655	NA	186.089	104.071	-0.838429
XLOC_043291	Lars2	54.1421	31.7281	-0.770992
XLOC_042541	Scn4b	6.89603	4.24921	-0.698573
XLOC_009051	"Mir134,Mir3072,Mir369,Mir377,Mir382,Mir410,Mir412,Mir485,Mir668,Mirg,mmu- mir-409"	109.997	189.122	0.781857
XLOC_020566	Acta2	11.2751	20.0289	0.828945
XLOC_016056	Myh11	1.39237	2.50618	0.847945
XLOC_040079	Igf2	23.9839	43.4464	0.85717
XLOC_035502	Slc13a4	4.54425	8.43836	0.892921
XLOC_035046	Slc6a13	3.03417	6.57791	1.11633
XLOC_031722	"RP24-299A7.2,Spp1"	2.94455	6.57757	1.15951
XLOC_022831	Itih2	0.527077	1.41824	1.42802
XLOC_035503	Fam180a	0.31679	0.878295	1.47118
XLOC_042756	Aldh1a2	1.71598	4.9266	1.52156
XLOC_022973	Ptgds	222.968	657.018	1.5591
XLOC_019840	Slc22a6	0.602158	1.8997	1.65756
XLOC_025417	Crabp2	0.670912	2.40458	1.84159
XLOC_018170	Capn11	0.260244	0.983628	1.91825
XLOC_042537	Mpzl2	0.135587	0.622483	2.19882
XLOC_001093	Fmod	1.32626	6.26375	2.23966
XLOC_035047	Slc6a12	0.282123	1.415	2.32641
XLOC_017192	H2-Q1	0.0610413	0.386116	2.66118
XLOC_011116	"Gm24991,Hist1h2ac"	0.595305	4.09392	2.78178
XLOC_001875	Gm29216	5.47729	120.998	4.46537

Table 5: Table of the 32 genes by Cuffdiff position (XLOC) that had a multipe-test correction $\alpha < 0.05$, sorted by $\log_2$(fold change)

Appendix D: Pipeline scripts

md5sum

The quality of transferred files was verified using md5sum, with a combination of a bash script and command line arguments.

The bash scripts md5sum_wf.sh:

```
#!/usr/bin/bash
# This script collects all the mdsums from the working directory and writes
    ↪ thems to a file called working_md5sum.txt
FILES=/middgenpilot/RNA_Seq_Data/Working_Directory/*.fastq.gz

for f in $FILES
do
        md5sum $f >> /middgenpilot/RNA_Seq_Data/Working_Directory/
            ↪ working_md5sum.txt
done
```

Following this script, the md5sums provided by the NYGC were collected into a single text file and compared to the sample files that had been copied into the working directory.

```
 $ cat Sample_219*/fastq/checksum/*.md5 >> NYGC_checksums.txt
$ paste NYGC_checksums.txt working_md5sum.txt | awk '$1!=$3'
```

Quality control

```
 FastQC script: fastqc\_mei2.5\_hippo1.sh
#!/usr/bin/bash

#Run fastqc for all compressed sample files
fastqc /middgenpilot/RNA_Seq_Data/mei2.5_hippo_1/2197-3
    ↪ _AGCGATAG_BC92MLANXX_L002_001.R1.fastq.gz
fastqc /middgenpilot/RNA_Seq_Data/mei2.5_hippo_1/2197-3
    ↪ _AGCGATAG_BC92MLANXX_L002_001.R2.fastq.gz
fastqc /middgenpilot/RNA_Seq_Data/mei2.5_hippo_1/2197-4
    ↪ _TCTCGCGC_BC92MLANXX_L002_001.R1.fastq.gz
fastqc /middgenpilot/RNA_Seq_Data/mei2.5_hippo_1/2197-4
    ↪ _TCTCGCGC_BC92MLANXX_L002_001.R2.fastq.gz
fastqc /middgenpilot/RNA_Seq_Data/mei2.5_hippo_1/2199-1
    ↪ _TCCGGAGA_BC92MLANXX_L002_001.R1.fastq.gz
fastqc /middgenpilot/RNA_Seq_Data/mei2.5_hippo_1/2199-1
    ↪ _TCCGGAGA_BC92MLANXX_L002_001.R2.fastq.gz
fastqc /middgenpilot/RNA_Seq_Data/mei2.5_hippo_1/2199-6_GAGATTCC-
    ↪ ATAGAGGC_BC8W0JANXX_L002_001.R1.fastq.gz
fastqc /middgenpilot/RNA_Seq_Data/mei2.5_hippo_1/2199-6_GAGATTCC-
    ↪ ATAGAGGC_BC8W0JANXX_L002_001.R2.fastq.gz
fastqc /middgenpilot/RNA_Seq_Data/mei2.5_hippo_1/2199-7_CGCTCATT-
    ↪ ATAGAGGC_BC8W0JANXX_L002_001.R1.fastq.gz
fastqc /middgenpilot/RNA_Seq_Data/mei2.5_hippo_1/2199-7_CGCTCATT-
    ↪ ATAGAGGC_BC8W0JANXX_L002_001.R2.fastq.gz
```

```
  And for resequenced reads: fastqc\_run2\_mei2.5\_hippo1.sh
 #!/usr/bin/bash

fastqc /middgenpilot/RNA_Seq_Data/mei2.5_hippo_1/2199-6
    ↪ _GAGATTCC_BC92MLANXX_L002_001.R*
fastqc /middgenpilot/RNA_Seq_Data/mei2.5_hippo_1/2199-7
    ↪ _CGCTCATT_BC92MLANXX_L002_001.R*
```

```
 Trimmomatic:   trimmomatic_mei2.5_hippo1.sh
#!/usr/bin/bash

#Specifying the directory that contains the files
FILES=/middgenpilot/RNA_Seq_Data/mei2.5_hippo_1/*R1_001.fastq.gz
OUTFILEPATH=/middgenpilot/RNA_Seq_Data/mei2.5_hippo_1/Trimmomatic_Out/
TRIMTAG=_Trimmomamtic_Filtered.fastq.gz

# Move into the Trimmomatic directory to gain access to the trimmomatic .jar
cd /home/whenriques/Trimmomatic-0.36

for f in $FILES
do
        #TEMP is the slice of the full file path that contains the sample ID #
        TEMP=${f:42:6}
        # FILEPATH and TEMP can be concatenated together with the following
            ↪ command
        #echo $FILEPATH$TEMP$TRIMTAG

        java -jar trimmomatic-0.36.jar PE -trimlog /home/whenriques/mei2.5
            ↪ _hippo_exc_dir/run_logs/Trimmomatic/trimmomatic_mei2.5_hippo2_"
            ↪ $TEMP"trimlog.log -basein $FILE -baseout
            ↪ $OUTFILEPATH$TEMP$TRIMTAG ILLUMINACLIP:adapters/TruSeq3-PE-2.fa
            ↪ :2:30:10 MAXINFO:40:0.5
```

Merging resequenced files

```
Merging resequenced samples with SamTools:  samtools\_merge\_mei2.5\_hippo2.sh
#!/usr/bin/bash

outpre=/middgenpilot/RNA_Seq_Data/Working_Directory/mei2.5_hippo_2/
    ↪ Samtools_Merge/stm_2199
in1pre=/middgenpilot/RNA_Seq_Data/Working_Directory/mei2.5_hippo_1/Tophat_Out
    ↪ /2199-
in2pre=/middgenpilot/RNA_Seq_Data/Working_Directory/mei2.5_hippo_2/Tophat_Out
    ↪ /2199-

samtools merge $outpre"-6/accepted_hits.bam" $in1pre"6_thout/accepted_hits.bam
    ↪ " $in2pre"6_2_thout/accepted_hits.bam"

samtools merge $outpre"-7/accepted_hits.bam" $in1pre"7_thout/accepted_hits.bam
    ↪ " $in2pre"7_2_thout/accepted_hits.bam"
```

Annotation-based quality control

```
   RSeQC: read_distribution.py:
   script: RSeQC_read_dist_hippo2.sh
   #!/usr/bin/bash
# This script runs read_distributions.py from the RSeQC packages on the mei2.5
    ↪ _hippo2 samples, and writes the output to a textfile.

#Tophat output directories for samples with one fastq file
topout=/middgenpilot/RNA_Seq_Data/Working_Directory/mei2.5_hippo_1/Tophat_Out
    ↪ /*_thout/
# Tophat output directory for samples with merged accepted_hits.bam
topout2=/middgenpilot/RNA_Seq_Data/Working_Directory/mei2.5_hippo_2/
    ↪ Samtools_Merge/stm*/

bed=/middgenpilot/RNA_Seq_Data/Reference_Genome/Mus_musculus/bed/
    ↪ mm10_Ensembl_mod.bed

outdir=/middgenpilot/RNA_Seq_Data/Working_Directory/mei2.5_hippo_2/RSeQC/
    ↪ read_distribution/

for f in $topout
do
        # Give the num of char of topout
        #echo ${#topout}
        # Pulls the id number from the file name
         idnum=${f:71:6}
        if [ "$idnum" == "2199-6" ] || [ "$idnum" == "2199-7" ]
                then
                        continue
        fi

        echo "Inputs"
        echo $idnum
        echo $bed
        echo $f"accepted_hits.bam"
```

```
        echo "Output Files"
        echo $outdir$idnum
        #ls $f
        #read_distribution.py -r $bed -i $f"accepted_hits.bam" >>
            ↪ $outdir$idnum"_RSeQC.txt"

done
echo "Between the loops"

for f in $topout2
do
        idnum=${f:79:6}
        echo "Inputs"
        echo $idnum
        echo $bed
        echo $f"accepted_hits.bam"
        echo "Output Files"
        echo $outdir$idnum_RSeQC.txt
        read_distribution.py -r $bed -i $f"accepted_hits.bam" >>$outdir$idnum"
            ↪ _RSeQC.txt"

done
echo "Done"
```

```
RSeQC: geneBody_coverage.py
script: RSeQC2_geneBody_cov_hippo2.sh
#!/usr/bin/bash
# This script runs geneBody_coverage.py from the RSeQC packages on the mei2.5
    ↪ _hippo2 samples, and writes the output to a textfile.

#Tophat output directories for samples with one fastq file
topout=/middgenpilot/RNA_Seq_Data/Working_Directory/mei2.5_hippo_1/Tophat_Out
    ↪ /*_thout/
# Tophat output directory for samples with merged accepted_hits.bam
mergeout=/middgenpilot/RNA_Seq_Data/Working_Directory/mei2.5_hippo_2/
    ↪ Samtools_Merge/stm*/

bed=/middgenpilot/RNA_Seq_Data/Reference_Genome/Mus_musculus/bed/
    ↪ mm10_Ensembl_mod.bed

outdir=/middgenpilot/RNA_Seq_Data/Working_Directory/mei2.5_hippo_2/RSeQC/
    ↪ geneBody_coverage/



for f in $topout
do
        # Give the num of char of topout
        #echo ${#topout}
        # Pulls the id number from the file name
        idnum=${f:71:6}
        if [ "$idnum" == "2199-6" ] || [ "$idnum" == "2199-7" ]
                then
                        continue
        fi

        #echo $f"accepted_hits.bam"
        echo "Inputs"
        echo $idnum
```

```
        echo $bed
        echo $f"accepted_hits.bam"
        echo "Output Files"
        echo $outdir$idnum
        #ls $f
        geneBody_coverage.py -r $bed -i $f"accepted_hits.bam" -o $outdir$idnum

done
echo "Between the loops"

for f in $mergeout
do
        idnum=${f:79:6}
        echo "Inputs"
        echo $idnum
        echo $bed
        echo $f"accepted_hits.bam"
        echo "Output Files"
        echo $outdir$idnum
        geneBody_coverage.py -r $bed -i $f"accepted_hits.bam" -o $outdir$idnum

done
echo "Done"
```

Cufflinks Assembly

```
script: cufflinks_mei2.5_hippo2_2.sh
#!/usr/bin/bash

#Tophat output directories for samples with one fastq file
topout=/middgenpilot/RNA_Seq_Data/Working_Directory/mei2.5_hippo_1/Tophat_Out
    ↪ /*_thout/
# Tophat output directory for samples with merged accepted_hits.bam
mergeout=/middgenpilot/RNA_Seq_Data/Working_Directory/mei2.5_hippo_2/
    ↪ Samtools_Merge/stm*/

#Each sample will have a cufflinks output directory at this path
cuffout=/middgenpilot/RNA_Seq_Data/Working_Directory/mei2.5_hippo_2/
    ↪ Cufflink_Out-2/

#To iterate through each Tophat output file and find the accepted_hits.bam
    ↪ file for each sample
for f in $topout
do
        # Give the num of char of topout
        #echo ${#topout}
        # Pulls the id number from the file name
        idnum=${f:71:6}
        if [ "$idnum" == "2199-6" ] || [ "$idnum" == "2199-7" ]
                then
                        continue
        fi

        #echo $f"accepted_hits.bam"
        echo "Processing sample"
        echo $idnum
        #ls $f
        cufflinks -p 8 -o $cuffout$idnum"_clout" $f"accepted_hits.bam"

done
```

```
echo "Between the loops"

for f in $mergeout
do
        idnum=${f:79:6}
        echo "Processing sample"
        echo $idnum
        cufflinks -p 8 -o $cuffout$idnum"_clout" $f"accepted_hits.bam"
done
echo "Done"
```

Differential Expression Analysis

```
Cuffmerge, Part 1: Gather paths to *.gtf files
script: assemblies_mei2.5_hippo2-2.sh
#!/usr/bin/bash

#Cufflinks output files
clout=/middgenpilot/RNA_Seq_Data/Working_Directory/mei2.5_hippo_2/Cufflink_Out
    ↪ -2/*_clout/

for f in $clout
do
        #For each sample, writes the path to the transcripts.gtf file into
            ↪ assemblies.txt
        # Which is a text file in the executing directory
        # This file will be used for Cuffmerge
        echo $f"transcripts.gtf" >> /home/whenriques/mei2.5_hippo_exc_dir/
            ↪ assemblies_hippo2-2.txt
done
```

```
Cuffmerge, Part 2: Run Cuffmerge
script: cuffmerge_mei2.5_hippo2-2.sh
#!/usr/bin/bash
#Run Cuffmerge
# Note: this may be combined with the assemblies script to streamline this
    ↪ process, I think
cuffmerge -o /middgenpilot/RNA_Seq_Data/Working_Directory/mei2.5_hippo_2/
    ↪ Cuffmerge-2_Out/ -g genes.gtf -s genome.fa -p 8 assemblies_hippo2-2.txt
```

```
Running Cuffdiff
script: cuffdiff_mei2.5_hippo2-2.sh
#!/usr/bin/bash

# This script uses the merged reads for samples 2199-6 and 2199-7
sing_read_pre=/middgenpilot/RNA_Seq_Data/Working_Directory/mei2.5_hippo_1/
    ↪ Tophat_Out/
merged_read_pre=/middgenpilot/RNA_Seq_Data/Working_Directory/mei2.5_hippo_2/
    ↪ Samtools_Merge/
mergeout=/middgenpilot/RNA_Seq_Data/Working_Directory/mei2.5_hippo_2/Cuffmerge
    ↪ -2_Out/
outdir=/middgenpilot/RNA_Seq_Data/Working_Directory/mei2.5_hippo_2/diff_out_2/

cuffdiff -o $outdir -b genome.fa -p 8 -L Mut,WT -u $mergeout"merged.gtf"
    ↪ $sing_read_pre"2197-1_thout/accepted_hits.bam",$sing_read_pre"2197-3
    ↪ _thout/accepted_hits.bam",$sing_read_pre"2197-4_thout/accepted_hits.bam"
    ↪  $sing_read_pre"2199-1_thout/accepted_hits.bam",$merged_read_pre"
    ↪ stm_2199-6/accepted_hits.bam",$merged_read_pre"stm_2199-7/accepted_hits.
    ↪ bam"


echo "Done"
```