AN IN SILICO ASSESSMENT OF ALTERNATIVELY SPLICED ISOFORMS IN THE MOUSE BRAIN USING RNA-SEQ

By

Daniel W. Bottomly

A THESIS

Presented to the Department of Medical Informatics and Clinical Epidemiology
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Master of Science

June 2009

School of Medicine

Oregon Health & Science University

Certificate of Approval

This is to certify that the Master's Thesis of

Daniel W. Bottomly

"*An In Silico Assessment of Alternatively Spliced Isoforms in the Mouse Brain using RNA-Seq*"

Has been approved

_____

Thesis Advisor – Shannon McWeeney PhD

_____

Committee Member – Robert Hitzemann PhD

_____

Committee Member – Aaron Cohen MD

TABLE OF CONTENTS

FIGURES

TABLES

ABBREVIATIONS

RPKM	Reads per Kilobase of Exon Model per Million Mapped Reads
RMA	Robust Multi-chip Average
AEU	Alternative Exon Usage
AE	Alternative Event
PTC	Percent Trans-read Contribution
GLM	Generalized Linear Model
CE	Cassette Exon Event
IR	Intron Retention Event
ME	Mutually Exclusive Exon Event
ATSS	Alternative Transcription Start SIte
APA	Alternative Poly Adenylation
FDR	False Discovery Rate

ACKNOWLEDGEMENTS

Shannon McWeeney PhD

Robert Hitzemann PhD

Aaron Cohen MD

Ted Laderas

Nicole Walter

Portland Alcohol Research Center

Department of Medical Informatics and Clinical Epidemiology

ABSTRACT

With the advent of next generation sequencing techniques like RNA-Seq, there is the potential for unbiased transcriptome-wide analysis of gene expression and alternative splicing irrespective of abundance class of the transcript. One of the potential uses of this technology is to help us understand the role that alternative splicing plays in brain region-specific differences. In this study we focused on two RNA-Seq datasets derived from the mouse brain, one from the striatum and the other from the whole brain. We first quantified the abundance of the different forms of alternative splicing events using a very conservative approach utilizing exon definition information from both the Ensembl and ASTD public databases. We then applied a measure that quantified transcript isoforms and examined whether biases existed in these quantities when stratified by overall gene expression measured using a microarray. Further, we explored whether there was concordance between alternative splicing events quantified using RNA-Seq and those measured using a statistical model for an Affymetrix exon array experiment. Finally we examined whether a simple model-based strategy could be pursued in order to detect alternative splicing in a high confidence dataset and explored some of the properties of this model using simulation. Overall, we found that a major confounder for many of our analyses was the lack of sample size. This is an issue that will be explored further in future work.

Chapter 1

Introduction

Background

It has been recognized that alternative pre-mRNA splicing is a common and important occurrence in mammalian genomes. Relatively modest differences in the number of predicted protein coding genes between different organisms have suggested that protein diversity is not due to large numbers of discrete gene regions, but selective utilization by the cellular machinery of the different transcribed structures (exons) within those units (1-2). Up until very recently it has been estimated that around 70% of genes are alternatively spliced between tissues in humans (3-4). However, using RNA-Seq it has been described as being a nearly universal occurrence for multi-exon genes (5). Alternative splicing is not strictly a beneficial phenomenon. Disruptions of exon splicing patterns through mutations are thought to result in disease, such as the neuromuscular disorder myotonic dystrophy (6) or certain cancers such as melanoma (7). Similar mutations have also been shown to influence disease and drug susceptibility, even providing targets for drugs (8). An example of such a target is the inhibition of a particular transcript variant of a gene, COX-1, by acetaminophen-type drugs (9). In order to fully understand the implications of alternative splicing, aberrant or otherwise, for disease or complex traits such as height or alcoholism, it is necessary to quantify transcript isoforms as well as gene expression levels for many samples and for many organisms. A technology that may allow this to be feasible is RNA-Seq.

Splicing occurs after the transcription of an initial RNA molecule from DNA when the intervening sequences or introns are removed from the RNA transcript allowing the remaining pieces known as exons to be joined together. Alternative splicing can provide different versions of transcripts that originate from a single gene (10). These species of transcripts can generally be

called transcript isoforms. According to the ASTD database (11), a collection of alternative splicing predictions based upon the Ensembl gene predictions (12), there are 5 different forms of alternative splicing events that commonly occur: intron and exon isoforms, cassette and mutually exclusive exons, and intron retention events (11). Intron and exon isoforms both involve a change in exon boundary relative to an exon in another transcript; the main difference being a requirement that the boundaries of an exon isoform be precisely defined (11). Cassette and mutually exclusive exons both involve the uneven incorporation of entire exons in one transcript isoform relative to another (11). Cassette exon events involve the skipping of one or more exons in one transcript isoform relative to another (11). Similarly, mutually exclusive events refer to transcript isoforms that each have a unique and non-overlapping set of skipped exons (11). Intron retention involves the complete incorporation of an intron in one transcript isoform when compared to another (11). Closely related to alternative splicing events are those defined to be alternative start sites and alternative polyadenylation sites (11). These respectively involve transcript isoforms that start at a later position in the genome or terminate earlier relative to another isoform (11). The preceding groups of transcript events can be referred to generally as alternative events (AEs). AEs for genes have been assessed for eukaryotic organisms in relation to developmental stage, tissue type or disease state—especially in the central nervous system (13). Regardless of context the most important and most difficult step in any AE analysis is the detection of all transcript isoforms present in an RNA sample. It is important because without knowledge of the types and quantity of exons being produced from one sample relative to another, it is hard to say with confidence whether the two samples produce significant quantities of different transcript isoforms. It is difficult because the ability to experimentally detect alternatively spliced exons and therefore AEs depends on several factors: resolution of the technology, sample size and expression level. The latter being the most

important limitation since the best estimate to date is that 86% of human genes produce measureable levels of two or more distinguishable transcript populations (5). Two main experimental methods have been traditionally used to assess genome-wide levels of AEs: sequencing and microarrays.

A direct way of measuring AEs is by sequencing mRNA or its reverse transcribed form—complementary DNA or cDNA (14). Large scale Sanger sequencing (15) of cDNAs was (and still is to a certain extent) labor intensive and expensive. To circumvent this, higher throughput tag-based methodologies for studying transcript populations were created. Expressed sequence tags or ESTs were formed by sequencing short fragments of cDNA that were still long enough to be reliably realigned to the genome (16). This made them faster and cheaper to produce than a whole cDNA sequence (16). Even today, these tags are considered to be relatively expensive, low throughput and susceptible to cloning bias (17). Serial Analysis of Gene Expression or SAGE (18) provided a faster alternative, though the tags were initially very short and limited to the 3' end of the transcribed region. They were later extended to the 5' end (19). Throughput was expanded by a proprietary SAGE-like method called massively parallel signature sequencing (MPSS) (20). However, it seems to be rarely used today. The current state of this technology is digital gene expression (DGE) using the Illumina GA platform which has been shown to be an effective way to assess expression (21). Cap analysis of gene expression or CAGE, similar to SAGE, produces sequences from the first 20 bases of each transcript (22). This procedure is immensely useful for defining the transcription start site and therefore alternative promoter usage but like SAGE and MPSS, cannot address any forms of alternative splicing (19). High throughput sequencing was first applied to cDNA sequencing with the polony multiplex analysis of gene expression or PMAGE (23). This technology produced tags of length 14 bases, which are

too small for accurate mammalian genome realignment for many situations. The Solexa/Illumina GA and the Applied Biosystems SOLiD platform were first applied to cDNA sequencing by several groups in 2008 (17, 24-27).

Prediction of alternative splicing events relies upon the utilization of spliced alignments of cDNAs and/or tag data to a genomic sequence using fast realignment tools such as BLAT (28), GMAP (29), SPA (30) or more recently Splign (31). Also, smaller regions have been assessed using programs such as SIM4 (32) or Spidey (33). This has also been attempted for Illumina GA output using the QPALMA program (34) which is an adaptation of the PALMA program (35) to handle short reads. Very recently Tophat was described, which promises to provide efficient production of spliced alignments for these datasets (36). However, spliced alignments do not actually provide predictions of alternative splicing events; other programs have to be used in conjunction with their output as part of a prediction pipeline. Many such pipelines exist for annotation of genes using cDNAs and ESTS. One that produces AE predictions exclusively is ASTD (11). The framework for this pipeline was initially described by Clark and Thanaraj in 2002. ASTD uses Ensembl (12) gene predictions and expands upon them by applying a combination of BLAT, BLAST and custom heuristics to look for different intron/exon boundaries than those annotated in Ensembl (11). CAGE and SAGE-like tags are constrained in the types of AEs they can detect because of the way they are derived experimentally. Multiple groups of CAGE tags occurring within a gene can be used to infer alterative promoter usage (22). Similarly, SAGE-like tags can only assess alternative splicing events at the opposite end of the transcript such as alternative poly-A events (21, 37-38).

In addition to measuring gene expression, microarrays can also be used to detect certain types of AEs (39). Splice-junction, exon and tiling arrays have all been successfully used to

detect such events. Splice junction arrays rely on synthesized probes that span annotated boundaries between exons—so called splice junctions (40-41). Using the signal from these probes, differences in expression between predefined exon combinations can be measured and used to infer splicing, alternative or otherwise (40-41). Tissue specific alternative splicing has been examined using variations on this approach for mammalian organisms (3, 42-49). It also has been addressed using a similar approach involving fiber optic microarrays (50). These experiments are constrained to cassette or mutually exclusive exon events that have to be defined *a priori* (39). Exon arrays (51-52), microarrays containing probes present at several positions within annotated exons, can be used for the detection of *de novo* splicing events (4, 51, 53-55). However, the ability to detect such changes in expression may be affected by quality and placement of the probe (55) and SNPs for some types of experiments (56). Genome level tiling expression arrays were created as an attempt to lessen the impact of probe placement by dramatically increasing the number of probes. AEs have been assessed using this approach for Drosophila (57) and yeast (58-59). A relatively new approach is the use of "whole-transcript" custom microarrays that can provide multiple sources of information regarding transcript expression and alternative splicing (60).

Although multiple strategies have been developed for high throughput sequencing, three have been successfully commercialized and implemented in the context of transcriptome sequencing. They are the 454 platform (61) now part of Roche, Illumina GA formerly Solexa and the Applied Biosystems SOLiD machine, an adaptation of a protocol by Shendure et al. 2005 (62). Roche's 454 platform is a high-throughput implementation of pyrosequencing, a procedure first described by Hyman in 1988 (63) and implemented by Ronaghi et al in 1996 (64). Pyrosequencing relies on the detection of released inorganic phosphate from DNA polymerase

activity using luciferase (64). Although first utilized in the context of *de novo* sequence assembly, it was extended to transcriptome sequencing of a prostate cancer cell line (65). The Solexa/Illumina and Applied Biosystem's platform both implement versions of cyclic reversible termination chemistry (66). For both, sequencing depends on the utilization of a base or probe combined with a fluorescent group that can be successfully removed at the end of each base-calling cycle (66). Solexa/Illumina initially gained popularity through studies of chromatin binding using a procedure called ChIP-Seq (67-70). In 2008 it was used to sequence the transcriptome of tissues from multiple organisms including: *A. Thaliana*, *S. Pombe*, *S. Cerevisiae*, mouse and human (17, 25-27, 71-72). SOLiD has been used to sequence the transcriptome of several mouse embryonic stages (24).

Most of the groups using either the Illumina or SOLiD platform have addressed AEs with variations on a single method. This method utilizes public gene prediction tools and/or databases to query for the presence of known alternative splicing events as well as to create *in silico* splice junctions. Two of the earliest experiments were carried out in yeast (25, 27). Wilhelm et al. 2008 utilized both Illumina sequencing and a tiling microarray to analyze the transcriptome of *S. pombe*. They identified both alternative start and poly-A sites, and examined splice junctions using what they referred to as "trans-reads"—reads that spanned exons (27). Using *S. Cerevisiae,* Nagalakshmi et al. 2008 examined gene boundary definitions and through this strategy they identified possible alternative poly-A usage. It was Mortazavi et al. 2008, applying the Illumina platform to a subset of tissues from the mouse that first directly addressed alternative splicing (17). This was done by mapping otherwise unmapped reads to a database of known splicing events from UCSC (17, 73). As the only group to apply the SOLiD platform, Cloonan et al. 2008 also implemented a unique approach to alternative splice

detection (24). In addition to querying a public database for exon splicing definitions, they adapted a *de novo* assembly algorithm, VCAKE (74), to assemble overlapping non-mapping reads and realigned them to the mouse genome using BLAT (24, 28).

The use of database-defined splicing events was extended to *de novo* splice site detection through the mapping of trans-reads to splice junctions found from the pair-wise connection of defined exon sequences. Sultan et al 2008 followed this strategy for human HEK293T and Ramos B cells (26); Marioni et al. 2008 followed an expanded version for human liver and kidney cells (75). Instead of pair-wise concatenation within a gene, Marioni et al. 2008 concatenated exons in a pair-wise manner for the entire genome, probing them for trans-reads (75). A simpler approach was taken by Rosenkranz et al. 2008 who looked for trans-reads that resulted from 1, 2 or 3 exon skipping events in mouse ES cells (72). Pan et al. 2008, choosing to implement their own annotation pipeline, aligned publically available mRNA and EST sequences to the human reference genome using a combination of BLAST (76) and SIM4 (32, 77). From this they created an exon junction database, searched for trans-reads and successfully used a logistic regression and a decision tree to discriminate real from false junctions (77). They showed that the percent inclusion (%in) could be used to assess changes in the inclusion level of cassette exons in transcript isoforms between tissues (77). The %in measure was then shown to correlate with their previous microarray experiments and GenASAP algorithm (77). The most impressive and complete examination of alternative splicing has been carried for a large number of human tissues by Wang et al. 2008 (5). Combining gene predictions from GENSCAN (78) and EXONIPHY (79) along with three databases containing exon definitions, they created a database of all possible exon junctions (5). Relying upon their exon predictions and Ensembl (12) gene boundaries for their junction database, they successfully mapped trans-reads to these junctions

(5). Wang et al. 2008 also made several unique contributions to the field. The first was the idea of the inclusion ratio, which estimated the proportions of the different transcript isoforms in the sample (5). The second was the percent spliced in or PSI(ψ) which was an estimate of the proportion of the transcripts that contained an exon, a construct similar to the %in concept discussed by Pan et al. 2008 (5). With these measures they examined several AE characteristics related to this study. They examined the extent and quantity of tissue-specific and individual specific AEs through the use of a Fisher's exact test (5). The test addressed whether the number of reads that were included and excluded from AEs between tissues (and individuals) were significantly different (5). They then estimated the difference using a measure termed the inclusion ratio (5). Both cassette and mutually exclusive exons were grouped by their change in ψ to look at the features that accompanied large changes between tissues (5).

Since the paper by Wang et al. 2008, other proposals have been made on how best to quantify and compare the transcript isoform levels of one or more samples. Jiang and Wong 2009 devised a model that simplified to the RPKM measure for the case of a single transcript, but used a Bayesian approach to model the isoform ratio for the cases where there were multiple transcripts (80). Similarly, Zheng et al. 2009 implemented a hierarchical Bayesian model to attempt to declare transcript isoforms differentially expressed between multiple samples (81).

Study Aims

For this study there were two main objectives. The first was to utilize C57BL/6J (B6) mouse inbred strain whole-brain and striatum RNA-Seq datasets to address the question whether alternative splicing differs between a heterogeneous and homogenous tissue within an

inbred mouse strain. The first step was to realign the reads from both datasets to a reference genome and a set of splice junctions formed using Ensembl (12) and ASTD (11) definitions. This was done using both reads that mapped solely to a unique location in the genome and combining the unique reads with multi-mapping reads that were mapped to their most probable position. This process is described in Chapter 2. From the reads that mapped to splice junctions, alternative splicing was inferred using the positional relationships of the verified splice junctions. The analysis of these events and comparison of alternative splicing between the mouse whole-brain and striatum is the subject of Chapter 3. The work on quantifying transcript isoforms carried out by Wang et al. 2008 (5) and Pan et al. 2008 (77), was of much use since it allowed a better assessment of the makeup of the transcriptome than could have been provided by microarrays. A comparison of the isoform ratio calculation (5) with data from exon arrays along with a comparison of a generalization of the %in (77) to an exon array alternative splicing detection platform is the subject of Chapter 4.

Most of these procedures have not drawn on the full potential of this technology. By limiting characterization of AEs to splice junctions, which can be examined using microarrays, they lose many of the advantages of the increased resolution that may be provided by RNA-Seq technology. A possible solution to all three of these problems would use the read depth measure of expression from these experiments to identify promising AEs. If a population of mRNA molecules is considered; for a given gene many types of transcripts can be produced via AEs. Each of these transcripts can have exons that differ in size and/or number from the exons in other transcripts. From an RNA-Seq experiment, if two exons from separate alternatively spliced transcripts overlap based upon genomic position, then both should contribute reads to the dataset at a level consistent with their expression. These reads, when realigned back to the

genome, should produce locally higher read depth in the overlap region compared to a region of non-overlap. A form of this idea has been successfully used by Wang et al. 2008 to identify tissue regulated AEs and compare the proportion of corresponding transcript isoforms between tissues and individuals (5). However, it was of interest to see if this could be extended to the detection of overlapping exon regions within a gene. The results of which, given a cDNA or genomic sequence, could predict whether non-annotated isoforms of that gene transcript existed and the proportion of the sample they make up. Before this can be done, the first step is to determine whether these local differences in read depth are both present and identifiable for well defined exons and AEs. This was the second objective of this study. The most obvious way of accomplishing this was through the use of generalized linear models to assess significance of the read count, which could be used as a rate. The results and implications of an exploratory analysis of this type is the subject of Chapter 5. An overall summary and concluding remarks is made in Chapter 6.

Chapter 2

Read Realignment Strategy and Implications

Introduction

In order to use RNA-Seq as a measure of expression, these data have to be mapped back to reference sequences. For sequencing performed by Illumina, this procedure has typically been done by ELAND, which is their proprietary realignment software. Since ELAND is proprietary, it is not readily available to groups who do not own their own sequencer. To deal with this problem, many open source short read mapping software programs have now been created. Examples of these include Maq (82), RMAP (83), and Bowtie (84). In addition to read malformations that can cause mis-mapping (85), the process of sequencing mRNA transcripts results in some reads that will fail to map because they are derived from transcript regions containing splice sites—areas where two transcribed exons were spliced together (17). In addition to providing better utilization of the data, mapping reads to the areas bordering splice sites, known as splice junctions, provides evidence for splicing and more interestingly alternative splicing (17). The standard way to look for these splice site crossing reads has been to form *in silico* splice junctions based upon exon definitions from gene predictions (68). Generally speaking this has been done in two ways: a guided approach or a combinatorial approach. The guided approach was first carried out by Mortazavi et al. 2008 and involved forming splice junctions based upon previously known splice sites (17). The combinatorial approach exemplified by Wang et al 2008 and Pan et al. 2008, involved using all known exons and concatenating the exons together in every possible manner for each gene (5, 77). The benefit to the latter approach was that novel alternative splicing isoforms could be discovered, whereas

the former approach only looked for and measured the existence of known events. More recently, spliced alignment programs have been developed especially for use with the Illumina GA and other high throughput sequencers. The purpose of these programs is to predict splice junctions from the genomic alignment of reads with large insertions or deletions (indels) in their sequence similar to the more traditional BLAT program (28). These new RNA-Seq spliced aligners include QPALMA (34) and the recently developed Tophat (36). Since the data from reads mapped to *in silico* splice junctions and from splice alignments would essentially be the same, it would be relatively easy to allow the utilization of either type of analysis in a software package.

One of the more interesting techniques employed by two of the first groups to apply high-throughput sequencing to mRNA was the use of a multi-mapping assignment strategy (17, 24). Briefly, this involved assigning reads that mapped to multiple locations to their most probable location based upon relative expression levels (17, 24). Mortazavi et al. 2008 showed that the use of the multi-reads increased a gene-level correlation with a microarray experiment from an R^2 value of .62 to a value of .69. However, they did not mention whether it was statistically significant (17). The approach they used was similar to a maximum entropy strategy where they assigned fractions of reads to locations based upon the unique read count (17). Of special interest was the utility of these multi-reads for alternative splicing analyses, which to my knowledge, has not been thoroughly addressed. The goal of this chapter was to examine the issues regarding splice junction formation and read assignment. This was done in relation to the use of multi-reads in the context of an analysis concerning alternative splicing in the mouse brain.

Methods

In order to verify and quantify the transcript population in the C57BL/6J inbred strain of mouse, a guided splice junction formation approach was taken using the exons defined in Ensembl (12) in conjunction with the alternative splicing events from the alternative splicing and transcript diversity database (ASTD) (11). To do this, the exon boundaries were retrieved from Ensembl build 53 (12) and the alternative splicing events were retrieved in BED format from the February 2008 release of ASTD (11). To provide global access to sequence, the 2bit compressed version of the mm9 mouse genome build (86) and an accompanying extraction tool was downloaded from UCSC (73). This extraction tool was used to pull out all of the reference strand sequence for both the Ensembl (12) and ASTD (11) exon boundary definitions. This sequence was reverse complemented if the gene was on the negative strand. From these sequences, only $k - 4$ bases were kept from each end, if bordering a splice site, where k was the read length. Starting from the first exon as defined in Ensembl (12), each exon boundary region was concatenated with the neighboring exon with whom it shared the intron. For the situation where exons were smaller than the read length, bases were used from the exon(s) downstream or upstream of the splice site in question. Bases were borrowed from the neighboring exons in this manner until either the desired splice junction size was met, or there were no more exons in the gene. The ASTD (11) events were prepared similarly. First, since the ASTD (11) data release was built using the older mm8 genome, the coordinates were converted to the newer mm9 genome build using LiftOver (73). These events were then merged with the Ensembl (12) exons, looking for any overlap between the Ensembl (12) exon definition and the alternative ASTD (11) definition. The successfully merged events were used to determine in what order the ASTD (11) exons were concatenated with the Ensembl (12) exons, supplementing the splicing information

provided by the event definition. For example, a cassette exon event would be guided both by the ASTD (11) internal boundaries and by the Ensembl (12) definitions for the external boundaries—the first and last exon of the event. Similarly, splice junctions involving exon isoforms would be formed using the bordering Ensembl defined exons (12). These sets of splice junctions from Ensembl (12) and ASTD (11) were then filtered to remove redundancy for the situations where the two covered the same splice junctions. This was expected to occur because ASTD (11) reported both the normal splice event, which generally was covered by Ensembl (12) and the alternative exon definitions which generally were not.

The Bowtie short-read realigner was used to map the Illumina RNA-Seq reads to both the mm9 genome and the collection of splice junctions. For this analysis the datasets consisted of the whole-brain dataset from Mortazavi et al. 2008 (17) and the striatum RNA-Seq dataset from the Portland Alcohol Research Center (PARC). Both of these were derived from the same strain of inbred mouse: C57BL/6J, also known as B6. Each alignment was carried out in four steps. First, the reads were mapped to the genome separating those that mapped uniquely from the rest. Second, the reads that did not map to unique positions were remapped, keeping those that were placed in fewer than 10 possible positions. Those that mapped to more than 10 locations were discarded and those that failed to map anywhere were placed in a separate file. The non-mapping reads from this file were realigned to the splice junction set, again keeping any that mapped uniquely or to fewer than 10 locations.

Reads that mapped equally well to multiple places, known as multi-mapping reads, were assigned based upon the strategies by Mortazavi et al. 2008 and Cloonan et al. 2008. The multi-mapping reads were assigned to a position based upon the read depth of the regions that were $2k$ in length for the case of genomic reads and $2(k\text{-}4)$ for splice junction reads. Reads were

assigned at a level proportional to the relative read depth, so that the multi-read expression would be adjusted in a manner similar to what has been termed a rich-get-richer approach (87). The original reason for using such an approach was to ensure that the data was kept as whole counts, so as not to interfere with any foreseeable downstream statistical tests. Summary statistics and plots were then generated using R (88). Correlation plots were generated to make some basic inferences about the properties of the mapped reads. Expression was measured using a measure of normalized read count referred to as the reads per kilobase of exon model per million mapped reads (RPKM) (17). However, this measure was calculated differently than the method proposed in Mortazavi et al. 2008 (17). It was calculated by averaging the read count that was normalized by the exon length, over the number of defined Ensembl transcripts (12). This was further normalized by the number of million mapped reads. It differed from the calculation of Mortazavi et al. 2008 mainly because it did not take into account the splice junction reads (17). This was done originally in an effort to make the results more comparable to common microarray platforms which also do not interrogate splice junctions. Three basic plots were formed. First, the RPKM (17) measures for each gene were plotted against expression summarized at the gene-level from exon array experiments. These experiments used existing microarray datasets from the PARC that were also derived from mouse whole-brain and striatum. These microarray datasets were renormalized and analyzed using the methods from the ExonModelStrain package (55) utilizing only the core probes. See chapter 4 for more details. The RPKM (17) measures from both datasets were plotted against one another to look for both biologically meaningful results and to identify any potential biases in the microarray datasets. Finally, the length and normalized read count for splice junctions were plotted against gene RPKM (17). The splice junction read counts were calculated individually for each splice junction and the read count was divided by the length of the splice junction, 62 or 42 in most cases and

the number of million mapped reads. Correlation coefficients were computed using Spearman's rank correlation and confidence intervals were computed through bootstrapping. Comparing confidence intervals was a straight-forward way to test for differences between Spearman's rank correlations.

Results

There were a total of 266,005 distinct splice junctions formed from the combined set from Ensembl (12) and ASTD (11). Of these there were 201,404 that were from Ensembl (12) and 64,601 from ASTD (11). From the total list, 9,094 were redundant and subsequently removed. For the striatum dataset 727,112 reads were uniquely assigned to a position. A total of 5,578 multi-mapping reads were present and could be assigned. All together, a total of 105,033 splice junctions were found with at least one unique read mapping to them. Similarly, there were 799,483 unique reads assigned from the whole-brain dataset, with 101,145 multi-reads that were present and successfully mapped. This led to a total of 100,048 verified splice junctions.

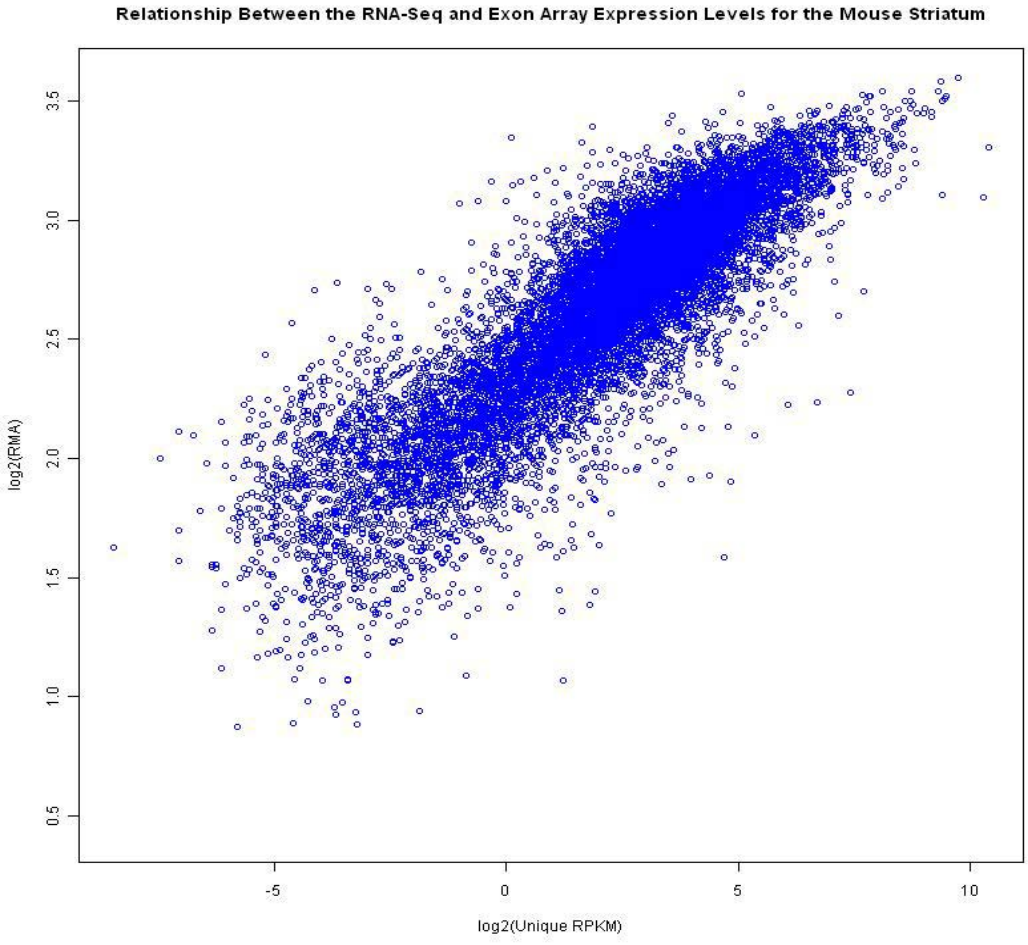
For the striatum dataset there were a total of 13,948,348 reads and 9,728,363 were assigned to a unique position in the genome. Bowtie reported 1,831,535 multi-mapping reads, 871,499 of these mapped equally well to multiple positions and 960,036 mapped to one position with a fewer number of mismatches than the other positions. These latter reads were flagged, but were included in the unique read counts. Of the true multi mapping reads, 643,955 could be assigned to one position with a greater probability than the others and were reassigned. The others mapped to regions where no uniquely mapping reads were present, and therefore could not be assigned to a position. This led to a total of 11,332,354 reads assigned to

the genome. Similarly for the whole-brain dataset, there were 31,116,663 reads, 12,552,370 mapped to a unique location with 1,660,481 of the 3,530,701 multi-reads mapping to a one position with fewer mismatches than the others. Of the remainder, 1,402,910 were reassigned to one position with the greatest probability. This resulted in a total of 15,615,761 reads. The genomic reads and splice junctions combined resulted in a grand total of 12,065,044 reads for striatum and 16,516,389 for whole-brain.

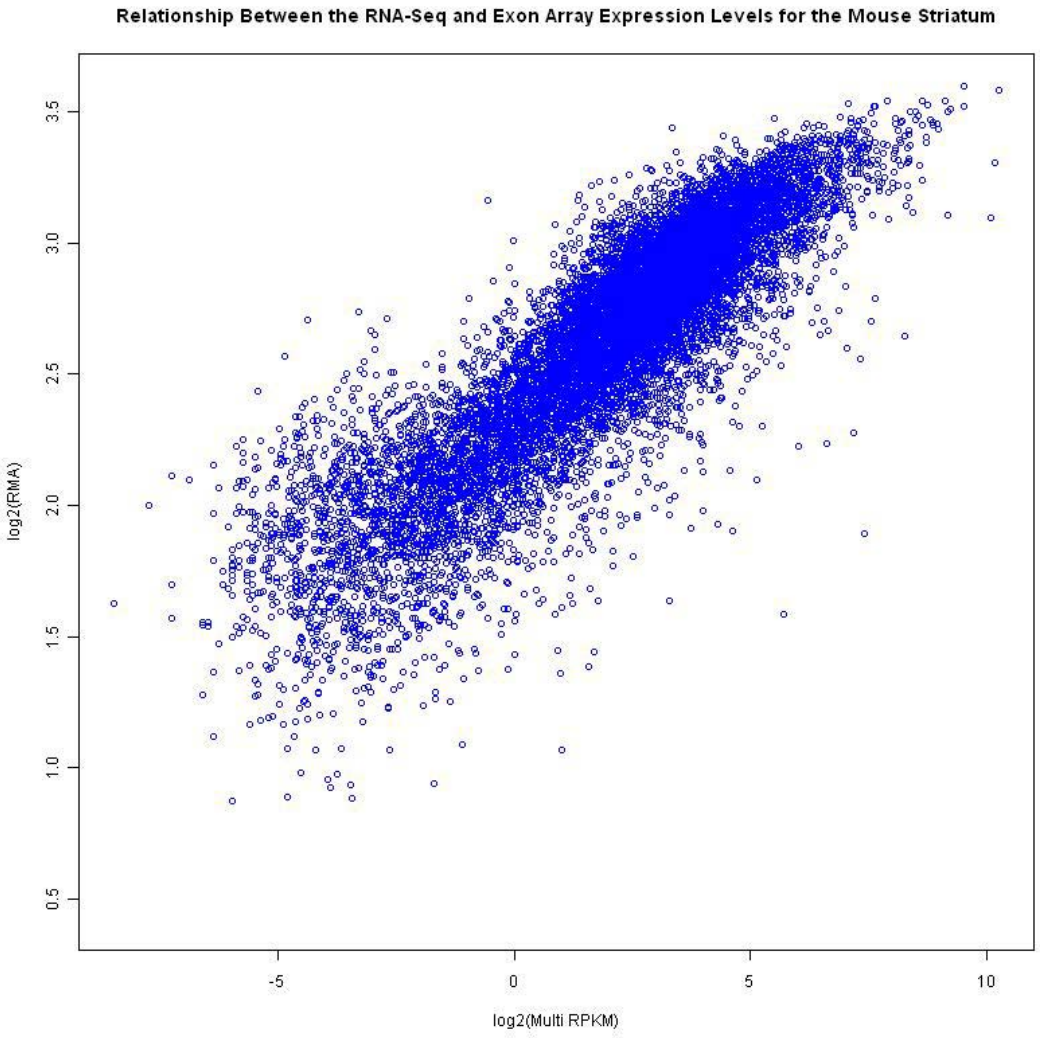
Fig. 1 shows the RPKM (17) calculated using both the unique (1a) and unique plus multi (1b) reads from striatum plotted against the RMA (89) from the corresponding microarray experiment. There was a high correlation between these datasets using both the unique and unique plus multi-mapping reads (rho= .893 95% CI: (0.888, 0.897) for unique reads and .898 95% CI (0.894, 0.902) for unique + multi-reads; Spearman's rank correlation, bootstrap CIs). The correlation was higher between the unique plus multi-mapping read RPKM (17) and the microarray experiment, than the corresponding unique-only mapping experiment, however the difference was not significant. Likewise, Fig. 1 c and d show the unique and unique plus multi-mapping RPKM (17) values for whole-brain plotted against the microarray experiment. Again, the correlation was higher, though non-significant for the unique plus multi-mapping read RPKM (17) than for the uniquely mapping version (Rho= 0.809 95% CI (0.802, 0.816) for unique reads vs. 0.812 95% CI (0.804, 0.819) for unique + multi-reads; Spearman's rank correlation). Shown in Fig. 2 were the RPKM (17) values for both whole-brain and striatum plotted against each other. Interestingly in this case, 2a which plots the uniquely mapping reads, had a significantly higher correlation than 2b, which plots the unique plus multi-reads (rho=0.908 95% CI (0.905, 0.911) for unique reads and 0.892 95% CI (0.889, 0.896) for unique + multi-reads; Spearman's rank correlation, bootstrap CIs). Figure 3 shows the relationship between the splice junction

normalized read depth and the overall gene expression as measured in RPKM (17) for both whole-brain and striatum. The correlation values were computed using Spearman's rank correlation and rho was estimated to be .886 (95% CI: 0.880, 0.892; bootstrap CI) and .871 (95% CI: 0.864, 0.878; bootstrap CI) for the uniquely and uniquely plus multi-mapping reads in the striatum and likewise .894 (95% CI: 0.888, 0.899; bootstrap CI) and .883 (95% CI: 0.877, 0.889; bootstrap CI) for the whole-brain.

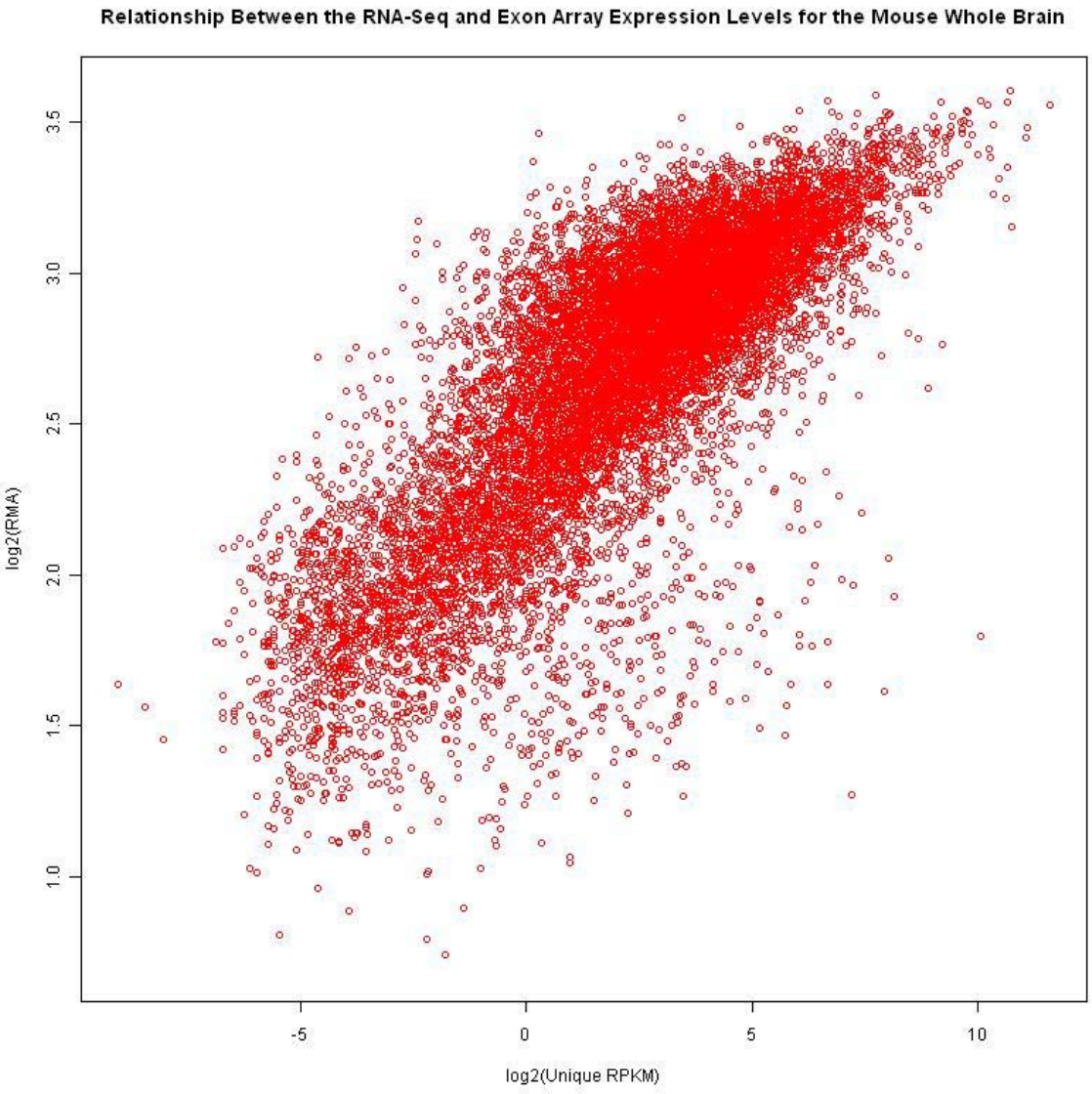
1a.)



1b.)


Relationship Between the RNA-Seq and Exon Array Expression Levels for the Mouse Striatum
log2(RMA)
log2(Multi RPKM)
0.5
1.0
1.5
2.0
2.5
3.0
3.5
-5
0
5
10

1c.)


Relationship Between the RNA-Seq and Exon Array Expression Levels for the Mouse Whole Brain
log2(RMA)
1.0
1.5
2.0
2.5
3.0
3.5
-5
0
5
10
log2(Unique RPKM)

1d.)

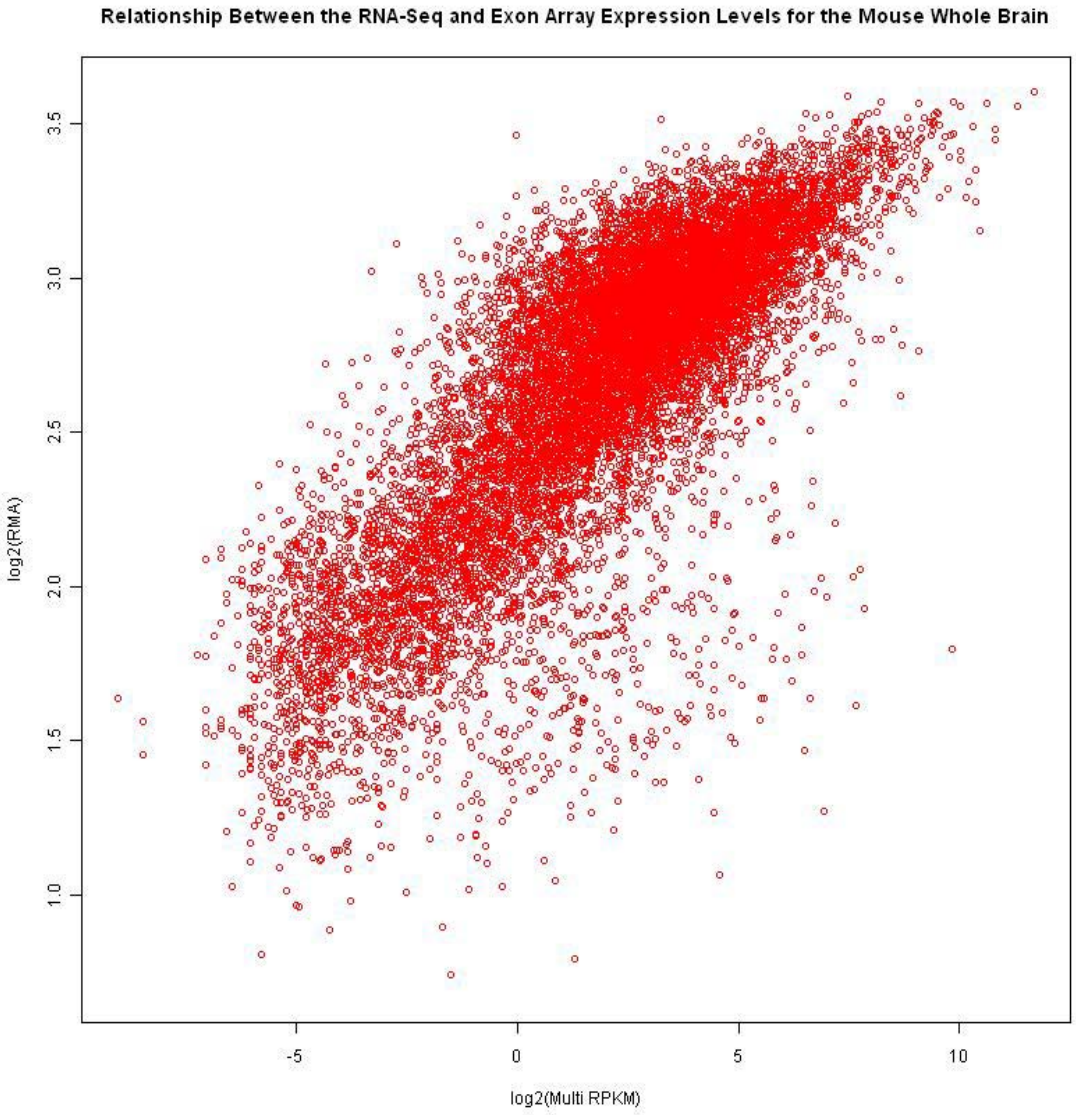
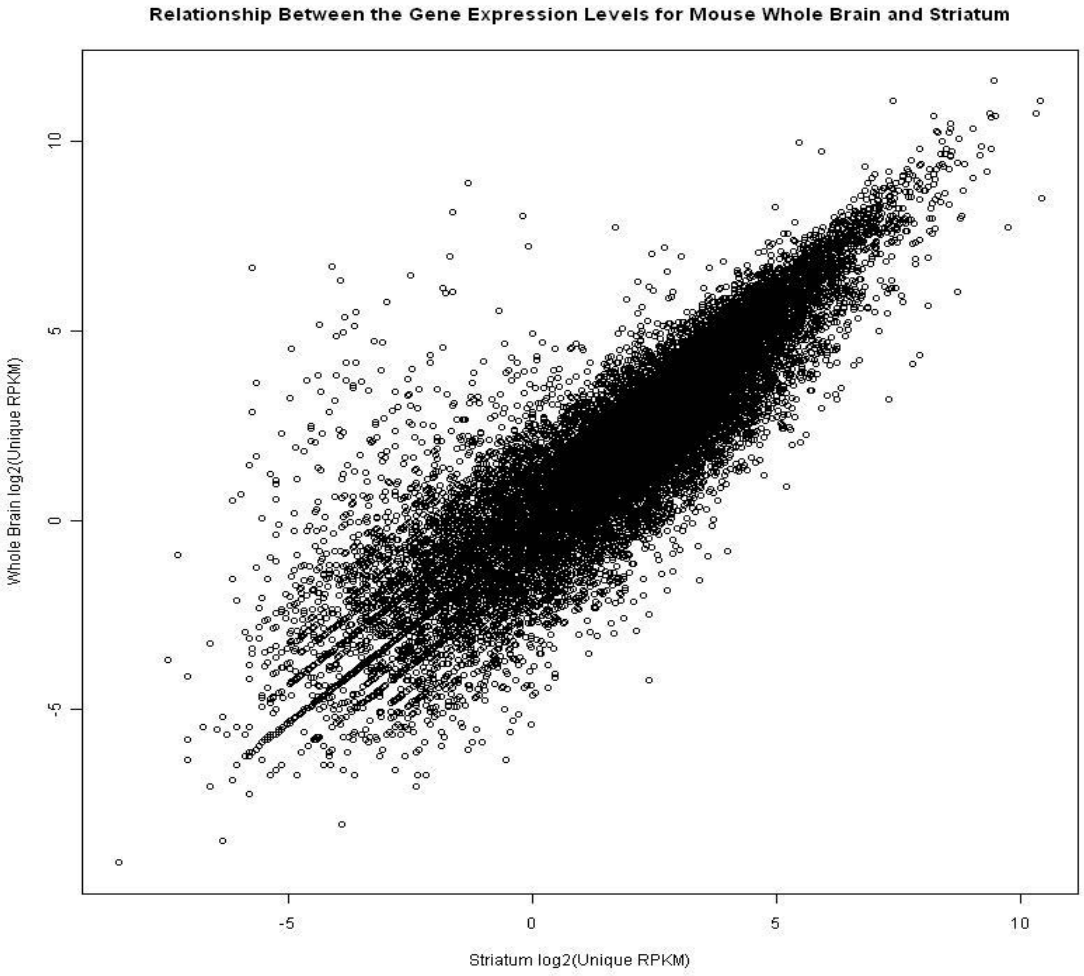


Figure 1. Relationship between RPKM and microarray RMA for mouse striatum and whole-brain. The log base 2 RPKM (17) was calculated using the uniquely (a, c) and uniquely plus multi (b, d) mapping reads (x-axis). The uniquely mapping reads were those assigned to one position best using the Bowtie (84) realignment program. Multi-mapping reads were assigned by determining their most probable Bowtie (84) assigned position based upon read depth for windows of length 2*k* for genomic reads or 2(*k*-4) for splice junctions. Where *k* was the read

length. These results were plotted against normalized gene-level striatum or whole-brain gene expression (89) from an exon array experiment (See chapter 4) also log base 2 transformed (y-axis). Correlation for (a) and (b): rho= .893 (95% CI: 0.888, 0.897; bootstrap CI) and .898 (95% CI: 0.894, 0.902; bootstrap CI) using Spearman's rank correlation. Correlation for (c) and (d): rho = .809 (95% CI: .802, .816; bootstrap CI) and .812 (95% CI: .804, .819; bootstrap CI) using Spearman's rank correlation.

2a.)



2b.)

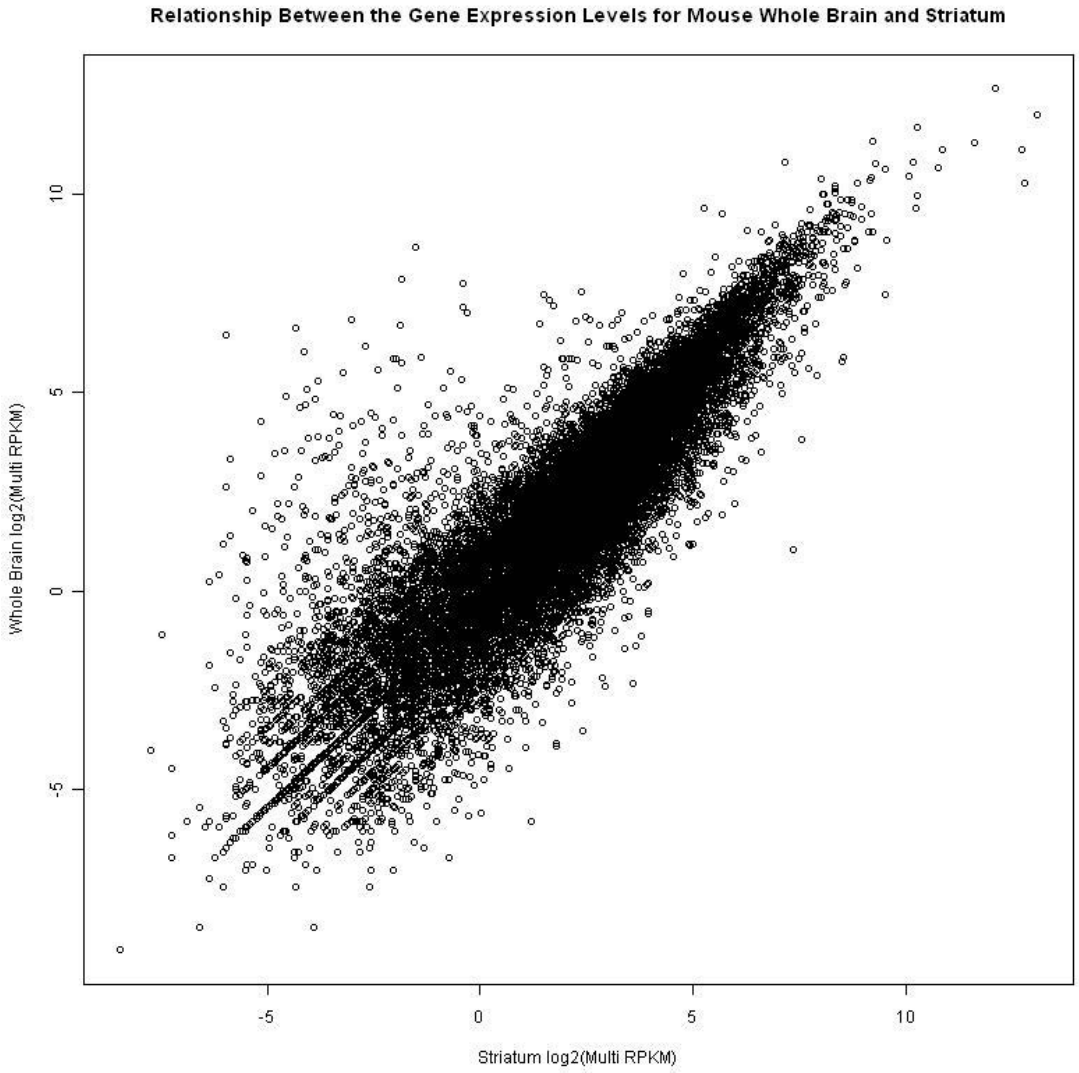
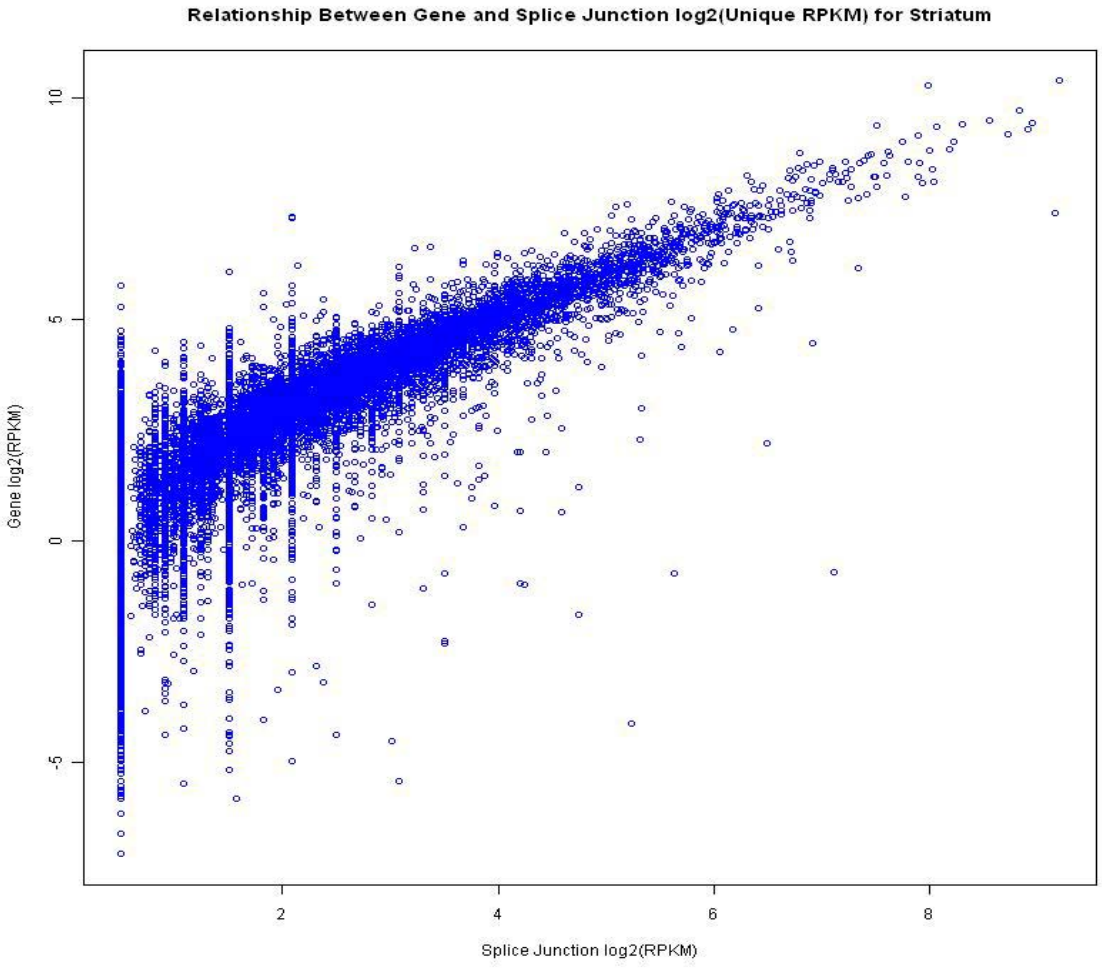


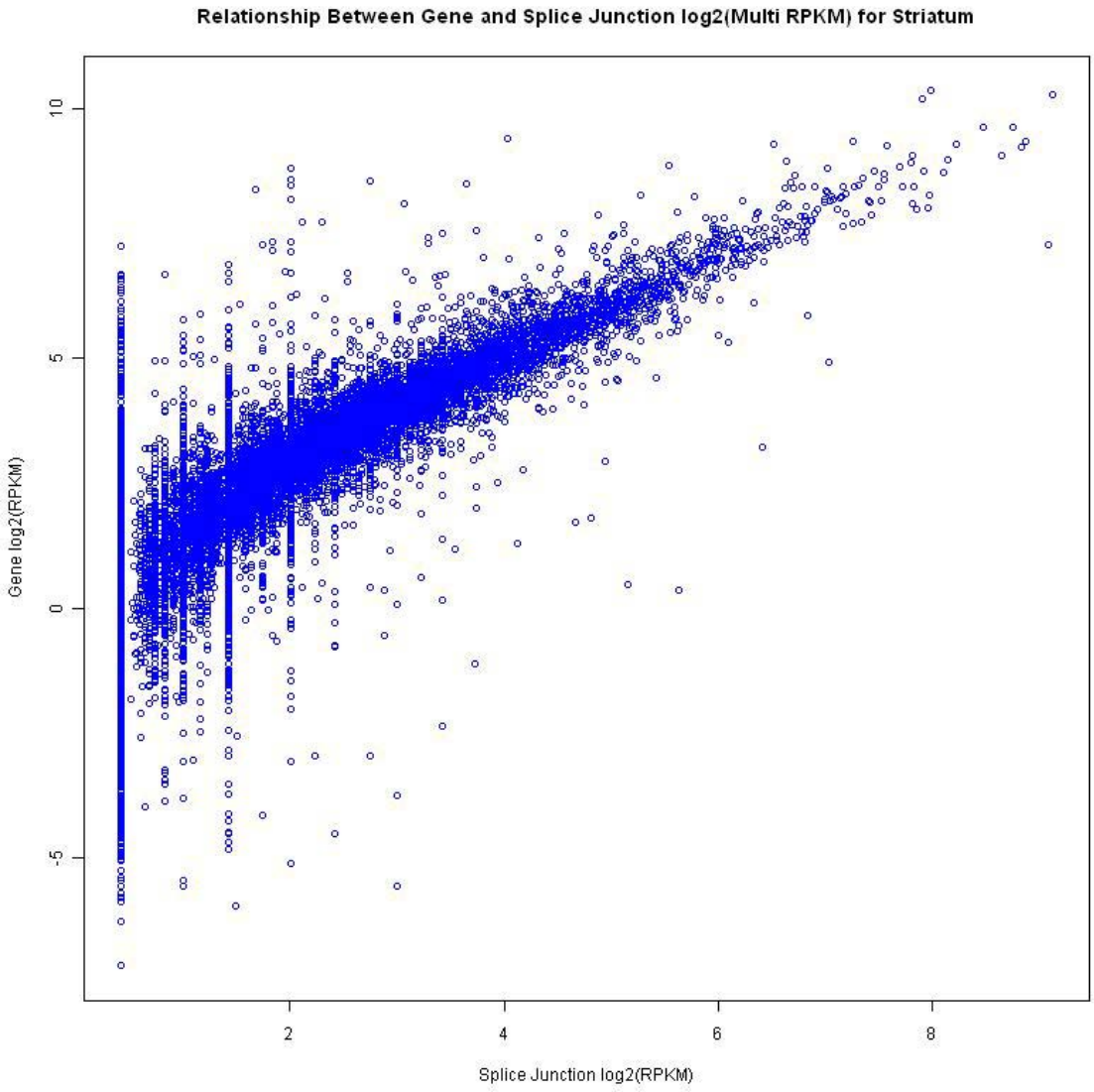
Figure 2. Relationship between mouse whole-brain and striatum RPKM expression. Whole-Brain RPKM (17) expression (Y-axis) was plotted against striatum RPKM (17) expression (X-axis). This was done for both the uniquely mapped reads (a) and uniquely plus multi-mapped reads (b). Unique reads were assigned to their position using the Bowtie realignment program (84). Multi-mapping reads were assigned to their most probable positions from Bowtie (84) using read depth from windows of 2*k* for genomic positions and 2(*k*-4) for splice junctions. Where *k*

was the read length. RPKM (5) values were log base 2 transformed. The correlation values were rho=0.908 (95% CI: .905, .911; bootstrap CI) for (a) and rho= 0.892 (95% CI: .889, .896; bootstrap CI) for (b) using Spearman's Rank Correlation.

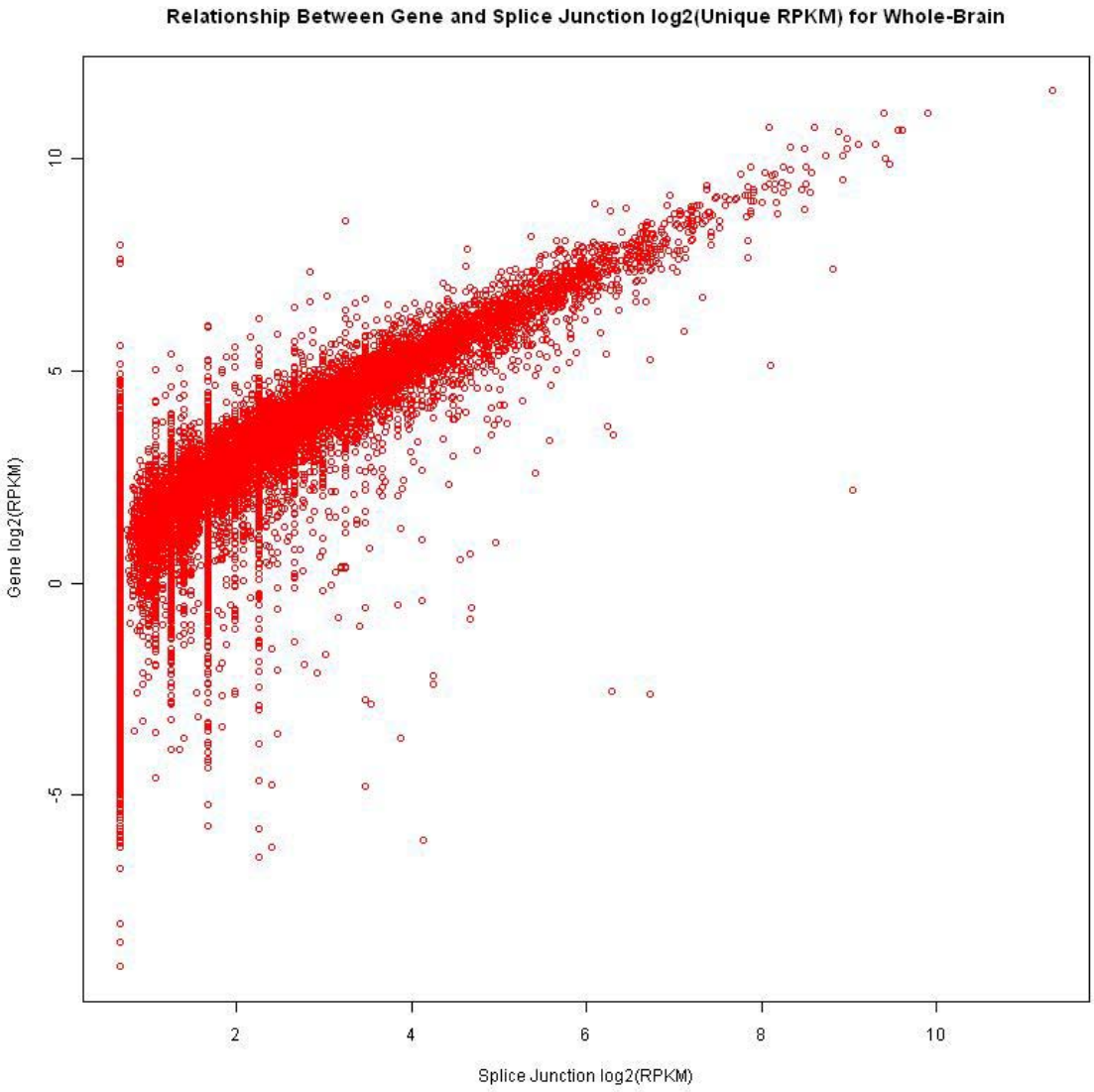
3a.)



3b.)


Relationship Between Gene and Splice Junction log2(Multi RPKM) for Striatum
Gene log2(RPKM)
10
5
0
-5
2
4
6
8
Splice Junction log2(RPKM)

3c.)


Relationship Between Gene and Splice Junction log2(Unique RPKM) for Whole-Brain
Gene log2(RPKM)
10
5
0
-5
2
4
6
8
10
Splice Junction log2(RPKM)

3d.)

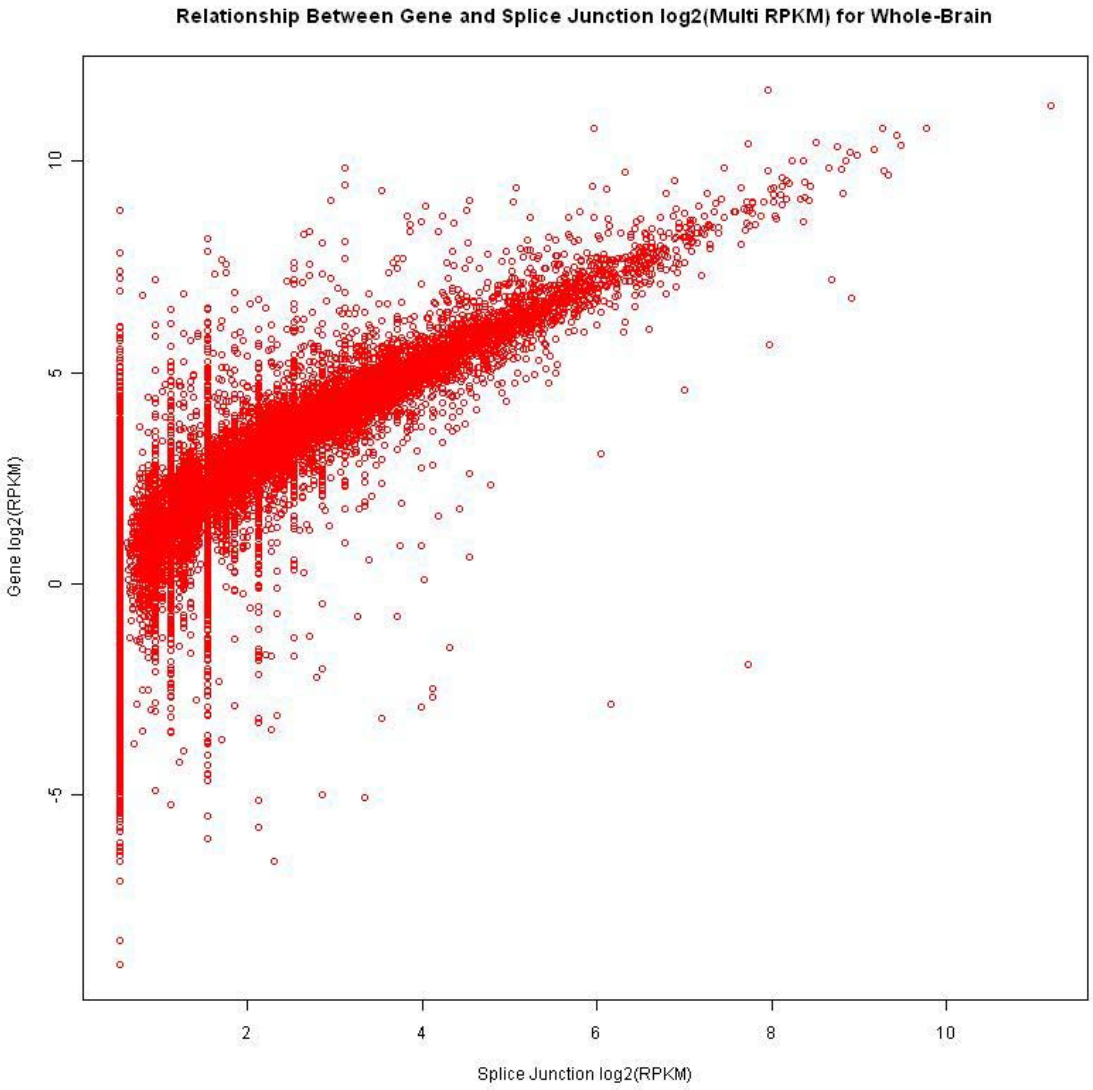


Figure 3. Relationship between gene and splice junction RPKM expression for mouse whole-brain and striatum. The RPKM (17) measure was calculated for each gene region (Y-axis) and compared to the normalized read count for each splice junction within the gene (X-axis). The read count was normalized by both the splice junction length in kilobases and the number of million mapped reads. Unique reads were assigned to the positions given to them by Bowtie

(84). Multi-mapping reads discovered by Bowtie (84) were assigned to their most probable positions based upon the read depth of genomic windows of size 2*k* and splice junction windows of size 2(*k*-4). Where *k* was the read length. Plots for both unique and uniquely plus multi-mapped reads for striatum are shown in (a) and (b), for whole-brain they are shown in (c) and (d). Correlation values were .886 (95% CI: .880, .892; 95% CI) and .871 (95% CI: .864, .878; bootstrap CI) for unique and multi-mapping reads for the striatum respectively. For whole-brain the correlation values were .894 (95% CI: .888, .899; bootstrap CI) and .883 (95% CI: .877, .889; bootstrap CI) for the unique and multi-mapping reads respectively using Spearman's rank correlation.

Discussion

Similar to the results given in Mortazavi et al. 2008, the inclusion of multi-mapping reads for the striatum dataset both increased the total read count from 10,455,475 to 12,065,044, which is about 86.5% of the total number of reads used as input and improved correlation with a microarray experiment. However, the 25 base whole-brain reads created for their experiment were not able to be assigned at an equivalent rate to the mouse genome. All together 53.08% total were successfully mapped or reassigned. This rather excessive failure rate was also mentioned by Mortazavi et al. 2008, since they reported that only 50.3% of their reads mapped uniquely to the genome—similar to the 45.6% achieved here using a different realignment algorithm. It seems likely that read length was the main cause of the difference in the percentage of mapped reads between the two datasets. As shown by Whiteford et al. 2005, the ability to place a read uniquely in a complex mammalian genome is related to read length. This problem was compounded even further by the presence of base-calling errors that increase in likelihood at the read ends (85). If a read had too many errors, it would not be realigned successfully to the genome.

The normalized gene expression levels stayed in approximately the same range for both the whole-brain and striatum datasets when compared to the microarray experiment, even though the number of reads successfully mapped was different (Fig. 1). However, the whole-brain dataset had more data points spread out and shifted upward, resulting in a lower correlation. This meant that the RMA (89) levels reported were higher than would be expected from the RPKM (17) levels. One possible explanation was that the relative failure to map reads, unique or otherwise, resulted in fewer reads mapping to the exons thereby lowering the reported expression level for the RNA-Seq dataset. This also may have caused more reads to be present for the highly expressed and/or longer transcripts (90) than would be expected, resulting in overestimation of the RPKM (17) measure as compared to the RMA (89), something that was also observed.

Comparing Fig. 2a and 2b, we saw that the whole-brain and striatum were very similar in gene expression levels, as would be expected. Again, the whole-brain seemed to have a group of genes that had higher than expected expression levels. Whether this was due to true biological processes or technical noise still needs to be determined.

The expression level of splice junctions as compared to entire genes is shown in Fig. 3. The two expression levels were highly correlated with the gene expression level containing a much greater dynamic range. All plots in Fig. 3 showed some under-expression of genes in relation to splice junctions at lower overall expression levels, something that would be expected given that the splice junctions were constrained in size. Interestingly, even though the inclusion of multi-reads decreases correlation, the outliers seemed to become more evenly distributed to both sides of the main body of points resulting in a more equal level of over and under estimation. Overall, the inclusion of multi-reads seemed to improve the homoscedasticity of the splice junction versus gene expression plots at the expense of the correlation.

Of special interest was the observation that although using multi-mapping reads for these samples seemed to not have an effect on correlation with a microarray dataset (Fig. 1), use of multi-mapping reads decreased the correlation between two similar datasets and between splice junction (Fig. 3) and gene-level expression (Fig. 2). This may have indicated that rather than improving the expression values, the inclusion of multi-reads may have caused them to be overestimated. Likewise, the microarray experiment may have overestimated gene expression as well, and lead to an improved correlation value when compared to the overestimated expression values. These results may have been unique to this analysis since the calculations used here were different than those performed elsewhere. One concern was that the use of this multi-mapping assignment approach may bias the read count in favor of the highly expressed splice junctions and would not improve our ability to detect rare splice variants. It may be beneficial to adapt this approach to use fractional count assignment for multi-mapping reads similar to that used by Mortazavi et al. 2008 as opposed to the whole-read assignment strategy. Since the read fractions would be distributed in a relative manner based upon unique read count, this strategy might reduce the overestimation currently seen in splice junctions when compared to gene expression. Further analysis is necessary to determine whether the use of multi-mapping reads provides a benefit to RNA-Seq experiments and may be addressed in future studies.

Chapter 3

Alternative Splicing in the Mouse Whole-Brain and Striatum

Introduction

The demonstrated uses of RNA-Seq have extended beyond simply quantifying gene expression to interrogating the diversity of the transcript populations themselves. The developers of the RNA-Seq methodology showed that some of the reads that failed to map to the genome could be realigned to sequences consisting of *in silico* splice junctions (17). These splice junctions were composed of a set of exons gathered from the knownGene collection from UCSC (73) that were concatenated together to simulate the sequence present in an mRNA transcript, but separated by intronic sequences in the genome (17). However, the Mortazavi et al. 2008 group did not infer alternative splicing isoforms but indirectly hinted at their presence by looking at the number of exons mapped to by each splice crossing read. This methodology was expanded to look for *de novo* splicing events by Sultan et al. 2008 and Marioni et al. 2008. Existing exon definitions were supplemented with predictions for Pan et al. 2008 and Wang et al. 2008. The splice site-centric view taken by the latter groups can be easily harnessed to identify putative splicing and alternative splicing events based upon relatively simple heuristics. In this manner alternative splicing and the conceptually similar phenomena of alternative transcription start sites (ATSS) and alternative poly-A sites (APA) can be detected. These ideas have been implemented to globally assess alternative splicing in both the mouse whole-brain and striatum using splice junctions whose creation was guided from Ensembl (12) exon definitions and the alternative splicing and transcript diversity database (ASTD) (11) events.

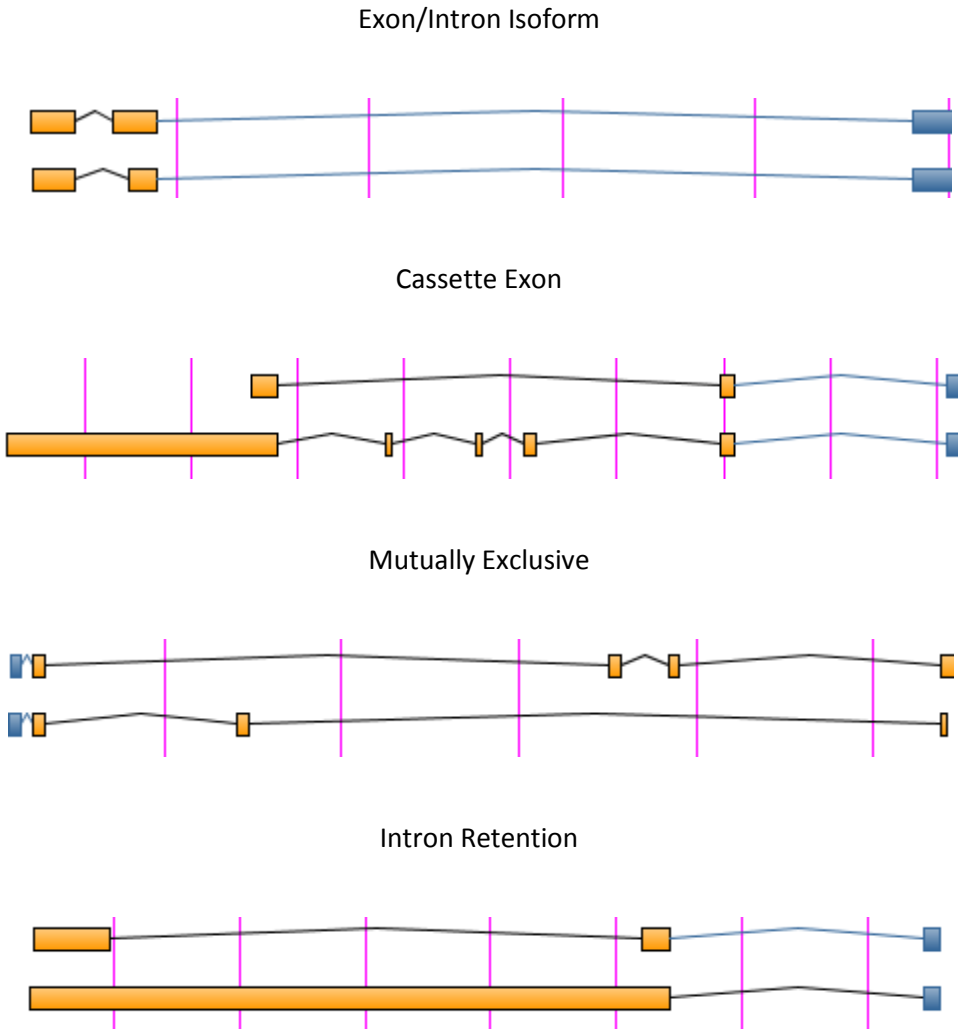


Figure 4. Graphical depiction of alternative splicing events. These pictures were retrieved from the alternative splicing and transcript diversity database (ASTD) (11). Exon/intron isoforms were used as one category to generally refer to any event where two different exons in separate transcripts overlapped, but had different boundaries (11). A cassette or skipped exon referred to any pattern where one or more exon was skipped in one transcript versus another (11). The mutually exclusive event was similar, with each transcript containing at least one skipped exon (11). Intron retention events referred to situations where there was a lack of splicing in one

transcript relative to another (11). In these representations, orange and blue boxes represented exons and the black or blue lines represented introns (11).

Methods

Two sets of splice junctions were formed and reads were mapped to them as described in Chapter 2 for whole-brain and striatum. Six types of events were searched for using the definitions given by ASTD (11). These events were the exon/intron isoform, cassette and mutually exclusive exons, intron retention and ATSS and APA events. The ASTD (11) depiction of these events is shown in Fig. 4. To detect these events all splice junctions with at least one unique read mapped to them were extracted and organized by chromosome. All distinct splice junctions from ASTD (11) or Ensembl (12) were associated with their full length exons and the coordinates of these exons were used to sort the splice junctions by starting genomic coordinate position. For those from genes on the negative strand, the start and end coordinates had to be reversed to maintain compatibility with those on the positive strand. These splice junctions were then stepped through and the genomic coordinates of the first and second exons of each consecutive junction were compared. Examples of these relationships are shown in Fig. 5. Note that the relationships shown in Fig. 5 are simplified and that the actual implementation was further generalized. Some of these generalizations included looking for overlap as opposed to exact matching based upon boundaries and procedures that could handle multiple alternative splicing events occurring within the same grouping of splice junctions. The described procedure in addition to judicious filtering to remove inaccuracies and redundant events was used to create a set of alternative splice junctions for both the whole-brain and striatum. Certain ATSS and APA events could also be determined in this manner. They are shown in Fig. 6.

Since putative ATSS and APA events were only supported by one splice junction instead of two, they were associated with two other types of data that could provide evidence for their existence: CAGE tags (22) and poly-A signals. For the whole-brain, CAGE tags from the FANTOM3 project (91) were retrieved from RIKEN (91). The tags used were of length 20 bases and derived from adult mouse whole-brain. They were mapped to the genome using Bowtie (84) allowing up to 2 mismatches. The count of these tags was taken from an area 50 bases upstream of the start of each exon to 50 bases downstream or the length of the exon, whichever was shorter. For the APA events, a collection of 13 poly A signals (Table 4) that were used in the AltPAS pipeline of ASTD (11) were searched for in all exons for the transcripts that had at least one uniquely mapping read assigned to each Ensembl-defined (12) splice junction. This search required an exact match to one of the sequences in Table 4.

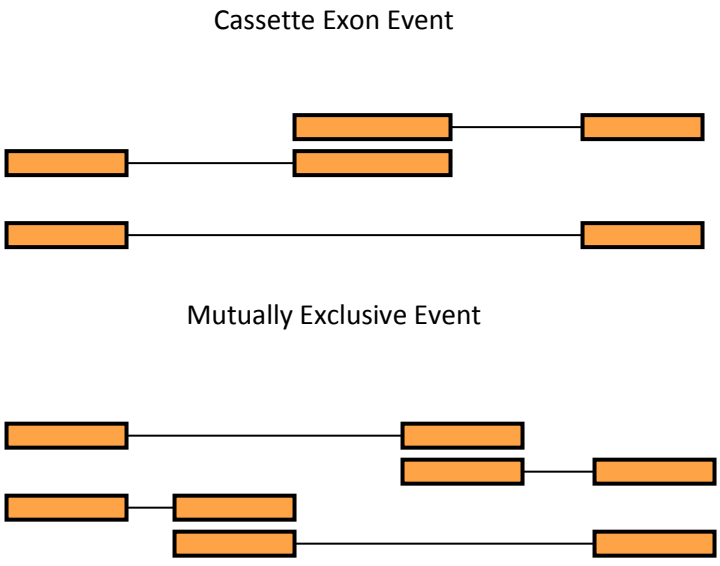


Figure 5. Splice junction spatial relationships for two alternative splicing events. Alternative splicing events were detected using the relationship of splice junction positions along a chromosome. These are examples of such splice junction layouts that would be classified as a

cassette exon (top) and mutually exclusive event respectively (bottom). Shown in orange are exons and the black lines separating them represent introns.

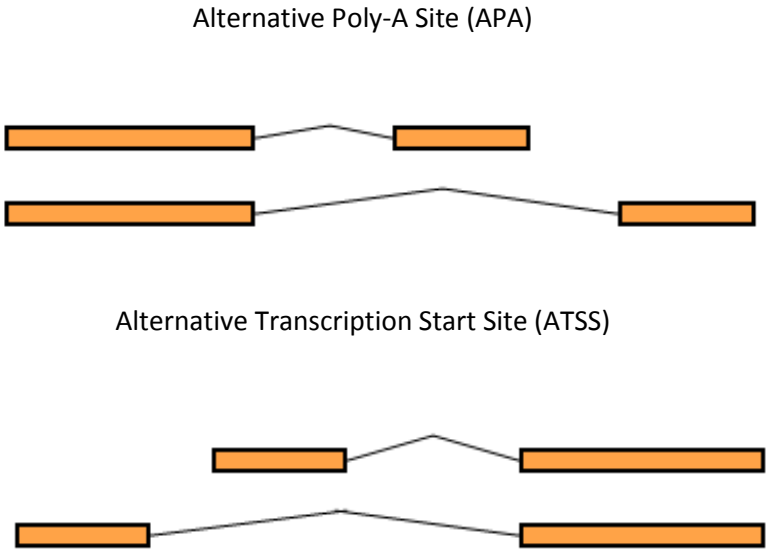


Figure 6. Graphical depiction of APA and ATSS events. The only detectable APA and ATSS events via splice junctions consisted of those where one exon was behind (trailing) of or in front of (leading) another at the start or end of a gene. The leading exon will be considered to be the unique outermost boundary exon. The trailing exon will be considered the unique exon following the leading exon. The orange boxes represent exons and the black lines represent introns.

Results

The total number of events found is shown in Table 1 categorized by type. The most common type of event was the cassette exon with a total of 1,398 events, followed by the exon/intron isoform with 1,024 events observed. At the other end of the spectrum there were only of 31 mutually exclusive exon events and 37 APA events. A total of 145 and 183 intron retention and ATSS events were also found. Some of the events were only seen in one sample,

the overall overlap for the two samples are depicted in Fig. 7 and are shown broken down by category in Table 2. Overall, the majority of events overlapped between tissue types and this held true for half of the individual categories with the exceptions of ATSS, APA and intron retention events. Possible reasons for lack of overlap of alternative splicing events were the different levels of sequencing (sample size) and potential differences in expression between the two samples. Since events were only counted if all splice junctions within the event were present it was possible that for the genes with lower expression some junctions within an event might have lacked a unique splice-crossing read and the event therefore would not have been counted. To examine this possibility, summary statistics for the splice junctions involved in the unique and common alternative splicing events were tabulated and are shown in Table 3. Although there were massive outliers for both, the mean and the median average read depth values for the common events were higher than those for the unique events. This was true when both uniquely and uniquely mapping plus multi mapping reads were counted. This indicated that the events may have only been unique in one tissue because of the failure to map any unique reads to one or more participating splice junctions in the other tissue.

Sample	Isoforms	Cass. Ex.	Mut. Ex.	Intron Ret.	Alt. TSS	Alt. PolyA	Total
Whole-Brain	533	661	15	78	101	16	1404
Striatum	491	737	16	67	82	21	1414

Table 1. Alternative splicing events seen in the whole-brain and striatum. This table shows the total count of each type of alternative transcript event for both the whole-brain and striatum requiring a minimum of one uniquely mapping read to be assigned to each splice junction. Shown are the exon/intron isoform (Isoforms), cassette exon (Cass. Ex.), mutually exclusive (Mut. Ex.), and intron retention splicing events (Intron Ret.). Also shown are alterative

transcription start sites (Alt. TSS) and the related alternative poly-A events (Alt. PolyA). The events are defined in Fig. 4 and Fig. 6.

Sample	Isoforms	Cass. Ex.	Mut. Ex.	Intron Ret.	Alt. TSS	Alt. PolyA
Both	302	369	10	39	45	8
Whole-Brain	231	292	5	39	56	8
Striatum	189	368	6	28	37	13
% Total	39%	48%	1%	5%	6%	1%
Wang. et al	19%	32%	.05%	.05%	32%	16%

Table 2. Common and unique events for each type of event. Shown was the number of each type of alternative splicing event seen in whole-brain or striatum uniquely as well as those seen in both regions. The event types shown are exon/intron isoforms (Isoforms), cassette (Cass. Ex.), mutually exclusive exon (Mut. Ex.), and intron retention (Intron Ret.) events. Alternative transcription start (Alt. TSS) and poly A (Alt. PolyA) events are also shown. In order for an event to be defined it had to have at least one uniquely mapping read assigned to each splice junction. The relative quantity of each category observed (% Total) was compared to a similar measure seen by Wang et al. 2008 from a large survey of a variety of human tissue types. Also note that Wang et al. 2008 also had data for another category which was not included in this analysis. Alternative events are defined in Fig. 4 and Fig. 6.

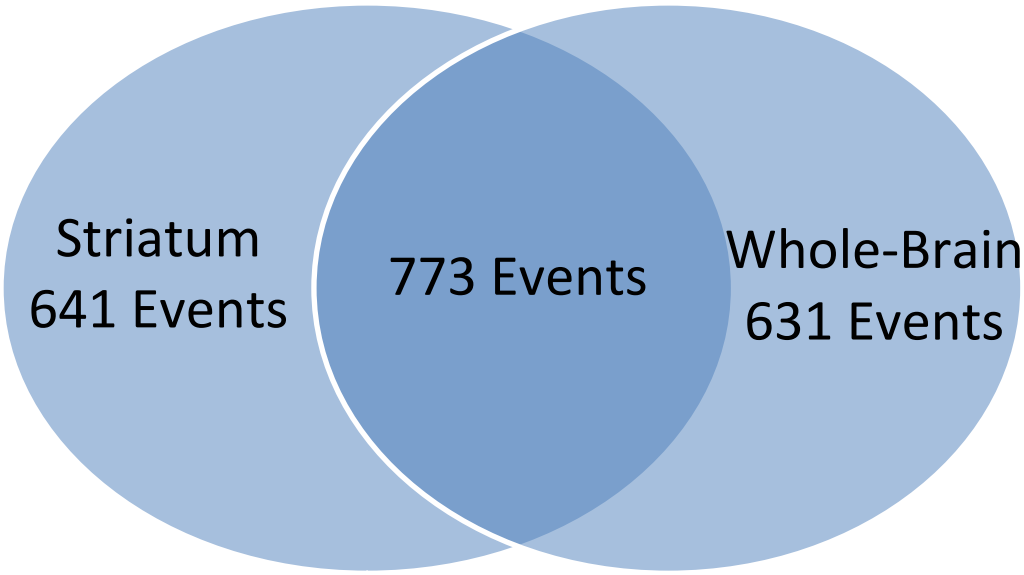


Figure 7. The number of alternative splicing events that were shared between the two tissues. Shown is the overlap between the total alternative splicing events seen in only striatum, only whole-brain or in both. In order for an event to be defined it had to have at least one uniquely mapping read assigned to each splice junction.

Uniquely mapping read depth for the common striatum events				
Minimum	Median	Mean	Max	Std. Dev.
1.00	4.00	10.86	520.00	27.06
Uniquely mapping read depth for the common whole-brain events				
Minimum	Median	Mean	Max	Std. Dev.
1.00	5.00	14.23	1129.00	45.85
Uniquely mapping read depth for events unique to striatum				
Minimum	Median	Mean	Max	Std. Dev.
1.00	2.00	4.30	130	7.08
Uniquely mapping read depth for events unique to whole-brain				
Minimum	Median	Mean	Max	Std. Dev.
1.00	2.00	7.032	445.00	18.91

Multi mapping read depth for common striatum events				
Minimum	Median	Mean	Max	Std. Dev.
1.00	5.00	11.21	523.00	27.73
Multi mapping read depth for common whole brain events				
Minimum	Median	Mean	Max	Std. Dev.
1.00	5.00	14.56	1194	47.14
Uniquely mapping read depth for events unique to striatum				
Minimum	Median	Mean	Max	Std. Dev.
1.00	2.00	4.43	130	7.25
Uniquely mapping read depth for events unique to whole-brain				
Minimum	Median	Mean	Max	Std. Dev.
1.00	2.00	7.31	457.00	19.42

Table 3. Summary statistics for alternatively spliced junctions. Summary statistics are presented for the number of reads mapped to alternative event junctions for both those events that were unique to whole-brain or striatum compared to those that were shared between the two. This was done distinguishing the counts of uniquely-mapping reads from the counts of unique reads plus multi mapped reads. Alternative events were detected as defined in Fig. 4 and Fig. 6, requiring a minimum of 1 uniquely mapped read.

For striatum, out of the 8,433 transcripts that were defined in their entirety with at least one read crossing each splice junction, 4,576 contained a poly-A signal at the last exon that met the defined criteria. Of the 21 candidate alternative poly-A events seen in striatum, 7 contained at least one signal in the leading exon (Fig. 6) and 4 contained a signal in the trailing exon. All of the leading exons and 14 of the trailing exons were defined to be the last exon in a transcript by Ensembl (12). Similarly, for whole-brain there were 8,978 transcripts defined with 5,043 of them containing the canonical consensus signals (11). Eight of the leading and 4 of the trailing exons contained a signal out of the 16 total sets of leading and trailing exons. Again, all of the leading and 14 of the trailing exons were considered to be the last exons by Ensembl (12). There were a total of 25,341 CAGE tags from the whole-brain library (91) and 11,551 were

successfully remapped to the genome at unique positions. Of the 8,978 whole-brain verified transcripts, 1,454 had at least one CAGE tag mapped to the defined first exon. All of the leading and 82 of the trailing exons were defined to be first exons in Ensembl (12). Of these 7 of the leading and 11 of the trailing contained one or more CAGE tags.

Discussion

The proportions of the different types of alternative splicing events were similar to what was seen in Wang et al. 2008, the largest such study to date (Table 2). They found that out of their set of 37,782 events seen between one or more human tissues, around 32% were cassette exons, .05% were intron retention events, 19% were exon/intron isoforms, .05% were mutually exclusive, 32% were ATSS and 16% were APA. Another group, tandem 3' UTRs, made up the rest (5). This compares to around 48% cassette, 39% exon/intron isoforms, 1% for both mutually exclusive and APA and 5 and 6% for intron retention and ATSS respectively for this experiment. Mortazavi et al 2008, who created the whole-brain RNA-Seq dataset, reported 1,516 alternatively spliced genes by defining an alternative splicing event as any time a read either started or ended on multiple exons. Here 1,404 events were found for whole-brain, the difference between the two numbers could have been caused by realignment strategy or the use of a different database to guide the creation of the splice junctions. Also Mortazavi et al. 2008 did not rely on the reconstruction of alternative splicing events using splice-crossing reads, but used the simpler method outlined above. All together, this seemed to indicate the proportions and numbers of alternative splicing events seen here were not unusual with regards to sample size and tissue of origin. Especially since alternative splicing is thought to occur frequently in the brain (13). Detection of alternative splicing is highly dependent on either expression and/or sequencing depth. This was shown much more elegantly in Wang et al. 2008,

but can be seen in Table 3, where differences between the average number of reads mapped to splice junctions may have lead to the failure to observe an event in one tissue. The inclusion of multi-mapping reads did not seem to change the relationship between the unique or common splice events. Also of interest was the massive outlier seen in Table 3, reaching a read depth of 520 for one splice junction in the striatum. This gene was the myelin basic protein or MBP and was massively expressed, for example one exon of size 234 bases had 5,295 striatum reads mapped to it. It was also alternatively spliced, containing a cassette exon event (Fig. 8). Although there were relatively few poly-A sites detected in alternative (trailing) APA exons, this could have been due to further degeneracy of the sites, which may have even been the cause of the events. Also, it was very possible that the 13 defined signals did not represent the entire population of such signals, as evidenced by the incomplete presence of these signals with respect to the defined transcripts ends for both whole-brain and striatum. A similar situation existed for the CAGE tags, with the presence of such tags being evident in a vast minority of events. It was expected here because of the relatively low coverage of such data.

There were limitations to detecting alterative splicing with this approach. False positives likely existed based upon the complexity and diversity of the transcriptome. To further limit this possibility more reads could be required to span splice junctions at the cost of sensitivity. The practicality of this requirement was also constrained by sequencing depth, sample size in this context, although this may be overcome in the future. There were also cases where genes overlapped or where very different transcripts were produced from the same locus. If alternative splicing-like events occurred in these cases, they were kept simply because not doing so ignores the complex reality of biology. However, further modifications and improvements to the alternative splicing pipeline can and will be done in the future. These will

include *de novo* alternative splicing detection based upon the strategies implemented by Wang et al. 2008 and Pan et al. 2008 and support for spliced-alignment output from programs such as Tophat (36).

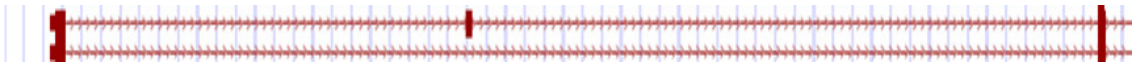


Figure 8. Cassette exon event of the MBP gene. The MBP gene was determined to be very highly expressed in both striatum and whole-brain and was also alternatively spliced. This is a cassette exon event for this gene that was annotated by Ensembl (12) and visualized using the UCSC Genome Browser (73).

AATAAA	CATAAA
ATTAAA	GATAAA
TATAAA	AATGAA
AGTAAA	TTTAAA
AAGAAA	ACTAAA
AATATA	AATAGA'
AATACA	

Table 4. Defined poly-A signals. These are the poly-A signals from ASTD (11) that were used as the query against all the exon sequences in the completely defined transcripts. Completely defined in this context refers to those containing a minimum of one read mapped to each splice junction.

Chapter 4

Quantifying Transcript Isoform Differences

Introduction

Knowing the number and type of alternatively spliced transcripts is useful; however reliable quantification of these transcripts is the ultimate goal. This quantification in conjunction with overall gene expression information will allow us to further analyze the contributing mechanisms behind disease states, inherited traits or tissue differentiation. Statistical methods (80-81) that are in development in conjunction with splice junction mapping data will provide estimates of the type and quantity of transcripts that are produced at a given time. The makeup of transcript isoform populations has been shown to be more different between tissues than between individuals or cell lines (5). In some instances proper cellular function is dependent on this isoform ratio such as in the case of MAPT where proper functioning of neurons depends on this ratio (92). RNA-Seq provides a very effective platform to assess transcript isoform abundance and methods have been put forward to measure it as was done by Pan et al. 2008, Wang et al. 2008 and more recently Jiang and Wong 2009 and Zheng and Chen 2009. Commercially available microarray platforms of increased resolution have also been developed. Exon arrays, microarrays that can probe individual exons, are now commonly used and can provide information on alternative splicing and isoform abundance as well (4, 51, 53-55). Pan et al. 2008 showed that an isoform ratio measure in an RNA-Seq experiment was highly correlated to that from their custom microarray. Although Jiang and Wong 2009 achieved correlation coefficients that were considerably lower when comparing their RNA-Seq isoform abundance measure against previous microarray experiments from the Pan et al. group

(49). The goal of this chapter was to examine the measures of isoform quantification put forward originally by both Wang et al. 2008 and Pan et al. 2008 and apply them to the mouse whole-brain and striatum RNA-Seq datasets. First addressed was the question of whether gene expression significant in one direction or another relative to tissue was associated with overall directional changes in isoform levels as measured by the inclusion ratio of Wang et al. 2008. Following this, a measure related to the %in derived from Pan et al. 2008 was then used to determine if this RNA-Seq measurement could agree with alternative splicing predictions from a commercially available exon array platform. For this commercial exon array, statistical methods have been implemented in ExonModelStrain (55) to detect alternative splicing by looking for significant changes in exon expression relative to the level of gene or transcript expression between two samples (denoted as strains in the package). This was termed alternative exon usage or AEU (55). It should, in theory, be possible to merge the RNA-Seq alternative splicing events involving whole exons to the exon determined by ExonModelStrain (55) to have the greatest change in exon expression.

Methods

The isoform ratios of the alternatively spliced junctions described in Chapter 3 for the mouse whole-brain and striatum were calculated as described in Wang et al. 2008. The inclusion ratio was calculated by taking of the ratio of the number of inclusion reads to the total number of inclusion, exclusion and common reads (5). Inclusion reads were defined as those that would be included in one isoform relative to another and exclusion reads were those that were mapped to exons bordering the included region of the other isoform (5). Common referred to those reads that were common to both (5). The inclusion, exclusion or common reads were counted using either unique reads or unique plus multi-mapping reads. Non-splice

junction reads were required to map completely within the common, inclusion or exclusion region (5). Significance of the isoform ratio was determined by a Fisher's Exact Test for a 2X2 contingency table testing for differences between the number of inclusion and exclusion reads for the whole-brain and striatum as described in Wang et al. 2008. The p-values resulting from these tests were converted to q-values (93) using the qvalue R package (94) to correct for the false discovery rate.

An extension of the %in defined by Pan et al. 2008, the percent trans-read contribution (PTC), was created to allow comparison of a greater number of events with microarray platforms. Trans-reads in this context referred to reads that cross splice junctions. The %in measure averaged the number of inclusion junction reads over the number of exclusion junction reads for single cassette exon events (77). Inclusion and exclusion reads are defined in Fig. 9. This was generalized to allow all forms of cassette, mutually exclusive, intron retention events, ATSS and APA events by taking the averages of all inclusion and exclusion reads for cassette, mutually exclusive exons, APA and ATSS events. Further, it averaged over the number of splice junction sized regions for intron retention events for the inclusion read measure. The exact definition of inclusion and exclusion in relation to splice junctions is shown in Fig. 10 for single cassette and mutually exclusive exon events. Differences between these ratios for the samples were measured by looking at the result of the striatum PTC divided by whole-brain PTC.

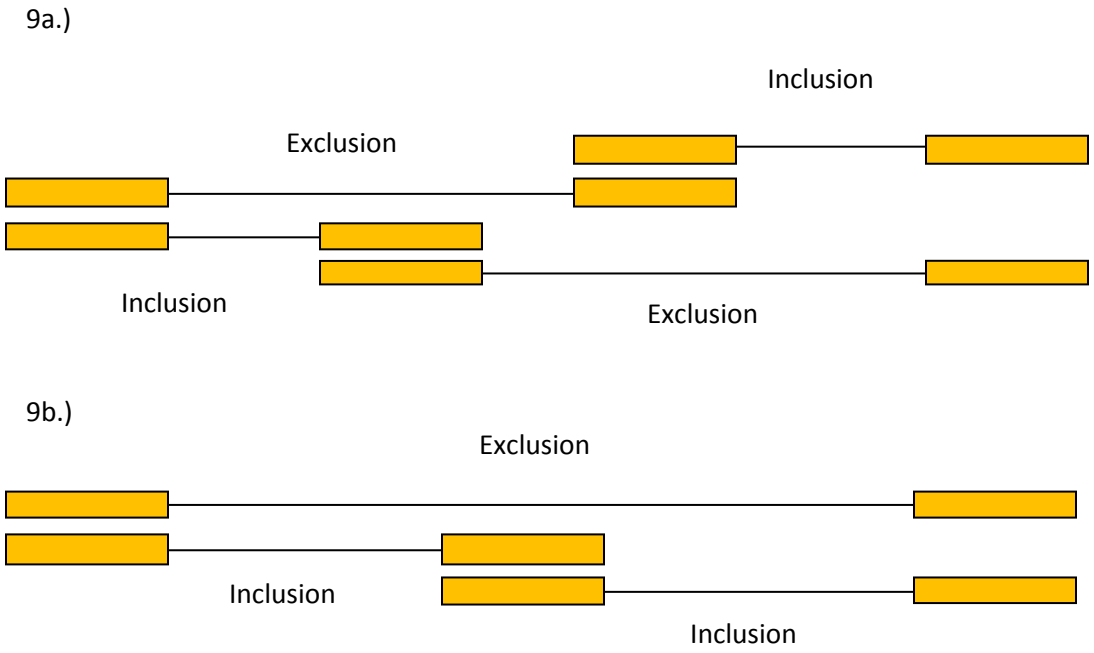


Figure 9. Definition of inclusion and exclusion splice junctions for the PTC calculations. Shown are the splice junctions defined as inclusion or exclusion for both a mutually exclusive event (a) and a cassette exon event (b). Reads that mapped to these splice junctions were included into the listed categories. The orange boxes are exons and the lines represent introns. The designation inclusion or exclusion refers to the reads that map to the splice junctions formed from the concatenation of the exons (See Chapter 2 for more details).

The microarray platform used was the Mouse Exon Array 1.0 ST from Affymetrix. Two experiments were previously run by the PARC, one using whole-brain and the other striatum. In order to compare the expression between the two, they had to be renormalized. Originally there were 12 whole-brain arrays, though a QC report indicated that two had scanner errors and were therefore not used in this analysis. Unfortunately there were also only 7 striatum arrays, so two experimental setups were used, one with a balanced number of arrays--7X7 with 7 of the whole-brain arrays chosen at random. This was the main experimental setup and was used for analysis unless labeled otherwise. For comparison the other setup was unbalanced, 10X7 which

compared all available arrays. This was done to determine if the strategy of choosing only 7 out of 10 arrays for whole-brain would produce overly different results. An estimate computed using 7 arrays could be less be precise with respect to the 10 array setup and using them both in a statistical model could therefore have a negative impact on the detection of differential expression. These experiments were RMA background corrected and normalized (89) and summarized at the probeset level using median polish (95). Two different programs were used to analyze the data, ExonMap (96) and ExonModelStrain (55).

A list of differentially expressed genes was determined using ExonMap (55) requiring a fold change > 1 and a t-test p-value <= .0001 at the probe-level which was subsequently extrapolated to the exon and gene levels (96). This package was used because the ExonMap program (96) could interrogate many more genes than the ExonModelStrain package (55) at that time, since it was not limited to the core probes. The list of differentially expressed genes from ExonMap (55) was used to stratify the inclusion ratio (5) by gene expression. Both the gene expression and inclusion ratio values were separated into three categories: striatum greater than whole-brain, striatum less than whole brain or neither. This was done for both significant isoform ratios at a q-value (93) of .05 or a p-value at .05 and for both sets including unique and those including unique and multi-mapped reads. Significance of the results was determined using a Fisher's Exact Test with a p-value simulated using monte carlo methods in R (88). The values reported were only for the balanced setup, though the unbalanced setup was used for comparison purposes.

A list of all exons with a max alternative exon usage (AEU) delta for each gene was generated using ExonModelStrain (55) for only those exons annotated as core. This list was then merged with a complete list of the affected exons from all CE, ME, APA and ATSS events

(See Chapter 2 for definitions). Affected exons in this sense referred to the exon(s) that were included in one isoform transcript relative to another. This was done for both the balanced and unbalanced microarray experiment setups. The PTC for both whole-brain and striatum as well as the change between the two for single cassette and mutually exclusive exon events were then calculated. This measure was compared to the delta AEU measure computed by ExonModelStrain (55). They were assessed first by looking at direction of expression change, then for significance—requiring a q-value <= .05 for the AEU delta and a change of 50% for the PTC delta (a measure put forward by Pan et al. 2008).

Results

Of the 24,104 genes annotated in ExonMap (96), 4,214 were reported to have greater expression in whole-brain relative to striatum, 641 had expression greater in striatum relative to whole-brain and 19,249 were reported to not be significantly different. Similarly for the unbalanced setup, 4,496 were greater in whole-brain, 772 were greater in striatum and 18,836 were not significantly different. Shown in Table 5 was the agreement between the balanced and unbalanced array setups with respect to gene expression differences using ExonMap (96). They generally agreed with each other. However, there were a large number of genes considered to have greater expression in whole-brain relative to striatum in the unbalanced setup, but were also considered to be non-significant in the balanced setup. Table 6 shows contingency tables describing the significant changes in isoform ratio as compared to changes in exon array gene expression for both isoform ratios significant at q or p-values of .05. There did not appear to be significant differences between the numbers of significant isoform ratios observed for each category between gene expression categories for any of the four tests

Unbalanced

Balanced	STR < WB	STR > WB	Neither
STR < WB	3,906	17	291
STR > WB	10	539	92
Neither	580	216	18,453

Table 5. Agreement between the balanced and unbalanced exon array analyses using ExonMap. The number of genes where one tissue was considered significantly expressed over another or otherwise is shown. This was done using the ExonMap (96) package in R (88). STR refers to striatum; WB refers to whole-brain.

6a.) Unique Read Count

Isoform Ratio

Balanced Array Expression	Isoform Ratio q-value <= .05				Isoform Ratio p-value <= .05		
	STR < WB	STR > WB	Neither		STR < WB	STR >WB	Neither
STR < WB	7	4	258		20	8	241
STR > WB	1	1	25		1	2	24
Neither	5	5	393		14	20	369
	p-value = 0.2464 Fisher's Exact Test				p-value = 0.08618 Fisher's Exact Test		

6b.) Unique + Multi Read Count

Isoform Ratio

Balanced Array Expression	Isoform Ratio q-value <= .05				Isoform Ratio p-value <= .05		
	STR < WB	STR > WB	Neither		STR < WB	STR >WB	Neither
STR < WB	8	5	256		19	9	241
STR > WB	1	1	25		1	2	24
Neither	5	7	391		14	21	368
	p-value = 0.2458 Fisher's Exact Test				p-value = 0.1446 Fisher's Exact Test		

Table 6. Relationship between isoform ratios and balanced array expression. Contingency tables comparing the isoform ratio (5) and gene expression as determined by ExonMap (96) are shown. Tables are represented for both p and q-values of less than or equal to .05 and were calculated using both unique reads and unique + multi-mapping reads. The test statistic was computed using a Fisher's exact test in R with a p-value computed by simulation (88). The q-values were computed using the qvalue package (94).
(p-values: .246, .086, .246 and .145 for q and p-values <= .05 for unique and q and p-values <= .05 for unique + multi reads; Fisher's Exact test with simulated p-value). The changes in expression category counts that would occur if the unbalanced array setup was used are shown in Table 7. The effect of using that setup would be to shift 34 genes from being non-differentially expressed to having differences in one direction or another and subsequently shift 17 other genes back into being non-differentially expressed.

		Unbalanced		
Balanced		STR < WB	STR > WB	Neither
	STR < WB	255	0	14
	STR > WB	0	24	3
	Neither	29	6	368

Table 7. Change in expression category for balanced vs. unbalanced array setup. This shows the shifts between gene expression categories that occurred when using a balanced or unbalanced microarray design with respect to the number of arrays. The exon arrays were analyzed using the ExonMap (96) package. The balanced setup involved seven arrays for both tissues and the unbalanced contained 10 arrays for whole-brain and 7 for striatum. These are the genes that were involved in the calculation for Table 6. STR represents striatum, WB represents whole-brain.

There were a total of 1,217 events consisting of the CE, ME, APA and ATSS events described in Chapter 2. Of these, 62 had at least one exon that successfully merged with the max AEU delta exon from ExonModelStrain (55). For the balanced analysis, 26 of these contained a q-value <= .05 as did 29 for the unbalanced analysis. These results are summarized in Table 8. The total shifts that would occur for each of the events common and unique to whole-brain and striatum when using the unbalanced experiment instead of the balanced experiment are shown in Table 9. These values largely mirrored what was seen in the ExonMap (96) analysis. But even considering the small number that successfully merged, few would have shifted expression categories. From Table 8, one event from each of the common and unique categories for both tissues that shifted from non-significance to significance is shown in Table 10. For two of the events, gene expression was determined to either be significant between whole-brain and striatum or became significant for the unbalanced experiment. One, ENSMUSE00000662850, which had an event unique to whole-brain actually became non-significant with respect to gene expression. However, it became significant with respect to the change in AEU. A possible reason for the lack of agreement between the ExonModel strain significant exons and those found using RNA-Seq may have been the effect size or read coverage of the events in question. Fig. 11 shows histograms of the log base 2 transformed PTC values for those alternative splicing events containing an inclusion exon that successfully merged with a significant AEU exon and those alternative splicing events that did not. The average log2(PTC) was actually estimated to be lower for those that merged (2.757) compared to those that failed to merge (4.128). Effect size may have been skewed by the presence of splice junctions containing very few reads. Also shown in Fig. 11 are boxplots representing the minimum read depth for splice junctions for those that merged and those that did not.

Balanced

	Common	Unique to Striatum	Unique to Whole-Brain	Total
Total	432	424	361	1217
# merged	24	22	16	62
q value <= .1	14	8	6	28
q value <= .05	13	8	5	26
p value <= .1	14	10	8	32
P value <= .05	14	10	6	30

Unbalanced

	Common	Unique to Striatum	Unique to Whole-Brain	Total
Total	432	424	361	1217
# merged	24	22	16	62
q value <= .1	16	9	6	31
q value <= .05	14	9	6	29
p value <= .1	16	11	9	36
P value <= .05	16	10	7	33

Table 8. Results from merging all CE, ME, APA and ATSS events with the AEU results from ExonModelStrain. Shown are the number of CE, ME, ATSS or APA events from the striatum and whole-brain RNA-Seq experiments that successfully merged with the AEU results from ExonModelStrain (55), also using whole-brain and striatum. Events were merged based upon the inclusion exon for the RNA-Seq experiments and the exon with the maximum delta for the AEU. The striatum and whole-brain RNA-Seq experiments were categorized based upon whether they were detected in only one (Unique) or both tissues (Common). In addition to the number merged, those successfully merged that also met the q and p value cutoffs of .1 and .05 for the AEU were also counted. This was shown for both the balanced and unbalanced microarray setups.

9a.)

Total

		Unbalanced								
		Common			Unique to Striatum			Unique to Whole-Brain		
Balanced		STR<WB	STR>WB	Neither	STR<WB	STR>WB	Neither	STR<WB	STR>WB	Neither
	STR<WB	18	0	0	9	0	1	11	0	0
	STR>WB	0	1	0	0	5	0	0	2	0
	Neither	0	2	3	2	0	5	0	0	3

9b.)

q <= .05

		Unbalanced								
		Common			Unique to Striatum			Unique to Whole-Brain		
Balanced		STR<WB	STR>WB	Neither	STR<WB	STR>WB	Neither	STR<WB	STR>WB	Neither
	STR<WB	10	0	0	6	0	0	4	0	0
	STR>WB	0	1	0	0	2	0	0	1	0
	Neither	0	1	1	0	0	0	0	0	0

Table 9. The shift in expression differences from balanced to unbalanced for unique and common events. Shown are the number of genes contained within each RNA-Seq alternative splicing category that would have shifted from one category of gene expression to another if a different experimental design was used. This was repeated for both (a) the total number of genes, and (b) those with a significant q-value at the .05 level for AEU. STR represent striatum, WB represents whole-brain.

ExonModelStrain

			Balanced		Unbalanced	
ASE		Exon	qStrain	qExonStrain	qStrain	qExonStrain
	Common	ENSMUSE00000352780	<.0001	0.075	<.0001	0.032
	STR	ENSMUSE00000435932	.090	0.145	.026	0.032
	WB	ENSMUSE00000662850	.074	0.059	.140	0.003

Table 10. Three ASEs that shifted from non-significance to significance based upon experimental setup. Shown are three alternative splicing event (ASE) exons that shifted from non-significant to significant based upon use of a balanced microarray setup as compared to an unbalanced. Each was from a different category of alternative splicing event: Common referred

to those events that were seen in both tissues, STR was from those events only seen in striatum and WB represented those only seen in whole-brain. The qStrain and qExonStrain variables represented the q-value for the test of differential expression and AEU respectively from the ExonModelStrain package (55).

A subset of size 357 from the 1,217 events mentioned above was extracted consisting of the single CE and ME events that were observed in both tissues. Of these, 12 were successfully merged with the maximum AEU delta exons. This was initially done using both unique reads and unique plus multi-mapped reads. However, the few that successfully merged had no multi-mapped reads so only the PTC results for the unique reads are reported. Only considering similar direction with respect to tissue type expression, 4 were concordant between the two measures while 8 were discordant. This data is summarized in Table 11. When filtered based upon significance measured by a 50% change in the delta PTC and a q-value of .05 for AEU, only four exons remained as seen in Table 11b, however they were all discordant. One of the concordant exons from Table 11a, ENSMUSE00000288945 was the affected exon in a CE event that was present in both whole-brain and striatum. It is highlighted in Fig. 10, which plots the expression levels obtained for each exon using ExonModelStrain (55). The average gene expression was estimated to be 11.16 for whole-brain and 10.54 for striatum. The PTC of this event was .786 striatum/whole-brain. This change was non-significant using the 50% cutoff, but the AEU delta was significant with a q-value of .05.

11a.)

AEU

PTC		Striatum > Whole Brain	Whole Brain > Striatum
	Striatum > Whole Brain	1	7
	Whole Brain > Striatum	1	3

11b.)

AEU

PTC		Striatum > Whole Brain	Whole Brain > Striatum
	Striatum > Whole Brain	0	4
	Whole Brain > Striatum	0	0

Table 11. Concordance of merged AEU and PTC events. (a) The change in PTC was computed using uniquely mapping reads and directionality was compared to the overall direction of gene expression for whole-brain and striatum computed using the ExonModelStrain package (55). For (b) the same comparisons are made in (a) but the results were additionally filtered requiring significant exon/strain interactions (representing a significant AEU event) for q-values at the .05 level and using a 50% delta cutoff for the PTC.

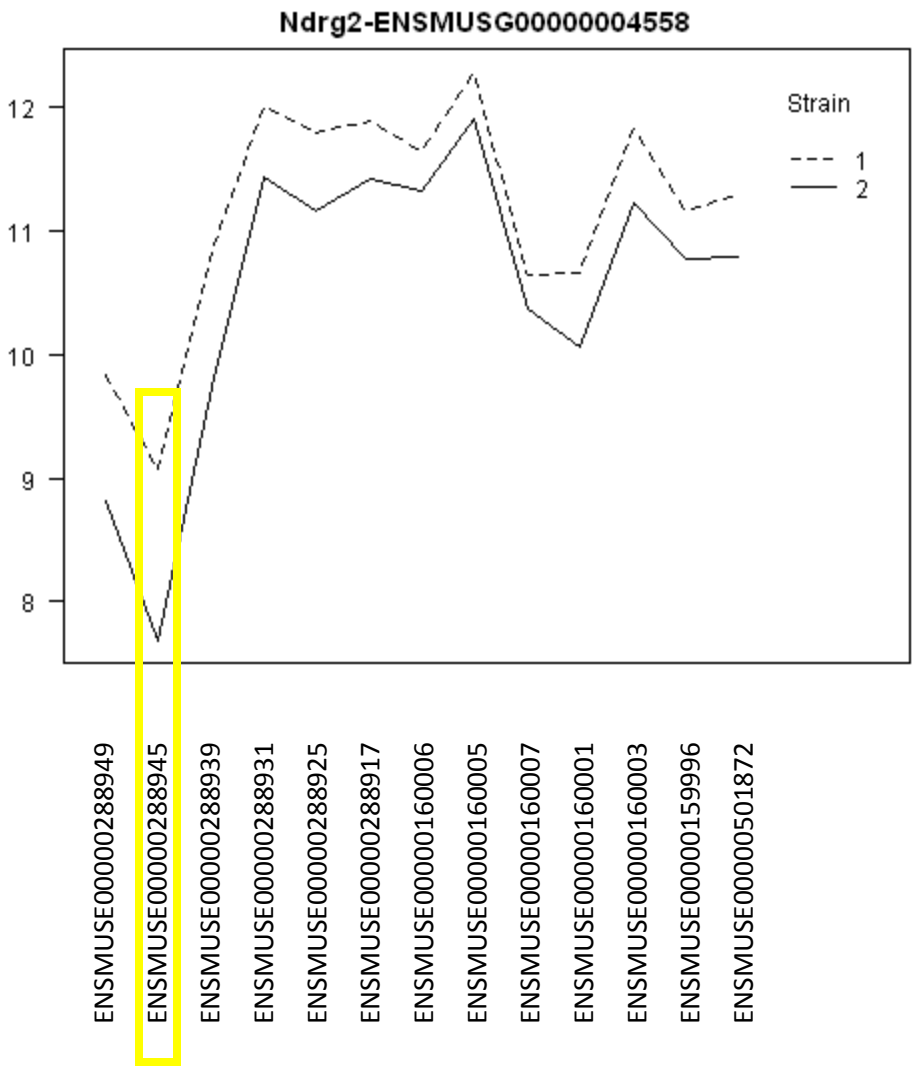


Figure 10. Concordant AEU and PTC exon for CE event. Shown is an interaction plot highlighting a concordant exon based upon the delta PTC and AEU. The Y-axis shows the RMA (89) values for each Ensembl (12) exon computed using the ExonModelStrain package (55). The Ensembl (12) exon identifiers are shown on the X-axis. Whole-brain expression is represented by the dotted line and striatum expression is represented by the solid line.

Discussion

With respect to the different microarray experimental setups, it seemed that the biases introduced from use of an unbalanced experiment were not that great. From the ExonMap (96) analysis it was determined that 580 more genes would be declared significant with whole-brain greater than striatum in unbalanced when compared to balanced. However, 291 genes were also shifted from this category into the non-significant one. This would overall have a minor impact on the comparisons between the isoform ratio and the gene expression results. Any impact would be lessened given the overabundance of counts in the "neither" category. Again, there was an impact on the full merge between the CE, ME, ATSS and APA events most notably on the three exons listed in Table 10, which became significant in the unbalanced setup.

There was no evidence from this rather crude measure of there being co-regulation of isoform and gene levels. It is possible that a network analysis approach could be successfully used to look for biological significance and may be implemented in the future. It was noted that both measures seemed to produce consistent results. Because of the similarity of the two tissues, very few differences between them could be reasonably expected in gene or transcript isoform expression. Overall, the alternative splicing events found using these RNA-Seq experiments did not agree well with the results from ExonModelStrain (55). Only about 5% of the CE, ME, ATSS and APA events successfully merged with 1% of these being usable to compare the change in PTC and AEU measure calculated by ExonModelStrain (55). This lack of agreement was not because of lack of overlap between the genes declared to be alternatively spliced in the RNA-Seq dataset versus those that had significant AEU in the exon arrays. Out of the 937 distinct genes present in the alternatively spliced gene set, 820 overlapped with the 13,852 genes in the balanced output from ExonModelStrain (55). To ensure that the presence of

missing exons within the exon array dataset did not affect the results, these missing exons were extracted using ExonModelStrain (55). Of the 13,852 genes, there were 1,829 missing exons, however only 19 of the 1,616 exons used in this analysis were in the missing exon list. So this too was not the reason. It was also interesting that in the 1% where the PTC could be calculated more were discordant than concordant. Effect size as measured by the PTC did not seem to be higher in those events that merged successfully with the AEU exons. There was a small amount of evidence for the idea that the alternative splicing events that did merge tended to have slightly greater expression (Fig. 11). Also, the large difference in sample size between these two comparisons, 62 events that merged successfully versus 1,155 that did not, made statistical testing of this result unhelpful. This may unfortunately mean that these two technologies are unlikely to be successfully compared in this manner. One of the major limitations of this study was the use of only the core probes in ExonModelStrain (55). These probes only cover the best defined exons so it is possible that some of these events may have been missed because of this incomplete coverage (55, 97-98).

Many of the disagreements between microarrays and RNA-Seq surrounding detection of alternative splicing might have been due to the relative lack of difference in expression or splicing between whole-brain and striatum. The greater the tissue differences, the easier it may be to detect changes in isoform and AEU levels between the datasets. This in turn should increase the agreement between the RNA-Seq and exon array technologies. It may suggest that even using cutting-edge high throughput technologies such as RNA-Seq and exon arrays, we still cannot accurately quantify relatively small changes in transcript isoform populations that could play a role in biological mechanisms. Although much work has been done by Wang et al. 2008 in the context of RNA-Seq, ongoing improvement in technology make it important to continuously

reevaluate techniques and previous results. This is especially true for detecting alternative splicing using RNA-Seq.

Supplementary Information

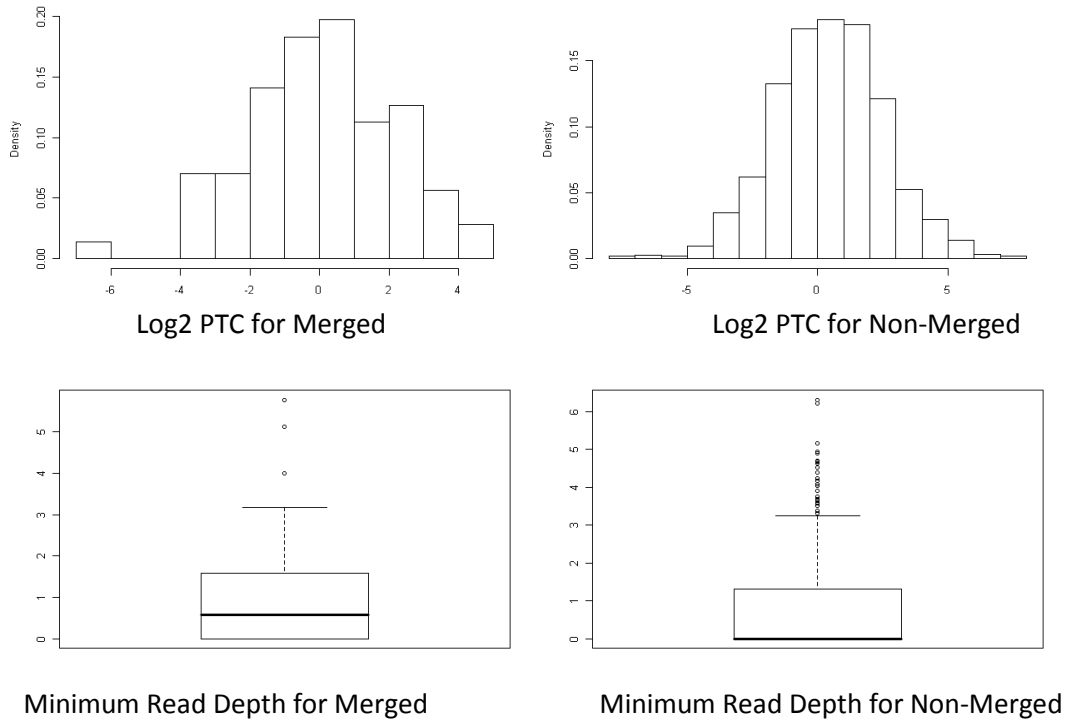


Figure 11. Properties of the merged RNA-Seq/exon array events. The distribution of log base 2 transformed PTC values are shown for those CE, ME, ATSS, APA events containing at least one read mapping uniquely to each splice junction (top). On the left are those events that had an inclusion exon successfully merged with an AEU exon. On the right are those events that did not contain a successfully merged exon. The PTC for each was calculated by dividing the average number of inclusion junction reads over the average number of exclusion junction reads. The boxplots shown on the bottom display the minimum number of reads assigned to each splice

junction for those alternative splicing events that had an inclusion exon that successfully merged with the AEU exons from ExonModelStrain (55) (left) and those that did not (right).

Chapter 5

Detecting Differential Expression of Exon Segments

Introduction

Traditional strategies to detect splicing and alternative splicing such as aligning reads to preformed splice junctions or the direct spliced alignments of sequences to a genome have been shown to be effective (17, 28). Furthermore, specialized tools that are readily available for performing these analyses for long sequences are becoming available for RNA-Seq experiments (36). However, in the case of alternative splicing, it was of interest to determine whether more novel approaches could be taken to look for differences in transcript splicing that would utilize the large amount of data generated from RNA-Seq experiments. One possible strategy would draw on the differences in the incorporation of exons in alternatively spliced transcripts and use significant read depth changes to infer the presence of alternatively spliced transcripts (Fig. 12). A similar idea was demonstrated using bayesian inference by Wang et al. 2008 for the detection of alternative poly-A sites and extended by Jiang and Wong 2009. Segments of (or whole) exons that were shared between multiple transcripts will be referred to as variably incorporated exon segments or simply exon segments. These exon segments can be thought of as categorical variables, each with a value equal to number of transcripts containing the segment. As a first step it was desired to test whether it was possible to detect differential expression between these categories of exon segments. The underlying future goal was to take the results from these explorations to reformulate the problem into a more complex machine learning context. These results would also serve as the initial baseline through which future, more complex, models would be compared against. For this analysis the results from two separate RNA-Seq

experiments from the mouse brain were used. One dataset was from the mouse whole-brain and one was from the mouse striatum. See Chapter 2 for more details regarding the alignment and splice junction formation.

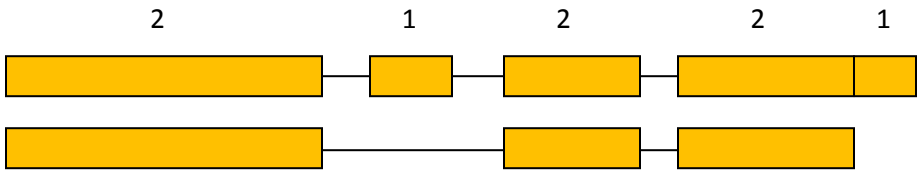


Figure 12. Levels of overlap for a theoretical gene. The theoretical gene shown above has two alternative splicing events: a cassette exon event and an alternative poly-A site. As can be seen, while the rest of the gene's exons are incorporated into two transcripts, the affected cassette exon and overhanging segment of the poly-A event are incorporated into only one transcript.

Methods

In order to determine which transcripts were produced by a given gene it was necessary to first provide evidence for the existence of the transcripts retrieved from the public databases in our datasets. Transcripts annotated in a database may not have been expressed in every tissue in the mouse body. These transcripts were only kept if they had a minimum of 5 unique reads mapped to each splice junction and every exon within the transcript had at least one read mapped to it. The cutoff of five was chosen because it translated to a mapping rate of ~ 80 reads per kilobase (RPK) for striatum and ~120 for whole-brain which, as will be shown later, was near the rate necessary for optimal detection using a statistical model. Mapping was defined as the presence of a read start position within a segment boundary—a measure consistent with a rate. Splice junctions were mapped to transcripts by matching the internal

boundaries of the formed splice junctions to the corresponding boundaries from the transcript. For the situations where several transcripts existed that had the same internal boundary definitions but different start or stop positions (i.e. alternative start or end site), they were examined further. In these cases the longer overhanging transcript ends were only kept if they had reads mapping to them. If these overhanging ends were smaller than the read length and had no reads mapping within the boundaries they were checked to see if any reads ended within the boundaries. If this was true, the ends were shortened to the length of the nearest segment containing mapped reads; otherwise the entire transcript was discarded. The remaining transcripts were divided into exon segments based upon the number of verified transcripts that included the particular exon segments for a given set of gene boundaries. Again, reads were assigned to these segments requiring that a read start within a segment. If this was not possible, for example the segment length was smaller than the read length, these segments were removed. Once the genes that only contained one transcript were removed, the set that remained consisted of those genes containing alternative splicing events.

Statistical models have been applied in the context of alternative splicing detection using microarrays such as the linear model implemented in ExonModelStrain (55). This model, for instance, can only work at the exon level because it is constrained by the placement of the probes it interrogates (55). However, RNA-Seq is not constrained by probe placement. In this manner RNA-Seq may be able to detect more complex events than those that affect an entire exon in one transcript relative to another (e.g. CE, ME, APA or ATSS events). If an exon segment is defined as any portion of or whole exon, then it would be natural to determine whether a statistical model could detect any differences between these portions and the rest of the gene. The theory behind this was that an alternatively spliced transcript should contain exon segments

that are incorporated into fewer transcripts than others. Since a segment that was incorporated into fewer transcripts should contribute fewer reads, detecting differential expression (i.e. alternative splicing) of these portions may be possible. One way to detect differential expression of exon segments would be to use a generalized linear model framework (99). Generalized linear models allow the linear regression framework to be extended to situations where the errors are distributed in a non-Gaussian manner (99). For count data like RNA-Seq expression the Poisson or negative binomial distributions have been successfully used (100). For this formulation, the response variable would consist of the read counts with the number of transcripts including the specific exon segment represented as categorical independent variables—similar to an ANOVA. For instance if it was determined that a segment was included as part of two transcripts then the category would be labeled 2 and the number of reads mapped to it would be the dependent variable. If there was at least one segment in another category, say 1, the read depth from category 2 would be compared to the depth in category 1. In this manner we could compare the read count of different categories within a gene to determine whether a gene was alternatively spliced. Note that this model is similar to the one put forward by Jiang and Wong 2009, but it serves a different purpose. There are several ways to approach the problem of detecting differences between categories. Most easily the categories could be examined globally to look for overall trends in expression. However, this strategy would not be useful since global statements about genes that would be in many cases very heterogeneous in structure and expression level would not be informative.

A better strategy to detect differential exon segment expression would be to fit a model to each gene separately to estimate expression differences between categories. Since the number and types of categories should vary from gene to gene, it was important to ensure that

a model be chosen that was robust enough to be valid for the majority of the examined genes. For this purpose three generalized linear models were created in an effort to find the most parsimonious model that would meet the error distribution assumptions. The three models were set up in a similar manner, but used different families for the error distributions—the Gaussian, Poisson and negative binomial. The general formulation of these models is shown below:

$$g(\mu) = \beta_0 + \sum_{i}^{n} \beta_i x_i + g(t) + E$$

Where *g* was the link function for which the identity function was used for the Gaussian model and a log function was used for the Poisson and negative binomial models. The variables *n* and *t* referred to the number of categories and the offset, which was the length of the exon segments in kilobases. *E* represented the error coefficient. The variable *n* was required to be greater than or equal to 2, otherwise the gene in question would not be alternatively spliced. The only difference between the formulations was that the counts were log transformed before being entered into the model for the Gaussian model in an effort to directly stabilize the variance. An important concern for these local calculations was the model assumptions, especially overdispersion in the case of the Poisson regression. The Poisson model estimated its variance directly from the mean, an assumption that was unlikely to be true. This can be corrected through the use of the Poisson distribution fit using a quasilikelihood function instead of the standard likelihood function (101). This allows the model to account for overdispersion through the use of the dispersion parameter σ^2 (101). While this procedure does not affect the estimates

for the β coefficients, it does result in more conservative measures of significance—wider confidence intervals and larger p-values (100).

A simple method to detect the appropriateness of the model assumptions was to extract the Pearson's residuals and plot them against the expected Gaussian probability (100). Since this transformation of the residuals should be approximately normally distributed, a linear relationship would be observed if the underlying model assumptions are met (100). Pearson's residuals are calculated as shown below (100):

$$r_i^P = \frac{(y_i - \hat{\mu}_i)}{\sqrt{V(\hat{\mu}_i)}}$$

Based upon this idea, a method was devised to extract the Pearson's residuals from each model run and compute the Pearson's correlation coefficient from a q-q plot calculation. An example of this is shown graphically in Fig. 13. The three models were run for all genes. Models were determined to fit best if their correlation coefficients were the greatest of the three.

To look at whether the models could successfully find differential expression between the sets of verified alternative splicing events, all useful comparisons were extracted. For example if there were three categories of exon segments, the comparisons between category 3 and 1, 2 and 1 and 3 and 2 for each model were extracted and the p-values recorded. Comparisons were made both across genes and within genes. To correct for the false discovery rate q-values (93) were calculated for each set of p-values from the model runs. Simulations were run using a theoretical gene to determine how a GLM (99) model would perform under varying conditions.

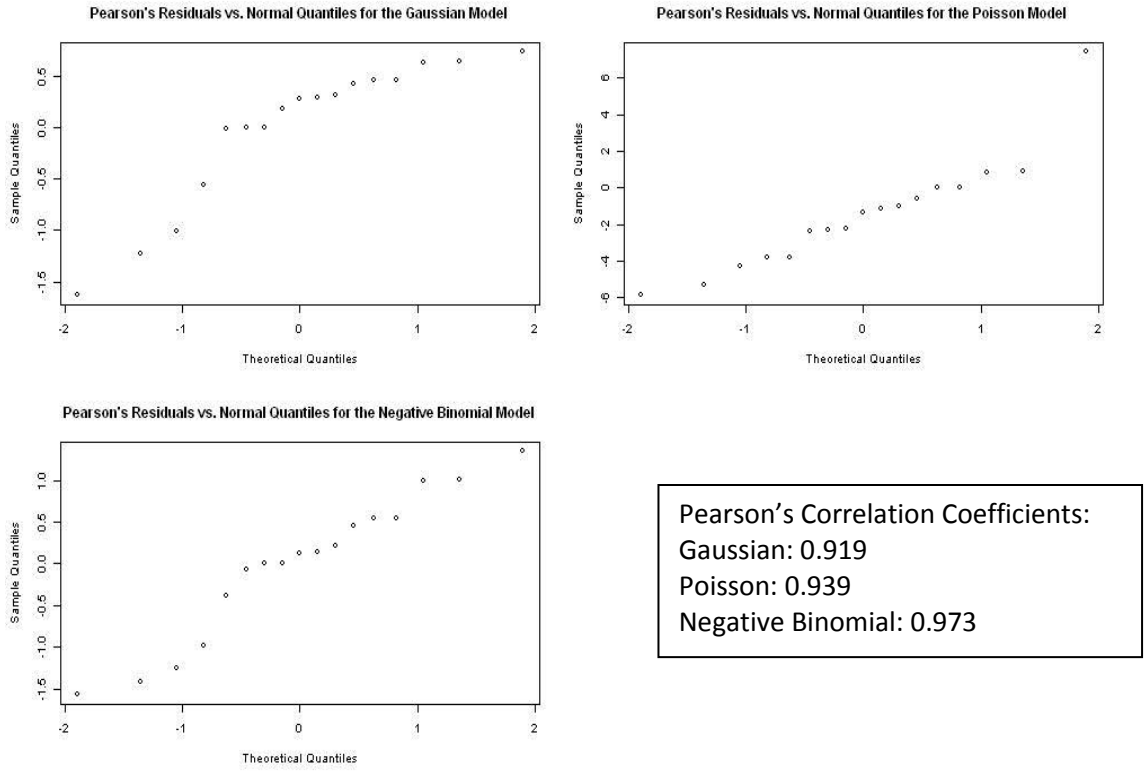


Figure 13. Q-Q plots for gene ENSMUSG00000010086 in striatum. An example of the q-q plot calculations used to test the models for validity of the underlying distributional assumptions. The quantiles from the Pearson's residuals were plotted against the normal quantiles. The straightest line fits the model assumptions the best. Linear relationships can be measured directly using Pearson's correlation coefficient. The plot from the Gaussian GLM (99) is shown in the upper left-hand corner, the plot for the Poisson is shown in the right-hand corner and the negative binomial plot is shown at the bottom left.

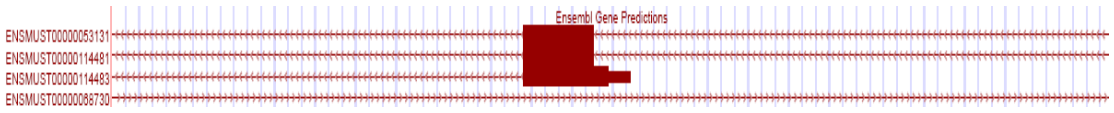
Results

Of the 1,184 striatum genes with all splice junctions present for at least one transcript, 6 genes still contained an exon segment larger than the read length where the count was zero. These were discarded so as to not affect the log transformations. These zero counts reflected rare situations that were not accounted for in the other filtering measures such as the gene

shown in Fig. 14a. Here the alternative poly-A event was an extension of an internal exon and not another terminal exon—a situation that was not handled by the script. Of the remaining genes, 1,039 had categories of only 1 after filtering, meaning they were either not alternatively spliced or the affected regions were smaller than the read length and were subsequently removed. Eleven had categories all equal to 2. Several of these were cases where two Ensembl (12) transcript exon boundaries were the same, but the coding sequences were different, as shown by the middle transcripts in Fig. 14b. The others were transcripts with small exon segments that contained no mapped reads where these segments were removed. The remaining 129 were valid with the exception of 20 negative binomial models that failed to converge after 25 iterations—leaving 109 genes to compare. Similarly the whole-brain sample had a total of 1,579 genes that met the criteria for inclusion, 9 contained a zero, 1,376 contained all ones and 15 contained segments that contained an equal number of underlying transcripts. Again, the negative binomial model fitting failed for 26 genes, leaving 157 remaining. To assess model fit, the maximum correlation values of the Pearson's residuals versus the normal quantiles across the different models for each of the 109 and 157 genes were counted. The negative binomial model seemed to fit best since it was the highest 48 out of the 109 striatum runs and 68 out of the whole-brain runs. For comparison the Gaussian and Poisson models were highest a total of 28 and 33 times respectively for striatum and 50 and 39 times for whole-brain. The overall success of the negative binomial distribution may have indicated overdispersion in a significant percentage of the genes surveyed. As can be seen in Fig. 15a and b, the best fitting negative binomial model runs seemed to have equivalent sample sizes and category numbers compared to the Poisson model. Also, the best fitting negative binomial and Poisson models both had larger sample sizes than the Gaussian. The success of the negative binomial model in the majority of these simple goodness-of-fit tests could be extrapolated to provide evidence for

it being the most appropriate distribution out of the three (100). However, since the MASS (102) implementation of the negative binomial generalized linear model failed to converge in numerous situations it would be better to use a more stable overdispersed Poisson regression model such as the quasipoisson model mentioned above.

14a.)



14b.)

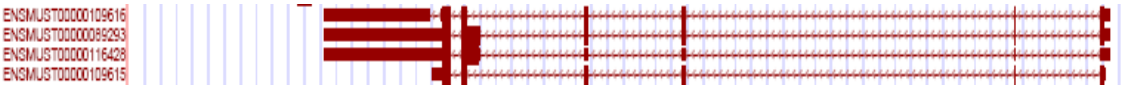
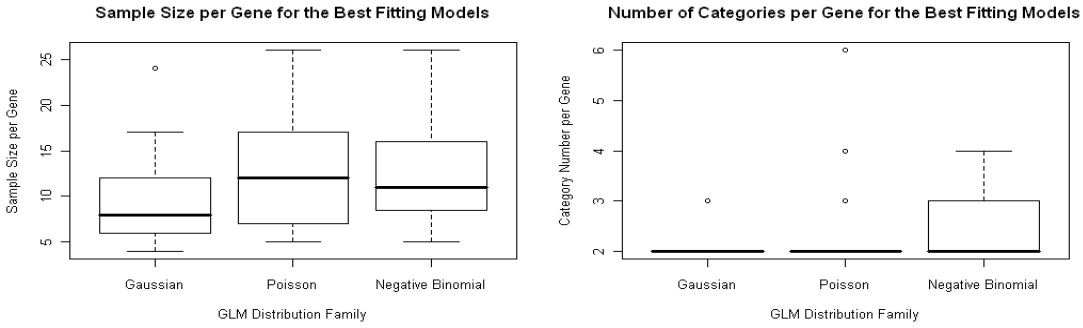


Figure 14. Representative transcripts structures of discarded genes. a.) shows an alternative poly-A site that overlaps with an internal exon. b.) the two middle transcripts have the exact same boundaries except they have a different coding sequence. This can be seen by comparing the thick bars, which represent the coding sequence to the thin bars which represent non-coding sequence.

15a.)



15b.)

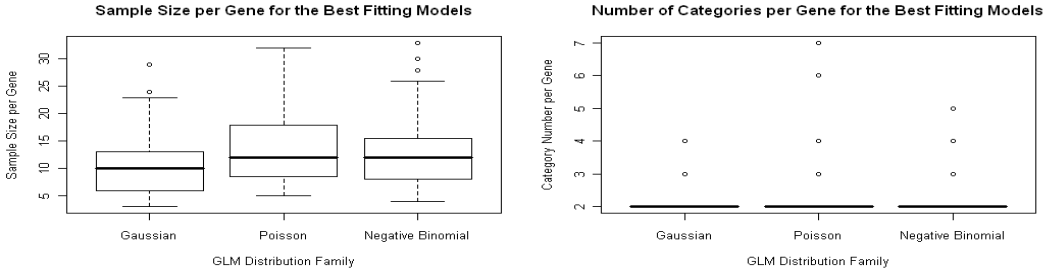
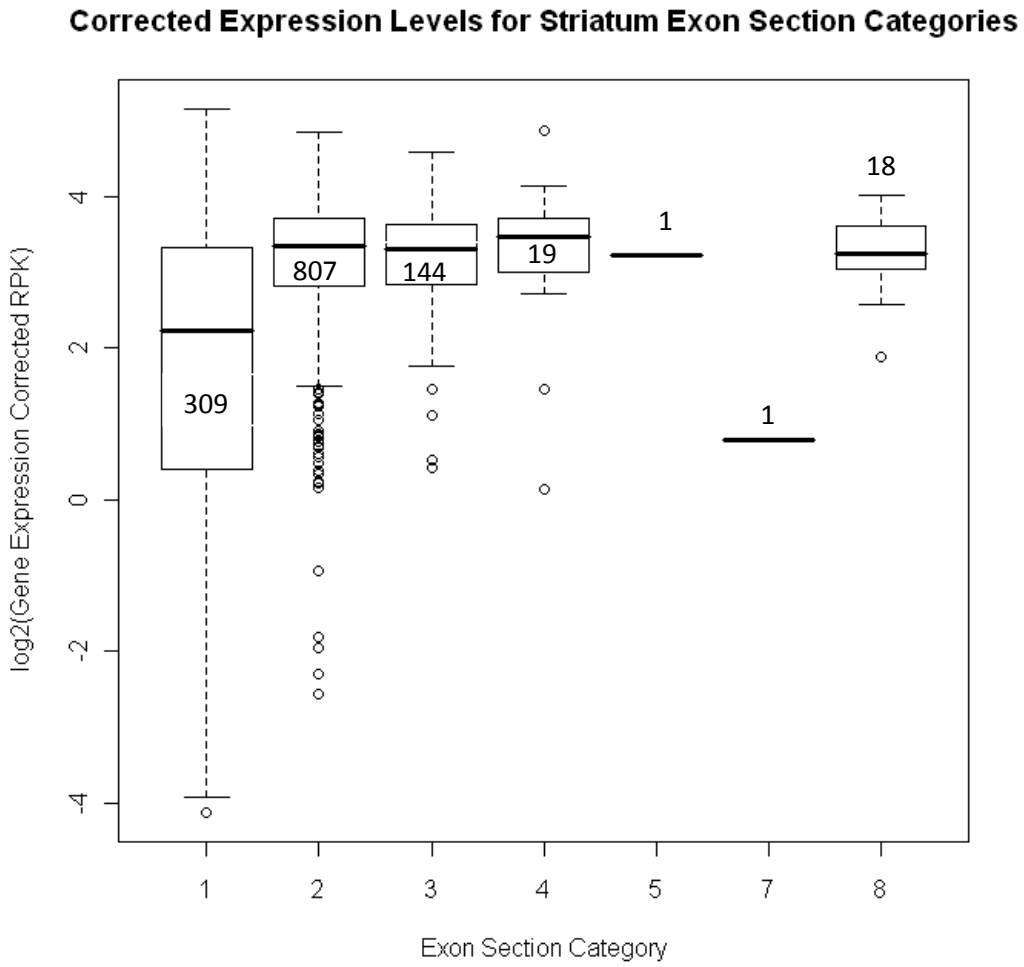


Figure 15. Sample size and category number for best fitting models. Boxplots plotting the distribution of sample size and category number for the best fitting models according to the Pearson's correlation coefficient of the Pearson's residuals vs. Gaussian quantiles tests are shown. Striatum is a.), whole-brain is b.). Categories in this context represent the number of transcripts each exon segment was incorporated into.

When the exon segment categories were compared globally without using a GLM (99) framework, an increasing trend in normalized expression was seen for those exon segment categories that had sufficient sample size (Fig. 16). This was true for both striatum and whole-brain. Since these categories came from different genes with different levels of expression, the

RPK was further normalized by the gene-level RPKM (17) computed in Chapter 2 and log base 2 transformed. The majority of exon segments fell into either category 1 or 2. The difference between the two categories also showed the greatest increase in median corrected expression level.

16a.)



16b.)

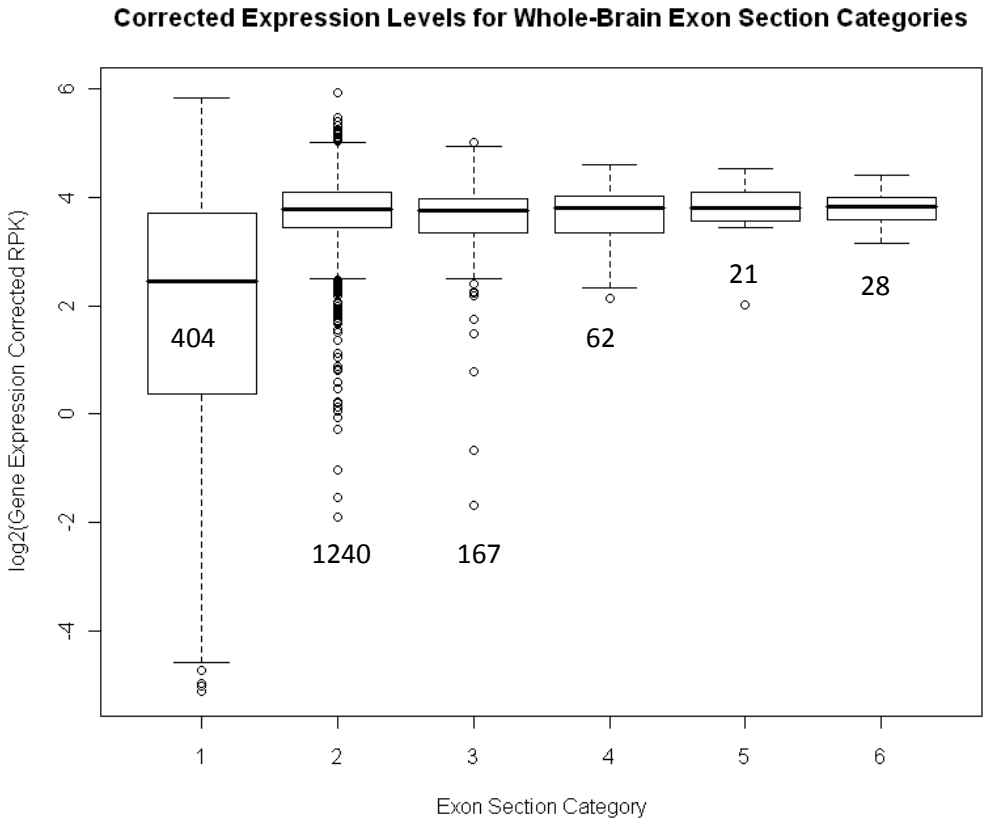


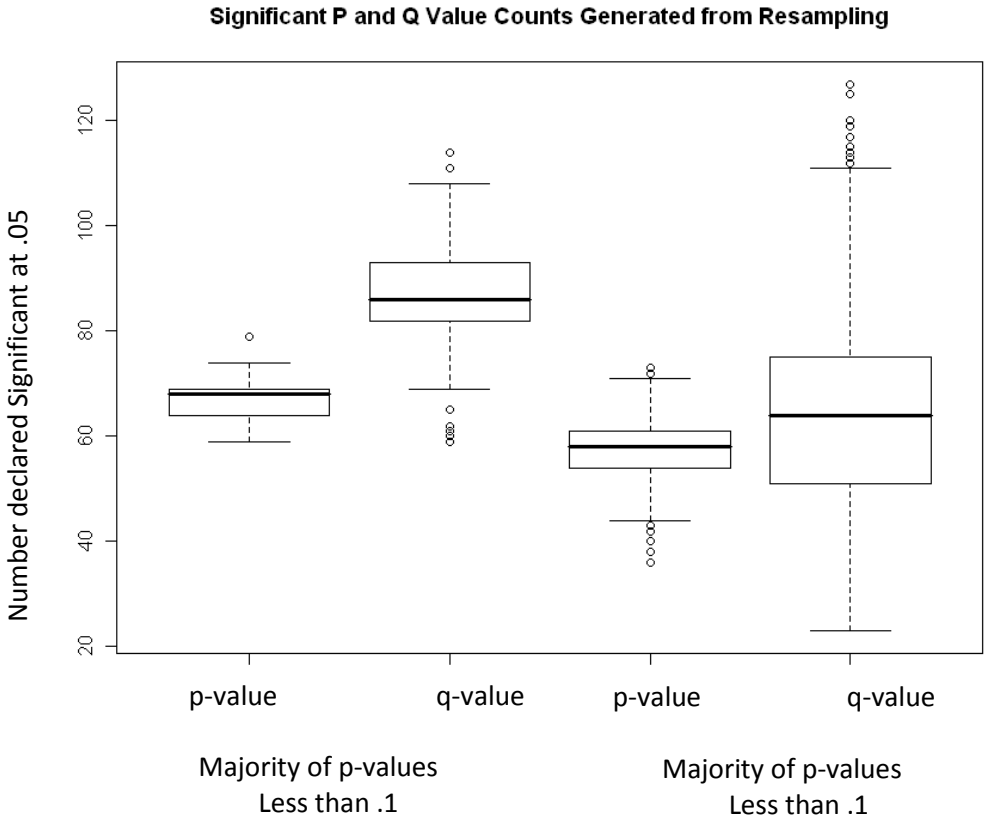
Figure 16. Overall distribution of expression levels for exon segment categories. Corrected read count for the different overlapping exon segment categories for striatum and whole-brain is shown. a.) shows the log base 2 corrected read count that was normalized by both exon size and gene expression for striatum for each observed exon segment category. b.) shows the same information but for whole-brain.

Significance of the 109 striatum and 157 whole-brain genes was judged by the pairwise comparison of the q-values calculated for the categories across genes. All together there were 167 category comparisons made within the 109 striatum genes and 253 in the 157 remaining whole-brain genes. The q-value is a popular method of correcting for the false discovery rate.

However, for this study it provided unusual results. As can be seen from Table 12, there were more significant q-values than p-values at the .05 level. The reason for this seemed to be the underestimation of the q-value due to the large number of low p-values at the .05 cutoff level. In Fig. 17 it could be seen from a resampling of the p-values from the striatum quasipoisson model that when the majority of the p-values were less than .1, there tended to be more significant q-values than p-values at the .05 level. This was not true for the situation where the majority of p-values were greater than .1. This relationship was not seen for the .01 level. These highly skewed p-value distributions seemed to occur for these data as illustrated in Fig. 18 where the vast majority of the p-values were very low. The q-value cutoffs were kept at .01 to minimize false discovery rate problems resulting from the inaccurate q-value results. However, the negative binomial model performed especially poorly in this area and even at this stringent cutoff still produced more significant q-values than p-values. Further results pertaining to the negative binomial GLM (99) are presented but should be viewed with caution. The Poisson model also performed poorly with respect to p and q values. As can be seen in Fig. 18a, the p-value histograms were highly skewed for the Poisson model. This was true to such an extent that the q-value could not be computed for whole-brain Poisson run. To remedy this problem, the Poisson model was replaced with the quasipoisson producing much more conservative estimates (Fig. 18a and b).

	.05 q and p-value level					.01 q and p-value level			
	Striatum		Whole-Brain			Striatum		Whole-Brain	
	q-value	p-value	q-value	p-value		q-value	p-value	q-value	p-value
Gaussian	84	71	140	113		35	37	72	72
Q. Poisson	63	58	98	84		20	32	28	38
Neg. Bin.	92	88	140	129		64	64	101	97

Table 12. Overall number of significant exon segment categories for each model. Shown above was the number of significant tests for each model using either a cutoff of .05 for the p or q values for the left side of the table and .01 for the right as the level of significance. Significance was determined using a Gaussian, quasipoisson and negative binomial distribution in a GLM (99) framework. The q-values (93) were determined using the qvalue package (94).



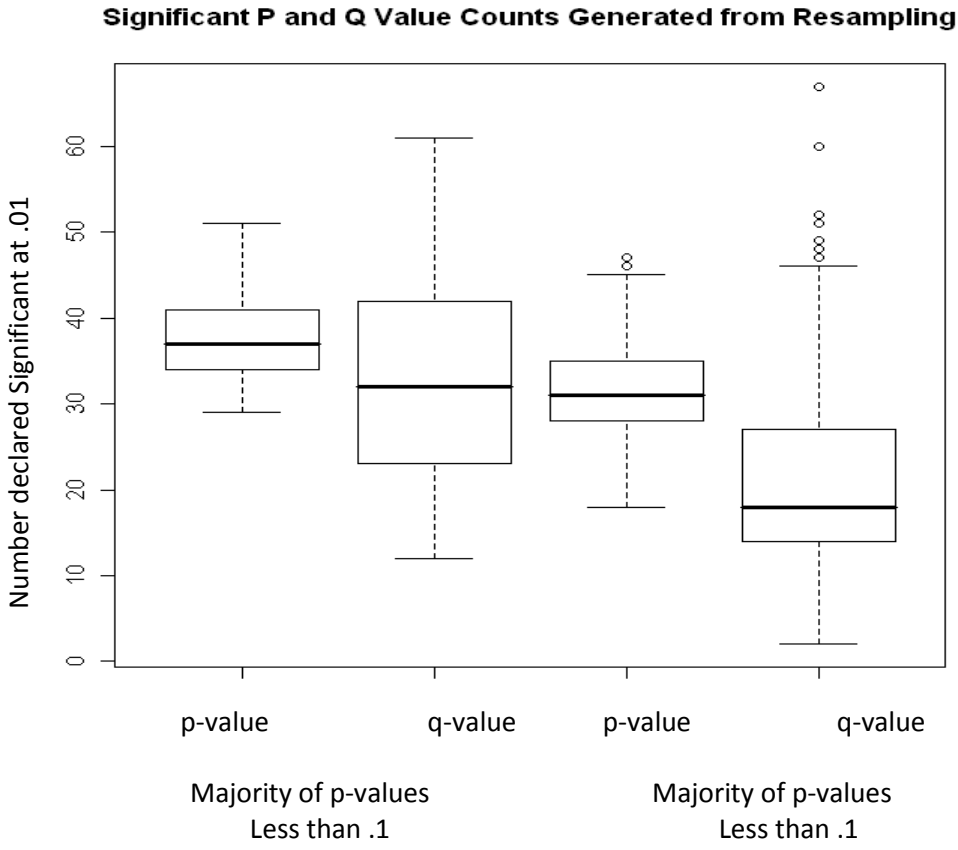
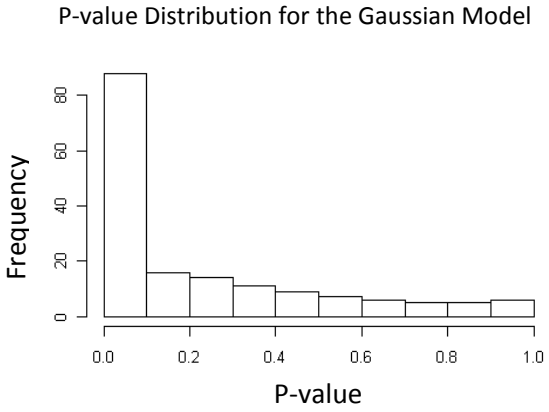
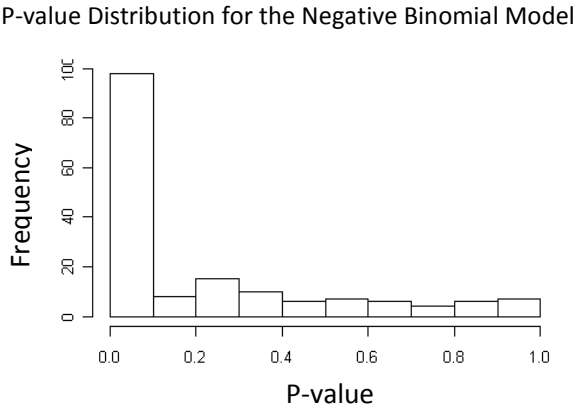
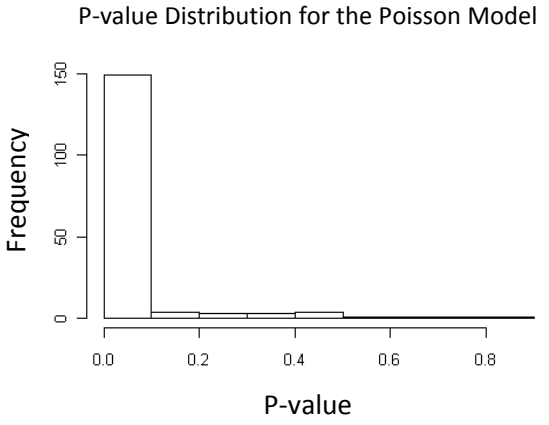


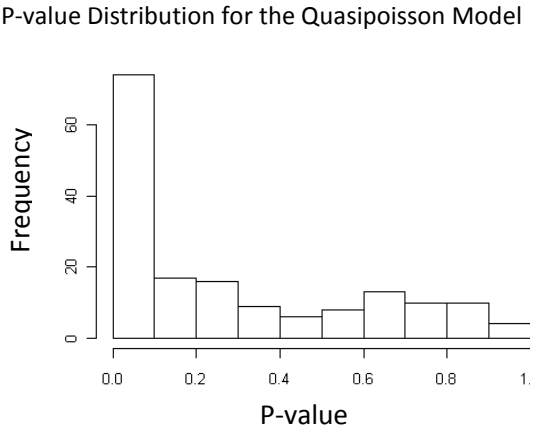
Figure 17. Relationship between significant p and q values as a function of p-value composition. Shown is the number of significant p and q values in relation to the makeup of p-values resampled from the striatum quasipoisson model. One thousand samples of p-values of length 167 were taken with replacement from the set of 167 p-values of the striatum quasipoisson model. This set was separated into two groups: those where the majority of the sampled p-values were less than .1 and those that were otherwise. From these categories, the q-values (93) were calculated using the qvalue package (94) and both the p and q values that were significant at the stated level were counted. The significance levels used for this example were .05 and .01

18a.)


P-value Distribution for the Gaussian Model
Frequency
0
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0
P-value


P-value Distribution for the Negative Binomial Model
Frequency
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
P-value


P-value Distribution for the Poisson Model
Frequency
0
50
100
150
0.0
0.2
0.4
0.6
0.8
P-value


P-value Distribution for the Quasipoisson Model
Frequency
0
20
40
60
0.0
0.2
0.4
0.6
0.8
1.
P-value

18b.)

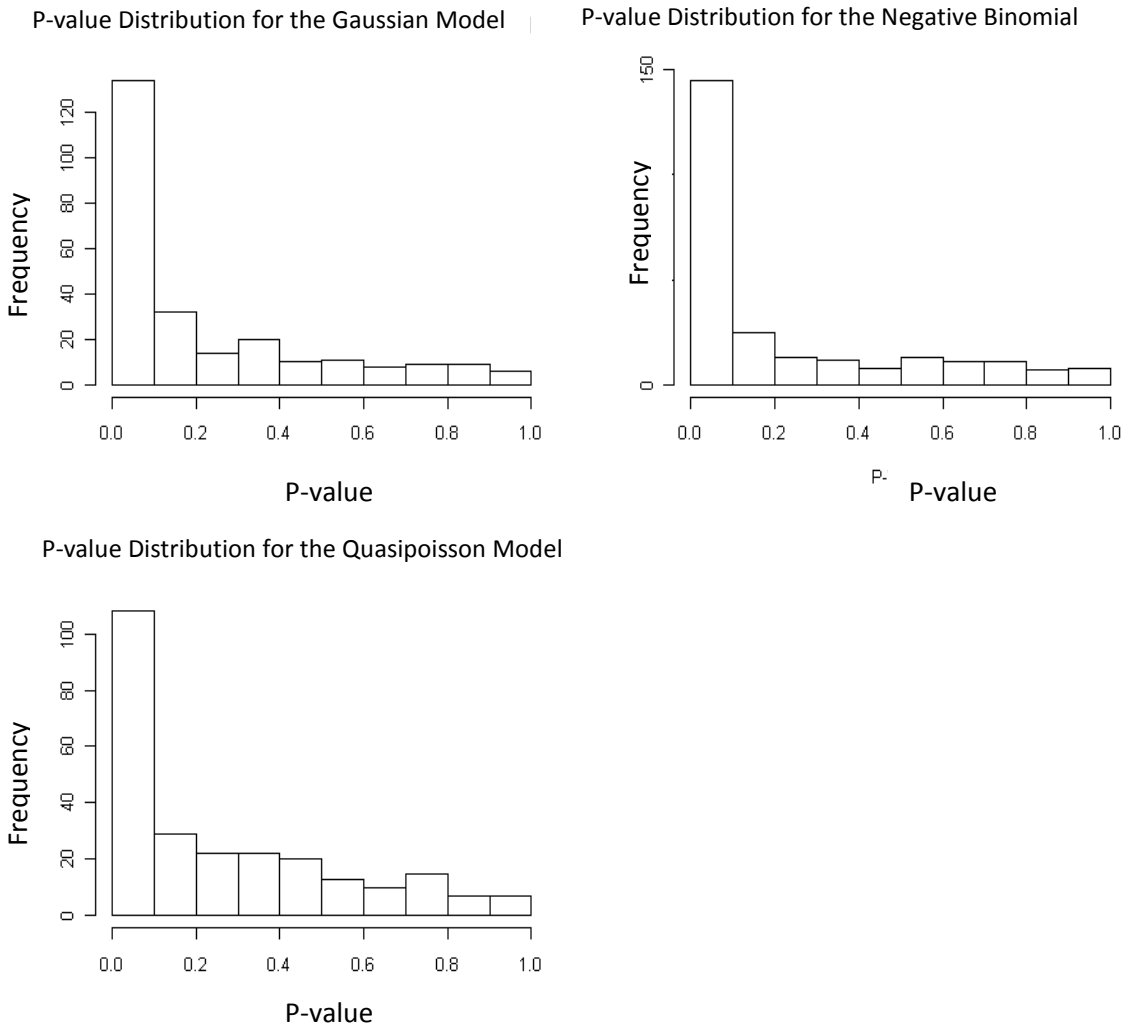


Figure 18. Distribution of p-values for the three models. Striatum (a), whole-brain is (b). For a.), the Gaussian, Poisson, quasipoisson and negative binomial p-value distribution are shown. b.) shows only the p-value distributions for the Gaussian, negative binomial and quasipoisson since the Poisson could not be calculated.

Overall, per category the whole-brain and striatum produced similar results with respect to GLM (99) family (Table 12). Both the Gaussian and quasipoisson produced similar numbers of significant values, with respect to q-values at the .01 level and as expected showed an increased

number of successes using uncorrected p-values. If we looked across genes for each type of segment category, we saw that the majority of comparisons took place between two overlapping transcript segments compared to one (2-1) or three compared to two (3-2) or one (3-1) (Table 13). Within these categories the negative binomial seemed to perform the best as compared to the quasipoisson or Gaussian models for both whole-brain and striatum. For each gene, the effectiveness of the models were compared by looking for the number of times each model counted a gene as having the majority of its categories declared significant. Again the negative binomial models performed the best containing 44 significant genes for striatum and 63 for whole-brain. The Gaussian and quasipoisson contained 24 and 15 for striatum and 47 and 20 for whole-brain. Of these, 19, 12 and 31 were significant for every exon segment comparison in striatum for the Gaussian, quasipoisson and negative binomial models. Similarly there were 43, 18 and 54 significant for whole-brain Gaussian, quasipoisson and negative binomial distributions. Significance for the majority of gene categories did not seem to depend on the type of event. For the negative binomial model applied to the striatum dataset, 45% of those genes that were considered to be mostly significant contained both differences affecting entire exons (e.g. CE or ME events) and exon isoforms, 25% contained only events that involved whole exons and 30% contained events with only exon isoforms. This compared to 57%, 17% and 26% for those genes where the majority of negative binomial category comparisons were non-significant. For whole-brain, the composition of event types for those genes where the majority of comparisons were non-significant was about the same as it was for striatum: 56%, 12% 32%. However, the categories contained about an even proportion of event types for those genes that were significant. Interestingly, an estimate of effect size for the genes containing a 2-1 comparison showed that the genes where the majority of comparisons were significant had a larger difference than those genes that were non-significant for both whole-brain and striatum

(Fig. 23). All GLM (99) models agreed upon significance for 10 genes in striatum and 12 for whole-brain. Considering that these numbers were out of a possible 109 and 157, they were rather low, though this was to be expected considering the stringent significance criteria. A possible factor that may have hindered local detection of differential expression was the overall gene expression. To examine this possibility, boxplots showing the log2(RPKM) (17) for each of the 109 and 157 genes were created for both those genes that had the majority of their exon segment categories declared significant by all GLM (99) models and those genes that did not meet this criteria (Fig. 19). Although the median expression level for those genes that were considered significant was higher in both striatum and whole brain, the difference was significant for striatum (one sided p-value = 0.03319; wilcoxon rank sum test) but not for whole-brain (one sided p-value = 0.2488; wilcoxon rank sum test).

13a.)

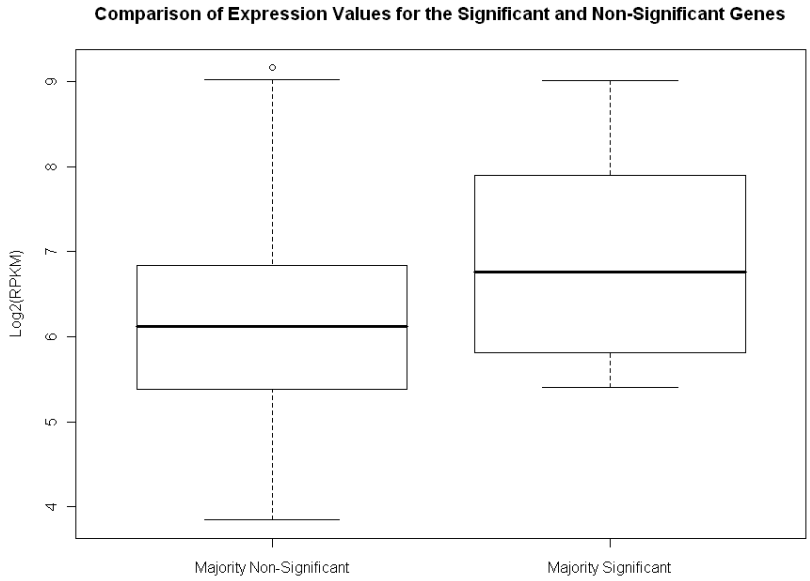
Whole-Brain				
Comparison	Gaussian	Q. Pois.	Neg. Bin.	Total
2-1	49	20	60	153
3-1	9	5	14	25
3-2	4	1	9	27
4-1	4	0	7	10
4-2	1	0	3	9
4-3	0	0	1	9
5-1	1	0	2	3
5-2	1	0	1	2
5-3	0	0	0	3
5-4	0	0	1	3
6-1	2	1	2	2
6-2	0	0	0	2
6-3	0	1	1	2
6-4	0	0	0	2
6-5	1	0	0	1

13b.)

Striatum				
Comparison	Gaussian	Q. Pois.	Neg. Bin.	Total
2-1	23	16	38	106
3-1	7	4	14	19
3-2	2	0	5	20
4-1	2	0	3	4
4-2	0	0	2	3
4-3	0	0	1	3
5-1	0	0	0	1
5-2	0	0	0	1
5-4	0	0	0	1
7-1	0	0	0	1
7-2	0	0	0	1
7-4	0	0	0	1
7-5	0	0	0	1
8-1	1	0	1	1
8-2	0	0	0	1
8-4	0	0	0	1
8-5	0	0	0	1
8-7	0	0	0	1

Table 13.) Number of significant tests for each exon segment category. Shown are the number of significant tests for the Gaussian, quasipoisson (Q. Pois) and negative binomial models (Neg. Bin.) at a q-value of .01 for whole-brain (a) and striatum (b) for each category comparison type (Comparison) along with the total number of comparisons (Total).

19a.)


Comparison of Expression Values for the Significant and Non-Significant Genes
Log2(RPKM)
4
5
6
7
8
9
Majority Non-Significant
Majority Significant

19b.)

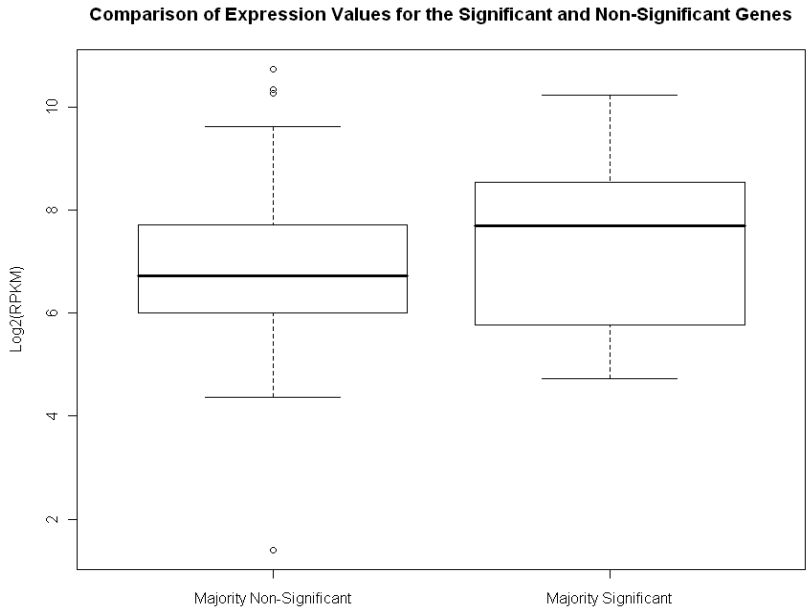

Comparison of Expression Values for the Significant and Non-Significant Genes
Log2(RPKM)
2
4
6
8
10
Majority Non-Significant
Majority Significant

Figure 19. Relationship between expression and significance. Boxplots comparing the log2(RPKM) (17) expression levels of the genes that had greater than half of the exon segment category comparisons declared significant (Majority Significant) to those that had fewer than half (Majority Non-Significant). Striatum is shown in a.) and whole-brain is shown in b.).

The next goal was to determine under what conditions differential expression could be reasonably detected using a representative generalized linear model. For this analysis a theoretical gene was created consisting of two transcripts, one of which contained a cassette exon and an alternative poly-A site. Fig. 12 shows a representation of this gene. Reads were assigned to the gene at rates starting from 10 to 150 reads per kilobase (RPK) and increasing in units of 10 RPK. The reduction in read depth between the two exon segment categories was varied from 10% to 50% in increments of 10%. Significance of the quasipoisson model measured by the p-value was used as the response variable. As can be seen in Fig. 20, the ability to detect differential expression of the exon segments was dependent on both the delta, and expression or sequencing depth—both were represented by the RPK read assignment rate. From this simulation you would need a minimum of 50 reads per kilobase of expression to comfortably detect a 10% drop in expression between the two categories. To determine how well this test would perform when the read quantity was subject to random fluctuations in read assignment, another simulation was run. Here we were looking for how robust the model would be to random mapping errors. The number of significant tests was counted after each iteration of 100 runs allowing for fluctuations up to 10% percent of the normal read assignment rate. The possible values were chosen uniformly at random to attempt to simulate a more realistic situation. This was done across a RPK range of 0 to 1,000, in steps of 10. Interestingly, in order to achieve at least a 60% success rate it was necessary to have a minimum of 20 RPK and a delta

of .3 (green line Fig. 21). Having a delta of .4 or .5 produced a very high success rate for a minimum of 20 reads per kilobase. Assuming that transcript isoform levels scale with expression or sequencing, the ability to detect alternative splicing in this manner is more or less independent of sequence assignment rate after about 20 RPK and relies mainly on differences in expression between the exon segments.

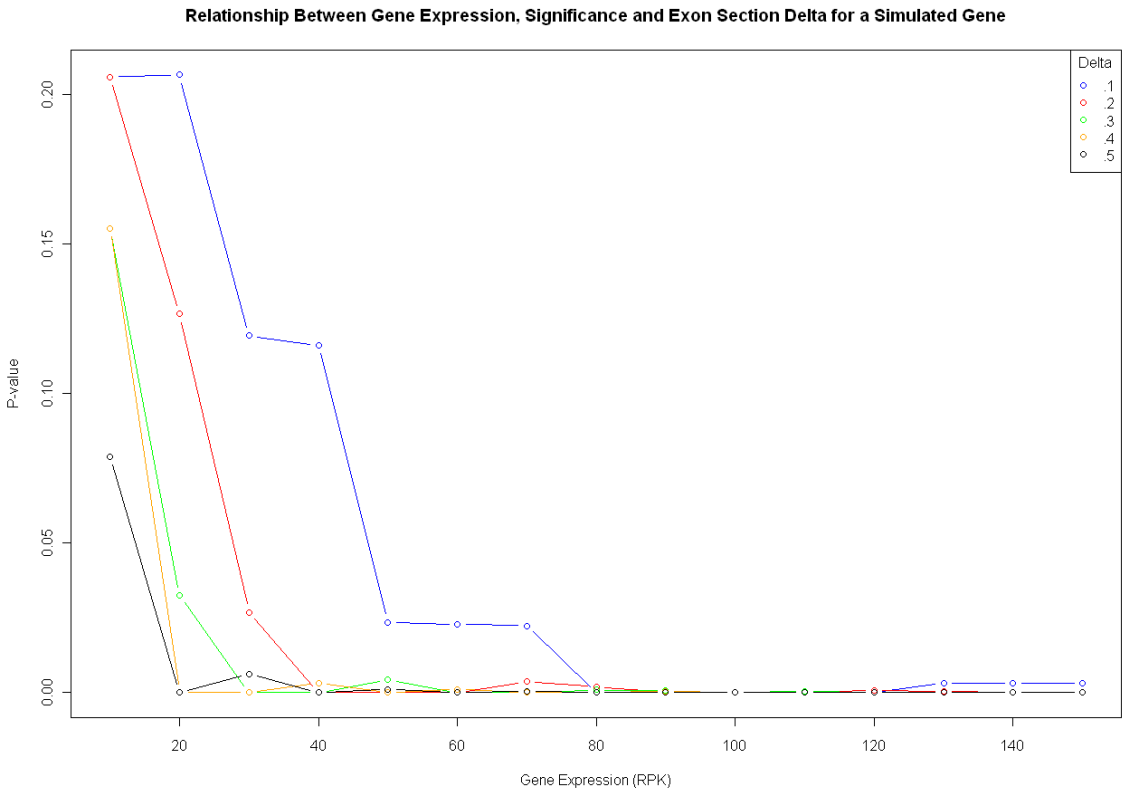


Figure 20. P-value levels for different exon segment deltas and sequencing/gene expression categories. Shown above was the relationship between gene expression, exon segment delta and p-value from the test of exon segment significance using the quasipoisson distribution for the simulate gene pictured in Fig. 12. The deltas levels ranged from a 10% drop to a 50% drop from one segment to another. Gene expression is measured in reads per kilobase (RPK).

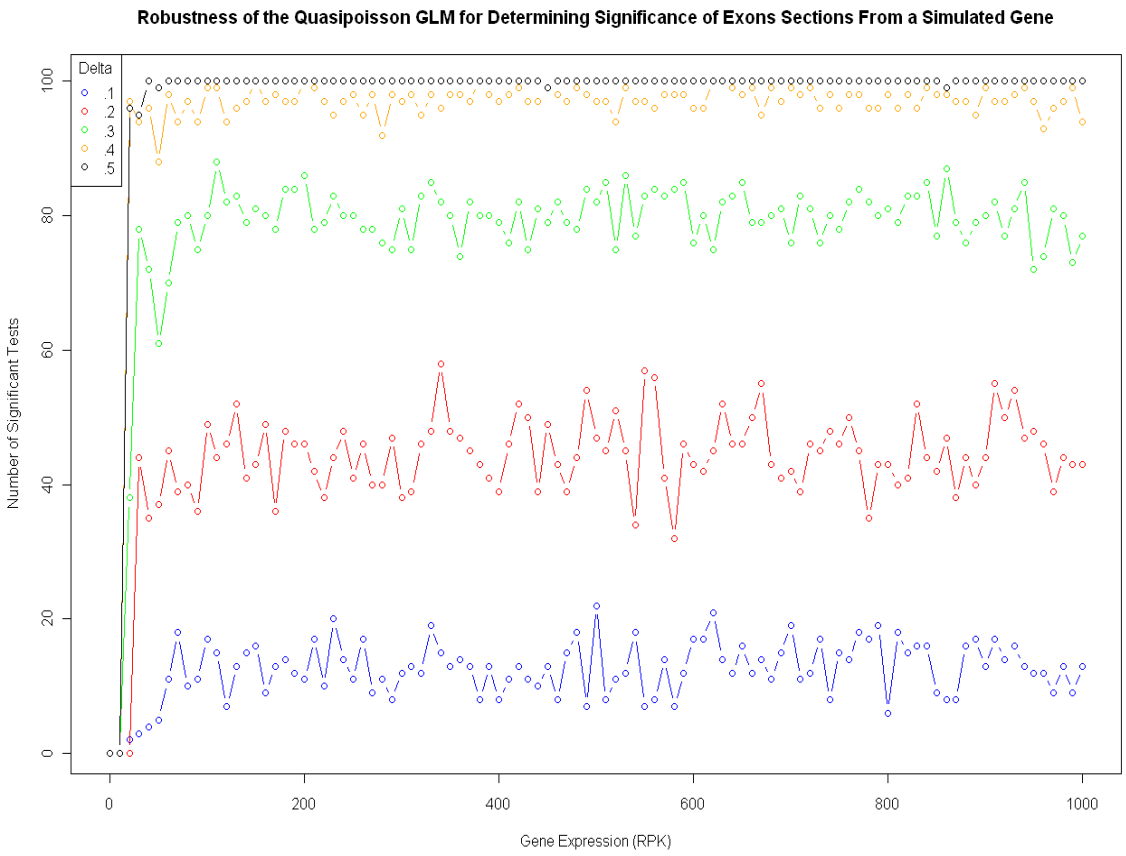


Figure 21. Robustness of quasipoisson model to random errors. This figure illustrates the number of times the quasipoisson model declared a comparison between two exon segments significant when exposed to random errors. A simulation similar to the one depicted in Fig. 20 was carried out except random mapping errors affecting at most 10% of the expected reads were introduced and the tests were carried out a 100 times each with the number of successes recorded. The read assignment rate in RPK was varied from 0 to 1,000 in increments of 10.

Discussion

The generalized linear model framework was in theory a natural fit for this type of analysis because of its flexibility and the fact the calculations are incorporated into many statistical environments such as R (88). However, because of the variability of these exon

segments in genes it was hard to pick a model that would be robust in all situations. In general all models succeeded in detecting alternative splicing, though the negative binomial seemed to be dominant both in model fit and significance tests. However, because this implementation of the negative binomial model was relatively unstable in conjunction with the difficulty in estimating accurate q-values at both the .05 and .01 level made this model less attractive. Interestingly, the Gaussian model seemed to perform better than the quasipoisson. This is likely to be because of the conservative nature of the dispersion parameter estimation procedure for the latter. For use in detecting differential expression of exon segments, the quasipoisson model seemed to be a good choice even though it performed worse than the Gaussian model because of our desire for conservative estimates. The highly variable nature of tests of these kinds warrants a conservative treatment of their results. It is also true that interpretation of the quasipoisson model's coefficients, which translate to a rate ratio (100)— could work nicely as a measure of isoform abundance. Although these models, as setup, would likely produce marginal results.

Especially interesting were the issues surrounding calculation of the q-values. These problems were likely to be due to lack of robustness of the q-value calculation in situations where the p-value distribution was highly skewed toward low p-values. Although the other commonly used procedures for accounting for multiple hypothesis testing did not show a similar effect (Table 14). The estimation of the q-values hinges on determining the parameter $\widehat{\pi_0}$ which represents the proportion of truly null features (103). This is done by default through a curve fitting process whereby the curve is fit to the equation from Storey and Tibshirani 2003:

$$\widehat{\pi_0}(\lambda) = \frac{\#\{p_j > \lambda\}}{m(1-\lambda)}$$

Where $p_1 \ldots p_m$ is an ordered list of p-values.

The problem undoubtedly arises when the distribution of the *m* p-values are highly skewed with many low values. In this case $\widehat{\pi_0}(\lambda)$ can no longer be reliably estimated through a cubic spline (Fig. 22). However, with the exception of the negative binomial distribution for whole-brain, requiring a small significance cutoff seemed to produce acceptable results when considering the relationships between the p and q values.

As was shown, a strategy based on relatively simple linear models could detect differences between exon segments that were incorporated into different numbers of transcripts. However, it was important to note that we were not able to determine the specificity of these models because of the lack of a high quality dataset containing genes with no alternative splicing. Although only around 40% of the alternatively spliced genes were detected using the negative binomial model, firm statements about the efficacy of these models could not be made without power estimates. From Fig. 20 it was determined that the change in read depth between two exon segments was more important than quantity of reads. More specifically for reliable detection of differences between these segments it was necessary to have at minimum a 30% (Fig. 21) difference between them. Wang et al. 2008, in one of the largest and most thorough experiments of this type performed, showed that about 90% of genes are estimated to have minor isoform frequencies of 20% or greater. Their main statement from this was that the majority of alternatively spliced transcripts from a given gene in humans are expressed at levels different enough to be detected (5). A minor isoform

frequency of 20% could be represented by 4 reads mapping to one segment and 16 mapping to another (5). This would be equivalent to a delta of .75 in a scenario in which the exon segment comparisons were made solely between the major and minor isoforms. For example this would be the case for the simulated gene in Fig. 20 if the top transcript was the minor isoform. Based upon the prediction from Wang et al. 2008, events for human tissue could be detected with near 100% accuracy at any combination of sequencing or expression that resulted in sequence assignment rate greater than 30 reads per kilobase for the simplest type of comparison between a major and minor isoform (5). However, the framework presented here only looks at the number of transcripts that the exon segments would be incorporated into. For complex alternative splicing situations, there is likely to be a mixture of isoforms represented in these exon segments. This makes detection of the events much more difficult as seen by the relatively low rate of success of these models in differentiating between exon segments for a set that are all alternatively spliced. The exact quantity of these mixtures is unknown since the extent of alternative splicing has not been quantified in its entirety. How best to incorporate individual isoform expression information into the model will be examined in future studies.

Supplemental Data

	Striatum				Whole-Brain			
	p-value	q-value	BH-FDR	Bonferroni	p-value	q-value	BH-FDR	Bonferroni
Gaussian	71	84	46	13	113	140	77	30
Q. Poisson	58	63	32	10	84	98	35	3
Neg. Bin.	88	92	70	43	129	140	107	63

Table 14. Significance of the GLM models for the most common multiple hypothesis corrections. The p-values were generated using three GLM (99) models: the Gaussian, quasipoisson and negative binomial fit to the exon segment comparisons for each gene. All such comparisons were extracted and the q-value, Benjamini and Hochberg FDR (BH-FDR) (104), and

Bonferroni correction were calculated. The q-value was computed using the qvalue R package (94), and the BH-FDR and Bonferroni corrections were performed using the multtest R package (105).

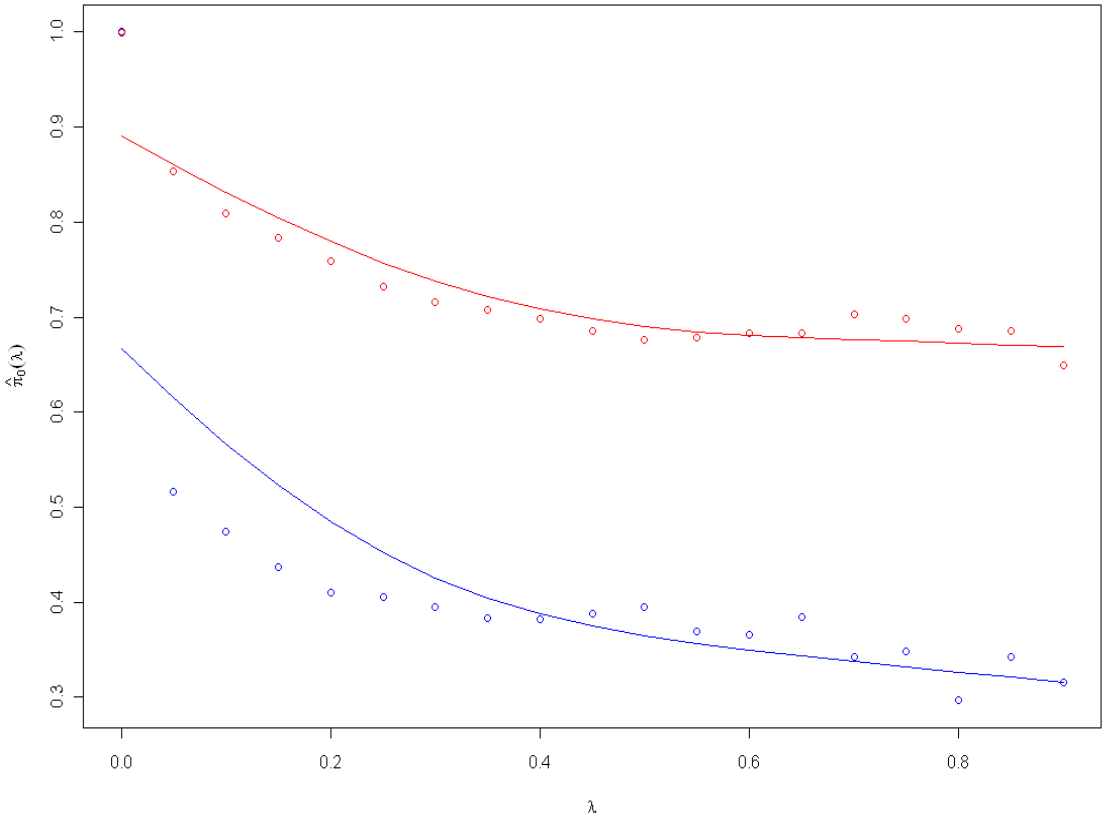
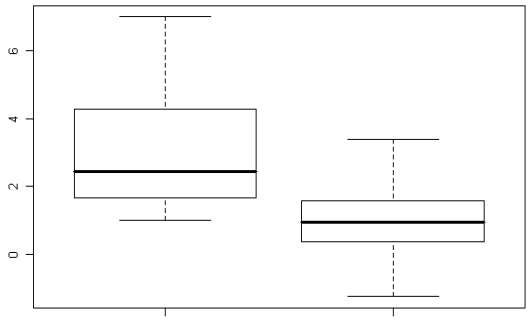


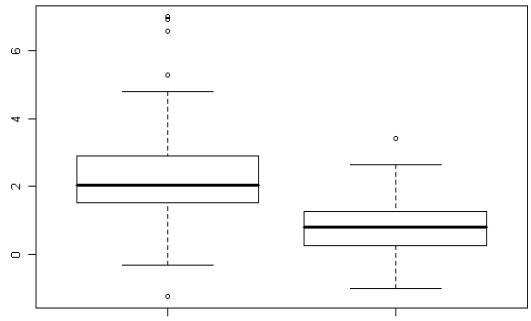
Figure 22. Cubic splines fit to whole-brain negative binomial p-values and the p-values from Storey and Tibshirani 2003. P-values were extracted from both the total negative binomial model run (blue)and microarray data reported in Storey and Tibshirani 2003 originally from Hedenfalk et al. 2001 (red) (106). This data was retrieved from: http://genomine.org/qvalue/results.txt and the plot was generated using a modified version of the qplot function from the qvalue package (94). The lines are from cubic spline fits.

23a.)

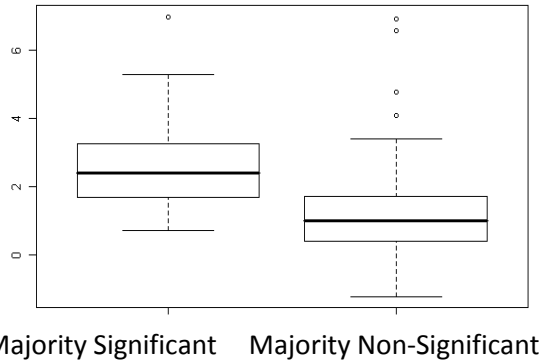
Gaussian



Negative Binomial



Quasipoisson



23b.)

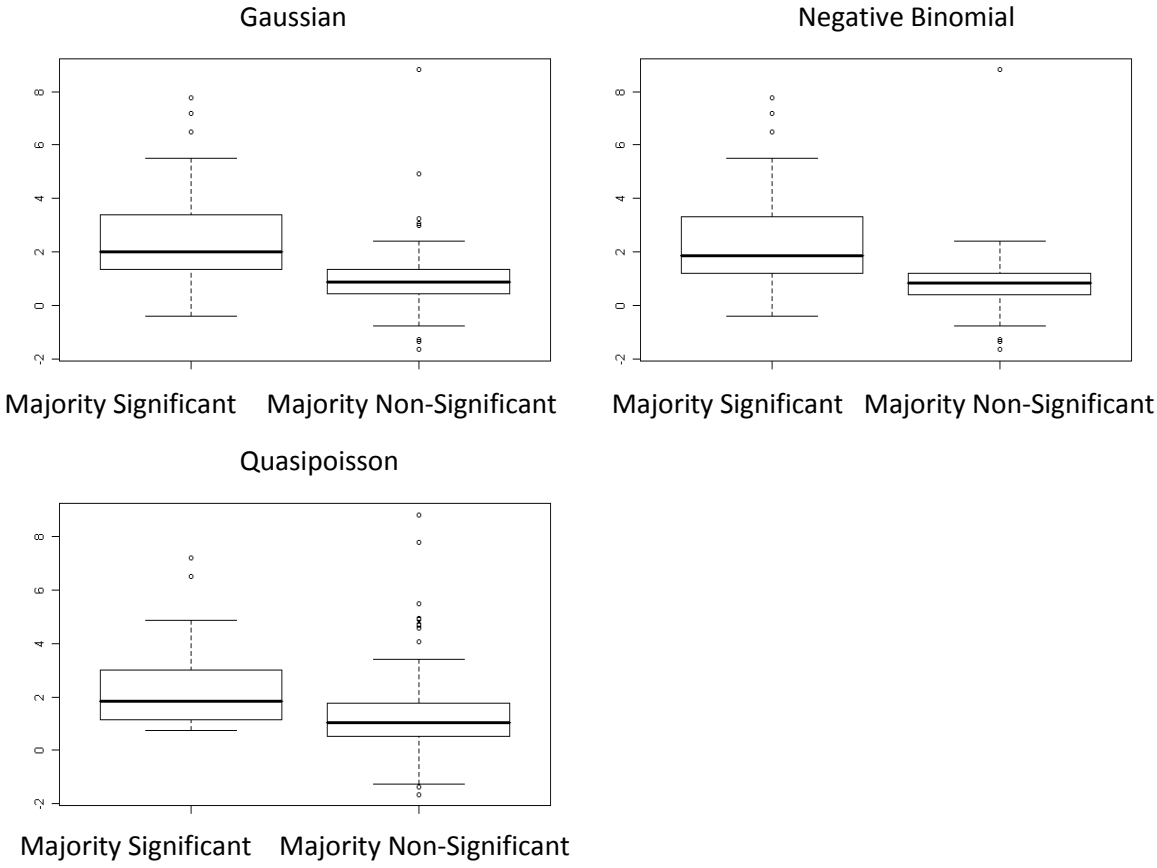


Figure 23. Log2 ratio of 2-1 comparisons for the two gene significance categories. Presented are boxplots showing the log base 2 transformed ratios consisting of the mean RPK from exon section category 2 comparisons divided by the mean RPK from the category 1 comparisons for each GLM (99). This was done for both those genes that had the majority of their category comparisons declared significant and those that did not. Striatum is shown in (a), whole-brain in (b).

Chapter 6

Conclusion

Overall the whole-brain and striatum from inbred C57BL/6J mice exhibited very similar patterns of expression and alternative splicing. Both had expression levels that correlated highly with exon array experiments. Considering that they were derived from two separate experiments from two separate labs, once normalized, they exhibited a very high correlation with each other. Multi-reads that were reassigned to their most likely position increased the correlation with the exon array experiment, though the increase was non-significant. This was similar to what was shown in Mortazavi et al. 2008. No improvement of the correlation was seen for comparisons of the splice junction and gene expression levels or for a comparison of expression values between the samples. The differences seen from the inclusion of multi-reads were relatively minor and, at least for the splice junction and gene comparison, seemed to create a more even spread of the points around the mean. This was an improvement from the consistent underestimation that occurred from only using the unique reads.

The different types of alternatively spliced isoforms were seen at similar levels to what had been reported previously. This, of course, takes into the differences in sample size (5) (15 RNA-Seq experiments from many tissues vs. 2 from the brain) and that the brain is known to have a high level of alternative splicing (13). Alternative splicing events that were unique to a single tissue were not observed at a higher rate in one tissue type relative to the other. Since the unique events seemed to have less read coverage than the common events, the loss of one or more splice junctions which lacked uniquely mapped reads likely contributed to the number of observed unique events. The presence of multi-mapping reads within these junctions likely

exacerbated this problem. However, within this set of unique alternative splicing events there were likely situations where a transcript was alternatively spliced in one tissue type, but not in the other. This was more likely to occur in whole-brain since the whole-brain should contain all of the alternative splicing events within the striatum. Promising candidates still need to be confirmed experimentally to provide evidence for this situation.

When the transcript isoforms from alternative splicing events were quantified using the isoform ratio from Wang et al. 2008, there was no significant global relationship between gene expression and isoform regulation. In other words genes that were up-regulated in one tissue compared to the other using the exon arrays did not have isoforms that were significantly regulated in either direction. This seemed to indicate that many genes in these two samples did not have expression levels that were different enough to discover relationships of this kind. Overall, the comparison of RNA-Seq-derived alterative splicing events with those found from the ExonModelStrain (55) package did not go well. Very few of the exons determined to be affected by alternative splicing were predicted to have the maximum difference in expression by ExonModelStrain (55), taking the overall gene expression into account. Even loosening this requirement to include multiple candidate exons from ExonModelStrain (55) did not appreciably help (data not shown). The sample sizes were too different to be able to tell if effect size measured by the PTC or the minimum number of reads mapping to a splice junction event was significantly associated with the successfully merged events. When a related measure to the %in measure put forward by Pan et al. 2008 (the PTC) was compared to a measure of alternative splicing using exon arrays, the two generally did not agree. The lack of concordance between the two measurements could have easily been a byproduct of the overall lack of compatibility of the comparisons. Using a microarray analysis that was unbalanced in the number of replicates

would have had a minor impact on the results. Bias from using an unbalanced array analysis may have been more noticeable if there was better agreement between the microarray and RNA-Seq datasets for alternative splicing.

Detection of alternative splicing events using differences in incorporation of exon sections into transcript isoforms was highly dependent on the magnitude of the actual differences between the segment categories. After we carried out exploratory data analysis it was determined that a generalized linear model framework (99) could detect differences between exon segments, though this process was fraught with difficulties. The negative binomial model framework seemed to provide the best results out of the three models compared using a set of verified alternative splicing events. However, the issues with model convergence and the creation of p-value distributions that were highly irregular made this model seem less attractive. The Poisson model, fit using quasilikelihood as opposed to the standard likelihood function made more sense because of ease of interpretation of the results and its relatively conservative estimation of the regression coefficients. Probably the best strategy to perform *de novo* prediction of alternative splicing would be to utilize the recent work on this problem implementing a Bayesian methodology (80-81). Although the exact details have yet to be worked out regarding how this is to be improved

Even though it has been around for a relatively short period of time, RNA-Seq has been shown to be useful to be able to address many of the questions that have been asked of the field of transcriptomics. Quantification of gene expression through the reads per kilobase of exon model per million mapped reads (RPKM) measure has been put forward by Mortazavi et al. 2008. Alternative splicing and quantification of transcript isoforms in the human genome has been examined thoroughly by Wang et al. 2008. However, the techniques from these landmark

papers are still being refined. Transcript isoform expression, such as the work pursued in Chapters 4 and 5 has become an active area of research for statisticians with groups putting forward sophisticated methods to attempt to provide a solution to this problem (80-81). Differential expression, whether it is at the gene or transcript level, is still unable to be reliably determined. This is because the role variability, both biological and technical, plays in these analyses is still being addressed. This study provided a glimpse into possible issues that occur when attempting to compare different datasets from different labs. Sequencing depth and read length are major factors in RNA-Seq analyses and can influence the outcome of an experiment comparing gene expression between two or more samples. Concerns involving sequencing errors (85) for Illumina sequencing and biases in expression measurements resulting from differences in transcript length (90) are other issues that persist. As with microarrays, these issues will be resolved once more data can be generated and analyzed, revealing the strengths and limitations of this technology.

The work presented in the preceding chapters addresses some important questions concerning the viability of RNA-Seq to detect and quantify alternative splicing between two very similar tissues—whole-brain and striatum. Because of the similarity of the two tissues the differences in effect size, in this case the differences in the population of transcripts, was very small. The easiest way to increase observed effect size is to compare alternative spicing in different types of tissues. As more RNA-Seq datasets become available it will become feasible to quantify these events for multiple tissues and time points. It has been suggested from this work that RNA-Seq can detect alternative splicing events, however this technology seemed highly dependent on the differences in the makeup of the transcript populations. For example, Wang et al. 2008 estimated, based on simulation, that an event with a minor isoform frequency of 10%

could be detected reliably (.79 power) at 100 RPKM. An event with a minor isoform frequency of 1% would need ten times that amount or around 1000 RPKM (5). The minor isoform, as defined by Wang et al. 2008, referred to the lesser expressing transcript of the two compared for a given alternative splicing event. This dependence on effect size may not have been overcome through more sequencing for some lowly expressed genes. Mortazavi et al. 2008 estimated that for their dataset consisting of 40 million reads a splice junction could be detected with 95% confidence at expression values greater than 11 RPKM. The RPKM measure by definition takes into account the total number of reads so if we assume that the values would remain constant between datasets then about 27.2% of the expressed genes (those that had at least one read mapping to their exons) in the striatum dataset met this threshold. A rough calculation shows that in order to lower this requirement to include those genes that had expression values of 5 RPKM or greater (45.5% of the striatum dataset) we needed around 44 million mapped reads, equivalent to about 7-8 lanes. This calculation assumes that we would need 220 reads per kilobase for accurate splice junction detection since only ~50% of the reads from Mortazavi et al. 2008 mapped uniquely. However, in order to truly determine the necessary sample size for these experiments larger datasets are needed to accurately estimate the effectiveness of alternative splicing detection at saturation. It was clear that this analysis had insufficient sample size for both tissues. This was especially true since the two tissues it focused on were closely related, which limited the achievable effect size and likely made the minor isoforms more difficult to reliably detect. Future work in the short term will need to first focus on accurately estimating the sample size needed for these experiments and on quantifying technical variation. Long term future work should focus on applying more advanced statistical and machine learning approaches to aspects of RNA-Seq data analysis including the detection of alternative splicing.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.
2. Modrek B, Resch A, Grasso C, Lee C. Genome-wide detection of alternative splicing in expressed sequences of human genes. NuclAcids Res. 2001;29(13):2850-9.
3. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science (New York, NY). 2003;302(5653):2141-4.
4. Clark T, Schweitzer A, Chen T, Staples M, Lu G, Wang H, et al. Discovery of tissue-specific exons using comprehensive human exon microarrays. Genome biology. 2007;8(4):R64.
5. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456(7221):470-6.
6. Philips AV, Timchenko LT, Cooper TA. Disruption of Splicing Regulated by a CUG-Binding Protein in Myotonic Dystrophy. Science. 1998;280(5364):737-41.
7. Ge K, DuHadaway J, Du W, Herlyn M, Rodeck U, Prendergast GC. Mechanism for elimination of a tumor suppressor: Aberrant splicing of a brain-specific exon causes loss of function of Bin1 in melanoma. Proceedings of the National Academy of Sciences of the United States of America. 1999;96(17):9689-94.
8. Wang GS, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. Nature reviewsGenetics. 2007;8(10):749-61.
9. Chandrasekharan NV, Dai H, Roos KLT, Evanson NK, Tomsik J, Elton TS, et al. COX-3, a cyclooxygenase-1 variant inhibited by acetaminophen and other analgesic/antipyretic drugs: Cloning, structure, and expression. Proceedings of the National Academy of Sciences of the United States of America. 2002;99(21):13926-31.
10. Gilbert W. Why genes in pieces? Nature. 1978;271(5645):501-.
11. Koscielny G, Texier VL, Gopalakrishnan C, Kumanduri V, Riethoven J-J, Nardone F, et al. ASTD: The Alternative Splicing and Transcript Diversity database. Genomics. 2009;93(3):213-20.
12. Hubbard TJP, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, et al. Ensembl 2009. Nucleic acids research. 2009;37(suppl_1):D690-7.
13. Grabowski PJ, Black DL. Alternative RNA splicing in the nervous system. Progress in Neurobiology. 2001;65(3):289-308.
14. Rougeon F, Kourilsky P, Mach B. Insertion of a rabbit {beta}-globin gene sequence into an E.coli plasmid. NuclAcids Res. 1975;2(12):2365-78.
15. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 1977;74(12):5463-7.
16. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. Science (New York, NY). 1991;252(5013):1651-6.
17. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods. 2008;5(7):621-8.
18. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science (New York, NY). 1995;270(5235):484-7.
19. Harbers M, Carninci P. Tag-based approaches for transcriptome research and genome annotation. Nature methods. 2005;2(7):495-502.

20. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nature biotechnology. 2000;18(6):630-4.
21. t Hoen PAC, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RHAM, de Menezes RX, et al. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. NuclAcids Res. 2008;36(21):e141.
22. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(26):15776-81.
23. Kim JB, Porreca GJ, Song L, Greenway SC, Gorham JM, Church GM, et al. Polony Multiplex Analysis of Gene Expression (PMAGE) in Mouse Hypertrophic Cardiomyopathy. Science. 2007;316(5830):1481-4.
24. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nature methods. 2008;5(7):613-9.
25. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science (New York, NY). 2008;320(5881):1344-9.
26. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science (New York, NY). 2008;321(5891):956-60.
27. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 2008;453(7199):1239-43.
28. Kent WJ. BLAT--the BLAST-like alignment tool. Genome research. 2002;12(4):656-64.
29. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21(9):1859-75.
30. van Nimwegen E, Paul N, Sheridan R, Zavolan M. SPA: A Probabilistic Algorithm for Spliced Alignment. PLoS Genetics. 2006;2(4):e24.
31. Kapustin Y, Souvorov A, Tatusova T, Lipman D. Splign: algorithms for computing spliced alignments with identification of paralogs. Biology Direct. 2008;3(1):20.
32. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. A Computer Program for Aligning a cDNA Sequence with a Genomic DNA Sequence. Genome research. 1998;8(9):967-74.
33. Wheelan SJ, Church DM, Ostell JM. Spidey: A Tool for mRNA-to-Genomic Alignments. Genome research. 2001;11(11):1952-7.
34. De Bona F, Ossowski S, Schneeberger K, Ratsch G. Optimal spliced alignments of short sequence reads. Bioinformatics. 2008;24(16):i174-80.
35. Schulze U, Hepp B, Ong CS, Ratsch G. PALMA: mRNA to genome alignments using large margin algorithms. Bioinformatics. 2007;23(15):1892-900.
36. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105-11.
37. Jongeneel CV, Iseli C, Stevenson BJ, Riggins GJ, Lal A, Mackay A, et al. Comprehensive Sampling of Gene Expression in Human Cell Lines with Massively Parallel Signature Sequencing. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(8):4702-5.

38. Pauws E, van Kampen AHC, van de Graaf SAR, de Vijlder JJM, Ris-Stalpers C. Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. NuclAcids Res. 2001;29(8):1690-4.
39. Moore MJ, Silver PA. Global analysis of mRNA splicing. RNA (New York, NY). 2008;14(2):197-203.
40. Black DL. Protein Diversity from Alternative Splicing: A Challenge for Bioinformatics and Post-Genome Biology. Cell. 2000;103(3):367-70.
41. Clark TA, Sugnet CW, Ares M, Jr. Genomewide Analysis of mRNA Processing in Yeast Using Splicing-Specific Microarrays. Science. 2002;296(5569):907-10.
42. Fehlbaum P, Guihal C, Bracco L, Cochet O. A microarray configuration to quantify expression levels and relative abundance of splice variants. NuclAcids Res. 2005;33(5):e47.
43. Boutz PL, Stoilov P, Li Q, Lin C-H, Chawla G, Ostrow K, et al. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. Genes & development. 2007;21(13):1636-52.
44. Ule J, Ule A, Spencer J, Williams A, Hu J-S, Cline M, et al. Nova regulates brain-specific splicing to shape the synapse. Nature genetics. 2005;37(8):844-52.
45. Srinivasan K, Shiue L, Hayes JD, Centers R, Fitzwater S, Loewen R, et al. Detection and measurement of alternative splicing using splicing-sensitive microarrays. Post-transcriptional Regulation of Gene Expression. 2005;37(4):345-59.
46. Le K, Mitsouras K, Roy M, Wang Q, Xu Q, Nelson SF, et al. Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. NuclAcids Res. 2004;32(22):e180.
47. Sugnet CW, Srinivasan K, Clark TA, O'Brien G, Cline MS, Wang H, et al. Unusual Intron Conservation near Tissue-Regulated Exons Found by Splicing Microarrays. PLoS Computational Biology. 2006;2(1):e4.
48. Fagnani M, Barash Y, Ip JY, Misquitta C, Pan Q, Saltzman AL, et al. Functional coordination of alternative splicing in the mammalian central nervous system. Genome biology. 2007;8(6):R108.
49. Pan Q, Shai O, Misquitta C, Zhang W, Saltzman AL, Mohammad N, et al. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Molecular cell. 2004;16(6):929-41.
50. Yeakley JM, Fan J-B, Doucet D, Luo L, Wickham E, Ye Z, et al. Profiling alternative splicing on fiber-optic arrays. Nat Biotech. 2002;20(4):353-8.
51. Gardina P, Clark T, Shimada B, Staples M, Yang Q, Veitch J, et al. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. BMC Genomics. 2006;7(1):325.
52. Xing Y, Kapur K, Wong WH. Probe Selection and Expression Index Computation of Affymetrix Exon Arrays. PLoS ONE. 2006;1(1):e88.
53. Purdom E, Simpson KM, Robinson MD, Conboy JG, Lapuk AV, Speed TP. FIRMA: a method for detection of alternative splicing from exon array data. Bioinformatics. 2008;24(15):1707-14.
54. Thorsen K, Sorensen KD, Brems-Eskildsen AS, Modin C, Gaustadnes M, Hein A-MK, et al. Alternative Splicing in Colon, Bladder, and Prostate Cancer Identified by Exon Array Analysis. Mol Cell Proteomics. 2008;7(7):1214-24.
55. Laderas T, Walter N, Mooney M, Vartanian K DP, Buck K, Harrington C, et al. Unpublished. 2009.

56. Walter NA, McWeeney SK, Peters ST, Belknap JK, Hitzemann R, Buck KJ. SNPs matter: impact on detection of differential expression. Nature methods. 2007;4(9):679-80.
57. Manak JR, Dike S, Sementchenko V, Kapranov P, Biemar F, Long J, et al. Biological function of unannotated transcription during the early development of Drosophila melanogaster. Nature genetics. 2006;38(10):1151-8.
58. Zhang Z, Hesselberth JR, Fields S. Genome-wide identification of spliced introns using a tiling microarray. Genome research. 2007;17(4):503-9.
59. David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, et al. A high-resolution map of transcription in the yeast genome. 2006;103(14):5320-5.
60. Castle JC, Zhang C, Shah JK, Kulkarni AV, Kalsotra A, Cooper TA, et al. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nature genetics. 2008;40(12):1416-25.
61. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376-80.
62. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, et al. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. Science. 2005;309(5741):1728-32.
63. Hyman ED. A new method of sequencing DNA. Analytical Biochemistry. 1988;174(2):423-36.
64. Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P. Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. Analytical Biochemistry. 1996;242(1):84-9.
65. Bainbridge M, Warren R, Hirst M, Romanuik T, Zeng T, Go A, et al. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics. 2006;7(1):246.
66. Metzker ML. Emerging technologies in DNA sequencing. Genome research. 2005;15(12):1767-76.
67. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Meth. 2007;4(8):651-7.
68. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science. 2007;316(5830):1497-502.
69. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, et al. High-Resolution Profiling of Histone Methylations in the Human Genome. Cell. 2007;129(4):823-37.
70. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. 2007;448(7153):553-60.
71. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008;133(3):523-36.
72. Rosenkranz R, Borodina T, Lehrach H, Himmelbauer H. Characterizing the mouse ES cell transcriptome with Illumina sequencing. Genomics. 2008;92(4):187-94.
73. Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, et al. The UCSC Genome Browser Database: 2008 update. NuclAcids Res. 2008;36(suppl_1):D773-9.
74. Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, et al. Extending assembly of short DNA sequences to handle error. Bioinformatics (Oxford, England). 2007;23(21):2942-4.

75. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome research. 2008;18(9):1509-17.
76. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215(3):403-10.
77. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature genetics. 2008;40(12):1413-5.
78. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. Journal of Molecular Biology. 1997;268(1):78-94.
79. Siepel A, Haussler D. Combining Phylogenetic and Hidden Markov Models in Biosequence Analysis. Journal of Computational Biology. 2004;11(2-3):413-28.
80. Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics. 2009;25(8):1026-32.
81. Zheng S, Chen L. A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. Nucleic acids research. 2009(Journal Article).
82. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research. 2008;18(11):1851-8.
83. Smith AD, Xuan Z, Zhang MQ. Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC bioinformatics. 2008;9(Journal Article):128.
84. Langmead B, Trapnell C, Pop M, Salzberg S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology. 2009;10(3):R25.
85. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. NuclAcids Res. 2008;36(16):e105.
86. Mouse Genome Sequencing C, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420(6915):520-62.
87. Faulkner GJ, Forrest ARR, Chalk AM, Schroder K, Hayashizaki Y, Carninci P, et al. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. Genomics. 2008;91(3):281-8.
88. CoreTeam RD. R: A language and environment for statistical computing. 2.8.1 ed2008.
89. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249-64.
90. Oshlack A, Wakefield M. Transcript length bias in RNA-seq data confounds systems biology. Biology Direct. 2009;4(1):14.
91. Genome Exploration Research G. FANTOM3: Functional Annotation of Mouse 3.
92. Garcia-Blanco MA, Baraniak AP, Lasda EL. Alternative splicing in disease and therapy. Nature biotechnology. 2004;22(5):535-46.
93. Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002;64(3):479.
94. Dabney A, Storey JD, Warnes wafGR. qvalue: Q-value estimation for false discovery rate control. Anonymous, editor.
95. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy---analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307-15.
96. Okoniewski M, Yates T, Dibben S, Miller C. An annotation infrastructure for the analysis and interpretation of Affymetrix exon array data. Genome biology. 2007;8(5):R79.

97. Robinson MD, Speed TP. A comparison of Affymetrix gene expression arrays. BMC bioinformatics. 2007;8(Journal Article):449.
98. Affymetrix. 2005.
99. Nelder JA, Wedderburn RWM. Generalized Linear Models. Journal of the Royal Statistical SocietySeries A (General). 1972;135(3):370-84.
100. Der G, Everitt BS. Statistical Analysis of Medical Data using SAS. Anonymous, editor. Boca Raton, FL: Taylor & Francis Group; 2006.
101. McCullagh P, Nelders JA. Generalized Linear Models. Anonymous, editor. Boca Raton, FL: CRC Press; 1989.
102. Venables WN, Ripley BD. Modern Applied Statistics with S. Anonymous, editor. New York: Springer; 2002.
103. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(16):9440-5.
104. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical SocietySeries B (Methodological). 1995;57(1):289-300.
105. Pollard KS, Ge Y, Taylor S, Dudoit S. multtest: Resampling-based multiple hypothesis testing. Anonymous, editor.
106. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, et al. Gene-Expression Profiles in Hereditary Breast Cancer. The New England journal of medicine. 2001;344(8):539-48.