Identifying developmentally regulated exons in the human brain

Alain Bateman

Master of Science

Department of Human Genetics

McGill University

Montreal, Quebec

2016-8-15

A thesis submitted to McGill University in partial fulfilment of the requirements of the degree of Master of Science

Acknowledgements

I would like to thank and acknowledge the contributions of my supervisor Dr. Claudia Kleinman for her help in the design and execution of this project. I would like to thank and acknowledge the contributions of my supervisory committee: Dr. Guillaume Bourque and Dr. Celia Greenwood for their feedback and support.

I would like to thank and acknowledge the support and contributions of the Kleinman lab. In particular, Dr. Rodrigo Sieira for his assistance in the primer design, Karine Choquet for her assistance in overseeing the qtRT-PCR experiments and for acquiring and preparing the biological samples. I'd like to thank Steven Hebert and Nicolas De Jay for the use of their, and their assistance with, RNAseq pipeline tools.

I would furthermore like to thank our collaborators Dr. Nada Jabado and Dr. Gustavo Turecki for their contribution of human brain samples that enabled our validation testing. I'd like to thank Dr. Marie Forest and Dr. Aurélie Labbe for their helpful discussions around establishing distances between parameterized curves. I'd also like to thank Dr. Santiago Costantino for his input on initializing parameters in solving non-linear equations.

Finally, I'd like to thank the Genomics Platform of the Institute for Research in Immunology and Cancerology (Montreal, Canada) for performing qtRT-PCR experiments and the FRSQ for partly funding this work.

Abstract

Alternative splicing generates variety in the exonic content of mature RNA. Exonic content has been shown to vary over the course of the maturation of the human brain. Current methods for the detection of these isoforms are limited by compounded uncertainty in the case of isoform reconstruction or limited dimensionality in the case of analysis of variant studies. Here we present a parameterized approach to summarize as trajectories the changing expression of exonic content over the course of the development of the human brain. We apply our methods to the Brainspan dataset and present several analyses of the exonic trajectories inferred by our methods to predict differentially expressed exons. We identify approximately 4000 developmentally regulated exons among 2500 genes in the brain. We validate our predictions in a novel validation dataset and some select candidates in mice. These methods provide evidence of previously unknown temporal bounds on the expression of these exons in-vivo. Thus we advance our knowledge on the evolving transcriptome of the human brain and increase our toolset for the analysis of temporally regulated exonic content.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AS	*Alternative Splicing*	9
5'ss	*5' splice site*	13
ipb	intron branch point	13
3'ss	3' splice site	13
snRNP	small nuclear Ribonucleic Proteins	15
SR	protein serine/arginine protein	15
RRM	RNA recognition motif	15
ESE	Exonic Splicing Enhancer	15
ESI	Exonic Splicing Inhibitors	15
ISE	Intronic Splicing Enhancers	15
ISI	Intronic Splicing Inhibitors	15
NMD	Nonsense-Mediated Decay	20
NGS	Next Generation Sequencing	22
RPKM	Reads per kilobase per million reads	33
bp	base-pair	36
LM	Levenberg-Marquardt	48
RMSE	Root Mean Standard Error	54
NRMSE	Normalized Root Mean Standard Error	54

1 Introduction

1.1 Motivation

Alternative splicing and the related selection of alternate transcription start sites are pervasive phenomena at the heart of molecular biology.[1] The varied inclusion of transcribed segments of a gene into the gene product greatly increases the diversity of mature transcripts produced by the cell.[2] Mutations in the factors regulating this complex process are a source of human disease[3–5] as is the aberrant inclusion of exonic content outside of temporal and tissue appropriate context.[6]

The motivation for developing methods of mapping the changes in exonic content over time is therefore twofold. Firstly, it sheds some light on the regulation of the mature transcriptome. Secondly, it provides an investigative avenue into neuropathological etiologies. Here we adapt a set of methods to identify differential exon expression over the course of the development of the human brain.

1.2 An overview of topics germane to this study.

This thesis deals with the approximation of temporal trajectories of exons over the course of the development of the human brain. In this section we review contemporary background knowledge that will assist in the understanding of the work.

As the principal aim of the thesis remains in identifying biological instances of alternative exon usage, we provide an overview of the phenomenon of alternative splicing (AS) in eukaryotes. While the detailed architecture of the molecular complexes involved are beyond the scope of this study, we strive to provide the required understanding of the different types of alternative splicing that occur, where they occur in the genome, and the machinery involved in their generation. Being concerned with the development of the human brain, we then proceed to layout some current examples of the regulation of alternative splicing in the development of the mammalian nervous system and some pathologies that result from aberrant alternative splicing events.

Our elucidation of temporal exonic trajectories relies on the analysis of RNAseq data. While, to the best of our current knowledge, the methods of analyses here are novel, the use of next generation sequencing to identify differential exonic expression is not. In the second part of this literature review we provide an overview of common modern approaches to this line of inquiry and we focus on two widely used approaches the first being isoform quantification and the second being variance based analysis of individual exon expression.

1.3 An overview of alternative splicing

The definition of an exon is tightly related to the process of splicing; it refers to those segments of transcribed RNA that are included in the matured transcript. It is worth noting that not all exons in the matured transcript will be translated; the regions that are not translated into the final product are called UnTranslated Regions (UTRs). In most cases exons are interleaved, along the loci, with relatively longer segments of regions termed introns. Most introns are removed over the course of the transcription of genomic DNA to RNA by a complex aggregation of proteic and ribonucleic sub-units termed the spliceosome. The association of regulators with this same ribonucleic complex may influence the selection of exons to excise from the finished product. In some cases, intronic regions are retained.

This alternative inclusion of RNA segments is an early target for the regulation of the cell's set of expressed gene products, i.e. the cell’s expression profile. By varying matured transcripts, AS events can modify the translated protein: varying its structure and therefore influencing its function. Inclusion of exons is not limited to function, AS events can also affect the localization of the mRNA or label transcripts for nonsense-mediated decay. Alternative splicing is frequent and the varied isoforms produced from AS events can produce disparate biological outcomes.

1.4 On the frequency of alternative splicing events in human beings

The initial sequencing of the human genome in 2001[7,8] confounded the expectations of many in its low estimation of the number of protein-coding transcripts: 26,000. This number has been adjusted downwards today.[9] The diversity of mature protein-coding transcripts is estimated at an order of magnitude higher and the number of mature gene products higher still. (One conservative estimate puts the number of protein-coding transcripts at 60,000[9] however, this number is less robust as it is more sensitive to genomic feature annotation.[10]) These numbers suggest a mean of 3 isoforms per gene product and indeed the vast majority of genes experience at least one AS event. Recent studies estimate that approximately 95% of multi-exon genes in human beings undergo some form of AS event although these events are scarcer in genes with fewer than 4 exons.[1] It is therefore not surprising that a consensus has developed around the suggestions that AS helps resolve the apparent discrepancy between the number of protein-coding loci and the diversity of human proteome.[11]

This consensus is further supported by evolutionary studies linking increased AS frequency with perceived organism morphological and behavioral complexity.[12] The greater frequency of AS events and the corresponding greater diversity of transcripts in human beings when compared to *Caenorhabditis elegans* (25% of multi-exon protein-coding genes undergo alternative splicing), or to Drosophila (45%,) or to mice (63%) suggest that AS may help resolve the apparent paradox of observed species diversity with conserved genomic sequences.[13] While human beings and chimps share 99% of genomic sequences, differences in splicing events range between 6-8% based on tissue.[14] Thus AS further provides a plausible mechanism for the range of observed biological phenomena, including the structural differences between the organs of cognition[15]. It is perhaps not surprising given the documented complexity of the development processes surrounding the maturation of the human brain[16,17]

1.5 On the essentials of the splicing machinery and its regulation

At the core of AS events is the spliceosome, a very large ribonucleoproteic complex that carries out the ejection of RNA segments not selected for inclusion in the mature transcript. In human beings there exists two spliceosome complexes: the common U1/U2-mediated spliceosome, and the less common U12-mediated spliceosome, which assembles on less than 0.5% of introns in human beings.[18] U12-mediated exon selection is absent in some metazoans altogether.[19] In the interest of concision, the description of the molecular essentials of alternative splicing will focus on describing the mammalian U2-spliceosome. We first describe the layout of the essential features of splicing sites: locations of the genome where alternative splicing can occur and then discuss the known, central mechanisms of its regulation. We then discuss known examples of the regulation of alternative splicing in the human nervous system germane to this work on identifying alternative exons in the human brain.

1.5.1 A review of the genomic features of splicing

Three genomic features are of particular interest in elucidating the mechanism of splicing. They are the 5' splice site (5'ss), the intron branch point (ibp) and the 3' splice site (3'ss). Upstream of the 5'ss is a sequence retained in the maturing RNA transcript. Just downstream of the 5'ss, in metazoan U2 introns, is the degenerated consensus sequence GURAGU (R is a purine). Downstream from the 5'ss is the intron branch point (ibp). An adenine base located in this branch point is directly involved in the transesterification reactions. The consensus sequence in metazoans is YNCURAC (Y is a pyrimidine, N is any nucleotide, R is a purine). In mammals, this is typically followed by a polypyrimidine tract[20] (15-20 base pairs rich in pyrimidine nucleotides, especially uracil). Downstream from the ibp by 18-40 base-pairs is the 3' splice site (3'ss). It is preceded, in metazoans, by the short degenerated consensus sequence of YAG (Y is a pyrimidine). Downstream from the 3'ss is the second retained sequence in the maturing RNA transcript.[20]

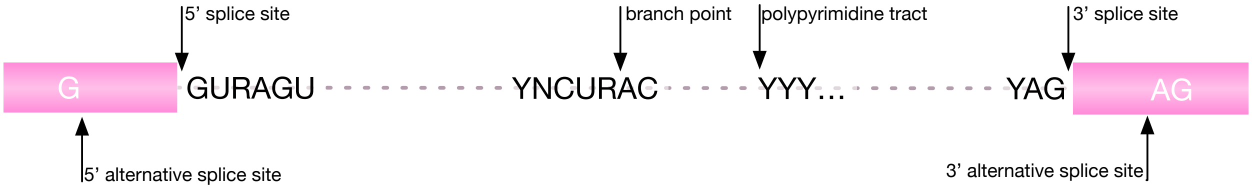

Figure 1 Semi-conserved genomic features of alternative splicing

Schematized layout of the semi-conserved genomic features involved in splicing. Above the maturing transcript indicate the essential elements recognized by the splicing machinery. The adenosine base of the branch point is involved in the first transesterification reaction. Below the transcript are indications of alternate 3'ss and 5'ss whereby a constitutive exon may be shortened or lengthened during inclusion by a repositioning of the splicing machinery.

1.5.2 An overview of the splicing reaction

The spliceosome complex, once assembled, will catalyze a pair of transesterification reactions that characterize precursor mRNA splicing. The 2' hydroxyl group of the ibp performs a nucleophilic attack on the 5'ss, cleaving the intron. The 3' hydroxyl group of the 5'ss then performs a nucleophilic attack on the 3'ss, ligating the 5'ss to the 3'ss. The circular intron (termed the lariat structure) is thus excised from the maturing mRNA.[21] There is some dispute as to the catalytic center of the spliceosome which includes RNA from the U2 and U6 snRNAs. The spliceosome's active site does contain RNA from a self-splicing group of introns[22] yet there is a substantial amount of evidence that indicate that proteins comprise part of the spliceosome's active site[23], in particular the spliceosomal protein Prp8 which contains a RNAse H-like domain[20].

1.5.3 An overview of the recognition of splicing features by the spliceosome

The U2-dependent spliceosome consists of the U1, U2, U4, U5, and U6 small nuclear ribonucleic proteins (snRNPs) and many non-snRNP proteins. Its assembly is regulated by some cis-elements located near these genomic splicing-related features discussed above which can promote or silence either these central snRNP components of the spliceosome. For instance, the U1 snRNA will recognize and base-pair with the 5'ss sequence while the U2 snRNA will recognize and base-pair with the ibp sequence[24].

1.5.4 On the regulation of splicing events by SR proteins

Cis-elements may also recruit proteins from the well characterized SR (serine/arginine) family to mediate alternative splicing events.[25] These proteins contain a RNA recognition motif (RRM) which will recognize and base-pair with the relevant elements on the maturing RNA transcript. In addition, SR proteins contain their eponymous domain, rich in serine and arginine residues. Cis elements such as exonic splicing enhancers (ESEs), exonic splicing inhibitors (ESIs), intronic splicing enhancers (ISEs) and intronic splicing inhibitors (ISIs), are sites that are often located within the exon or intron targeted for exclusion/inclusion and are generally recognized by at least one member of the SR protein family which will then interact with the core constituents of the splicing machinery as appropriate.[3]

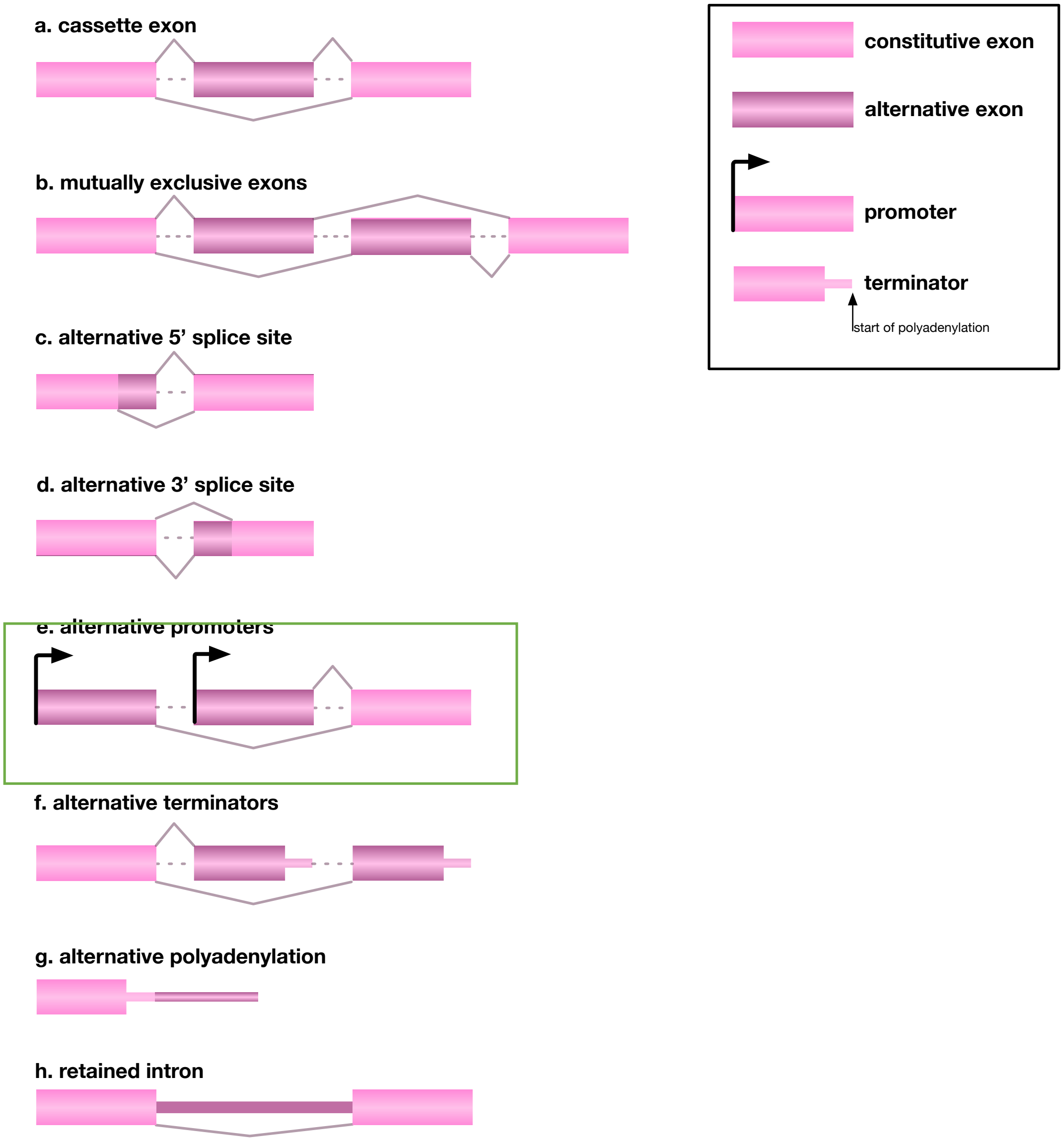
a. cassette exon
b. mutually exclusive exons
c. alternative 5' splice site
d. alternative 3' splice site
e. alternative promoters
f. alternative terminators
g. alternative polyadenylation
h. retained intron
constitutive exon
alternative exon
promoter
terminator
start of polyadenylation

Figure 2 Types of isoform generating events.

a) The alternately included cassette exon is flanked on either side by exons that will not be selectively excised, b) the mutually exclusive exon is a special case of the cassette exon in which the two alternately selected exons are generally not included in the same transcript, c) the alternative 5' splice site varies the length of the exon included in the transcript upstream from the splice site, d) the alternative 3' splice site varies the length of the downstream exon included in the transcript, e) alternative promoters are a particular case, they vary the start of transcription and can thus can be regulated quite differently to affect the start of the mature transcript, f) the varied inclusion of the last exon, g) the varied start of polyadenylation of the transcript, selectively includes different lengths of 3'UTR (untranslated region), h) the retained intron selectively includes the usually excised genomic region between 2 exons.

1.5.6 On cassette exons and exon skipping

Exon skipping occurs when the exon is selectively omitted from a transcript altogether. This is the most common AS event, accounting for 40% of AS events in eukaryotes.[26] Selective exon inclusion can insert whole functional domains or modifications to existing ones. A well-studied example is that of neuroligins transcripts: a set of membrane proteins that bind to human trans-synaptic proteins called neurexins. Neuroligin's specificity for different forms of neurexins is partly determined by the inclusion of short alternative exons at well specified positions within their ectodomains[27]. Thus, the functionality of these proteins, i.e. connecting cells at the synapse, is modified via exon regulation which in turn modifies their connector specificity.

Mutually-exclusive exons a special case of case of exon skipping. This is a much rarer event perhaps due to required coordination as the splicing of each exon involved is no longer an independent event. The selection between sequences of similar length (obviously within the constraints of an existing translative reading frame) implies that a downstream protein may conserve not only its functional structure but its spatial structure. Constitutive segments that are translated into the protein's shape are preserved and smaller active sites are modified. An intuitive and confirmed example are transcripts for proteins that are involved in transmembrane ion transport channels.[28,29]

1.5.7 On alternative splice site selection

The molecular machinery of the spliceosome complex and its cis-elements can act to shorten or elongate an included exon. Modification of the length of the included exon can occur either upstream, at the 5′ exon or downstream, at the 3′ exon giving rise to alternative 5′ splice sites (5′ss) or 3′ splice sites (3′ss) respectively.

Sites where alternative splicing occurs in cases of aberrant splicing are termed cryptic 5' or 3′ splice sites. A relevant example is a well-documented hereditary case that assisted in the discovery of a principal cause of an inherited form frontotemporal dementia through the MAPT protein (also known as TAU).[30] This study demonstrated that mutations only 13-16 base pairs downstream of *TAU*'s tenth exon promote the inclusion of exon 10, resulting in a longer isoform which presents high-yield phosphorylation targets. Later studies confirmed the authors suspicions that the mutations impacted a secondary structure of the mRNA precursor affecting the formation of the splicing complex.[31] Aberrant splicing at such sites are referred to as cryptic 3' splice site selection.

1.5.8 On alternative polyA sites

Almost all protein-encoding transcripts in human beings, with the exception of some histone genes are poly-adenylated.[32] Prior to the polymerization of the adenylated tail, the transcripts must be cleaved and this often occurs at a site (often a CA element) between the highly conserved AAUAAA motif and a downstream element rich in U or GU. Varying the cleavage site in transcripts is another potential factor in generating diverse isoforms. Such cross-talk between splicing and the synthesis of the poly-A tail has been characterized in the IgM's heavy chain gene and in the calcitonin gene.[33]

1.5.9 On intron retention

Of the classes of alternative splicing phenomena so far discussed, intron retention is perhaps the least well understood though it is well reviewed in the current literature.[34] While intron retention is perhaps the most common class of alternative splicing in plants[35] and in unicellular eukaryotes[36], it is the least frequent in metazoans.[37] (Cassette exons are the most common class in human beings.) Intron retention has been shown to play a role in gene expression regulation by relegation of its targets to nonsense-mediated decay (NMD).[38] A more nuanced mechanism links intron retention in presynaptic proteins to the presence of the polypyrimidine-tract binding protein Ptbp1 in mice.[39] (These transcripts are destroyed by the cell in a pathway different from NMD.) Ptbp1 is not expressed in neuronal cells (more on this in humans later) and precursors without the introns are matured to generate the requisite proteins in its absence. Not surprisingly the deregulation of intron retention has also been linked to cancer. Aberrant intron retention has been found in many cases of tumor-suppression inactivation in human beings.[40]

1.5.10 On alternative promoters

Alternative promoters provide different genomic positions for the initiation of transcription and is therefore quite distinct from the selection of different cassette exons for inclusion. These sites promote assembly of the transcription machinery and are regulated by the usual regulators of transcription: transcription factors[41], chromatin remodelers[42], histone marks[43] and DNA methylation[43] for example. Alternative start sites for transcription of the same loci have been well elucidated since the early 1980s when traditional biochemical methods investigated mutually exclusive starts of transcription in isoforms of the myosin light chain.[44]

Many genes have multiple promoters and each is subjected to its own set of transcription regulation factors and cofactors.[45] As for other cases of alternative isoform expression, events relevant to brain biology are well known. For instance, alternative promoters generate transcript diversity in neurexis, a family of protein that is thought to mediate trans-synaptic interactions.[46] Two alternative promoters give rise to a longer and shorter isoform that include up to five sets of splicing features creating hundreds of possible transcripts.

1.5.11 On the regulation of alternative splicing

This study is devoted to developing methods that identify events of developmentally regulated differential exon inclusion. While the methods are tissue agnostic, their development has been tested in a set of transcriptomes of the human brain. This organ has been selected because of the requirements of its strict development program and the long interest in the phenomena associated with this development. Alternative splicing in the human nervous system has been identified for many years and its elucidation has been pursued for as long[47–49]

One of the better studied cases of alternative splicing interactions in the mammalian brain is the set of interactions between PTBP1 and PTBP2 proteins. PTBP2 is thought to promote neuronal differentiation and its expression is limited, having been observed in neurons, myoblasts and spermatocytes[50]. On the other hand, PTBP1 is broadly expressed in many cell types but absent from neurons and muscle cells despite being abundant in their progenitors[50]. Recent dramatic evidence of the role of PTBP1 in repressing the neuronal differentiation program is the direct conversion of fibroblasts to neurons by removing PTBP1 from the cells[51]. The neuronal program is held in check, at least in part, through the omission of exon 10 in *PTBP2* transcripts through the activity of PTBP1. This leads to the nonsense-mediated decay of *PTBP2* transcripts.[52] The expression of *PTBP1*, in turn is repressed by the brain-specific microRNA miR-124, an indicator of neuronal differentiation. [53] A further indication of the importance of PTBP1 as a regulator of splicing events is the repression of a particular exon in *PBX1* transcripts. Experiments have shown that isoforms with the regulated exon lead to the expression of neurogenic genes[54].

A second form of regulation occurs via overlapping cis-elements that promote different splicing events. The expression of Srrm4, an SR-related protein has shown to be brain-specific in mice[55]. It binds to sequences rich in UGC between the poly-pyrimidine tract and the 3'ss of exons, likely antagonizing PTBP by overlapping nearby PTBP binding elements[56]. The preferential binding of the SRRM4 SR-related protein in place of its antagonist PTBP for example, promotes the excision of cassette exons from the *REST4* gene.[57] The resulting *REST4* transcripts code for a protein with 4 fewer zinc fingers (the reference isoform of Rest4 has 9 zinc fingers) reducing the transcription activity of the *REST4* gene[58].

1.6 An overview of large scale alternative splicing quantification through RNAseq

While the phenomenon of AS is not new, the advent of next-generation sequencing (NGS) technologies offers new tools to perform large scale analyses of their occurrence, identify a potentially large number of previously unknown events, and gain new insights into general underlying mechanisms. Among the NGS tools being developed is RNAseq, a technical approach to the reconstruction of the set of RNA transcripts (the cell's transcriptome). A full review of all the steps involved in the processing of RNAseq is beyond the scope of this paper but a short overview will be provided here. Many modern reviews exist for this purpose such as Conesa's 2016 paper.[59]

RNA is extracted from the samples considered by following an extraction protocol that, in addition to isolating the RNA from the remainder of the cell's content, deals with the problem of the overall abundance of ribosomal RNA (rRNA). rRNA is very abundant in the cell but in many cases it is not of singular interest. Therefore, it is often removed either by targeting rRNA directly (which requires high RNA integrity) or by selecting messenger RNA (mRNA) by their polyA tails (which results in a bias towards quantifying the 3' end of isoforms).

The NGS machinery fragments and produces millions of sequencing assays termed reads. These reads undergo a set of quality assurance filters and the remaining reads are assigned positions on a reference genome. The number of reads assigned to a specific region of the genome, after a normalization that accounts for the length of the region and library size, is assumed to be proportional to the level of expression of this genomic region. However, the problem of aligning a very large number of fragmented reads to a very large genome is a significant challenge.

In order to infer alternative splicing events, the ideal technology would simply sequence each mature RNA molecule from beginning to end and produce a report on each molecule thus evaluated. Human transcripts vary in length from *TTN*, expressed in isoforms that near 100kb to the much smaller micro RNAs of just about 20 nucleotides in length. Each edge case (in terms of length) present unique challenges. For the purposes of this work, the primary consideration is read length. While the technology that allows for longer read lengths is being deployed, there are substantial considerations such as increased error rate, reduced molecular throughput, sensitivity to molecular degradation and the requirements in terms of capital and time to overcome these obstacles. They have thus far been reserved for targeted analyses.

The current state of popular RNAseq technologies can generate reads on the order of 75-150bp often at both ends of the molecule being sequenced, leaving an un-sequenced region for each pair of reads that correspond to that middle part of the sequenced molecule which is unread. This is termed paired end sequencing.

This concept is introduced here because while most of the data considered in this work comes from an earlier technology in which reads were not paired, so-called single-ended reads, our validation dataset uses paired-end data. We discuss both in the framework of the quantification of isoform expression and their relative performance.

RNAseq data provides 2 types of information that assist in isoform inference. The first is simply a count of the reads that align along exon coordinates. The second type of data comes from a subset of the reads which span exon boundaries. In the case of both paired end and single ended sequencing, a sequenced fragment of RNA can begin in one exonic region and extend past the end of this exon, continuing into another exon that was spliced in by the splicing machinery discussed previously. These reads that span exon boundaries (in both single-ended and paired-ended data) will be referred to here as splice junction reads. They provide direct evidence of exon inclusion/skipping.

Methods applied to RNAseq technology that leverage these sets of reads in order to identify exon expression can be grouped into two broad categories. The first attempts to infer full-length transcripts from short(er) sequences that are generated from the NGS platforms. This allows for a quantification of isoforms that is prior to and central to any further analyses. The second category of exon expression analysis does not offer informative estimates about full-length isoform populations but instead reports differential exon use for specific genomic coordinates across a categorical variable through some variance analysis (such as ANOVA). Refraining from isoform reconstruction restricts the cumulative burden of uncertainty present in the first category but does not offer information on reconstructed transcripts. In this section, we will review both categories and provide some details around the lead tool in each category: cufflinks and dexseq.

1.7 An Overview of isoform construction: cufflinks and related tools

Many tools have been developed that generate and quantify transcripts from reads generated by NGS technologies[59].These tools can be grouped into two broad categories depending on the existence of a step of isoform reconstitution. Category 1 regroups tools that estimate isoform expression first, then computes the differential expression of the estimated isoforms as a second step. Perhaps the most commonly used is the CuffDiff2[60–62] algorithm (cited over 7000 times as of 2016) which quantifies isoforms using Cufflinks,[62] prior to computing the difference between isoform expression in the samples. Finally, a new method: rSeqDiff has been proposed that uses a hierarchical likelihood ratio test to detect differential isoform expression simultaneously with differential gene expression.[63] This first category of tools is limited by the challenge of obtaining accurate information at the isoform level from short-read sequencing.[64] Category 2 omits the reconstruction of isoforms all-together and compares the distribution of reads to exonic features instead. DEXseq[65] and DSGSeq[66] aim to identify significant differences in read counts between samples using only exons and junctions. rMats[67] compares exon inclusion levels defined by junction reads. Both Category 1 and 2 tools depend on the mapping of NGS reads to a reference genome by an aligner such as TopHat[68] or STAR[69], while only Category 1 tools depend on an initial step of isoform quantification.

Isoform quantification, that initial step on which Category 1 tools depend, is addressed by many tools: MISO[70], RSEM[71], eXpress[72] and Sailfish[73] among others. Of these, the precocious and often-updated Cufflinks from the Trapnell lab[62] is perhaps the most widely used. Cufflinks generates a bipartite graph of aligned reads weighted by the frequency of the fragments that are complementary to specific isoforms. Fragments are matched to the maximal number of possible isoforms and the count of isoforms identified for a particular transcriptome is the minimum set of transcripts that accounts for all fragments. Then, given the distribution of the reads, the abundance of each isoform is computed. A likelihood function is maximized under the assumption all reads are generated in a uniform manner, dependent only on the length of the sequence and their availability.

Computational requirements notwithstanding, comparison studies of the isoform quantification step on which Category 1 tools depend, show that the variance in the precision and sensitivity of these tools lie in the in precursor steps to the analyses (the RNA separation protocols discussed in section 1.6 for instance) rather than in the tools themselves and in the assumption of the completeness of the annotation set provided[74]. Current reviews of these methods show that while they achieve comparable precision and accuracy rates[71,74–76], there are substantial gaps in the overlap when it comes to sets of predictions made by each tool. In these third party studies, cufflinks performs well relative to other tools in the same problem space (that of isoform identification with annotation set). If we define precision as the ratio of true positives to the sum of true positives and false negatives and we define sensitivity as the ratio of true positives to the sum of true positives and false positives as Angelini does in her review of these tools[74], then cufflinks achieves precision rates of 70-90% in 100bp paired-ended read sets. Its sensitivity to known isoform quantification is poor (although still better than competitors), ranging between 40-50% ratio in 100bp paired-ended read sets. Unfortunately, the performance of these tools in datasets that older, single ended technology is fails to outperform any meaningful tolerance threshold. The high level of false positives generated by these methods keeps the sensitivity below 40%[74]. Precision is similar to that in data sets generated from paired end reads.

As previously discussed Cufflinks performs well in identifying known isoforms although it tends to create a great deal of false positives. However, based on its methodology some biases are readily discoverable. Given two transcripts of different lengths, cufflinks will assign a greater likelihood to the shorter transcript. This is compounded by the known 3' bias of read generation methods that purify RNA by using the polyA tail for extraction. The cuffdiff paper itself notes the unsuitability of isoform creation for moderately expressed genes when the coverage is less than 40 million reads.[62] (Less than 40% of transcripts are recovered in the author's own results for genes expressed between 4 and 7.5 FPKM.)

1.8 On modelling individual exon expression: Dexseq

Unlike cufflinks and related tools, Dexseq[65] claims to avoid the unmeasured accumulation of uncertainties involved in the isoform inference methods discussed above and to detect differential exon use across discrete categories with high sensitivity. Essentially, Dexseq process compares the fits of a generalized linear model (GLM) to a reduced form and calculates a p-value for a model coefficient that corresponds to the distribution of reads at an exon.

Dexseq fits the following linear model to each exon: $\log\mu_{ijl} = \beta_i^G + \beta_{il}^E + \beta_{ip}^C + \beta_{ipl}^{EC}$ where *l* is the exon index, *i* the gene index, *j* is the sample index and *p* is the discrete condition category of sample *j.* $\mu_{ijl}$is the estimated count of reads linked to the estimated fragment counts via a parameter that accounts for exon length. The authors model the likelihood of fragment counts sequenced (and thus build a likelihood distribution for μ_{ijl}) by using the negative binomial distribution.

The authors justify the use of the negative binomial distribution as the probability distribution for counts of reads overlapping exonic coordinates by citing the arguments the works of Lu et al[77]. and Robinson and Smyth[78]. The argument in the Dexseq supplementary papers[65] begins by extending the Poisson distribution. Assuming a concentration of cDNA fragments to be sequenced, the probability of the sequencing of any given fragment is small and independent from the sequencing of other fragments, and depends only on the total number of fragments to be sequenced in solution. (One might quibble with asserting the second point across all library preparation methods and existing sequencing technologies.)

Factors linked to detection efficiency, such as GC content, feature length and secondary structure are not considered by the model as they are determined not to depend on sample under study. (Once again, one might ask if secondary structure can be shown to be independent of all sample categories, including, different alkaline conditions for example.) Starting from the equivalency of the variance and the mean of a Poisson distribution, the authors establish the negative binomial mean and variance relationship ($v = \tilde{\mu} + \alpha\tilde{\mu}^2$). Here *v* is the variance, *μ* is the estimated mean and *α* is the dispersion parameter. The authors then show that the dispersion of proposed NB distribution is proportional to the Poisson distribution within a asymptotic value thus validating their chosen distribution.

For each exon *i*, The above model $\log \mu_{ijl} = \beta_i^G + \beta_{il}^E + \beta_{ip}^C + \beta_{ipl}^{EC}$ is compared with the reduced model $\log \mu_{ijl} = \beta_i^G + \beta_{il}^E + \beta_{ip}^C$ and a likelihood ratio test is evaluated by means of a χ^2 distribution. Multiple covariates can be considered in which case the null hypothesis is rejected if any of the conditions influence exon read counts.

As can be inferred in the discussion above, Dexseq contends with a few challenges. Foremost, for our purposes is the limitation on the discrete binning of samples. Each categorization of samples requires an expansion of the model being applied with the usual negative consequences for statistical power, which must account for overfitting, and complexity (a new parameter is added to the model for each new bin). In addition, it is not specified in their presentation of the proposed algorithm that terms are included for interactions between specified conditions. With respect to this question both options are unfavorable. The inclusion of interaction terms generates more capacity in an increasingly complex the model but their removal may overlook interactions and over-estimate the significance of the differential exon term.

1.9 A review of trajectory representations of gene expression

We have so far discussed the quantification of alternative splicing events. Several analytical approaches exist to analyze expression changes over time, in particular for microarray transcriptomic data and at the whole gene level. These studies are relevant as this work will mostly be concerned with modeling trajectories of exon expression and to use these trajectories as an alternative means of inquiry to the approaches described above.

Most commonly, time-series analyses seek to cluster co-expressed genes along a temporal coordinate. Popular methods aggregate temporal trajectories through auto-regressive equations[79], hidden Markov models[80], dynamic Bayesian networks[81] and translational matrices[82] (singular value decomposition). Otherwise, efforts have been made to model individual features such as Chen's integration of transcription and translation kinetics into differential equations[83]. However, genomic datasets based on next-generation sequencing (NGS) data present characteristics that differ from those specific to microarray technology. NGS datasets have distinct biases and underlying error models; there are generally fewer samples due to cost, and the technical complexities and biases of reads and their alignment are not necessarily taken into account by microarray-based methods. The widely used Bioconductor R package which implements Dexseq detailed above[65] is currently used to address this vacancy, however it performs only static pairwise comparisons.

Here we end our discussion of the biological occurrence of alternative splicing and the tools currently available for its quantification. We have reviewed the prevalence of alternative splicing and some of its particularities including the special case of alternative promoters. We have discussed the special case of its regulation in the human brain. We then reviewed some of the contemporary means of assessing the content of RNA in biological samples and how this information may be used to examine differences between samples, either through the full reconstitution of isoforms or in variance analyses of exons. We will now review the goals and hypotheses of this study where we will discuss a new approach to examining differential exon expression overtime.

2. Introduction

2.1 Goal

The overall goal of our work is to identify and characterize previously unknown exons that are functionally important to the development of the human brain, while the specific aim of this thesis is to identify exons whose inclusion in their respective gene-product is temporally regulated and restricted to a developmental stage of the brain.

As mentioned in the introduction (in section 1.9), characterizing the temporal nature of gene expression has been useful in identifying interactions between gene-products[79,83] and defining networks of such interactions[81] and has led to the discovery and elucidations of the underlying biological mechanisms[79–81,83]. Here we aim to extend this avenue of research to characterize the temporal nature of exon inclusion within gene products.

To obtain this goal, we propose to develop a framework in which exon expression levels in the brain are modelled as a parameterized function of the brain's age. In applying these methods to a large-scale dataset of exon expression in the brain, we aim to reduce complex and high-dimensional expression data to a small set of biologically meaningful parameters. We propose to leverage dissimilarity between exon models in shared loci in order to identify exons whose expression is limited to particular developmental phases of the brain's development.

2.2 Hypothesis

The hypothesis of this report is threefold. We hypothesize that there exists a large number of cases of developmentally regulated exons in the brain that are currently unknown. We further hypothesize that the current sets of developmental transcriptomes studied over the course of this work have a sufficient sampling size and density as to produce results that are accurate and reproducible. We further hypothesize that the technology used to assay these samples has the requisite resolution to detect these changes in exon expression given the samples examined.

2.3 Data

This study leverages RNAseq data from three different datasets. We use one publicly available dataset: Brainspan, described below, to generate exon trajectories and the associated parameter estimates from which predictions are later made with regards to trajectory dissimilarity and developmentally regulated exons. We amalgamate samples from two other independent datasets into a combined validation dataset to test our parameterization and our predicted differential exon inclusions. Findings were further confirmed in mice samples.

2.3.1 The Brainspan Dataset

The Brainspan dataset[84] is public resource for the study of transcriptional activity in the human brain developed by a consortium which includes the Allen Institute for Brain Science, Yale University, The University of California, Los Angeles and the University of Texas Southwestern Medical Center, among others. The dataset provides expression counts for a large number of exons from a relatively large number of samples (524) and donors (42), that provide a well-sampled coverage over the development of the human brain from the fetal period into adulthood.

Brainspan quantifies expression levels for 309,223 exon features and 52,376 genes as annotated by Ensembl release 65[85]. The expression is normalized to account for both the length of the feature expressed and the number of reads in the overall assay. The unit of expression is reads per kilobase per million reads (RPKM). These samples spanned donors and brain substructures. Donors ranged in age from 8 post-conception weeks to 40 years of age.

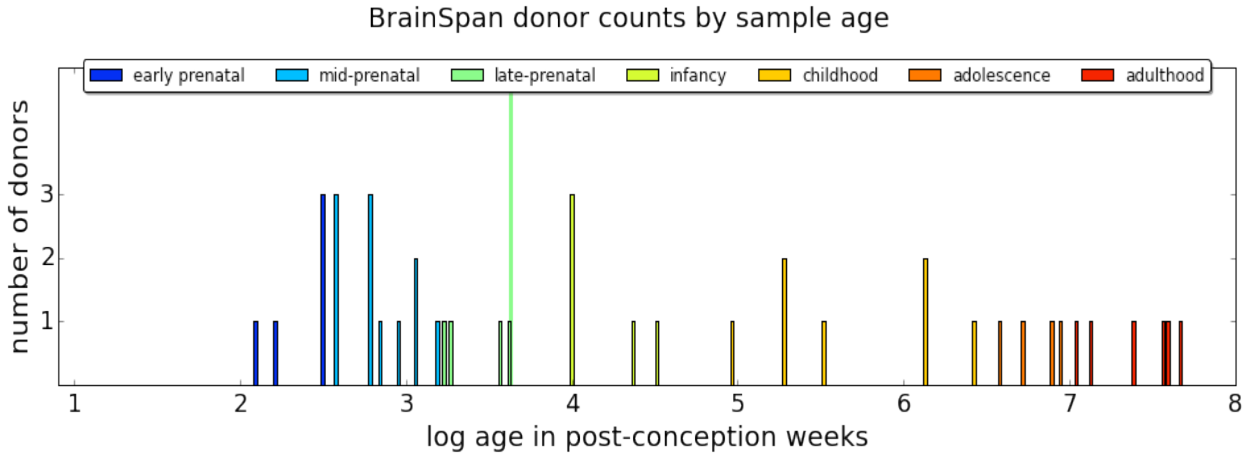

Figure 3 Brainspan donor distribution

Distribution of donors across the lifespan of human beings (age represented in log scale). The earliest samples are 8 post conception weeks (pcw). The most aged samples are 40 years of age. The light green line indicates birth. Early prenatal samples are 8-12 pcw. Mid-prenatal samples are 13-25 pcw. Late prenatal samples are 25-38 pcw. Infancy samples are of 0 to 18 months old. Childhood samples are 19 months to 11 years old. Adolescent samples are 12-19 years old. Adult samples are 20 to 60 years old.

In almost all cases, several transcriptomes were generated from each donor: one per brain substructure sampled. The number and regions of the brain sampled and assayed by donor varied based on the developmental stage of the donor and the integrity of the organ.

The Brainspan consortium group established guidelines on sample selection that exclude samples with insults to the human brain, visible or otherwise. Samples were excluded on evidence of large-scale chromosomal abnormalities detected by karyogram and/or Illumina Human Omni-2.5. Samples were excluded if drug or alcohol use by the mother was reported during pregnancy, if malformations or lesions were observed, or on positive tests of Hepatitis B, C or HIV. Samples were further excluded if drug or alcohol abuse was reported and on knowledge of neurological or psychiatric disorders, on the ingestion of neurotoxic substances, death by suicide, severe head injury, brain lesions, stroke or other signs of neural abnormalities or neurodegeneration. Dissections were documented and all dissections were performed by a single individual and reviewed for the above as well as tumors, infections demyelination and metabolic disease by a small team of pathologists.

mRNA library preparation was performed using a polyA approach and sequenced on Illumina's Genome Analyzer II generating 76-bp single-end reads. The Brainspan group use the RSEQtools[86] framework to align generated reads to the GRCh37 genome. Reads were filtered using a Phred-like quality threshold of B. Tophat[68] (version 1.3.1) was used to align reads. SAM tools[87] and mrfQuantifier[86] were used to quantify exon expression using Brainspan's composite model for exons.

Brainspan uses a composite model to summarize read counts for each exon, without resolving alternative, overlapping isoforms. One isoform may only include 100bp of an exon and excise a further 100bp that are retained in other isoforms, as seen in the following figure from The Brainspan Consortium's Technical White Paper.

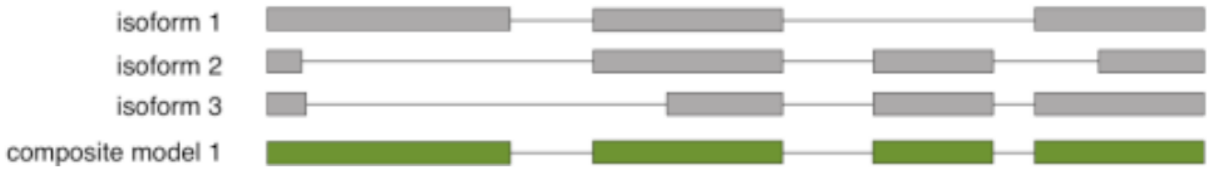

Figure 4 Collapsing Brainspan exon annotation

Theoretical set of three isoforms containing 4 exons collapsed to Brainspan style annotation. Taken from The Brainspan Consortium's Technical White Paper: Transcriptome Profiling by RNA Sequencing and Exon Microarray

This collapsing of exon genomic coordinates has implications for our inference and validation procedures that will be outlined in the section on predicting differential exon inclusion.

2.3.2 Validation Dataset

We assembled a second dataset in order to validate predictions made in analyses originating from the Brainspan dataset. The second dataset combines human brain samples from the laboratory of Dr. Nada Jabado at McGill and Montreal Children's Hospital with a second set of samples from the laboratory of Dr. Gustavo Turecki from McGill and The Douglas Mental Health University Institute. This second dataset will be referred to, in this paper, as the validation dataset.

Dr. Jabado provided transcriptomic data from 13 fetal brain samples from an indeterminate brain substructure. Twelve were aged from 6 to 18 post-conception weeks. (1 sample was of unknown age; it was not used in these analyses.) Dr. Turecki provided transcriptome data from 4 anterior cingulate cortex samples. These samples ranged in age from 15 years of age to 20 years of age.

Samples from both laboratories were prepared and sequenced by the Quebec Genome Center using the same protocols for library preparation, sequencing and bioinformatic analyses. The Quebec Genome Center uses Illumina's Ribo-Zero rRNA removal kit for ribosomal depletion and generates 100 base-pair, paired-end reads on Illumina's Hiseq 2000 platform. We received raw sequencing data for all validation samples and were thus able to control the analysis in all downstream processing steps.

Resulting read sets were trimmed for quality using a Phred33 score of 30 using a sliding window average. Reads under the required length of 30-bp (base-pair) meeting the quality score were discarded. Read trimming was done using Trimmomatic[88] (v0.32). Known adaptors and Illumina-specific sequences were removed from reads in palindrome mode. Reads thus processed were aligned using STAR[69] (v2.3.0e) to the UCSC hg19 genome.

We quantified expression levels for the annotation set used by the Brainspan group (above). Read counts were estimated by counting reads with featureCounts[89] (v1.4.4) using default settings. The minimum mapping quality score for a read to be counted is 3, and only primary alignments are counted. RPKMs are also calculated for each feature by dividing the calculated read count by the sample total read count and by the cumulative genomic length of each feature.

3. Results

3.1 Overview

In order to identify exons whose expression in the brain were restricted to a developmental period, we applied a modeling framework to a large scale dataset of brain transcriptomes. We sought to summarize exon expression over the development of the human brain. We fit curves that modelled the expression of exons in the brain as a function of brain age using the Brainspan dataset described in the previous section. We refer to these curves that summarize exon/gene expression as a function of brain age as exon/gene trajectories.

We tested parameterized models with 1. parameters interpretable in ways meaningful to biologists, 2. a small parameter set to avoid over-fitting the data, and 3. whose flexibility could accommodate unimodal peaks in expression. After testing several models, we settled on a function that resembles the Gaussian probability distribution function (in its shape and its flexibility). We generated trajectories using the Gaussian model for all exons and genes in the Brainspan dataset.

We then compared methods that establish a measure of distance between a gene's shared exons by taking into account the parameters that describe the exonic trajectories generated by the model. Each measure of distance considered suggests a set of exons which are dissimilar from the remainder of the exons that constituted its gene-product. In evaluating these methods to compare exon trajectories, within a shared gene, and to generate predictions about exons present only at particular stages of development, we leverage a second dataset. We tested that these exons are expressed only in particular developmental intervals using a novel validated dataset and, for a limited subset of exons, qtRT-PCR experiments among orthologues in mice. Figure 5 shows an overview of our methods.

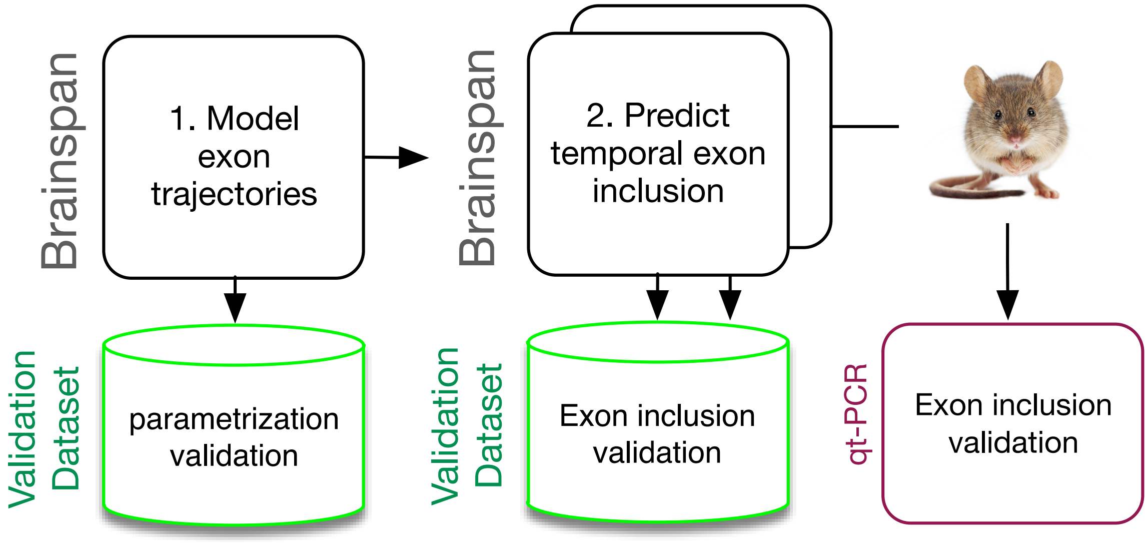
Brainspan
1. Model exon trajectories
Brainspan
2. Predict temporal exon inclusion
Validation Dataset
parametrization validation
Validation Dataset
Exon inclusion validation
qt-PCR
Exon inclusion validation

Figure 5 Overview of methods

Exon trajectories and predictions are made using the Brainspan dataset. The parameterization of the models and validation of differential exon inclusion is done in the validation dataset.

3.2 Modeling exon expression trajectories

We sought to summarize exon expression as curves, modelling expression (y-axis values) as a function of time or development over the lifespan (x-axis values). In particular, we define *developmentally regulated* trajectories as smooth functions of few parameters, with a single local maximum (i.e. a single expression peak). Reducing the expression data to curves that are described by a small set of parameters allows for comparison of the parameters that describe the exon expression patterns instead of the high-dimension expression data (524 samples per feature).

We sought to select a model in such a way as to reflect biologically relevant information in the values of the parameters. In this way, in knowing the values of the parameters that described the modelled pattern of expression of an exon, one could make biological inferences on the exon itself: such as its expression level for a particular point in the donor's lifespan.

We considered three families of well-known functions that fit these requirements: the exponential, the gamma family of curves and the Gaussian curve. All modeling strategies are of the form $y = f(x)$ where y is the expression of the feature and x is weeks since conception in a log scale. The justification for the log scale is simply to give a higher resolution to events at the beginning of the developmental time series. It is well known that the development of the brain undergoes largescale structuring prior to the maturation of the adolescent[90,91]. Neuronal proliferation and neuron migration in particular occur during gestation[91]. We utilize a log transform to ensure visibility of these early events.

The following discussion applies also to models of gene expression that were constructed in parallel to the exon models. However, our primary interest remains the characterization of exon expression.

3.2.1 The exponential model

The exponential family of curves takes the form

$$y = Ae^{cx} \qquad \text{Equation 1}$$

Where A is a scaling parameter and c serves the dual function of controlling the rate of descent from peak expression while the sign of c indicates the window of peak expression. This formulation limits the number of parameters estimated to 2. The first (A) is required to capture the variance in gene expression. The sign of the second (c) is useful in that it positions the peak expression of an exon in one of two developmental windows. Positive values of c indicate that the peak expression of an exon occurs at conception. Negative values of c indicate that the peak expression of an exon occurs at the end of a human being's lifespan. Figure 6 demonstrates the flexibility of the exponential curve for modelling exon trajectories.

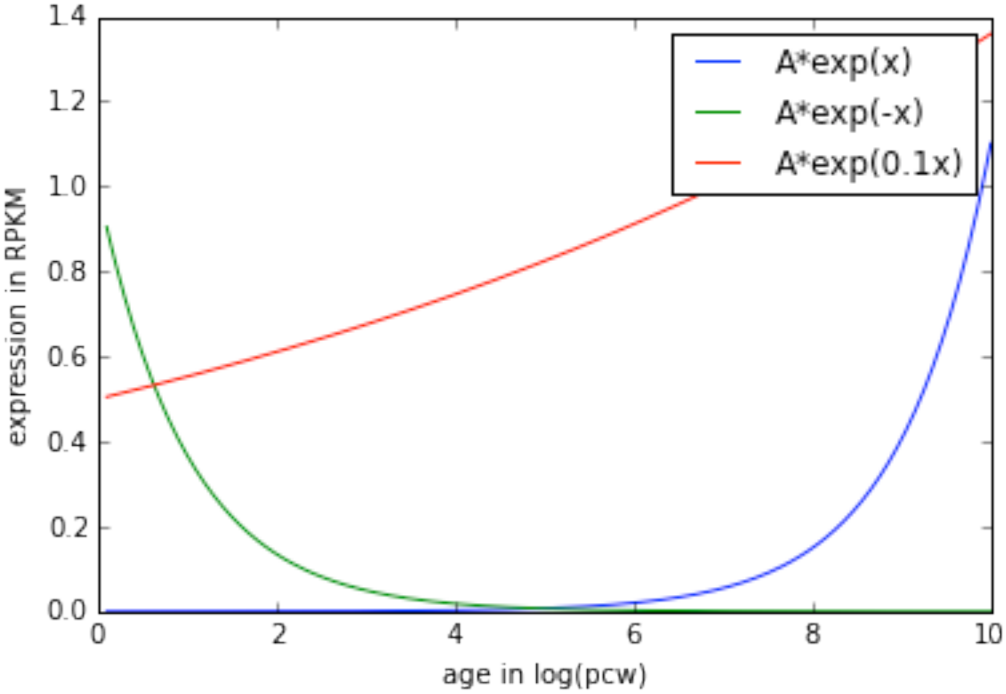

Figure 6 Flexibility of the exponential model

Each curve is modelled for different values of c in Equation 1. A is chosen to so as to plot similar peak expression values. The green and blue curves show the model's capacity to model monotonically increasing or decreasing expression trajectories. The red curve shows that values of c can influence the rate of change of expression.

The limitations of the model are apparent on inspection. Exons which reach a peak expression at an intermediate developmental stage will not be adequately captured by the model. This model was applied to all features: exon and genes (more details on the fitting process later) but was ultimately abandoned in favor of a more flexible model once we observed the frequency of features with a peak expression that did not lie at either extreme of the lifespan (conception or adulthood).

3.2.2 The gamma model

The gamma family of curves resembles the probability density function of the same name. However, only the flexibility of the curve of the non-cumulative probability distribution is of interest here. None of the characteristics of such a probability distribution function are conserved (or even desired). What is of interest is the curvature property which is a function of two parameters: the gamma shape (k) and the gamma scale (t). The function of the proposed model is put forth in Equation 2.

$$y = Ax^{k-1}\, e^{-\frac{x}{t}} \qquad \text{Equation 2}$$

A is a scaling parameter which accounts for the y unit scale. k is the gamma shape parameter, it affects the overall shape of the trajectory; t is the gamma scale parameter, it stretches out the curve along the age coordinate (x axis) as seen in Figure 7.

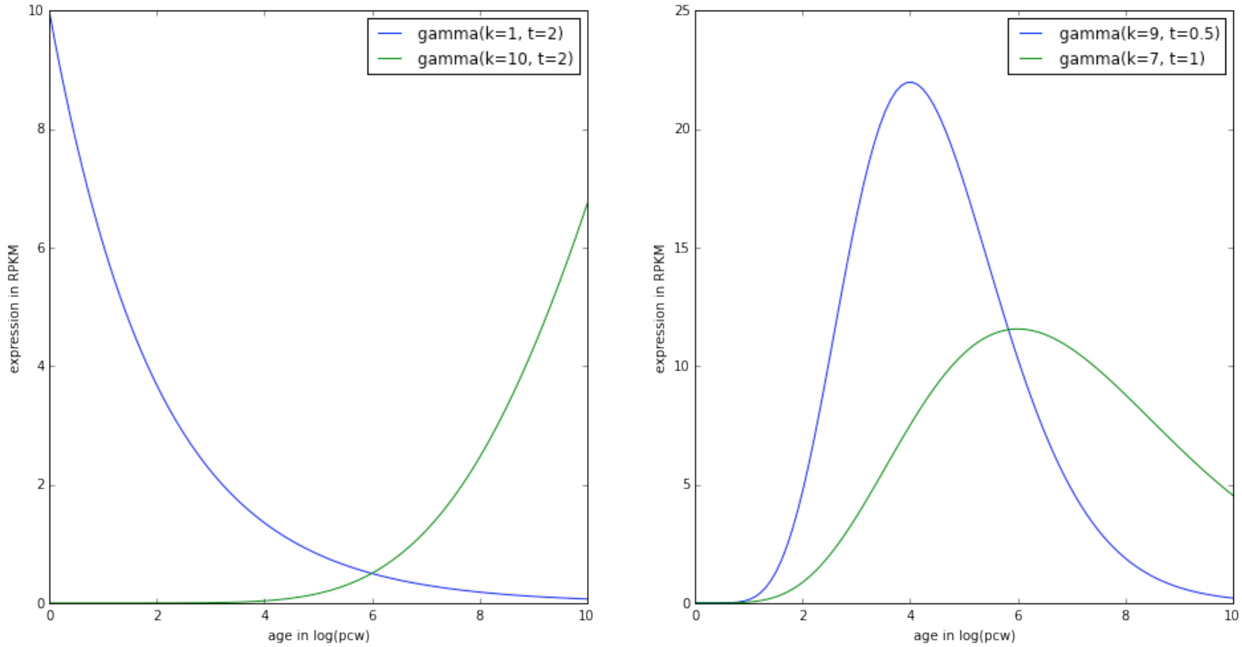

Figure 7 Flexibility of the gamma function.

A. The curve readily adapts to possible instances of peak feature expression at conception or in adulthood as in the case of the basic exponential function. B. Moreover, the gamma function adapts to peak expression at some intermediate developmental stage as shown. As can be seen, the gamma function allows for skew. The green curve shown has a higher rate of change when approaching peak expression than its rate of descent down from peak expression into adulthood

The gamma curves require 1 more parameter than the exponential model. It is thus more flexible and can accommodate more trajectories. Figure 3.2 makes the case for the adoption of the gamma function over the exponential function as the additional capacity afforded does capture biologically relevant cases. Genomic features such as genes are known to be highly expressed in developmental phases that are not simply extremities of the lifespan: conception or adulthood.

Nonetheless, the gamma distribution poses a challenge in the interpretation of its parameters. While the gamma scaling parameter can be readily communicated (it stretches the curve along the x-axis) the impact of the gamma shape parameter (k) is not as simple. One has to review a few cases to grasp the rightward migration of the function's peak and the corresponding skew.

A similar model to the gamma distribution was further tested: the so called Weibull distribution. The Weibull curve family shares the characteristics of the gamma curves and is so similar in fact that its discussion is redundant.

3.2.3 The Gaussian model

Like the gamma model, the Gaussian model is inspired by the shape of the eponymous probability distribution. As in the application of the gamma curves, only the shape of the curves are conserved in the model, properties that are attributed to the non-cumulative probability distribution are not relevant to this study. The Gaussian curves are described by Equation 3.

$$y_j = A_j exp\left(\frac{-(x - \mu_j)^2}{{\sigma_j}^2}\right) \qquad \text{(Equation 3)}$$

A_j is the estimated scaling parameter, μ_j is the estimated peak expression

and σ_j scales with the width of the peak expression.

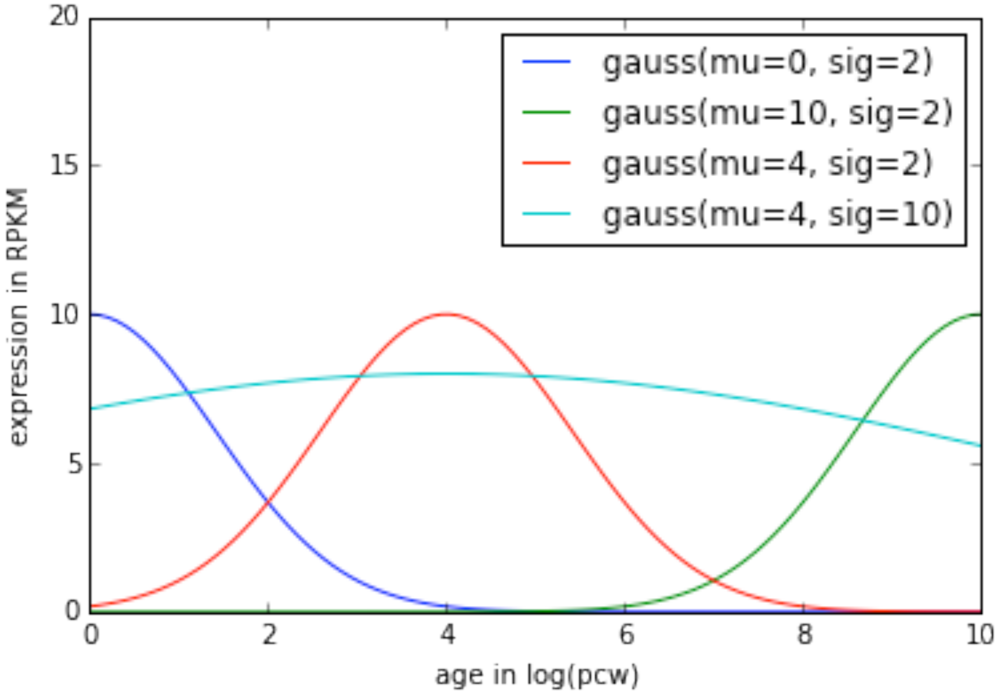

Figure 8 Flexibility of the Gaussian function

The Gaussian curves adapts to unimodal expression patterns. The blue and red curves model fetal and adult trajectories respectively. The red and teal curves peak midway through the developmental coordinate (log-scaled). The teal and red curves show the capacity of the Gaussian curves to adapt to different rates of change in expression from a pronounced peak (red) to a negligible difference in expression (teal) perhaps modelling a non-developmentally regulated trajectory.

Like the Weibull and gamma curves, the Gaussian curves restrict the cardinality of the parameter set to 3 parameters. Unlike the former two, the Gaussian parameters are inherently interpretive. They allow for a straightforward biological interpretation of the trajectories plotted. $\hat{\mu}_j$ corresponds to the age of peak expression of the j feature modelled and $\hat{\sigma}_j$ relates to the duration of the period of elevated expression of the *j*th feature. For this reason, we applied the Gaussian model to features in the Brainspan dataset and this representation of those time series is used in the remainder of the analysis.

In choosing to apply the Gaussian model to the data we have deliberately left out one of the advantages of the Weibull and gamma curves. The Gaussian model does not allow for any skew of the trajectories. The nth derivatives (rates of changes) that shape the curve up to the peak of expression are equivalent (in magnitude but not in sign) to the derivatives that follow the peak. We considered this tradeoff to be acceptable given that the skew of the trajectories is challenging to measure accurately given the noise of the dataset.

3.2.4 Parameter estimation

We created a Gaussian model framework by fitting Equation 3 to the expression data of each exon in the Brainspan dataset. Several non-linear parameter estimation techniques were attempted within a python 2.10 environment[92–94] using the well-known Scipy[95] library.

3.2.5 Objective function

We created an objective function L defined as the sum of the residuals across all samples and all donors for which the feature was quantified in the Brainspan dataset. L is defined in Equation 4.

$$L_j = \sum_{i=1}^{n_d} \sum_{k=1}^{n_i} \left(y_{jk}\text{-}\hat{y}_{jk}\right)^2 \qquad \text{Equation 4}$$

Where y_{jk} is the observed expression of the *j*th exon in the *k*th sample for donor *i*. $\hat{y}_{jk}$ is the value estimated by the Gaussian model (Equation 3) for the given parameter set.

3.2.6 Parameter initialization

All methods of parameter estimation implemented converge on local minima only and therefore are sensitive to parameter initialization. We estimated parameters for exon trajectories and gene trajectories in the same fashion. *A* was initialized to midway between the maximum and minimal expression reached by the feature across all donors and samples. (If expression was constant, A was assigned the mode). μ the median age of all values greater than A (or the age of the maximum expression if only a single value was greater than A). Like μ, σ was initialized based on the number of expression values greater than A. If there were more than 2, sigma was assigned half the spread of these values. If there was a single value, sigma was assigned a sixth of the spread of all values.

3.3 Optimization strategies

We tested several optimization strategies, beginning with unbounded parameters (no restrictions on values estimated) and later deciding on a strategy that bound parameters to values that did not greatly exceed the biological thresholds on life expectancy.

We first attempted the Scipy implementation of the common Levenberg-Marquardt (LM) algorithm to minimize the sum of the residuals objective function. Like all other methods attempted, LM converges only on local minima and does not support bounds on parameters although bounds can be enforced by changing the loss function so as to strongly penalize values that exceed the desired thresholds.

Convergence of the LM estimates was lackluster in that just over a thousand features did not converge in the computational window allotted: 8 hours on 20 i7 cores with 8 threads each. Additionally, the optimization method generated many estimates for μ that were either negative or greater than 10. As shown in Figure 8, negative values for μ indicate a monotonic decrease from peak feature expression at conception (blue line). Values for μ greater than 10 indicate a net monotonic increase in expression from conception to a peak expression in adulthood (red line). We therefore opted to set bounds on μ as the heterogeneity in the μ estimates could be adapted in the curve fitting by appropriately modifying σ and *A.* Rather than modifying the objective function, we utilized the Scipy implementation of Coleman and Li's Trust Region Reflective (TRR) algorithm[96] to estimate parameters for the feature sets.

3.4 Optimization results

We fitted the 309,223 annotated exons and 52,376 gene features using the Scipy implementation of the TRR algorithm. 309,152 exonic models and 52,369 genic models converged onto local minima within the same computational constraints. μ and σ are constrained to [0, 10] values, *A* is constrained to positive values.

Resulting parameter estimated fall into relatively well defined populations in both sets of features: genes and exons.

In order to verify that our modeling strategy was adequately capturing developmental expression patterns, we randomly selected sets of exon expression and the trajectories generated by our models for visualization. A subset of features randomly selected from the different parameter populations obtained after fitting are presented in Figures 9, 10 and 11. These support the abstraction of the data into the μ and σ parameters. As can be seen in these figures, the parameters estimated fall into relatively well defined populations in both sets of features: genes and exons.

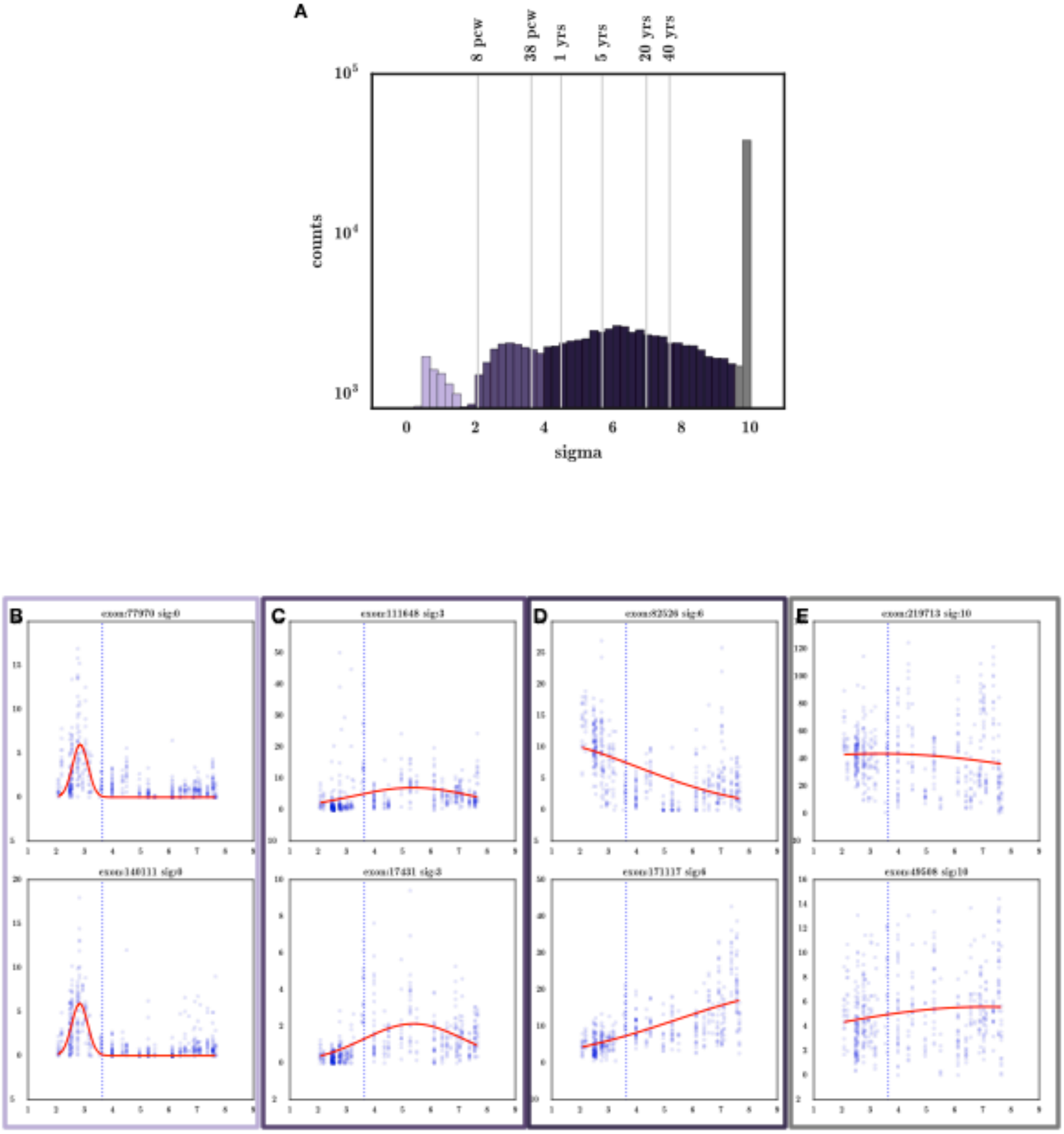

Figure 9 Randomly selected exon expression patterns by σ

The population of σ parameters estimated from the data fall into 4 populations with peaks at 0, 3, 6.5 and 10, the last peak consisting of exons with non-developmental trajectories (i.e. exons whose expression does not vary as a function of developmental phase). ***2B*** *Randomly selected trajectories with values of σ near 0 (random seed of 2016)* ***C*** *Randomly selected trajectories with values of σ near 3.* ***D*** *Randomly selected trajectories with values of σ near 6.* ***E****. Randomly selected trajectories with values of σ near 10.*

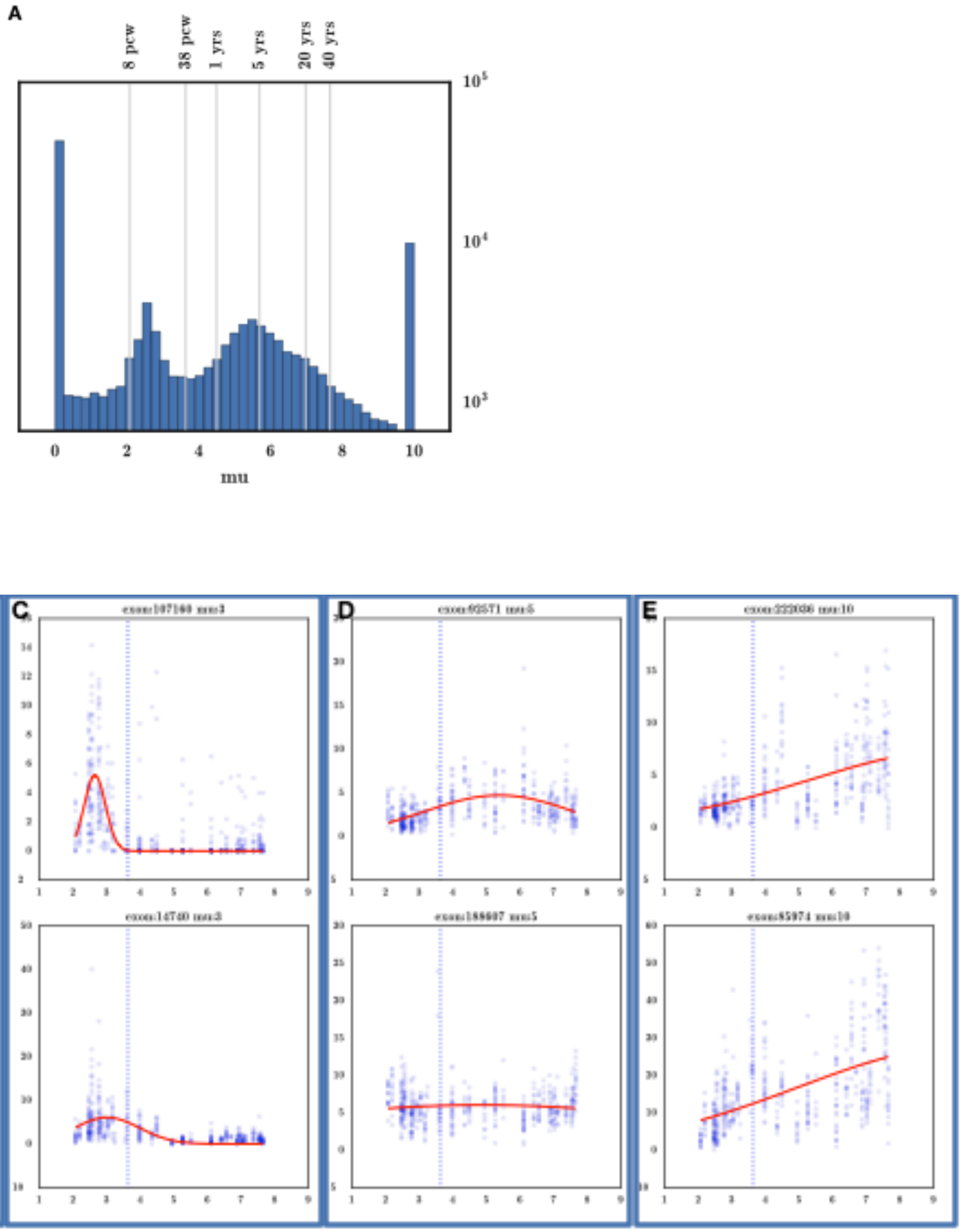

Figure 10 Randomly selected exon expression patterns by μ

*The μ (peak expression) values show a quad-modal distribution. One set of exons peak and decline in expression from conception, a second set peak before birth, a third nearing 5 years of age and a final set see peak expression in adulthood. 0.1**B** Randomly selected trajectories (random seed of 2016) with a μ at conception. **3C** Randomly selected exons with a μ near 3. **3D** Randomly selected exon trajectories with a μ near 5. **3D** Randomly selected exon trajectories with a μ near 10.*

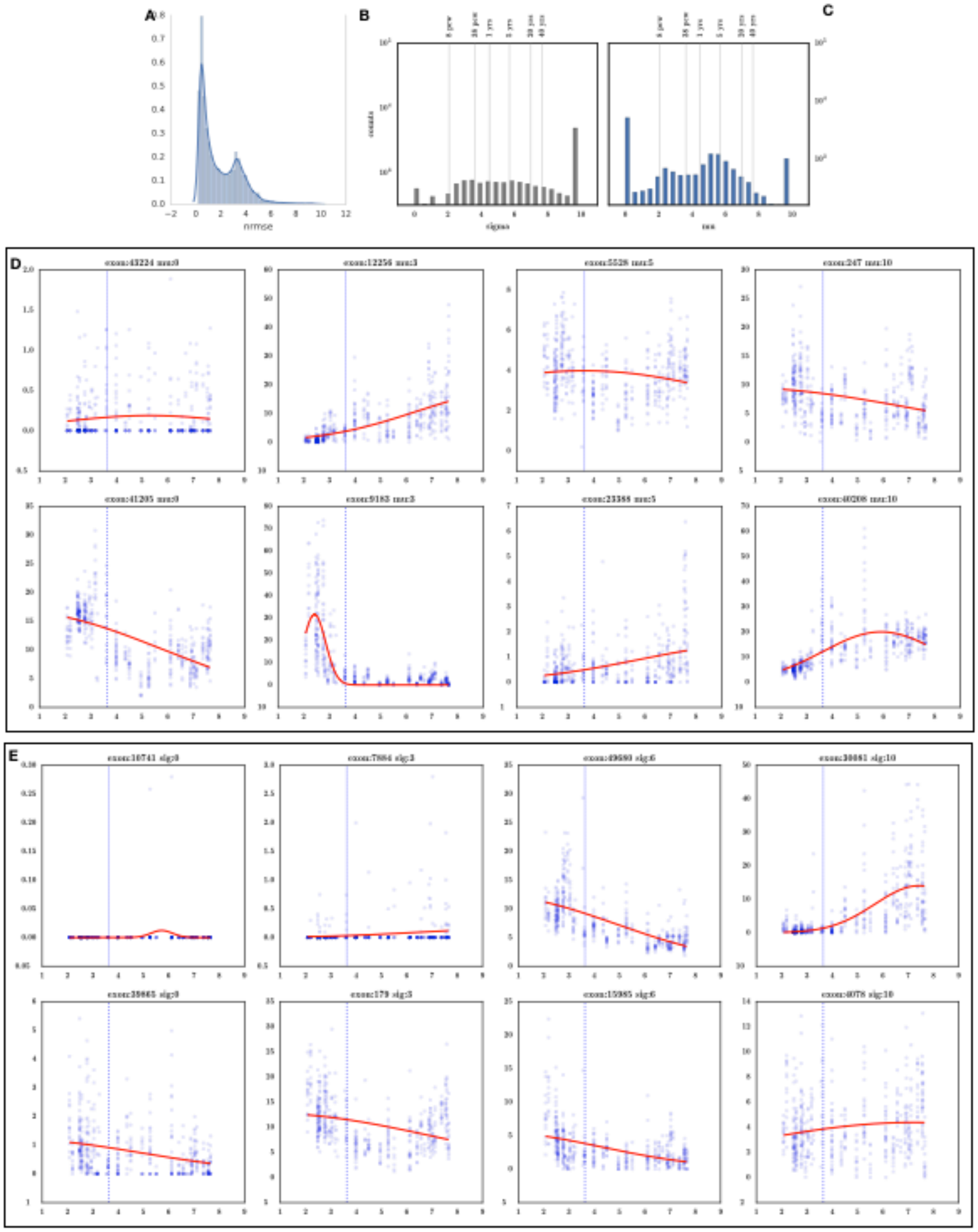

Figure 11 Randomly selected gene trajectories by σ and μ

*Distribution of the goodness of fit metric for gene trajectories. **2B** Distribution of estimated σ parameters for gene trajectories. **2C** Distribution of estimated μ parameters for gene trajectories. **2D** Random gene trajectories for increasing values of σ. **2E** Random gene trajectories for increasing values of μ.*

3.5 On the assessment of the goodness of fit

In order to assess the robustness of the parameter estimation, we used donor information to perform a leave-*p*-out cross-validation. The size (42 donors) and design of the Brainspan dataset (1-3 donors per time point, with samples originating from the same donor being highly correlated) precludes the use of a standard X-fold cross-validation procedure. Instead, leave-*p*-out calculations, for each exon, were performed as follows. In each iteration, all samples from one donor (*p* samples) were removed from the dataset, parameters were estimated using the remaining data, and the root mean standard error (RMSE) for the omitted observations was computed. The procedure was repeated for all donors, one at a time. The average RMSE obtained, normalized by the mean expression of the exon, constitutes the final leave-*p*-out score: NRMSE. Figure 3.7 shows the distribution of NRMSEs obtained for all exons.

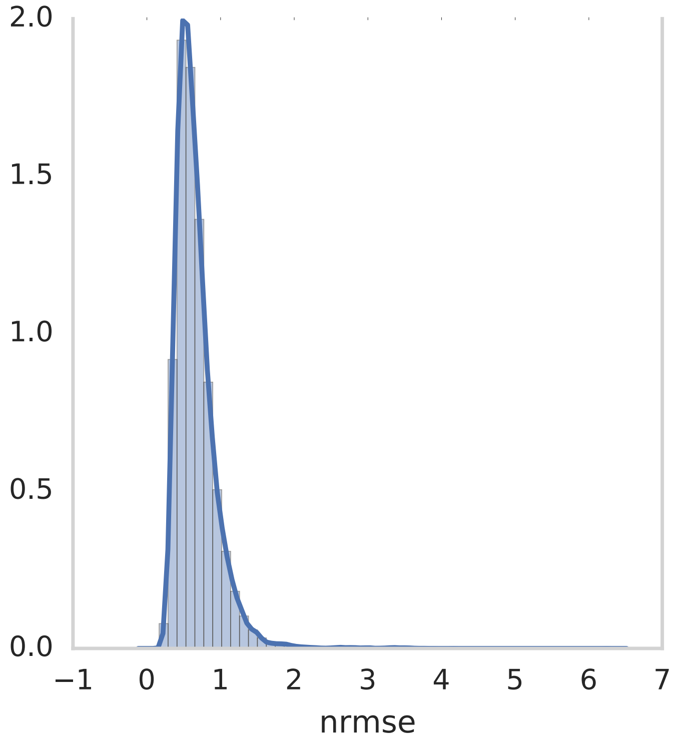

Figure 12 NRMSE distribution of exon trajectories

Density plot of the leave-p-out cross-validation metric: NRMSE (based on the average sum of residuals over mean expression) follows a unimodal distribution skewed to the right. A low NRMSE value indicates a robust fit. The long, sparse density right tail extends to a maximal NRMSE of 94.

3.6 Validation of parameterization

In order to assess the reliability of these estimated trajectories (particularly the predicted peak and width), we assembled a novel independent RNAseq dataset. This independent dataset presents many technical differences with respect to Brainspan, including sample collection and brain region, library preparation, sequencing technology and coverage, and thus represents a stringent test for predictions. Of primary concern was the accuracy of our estimated peak expression for each exon. Our validation samples fall into two broad developmental periods:

prenatal and young adult. Among our models, 127,286 predicted values for μ fell into the prenatal window and 64,051 fall into the adult window.

Since our validation dataset consists of samples belonging to two narrow developmental windows, we designed a binary test for validation. We validated exons whose predicted peak expression (μ) fell into developmental windows that were represented in our validation dataset. These windows are the fetal window, *F*, defined as conception to 38 weeks post conception, and the adult window *A*, defined as 780 weeks post conception (14 years of age) and onwards.

If we take f_j to be the mean expression of exon j in fetal validation samples and a_j to be the mean expression of exon j in adult validation samples, then the parameterization of an exon's peak expression is considered validated if either of the following conditions were met.

$\mu_j \epsilon F$ and $f_j > a_j$ (fetal peak predicted and corroborated),

$\mu_j \epsilon A$ and $f_j < a_j$ (adult peak predicted and corroborated)

Only exons with predicted expression within the two windows covered by the validation dataset were subjected to this test. We also omitted exons that were weakly expressed (less than 1 RPKM) and exons that overlapped another exonic feature (the latter is discussed in detail in section 4.2).

Validation rates are expected to vary depending on the values of the sigma parameter (width of peak expression) and NRMSE (robustness score). We thus binned exons by these two values, and then calculated the validation rate for each bin (Figure 13). Rates for each bin following randomized label shuffling hover around 50%, as expected for this binary test (Figure 14).

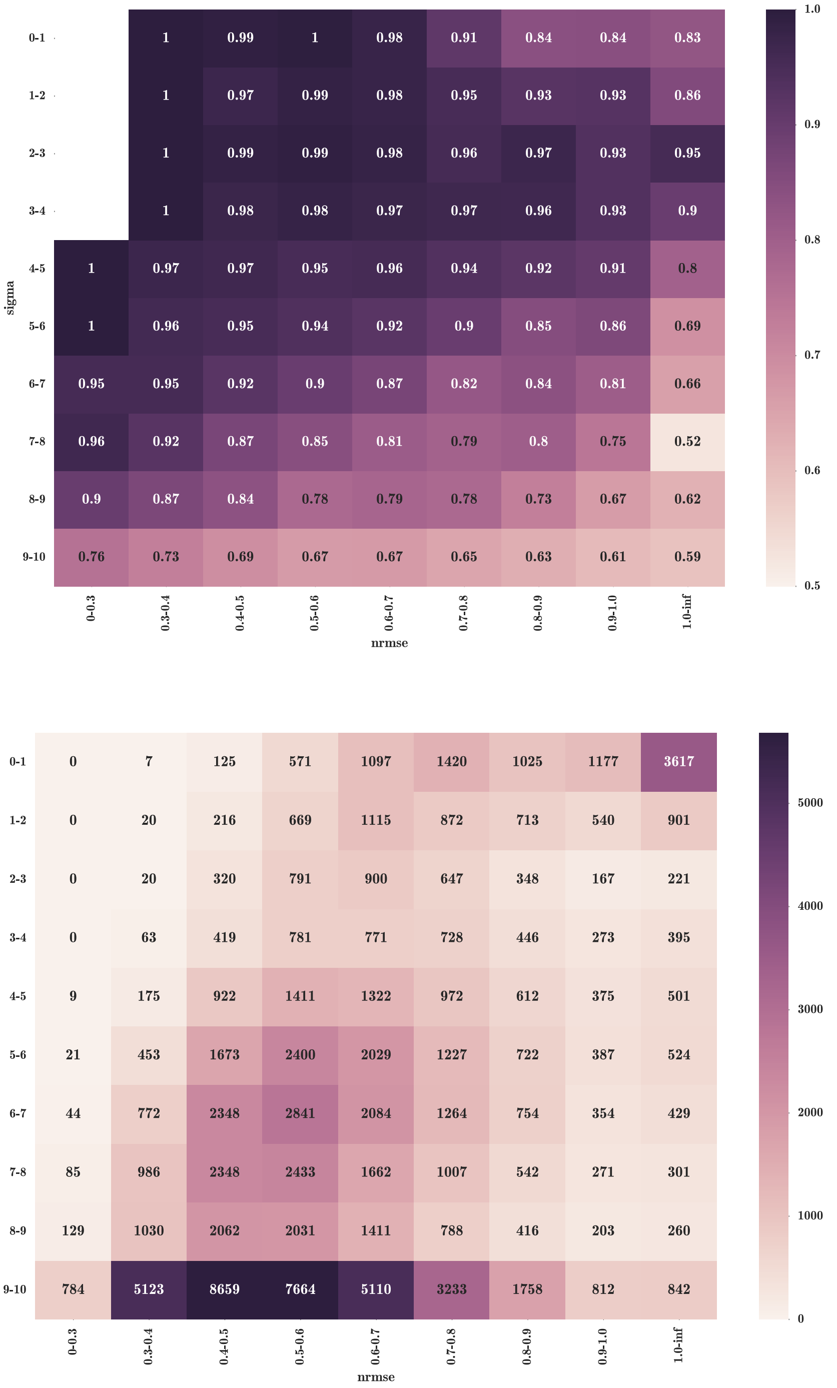

Figure 13: Validation of parameterization

Validation rate of peak expression concordance with validation dataset, binned by estimated value of sigma and NRMSE. ***B*** *Counts of exon trajectories for each bin.*

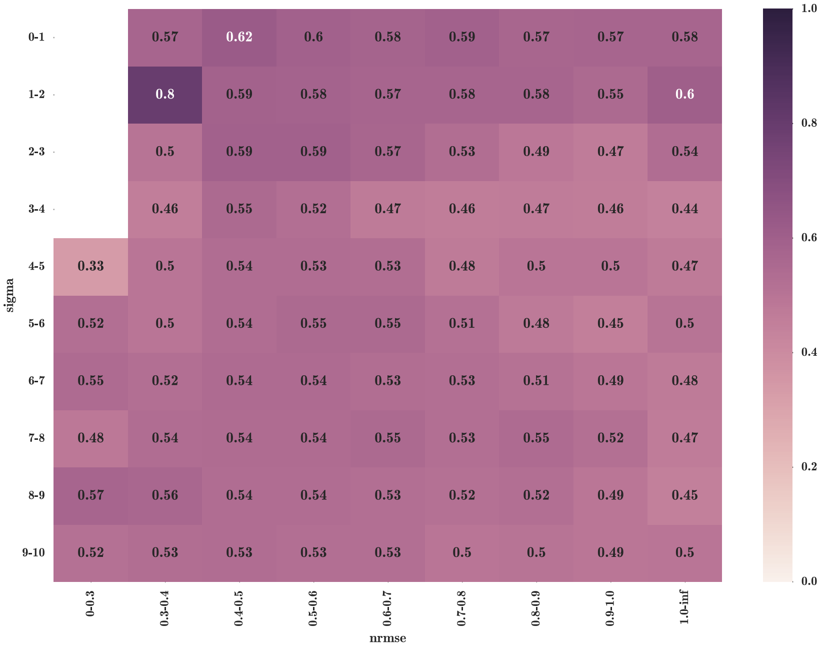

Figure 14 Validation of parametrization with shuffled labels

Validation rate of exons with shuffled labels, binned by estimated value of sigma and NRMSE.

As expected, the validation rate declines with higher estimated σ values and with higher NRMSE scores. The rate declines to below 80% with NRMSE values over 1 and σ values over 6, reflecting the anticipated uncertainty in the predictions of non-robust fits (high NRMS) and the low predictive value of the parameter μ for non-developmental (flat) trajectories (high values of σ). Very wide peak expression intervals (σ > 9) do not provide useful predictions for the exon's peak (μ) and are referred to as non-developmental trajectories in this work. Exons with these trajectories are sometimes referred to as non-developmental exons.

The bulk of exons fall into high σ bins and moderate NRMSE scores, indicating that the majority of exons display non-developmental trajectories that remain reasonably stable when substituting donors in the leave-one-out calculations.

In summary, our method reliably predicts peak exon expression for trajectories with σ value between 0 and 6 with an NRMSE value equal to or less than 1. Furthermore, the capacity of the Gaussian model is sufficient to categorize relatively flat, non-developmental trajectories which are relatively robust to leave-one-out calculations.

4. Predicting differential exon inclusion

4.1 Overview

The parameterized model described so far allows for the straight-forward comparison of parameters that summarize exon trajectories. Measures of robustness (NRMSE) and of indicators of developmental trajectories (σ) provide the remainder of the metrics required for making predictions about developmental exons by comparing exon trajectories within a shared gene product. For-instance, one may query the data for genes with robust, developmental exon trajectories whose estimated peak expressions (μ) fall into different developmental windows. We posit that at least one of these exons is differentially expressed between the 2 windows and predict these candidates as differentially and developmentally expressed. This is but one of a number of curve comparison methods can be applied for our purposes, from straightforward queries of parameters to more sophisticated methods of curve clustering and outlier detection. Each one of these methods will varies in sensitivity and specificity.

In what follows, we first describe the resolution of overlapping exonic features along the genome and the identification of categories of exons through the required annotation and curation steps in order to reduce false positive predictions.

We then design a validation test to test for the differential inclusion of exons between fetal and adult samples in our validation dataset. Cassette exons are tested separately from initial and final transcript exons because of the nature of the splice junction reads used in the test. We generate thresholds for this test by utilizing the query described above: detecting pairs of exons within a gene-product that peak in different development windows. We compare exons selected in this way to randomly selected exons.

We then present the exon trajectory comparison strategies we implemented, following a logical progression. We apply progressively more sophisticated distance measures to quantify dis-similarity of exon trajectories within gene products. Within a gene product we aggregate exons and compute mean, median and maximal distances between the estimated peak expressions of exons within the set. We also compute zscores of the distance of each exon's peak expression to estimates of the gene's peak expression. These scores are utilized to predict exons that are differentially included over the course of development. We increase the complexity of the distance computation and apply hierarchical clustering to the predicted peak expression of each exon within a gene product. Generally speaking, by increasing the complexity of the distance score we increase the number of candidates predicted by each approach however the validation rate for complex methods is lessened. Nonetheless, we feel that the incremental discovery of novel and valid candidates outweighed the lesser validation rates and so all results are presented here for exons with peak expression in the validation fetal and adult windows below. We also select some candidates for qtRT-PCR validation in mice.

4.2 Annotation curation

The human genome presents many instances of exons overlapping in genomic coordinates, many are located in opposite strands of the DNA. Brainspan collapses exon annotation for overlapping exons that share the same strand but have alternate 5'ss or 3'ss as described in section 2.3.4 (Figure 4). However, the Brainspan annotation does not contain strand information, only expression data for genomic coordinates and thus contains many instances of confounded expression signals. Moreover, the exon annotation of Brainspan, generated using the Ensembl 65 release, contains some outdated information, attributing some transcribed exons to the incorrect gene.

Therefore, some further curation was undertaken in order to remove these confounding factors in the comparison of exon trajectories: feature overlap identification and extra-genic exon. We intersected the genomic coordinates for the Brainspan exons with themselves using bedtools[83]. Any exon that overlapped more than 1 annotated feature (itself) was flagged and excluded from downstream analyses.

In order to reduce annotation errors that would induce false positive predictions, we compared Brainspan annotations with the latest release of Ensembl (release 82).[97] Brainspan exons that fell outside the gene coordinates in this version of Ensembl were removed from our analyses. We calculated gene coordinates in order to resolve the most encompassing genomic region for each gene by aggregating gene isoforms and computing minimal genomic starts and maximal genomic ends for the set of each gene's isoforms.

4.3 Identification of promoters, terminators and cassette exons.

To assign Brainspan annotated exons to a category of promoter, terminator or cassette, we utilized the Ensembl annotation set version 82.[97] The Ensembl data was filtered to contain only exons belonging to known protein coding genes and matured transcripts based on Ensembl's annotation of these properties. Informed by their position within the transcripts, annotated exons were marked as promoter (i.e. first exon of a transcript), 'terminator' (i.e. last exon of a transcript), 'cassette' or a combination of these labels if they existed in different roles among the different isoforms of a gene.

To assign Brainspan annotated exons to each category, the genomic coordinates of exons with unique labels (as promoters, terminators or cassettes) were then intersected with the existing Brainspan dataset annotation using pybedtools[98] a python wrapper around the bedtools[99] suite. Exons within that Brainspan annotation that overlapped with only one of the labelled exons were, in turn, labelled accordingly as a promoter, terminator or cassette exon.

4.4 Development of a validation test

To develop a validation test, we considered cassette exons separately from promoter and terminator exons since the method for the quantification of inclusion varies accordingly. Cassette exon inclusion can be calculated via the ratio of reads that bypass a given exon to the splice junction reads that support its inclusion in the transcript. Promoters and terminators are not skipped over by splice junction reads, we quantify their inclusion with respect to a set of exons with constitutive properties.

4.4.1 Cassette exons

In the case of cassette exons, we calculated an exon inclusion rate in the validation dataset as follows. Let u_{ji} be the number of splice junction reads that join exon j to an upstream exon in sample *i* (within the same gene). Let d_{ji} be the number of splice junction reads that join exon j to downstream exons in sample *i* (within the same gene). Finally, let k_{ji} be the number of reads that join exons upstream and downstream of exon j, passing over exon j (within the same gene, for sample *i*). The exon inclusion rate for exon *j* in sample *i* (Ψ_{ji}) is defined as follows:

$$\Psi_{ji} = \frac{max(d_{ji}, u_{ji})}{max(d_{ji}, u_{ji}) + k_{ji}} \qquad \text{(Equation 5)}$$

A visualization of these counts for a hypothetical exon j is shown in Figure 15.

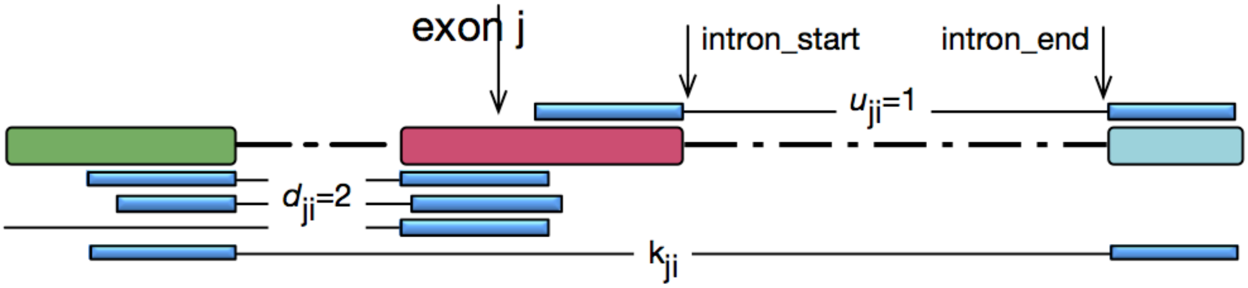

Figure 15 Exon inclusion rates

Visualization of hypothetical read counts for exon j. Exon j is situated on the plus strand and is part of a 3 exon gene. Reads that join exon j to an upstream exon number 2. (The third read is ignored as it originates before the gene begins.) Reads that join exon j to a downstream exon number 1. Reads that skip over exon j, joining exons upstream and downstream of it, number 1. Therefore the Ψ_j *calculation is max(2,1):[max(2,1)+1]=2/3.*

The magnitude of the difference of the means of fetal samples ($\overline{\Psi_f}$) and adult samples ($\overline{\Psi_a}$) in the validation dataset is reported as $\Delta\Psi$.

$$\Delta\Psi = \left|\overline{\Psi_f}\text{-}\overline{\Psi_a}\right| \qquad \text{(Equation 6)}$$

4.4.2 Promoter and terminator exons

Promoter and terminator exons are not always passed over by skipping reads that join exons upstream and downstream. Therefore, a different validation test is needed. On a per-sample basis, we normalize the number of splice junction reads supporting inclusion of the exon using splice junction reads from a set of constitutive exons belonging to the same gene. We select these constitutive exons based on 2 criteria: Equations 7 and 8.

$$\widetilde{\Psi_f} = \widetilde{\Psi_a} = 1 \qquad \text{(Equation 7)}$$

$$\widetilde{y_J} > \frac{1}{n_G}\sum_{l=1}^{n_G} \overline{y_l} \qquad \text{(Equation 8)}$$

Equation 7 states that the *median* inclusion rate for constitutive exons must be 1 for adult and fetal samples. Equation 8 states that the *median* expression (in RPKM) of constitutive exons ($\widetilde{y_J}$) across samples must be greater (and not equal) to the mean expression of exons ($\overline{y}$) in the gene across samples. Figure 16 shows a selection of constitutive exons for the *PHLDB1* gene.

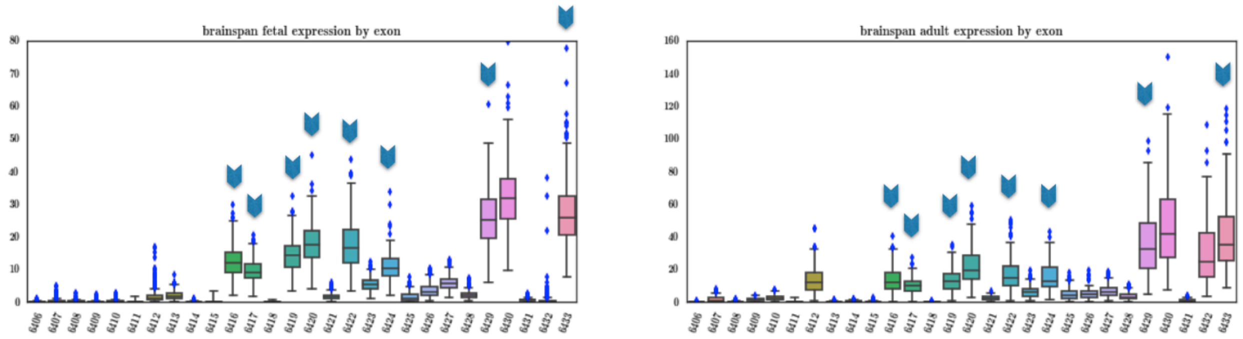

Figure 16 Constitutive exon identification

A boxplot of expression values among samples for each exon, arranged along the x-axis, of the PHLDB1 gene. Boxplots for fetal samples among the validation dataset are on the left. Boxplots for adult samples among the validation dataset are on the right. Arrows indicate the exons selected as constitutive by the algorithm outlined above. The exons have a $\Psi > 1$ *in both adult and fetal samples and are expressed greater than the mean overall expression for the gene.*

Using the set of constitutive exons for normalization, promoter and terminator inclusion is computed as a ratio of promoter/terminator splice junction reads to the median of the constitutive splice junction reads:

$$\rho_{ji} = \frac{max(d_{ji}, u_{ji})}{\tilde{r}_i} , \qquad \text{(Equation 9)}$$

where $\tilde{r}_i$ is the median of $max(d_{ki}, u_{ki})$ for all exons k in the constitutive exon set. We compute a log2 fold change between adult and fetal sample means and report this as ρ_j for exon j.

4.4.2 Comparison of initial dual-peak query with random exons

We queried our fitted models to identify genes that included exon trajectories whose peaks were in different developmental windows and whose developmental windows corresponded to those found in the validation dataset: fetal and adult.

We identified 104 genes with at least 1 adult and 1 fetal exon (cassette, promoter or terminator). In each case we considered only robust trajectories (NRMSE < 1) and informative values of sigma (sigma < 9).

4.4.3 Comparison with random exons

We defined a gene to be validated if one of its exons, predicted as being developmental, met a threshold for $\Delta\Psi$ in the case of cassette exons, or a threshold of ρ in the case of promoters or terminators. In order to identify a reasonable lower-bound for these thresholds, we constructed a test of validation rates that compared the candidates selected above in section 4.4 with a set of randomly selected exons.

We selected a set of random exons that were not predicted as developmental using the method outlined in section 4.4, that were expressed (RPKM > 1), whose peak expression was in a region covered by our validation dataset (fetal or adult) and which met prior set criteria for robustness (NRMSE < 1). We also ensured that these exons were not overlapping other features and that they respected the Ensembl 82 annotation set, as discussed above in the annotation and curation section (section 4.2).

We divided exons between promoter/terminator and cassettes. The results of the exon inclusion tests are shown in Figure 17. Only 10% of genes with randomly selected exons validate at a ρ of 1 or a $\Delta\Psi$ of 0.05; we set our thresholds accordingly.

Note, the log2 fold change is not computed in cassette exons. The delta of percentage inclusions measure applied to cassettes is intuitive whereas the ratio to constitutive exons, applied to promoter/terminators is not. For instance, this occurs in cases where a candidate exon exceeds the expression of the constitutive exons in one developmental window but not in the other. Furthermore, the delta threshold required for validating cassette exons (as seen in the following section) is small: a 5% difference between the percentage inclusions calculated in each developmental window disqualifies over 90% of random exons. Taking a log fold change further reduces this number further while obscuring its meaning.

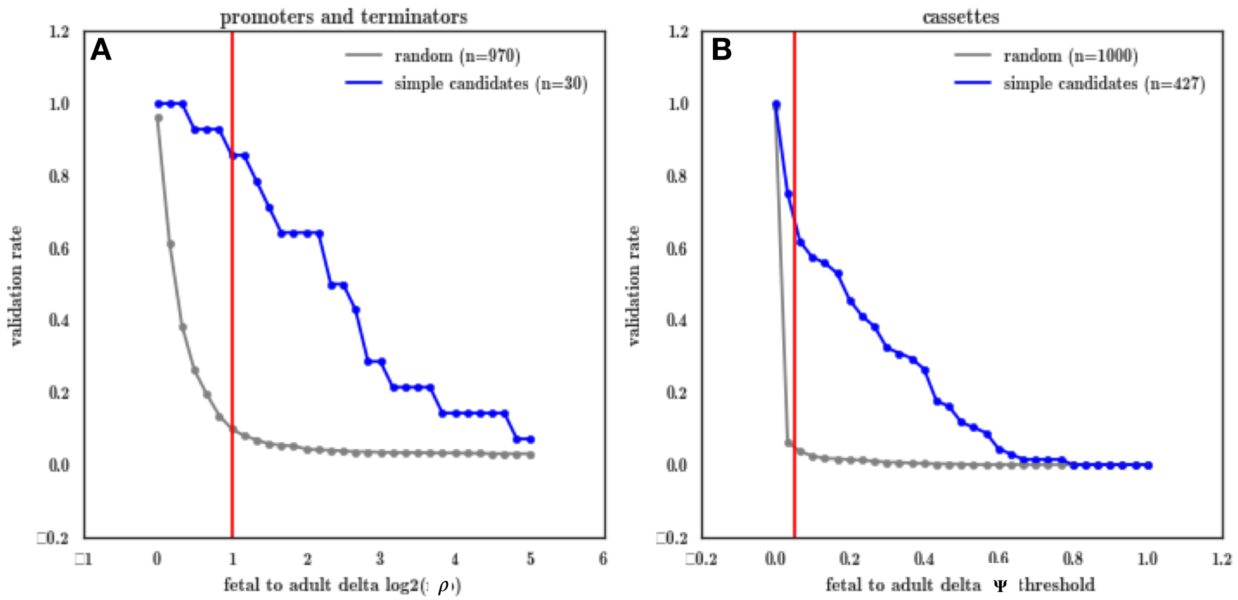

Figure 17 Validation rate of genes with fetal and adult exons

Validation rate of promoter and terminator exons. A candidate exon is considered to be validated if ρ is greater than the threshold along the x-axis (a constant of 1e-4 is added to avoid undefined values of log). 10% of random promoters/terminators validate at a threshold of 1 (the red vertical line). ***B*** *Validation rate of cassette exons as a function of ΔΨ along the x-axis. 10% of random cassettes validate at a delta psi threshold of 0.05 (corresponding to the red vertical line).*

4.5 Further methods for predicting temporal exon inclusion

Having implemented a parameterized modeling framework, verified the consistency of estimated the parameters, and implemented a validation test by analyzing the straight-forward query presented above, we now turn to other methods for predicting developmental exons.

Note in the elucidation of methods below we generally constrain ourselves to calculating measures of dis-similarity between trajectories based on their peak expression μ. Nonetheless, the other parameters inform the trajectories. A scales the trajectory to meet the expression unit scale. The result of our p-fold validation tests (NRMSE) while not strictly an estimated parameter, provides a measures of robustness. σ provides a measure on the accuracy of μ, small values of σ indicate developmentally sensitive trajectories. Prior attempts to integrate σ into measures of dis-similarity as discussed in the following sections produced less favorable validation rates when attempted.

4.5.1 Predictions Dmean, Dmedian, Dmax

We sought to extend the measure of distance between exon trajectories by leveraging dis-similarities in μ. Dmean calculates the mean distance from μ_j for an exon j to estimated values of μ for all other exons within the gene product (restricted to informative and robust trajectories).

Let j be an exon of interest and take *l* to index G: the set of informative ($\sigma_l < 9$) and robust $(NRMSE_l < 1)$ exon trajectories included within the gene product G. Then with n_G the size of the set *G*, d_j is the Dmean score for exon j (Equation 10).

$$d_j = \frac{1}{n_G} \sum_{l \in G} \left| \mu_j \text{-} \mu_{l \neq j} \right| \qquad \text{Equation 10}$$

Exons that are expressed (RPKM > 1) and developmental (σ < 9) with the greatest Dmean score among other exons in the same gene product are predicted as being differentially expressed if at least one other exon (not necessarily developmental) is found to be the opposing validation development window

We applied the same methods using the median of the distances. Dmedian. calculates the median distance from μ_j for an exon j to all other exons within the gene product with informative and robust trajectories. Finally, the Dmax measure takes the maximum distance between 1 exon and other exons within the same gene product.

Given the thresholds established above using the random exon trials, we validate 21 of the 29 exons predicted by Dmeans (72% accurate), 19 of the 27 exons predicted by Dmedian (70% accurate), and 17 of the 26 exons predicted by Dmax (65% accurate). There is a non-trivial amount of overlap between the methods as can be seen in Figure 18. In total, 31 exons are predicted by the union of the methods, of these, 21 validate (10 cassettes, 11 promoters, no terminators) for a 67% accuracy rate. The results are included in Table A1 in the appendix.

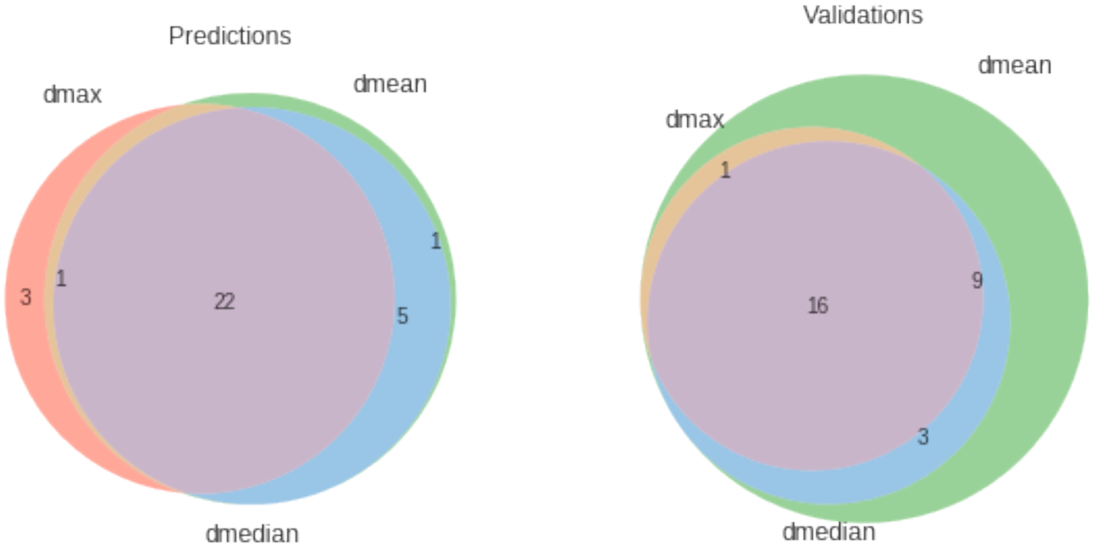

Figure 18 Intersections of predicted exons by Dmean, Dmedian and Dmax.

31 exons in total are predicted by the three methods and 21 validate. All three methods predict a set of 22 exons from which 16 validate. Dmax predicts three false positives and no validated exons that are not also covered by the other methods.

4.5.2 Predictions Zscore

We then attempted to integrate information about the peak expression of the gene trajectory parameter μ_G (estimated from whole gene expression) in order to predict candidate exons. A distance from the μ_G is calculated for each exon j and based on these distances, a zscore z_j is calculated for each exon j. If the distance between the exon and gene parameters are calculated such as in Equation 11

$$d_{jG} = |\mu_j - \mu_G| \qquad \text{Equation 11}$$

Then, the zscore for an exon j is

$$z_j = \frac{d_{jG} - \bar{d}_G}{\sigma_G} \qquad \text{Equation 12}$$

where $\bar{d}_G$ is the mean of the distances d_{iG} for all exons in in the set G and σ_G is the standard deviation of these distances.

This method predicts 30 developmental exons, 28 of which are predicted by the union of the methods in section 4.5.1. We validate 18 exons (8 cassettes and 10 promoters), all of which were validated by the methods above. Thus, incorporating explicit information from the gene-product's trajectory does not improve our prediction strategies, indicating that modeling the ensemble of individual exons may be sufficient to capture this information. These results are included in Table A1 in the appendix.

Table A1 presents all exons that are predicted by the above methods and that validate by our validation test: 130 exons distributed among 114 genes.

4.5.3 Hierarchical clustering

We then sought to determine dissimilar trajectories through agglomerative clustering methods. For each gene, we performed a hierarchical clustering analysis using the Euclidean distance between each exon's estimated value of u.

Hierarchical clustering[100] is the iterative assembly of elements into groups. In our case, the elements being grouped are exon trajectories. The distance between exon trajectories is the Euclidean distance between the estimated peak expression for each: u. Exons with the least distance between estimated values of u are grouped first. Distances are then recomputed between values of u for ungrouped exons. Distances between groups (also called clusters) and ungrouped exons are also computed. These distances are determined by taking the maximal distance between exons in the group and the ungrouped exon. (This so called distance function is referred to as a 'complete' distance). Distances between 2 groups, should that occur are taken as the maximal of the pairwise distances between the members of the disjoint groups.

This iterative clustering is done until all exons are grouped as a single set. We are primarily interested in the maximal distance of the cluster, that is the distance calculated between the last 2 groups or ungrouped exons or a mix, whichever collection remains. This final distance is indicative of the greatest degree of dis-similarity between groups of exons in a gene product.

By filtering clusters by maximal distance threshold (*t*) we identify groups of candidates based on this level of dis-similarity. Unlike the previous methods, the exons that are differentially expressed in different windows are not identified by this method as they can be in either of the dis-similar clusters. We thus test each of the robust, developmental trajectories using the validation dataset in order to validate at the gene level. At a threshold of t=1, only 13% of the genes validate, close to the random exon validation rate, as expected (the rate is slightly higher on account of the gene validating if one of its multiple exons validate.) Exons that validate (at a threshold of t=2 or higher, are tabulated in Table A2 in the appendix.

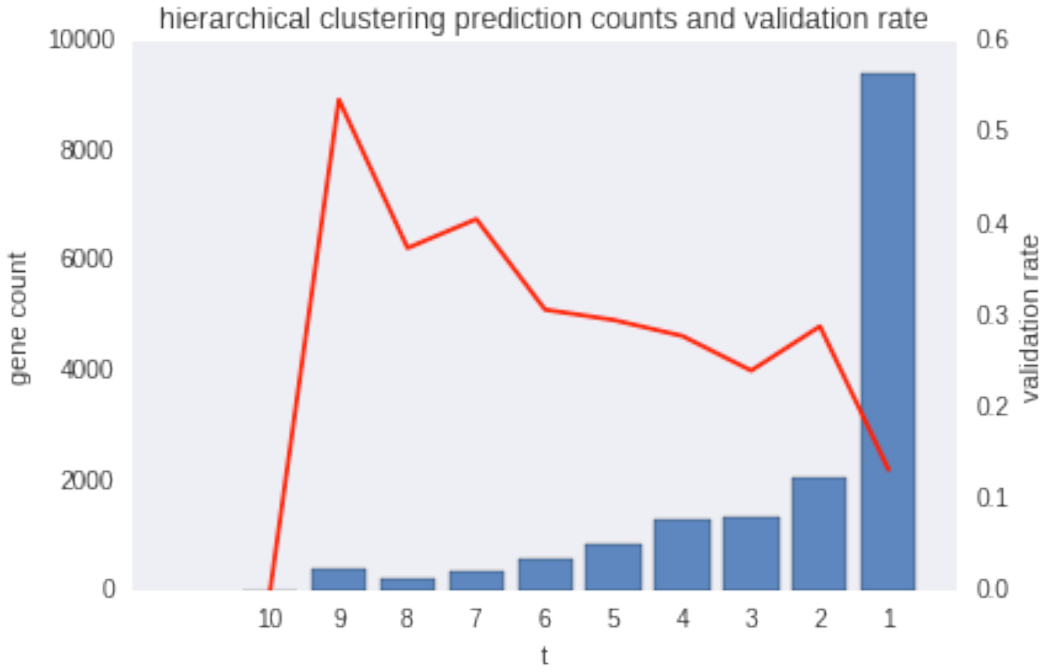

Figure 19 Validation rates for hierarchical clustering

The validation rate of genes predicted to have developmentally regulated exonic content as identified using the hierarchical clustering method for varied values of t (the cluster cutting threshold).

Hierarchical clustering produces a much higher number of candidates but at a greater false positive rate. Nonetheless, we can estimate the number of true positives identified using the final bin in Figure 19, if we assume that the false discovery rate of our validation test is uniform for all values of t, at approximately 13%. In this case, using a cut-off of t=2, hierarchical clustering methods identify 4096 exons among 2488 genes. In the interest of concision, we include the top 250 candidates (ranked by width of developmental peak) in Table A2. The remaining candidates are available through the corresponding author and in the supplementary material for the forthcoming publication.

4.6 Validation in mice

In order to ensure that our results were not simply technical artifacts, nor purely due to evidence of differential cellular composition of tissues, we selected some candidates with orthologues in mice for validation via qt-PCR. Orthologues were identified and verified using UCSC's BLAT tool[101] and genome browser[102]. Primers were designed to bind regions spanning splice junctions when possible, so as to obtain a product only when the candidate exon is included. Many candidate exons are of small size and thus, primers were designed to amplify regions that would only exist if the exon where included. Constitutive exons were used for normalization. Figure 20 shows the primer design.

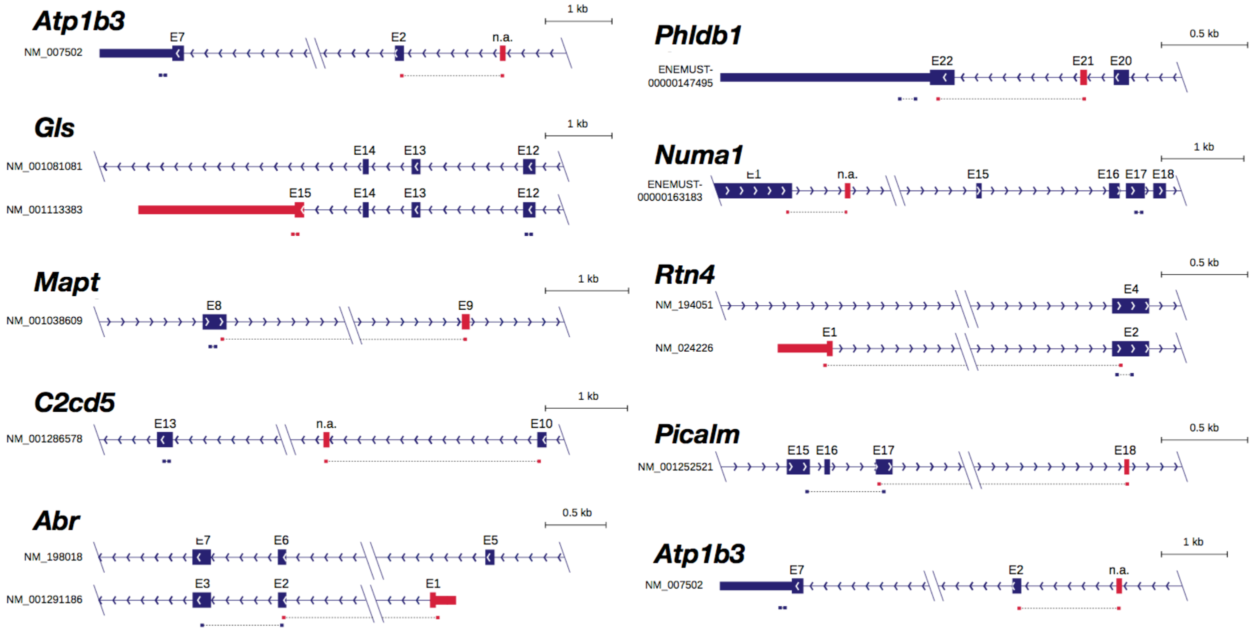

Figure 20 Primer design

Primers were selected to amplify their targets only if the developmentally regulated exons were present (depicted in red). For each gene-product a second set of primers was selected to span a region of the gene-product predicted to be constitutive. NCBI accession numbers are given for the developmental region when available, otherwise the Ensembl transcript identifier is given.

All 9 exons predicted as being developmentally regulated validated. (Figure 21).

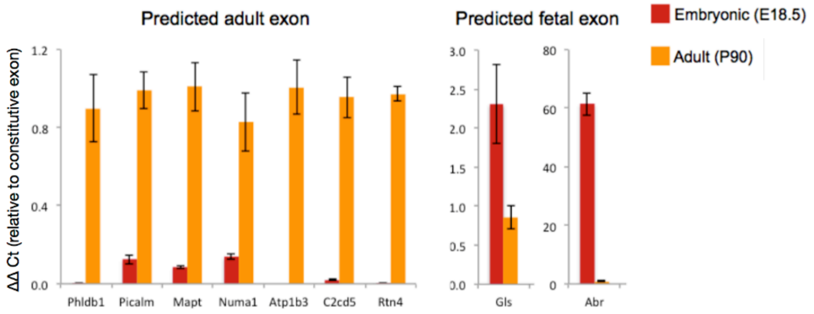

Figure 21 qtRT-PCR results

Bars plot the fold change of predicted developmental exons to the constitutive exon amplification target. Embryonic brain samples are in red, adult brain samples are in orange. All exon amplification products have a markedly different profile between fetal and adult mice samples and correspond to the predicted developmental window.

5. Discussion

Alternative splicing is an important biological phenomenon that greatly diversifies the population of transcripts and downstream gene products generated by the cell. This research approaches the temporal development of the brain's transcriptome through a framework of parameterized model building. Parameters are inferred from the time-series of individual exonic expression as assayed by next generation sequencing technologies and (the parameters) are tied to biologically meaningful properties of the developing tissue. Peak expression and the rate of its attainment are represented by the parameterization of the model. Thus high dimensionality data, equal to the number of observations in the time series (524 samples in this implementation), is reduced to a small set of parameters: 1 scaling parameter and 2 that shape the trajectory. The model has been chosen so that these parameters are immediately informative.

We measured dis-similarity between exon trajectories within the same gene product and predicted a differential of exonic expression among developmental periods. We tested these predictions by measuring exon expression by leveraging splice junction reads in a validation dataset. Our methods identify approximately 600 developmentally regulated exons among as many genes with a high degree of confidence (hierarchical candidates with t=9, section 4.5.3), using methods that predict differential inclusion in 1 dataset and validate the findings in an amalgamation of 2 more.

5.1 Differential expression of promoter exons

Our methods allow for the identification of promoters that are associated with developmental periods. Since activation of promoters is accompanied by large changes in chromatin states (chromatin accessibility, deposition of histone marks, changes in DNA methylation), predictions regarding temporal changes on promoter activation state can be validated using independent epigenomic data. Thus, using data from the Epigenome Roadmap Project[103] from the NIH we can further corroborate predictions made by the Brainspan dataset and corroborations made by our validation dataset. Here we present a novel promoter found by our methods: RTN4, an important regulator of neurite propagation.

Reticulons are a family of membrane-anchored proteins present in the endoplasmic reticulum (ER). There are 4 mammalian RTN genes, one of which is the 14 exon *RTN4*[104]*,* also known as *NOGO* for its role in the inhibition of CNS regeneration[105]. *NOGO* gives rise to 3 major isoforms: *NOGO*-A, *NOGO*-B and *NOGO* -C which share a common C-terminal of 188 amino acids. *NOGO*-A, the largest isoform is predominantly expressed by CNS myelin-forming oligodendrocytes in the CNS[106] and contains at least 2 domains that are strong inhibitors of neurite outgrowth (Nogo-66 and Nogo-40) proximal to the C terminal. Our findings suggest that the shorter isoform has a distinct temporal trajectory, unknown prior to this study. This isoform is transcribed from a distinct adult only-promoter that is most active in the adult development window. The clustering of the trajectories is shown in Figure 22.

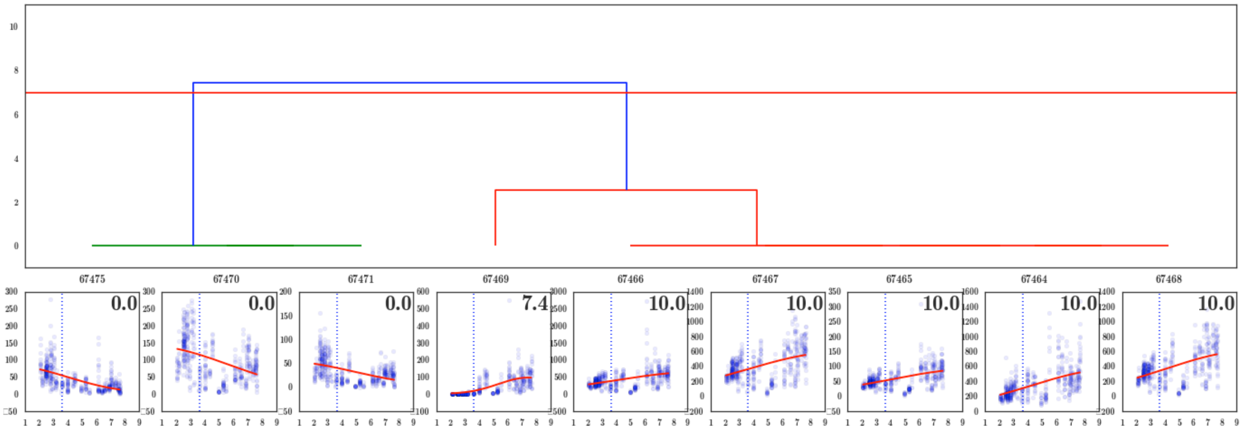

Figure 22 Hierarchical clustering of *RTN4* trajectories

RTN4 is flagged as a gene with developmentally regulated exons at a threshold of t=7. Informative, developmental exon trajectories separate into trajectories with adult peaks and fetal peaks. The three leftmost trajectories with a peak expression of 0 correspond to the RTN4/Nogo-A isoform known to be present in the brain. Of more particular interest is the promoter trajectory with a peak at 7.4 (exon 67469) signaling the increased presence of the shorter RTN4 isoform (NOGO-C) in the older human brains.

Using the Washington Epigenome Browser[107] to view epigenome information made available through the Epigenome Roadmap Project[103], we can see that chromatin marks often associated with promoter use are absent in the fetal sample yet are present in the adult samples.

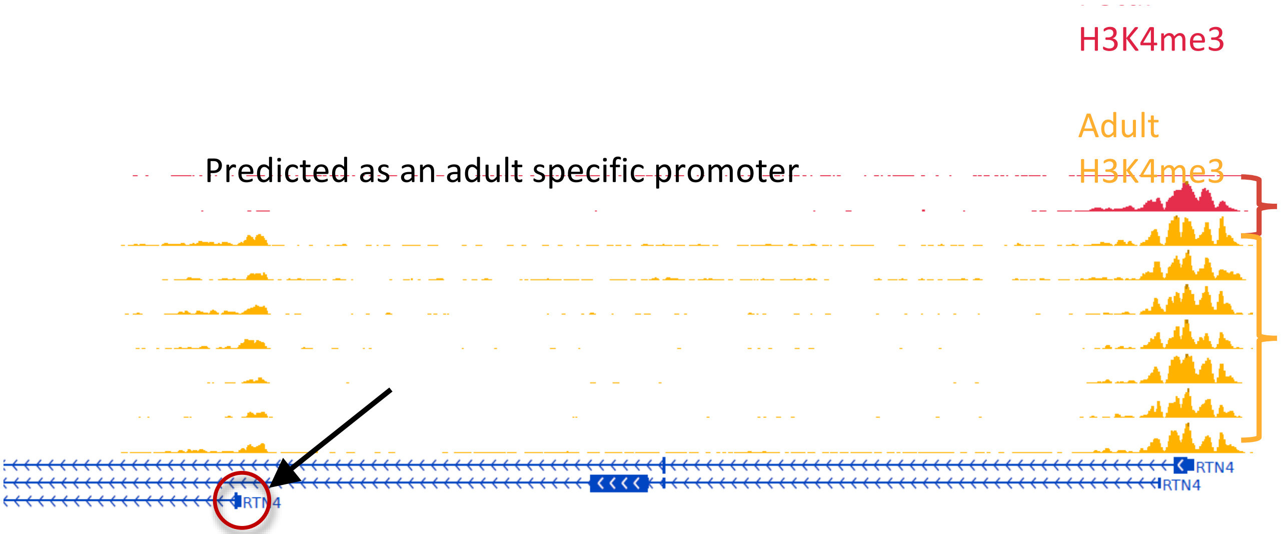

Figure 23 Histone modifications for *RTN4* exons.

Histone marks H3K4me3, associated with active promoter use, are present in both adult and fetal samples on the right at the distal promoter for the longer isoform (Nogo-A) however only adult samples show histone marks for the shorter isoform's promoter

Figure 24 demonstrates the corroboration of the validation dataset with the trajectory estimated using the Brainspan data.

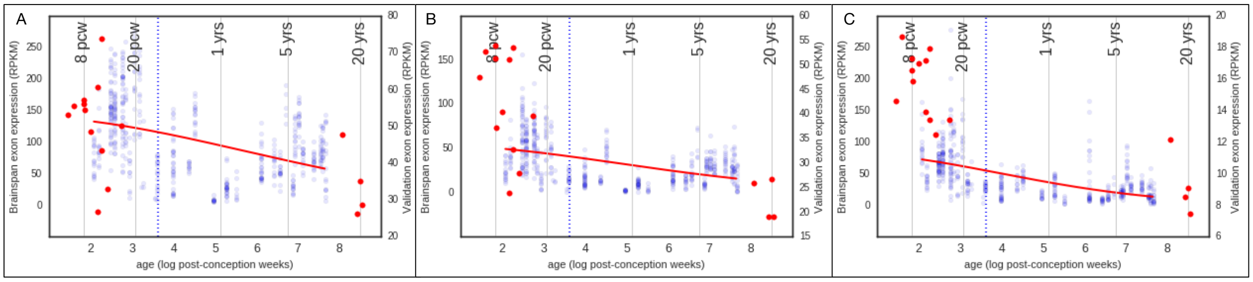

Figure 24 *RTN4* differentially expressed promoter exons

Our models traced as red curves predict developmental trajectories that peak in at conception. Brainspan samples are plotted in blue according to the left y-axis, greater saturation indicates a greater density of samples. Validation data is in red, plotted according to the right y-axis. Expression in validation samples declines from conception to adulthood.

Thus this developmental trajectory of these *RTN4* exons is corroborated by three different datasets: predicted by the Brainspan dataset and validated our aggregated validation dataset (provided by 2 different researchers). *RTN4* was also one of the candidates validated in mice orthologues (Figure 20).

5.2 Differential expression of *MAPT* cassette exons

Our methods allow for the identification of cassette exons that are included in transcripts only for a particular developmental window. Many of these exons code for quite short sequences and further research is required to determine their importance and functional consequences. Nonetheless, the disease relevance of some of these exons, in the case of previously identified genes, has already been proven. Our methods identify the differential inclusion of exons in the MAPT protein, also known as TAU that is associated with several forms of dementia.[31]

We predict differential inclusion of the third exon in *MAPT* transcripts as shown in Figure 25. Using data from three different data sources, we show that the incorporation of this exon occurs increasingly in adults. Its threonine rich sequence offers a greater number of the potential number of phosphorylation targets. Phosphorylation is known to reduce MAPT capability to bind to microtubules[108]. Phosphorylation of certain sites is normal in early development but hyper-phosphorylation in adults has been strongly linked to dementia.[109]

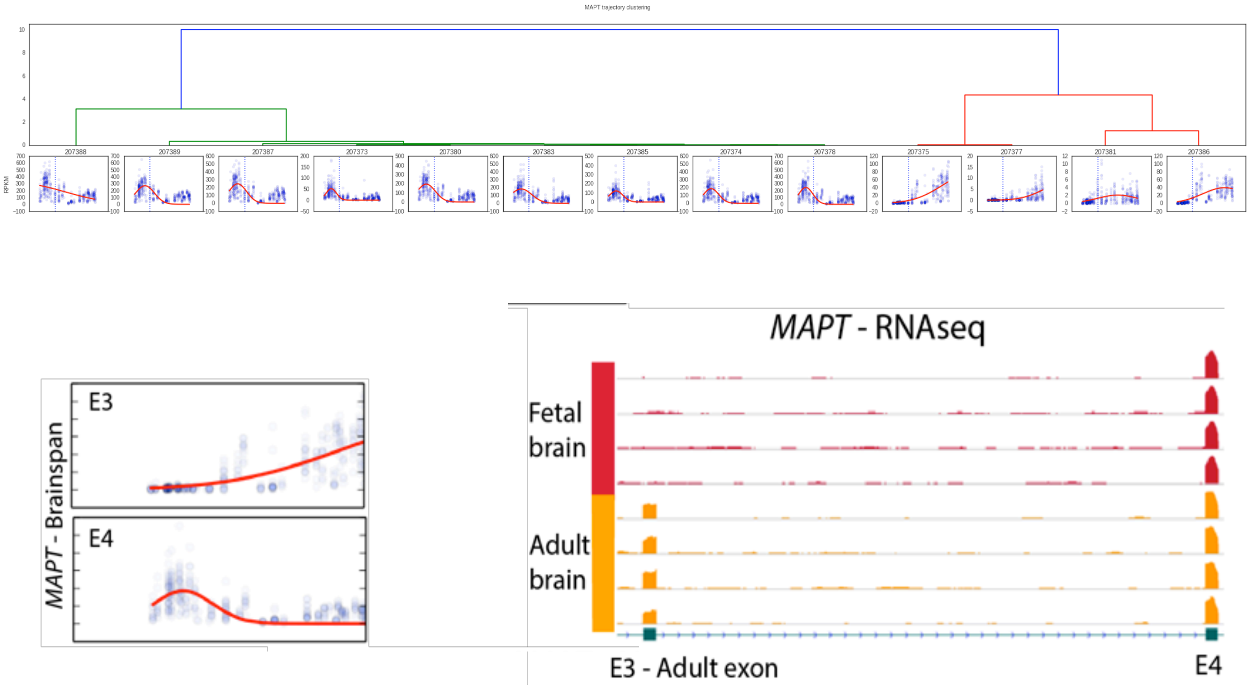

Figure 25 Developmental exons in *MAPT*

***A** Clustering of Brainspan trajectories for the MAPT exons. **B** Trajectories for E3 predicted developmental exon (circled in A) and the constitutive E4 MAPT exon. The E3 trajectory slopes upwards, peaking in adulthood whereas the constitutive E4 exon shows a peak in the prenatal period. **B**. RNAseq expression for the E3 and E4 exons in the validation datasets. E3 is expressed in all samples.*

5.3 Differential expression of *GLS* terminator exons

Our methods allow for the identification of alternative 3' ends to the transcript. As above we present a case study whereby our methods identify developmentally regulated exonic content for a particular gene important in the maturation of the brain.

Glutamine small charge neutral, polar amino acid that is an important source of carbon and nitrogen for the cell and the most abundant free amino acid in the blood.[110] In the nervous system, it is a precursor to glutamate, the primary excitatory neurotransmitter.[111] This pathway is mediated through the enzyme glutaminase which is coded by 2 paralogues: *GLS* and *GLS2*[112]. *Gls* is expressed as 2 transcripts (Figure 26) a long version (referred to as KGA) and a shorter isoform (known as GAC) that terminates on the 15th exon. The 15th exon is skipped in the KGA isoform. Recently, in vitro studies have suggest that both the GAC and KAG are upregulated during neuronal differentiation.[113]

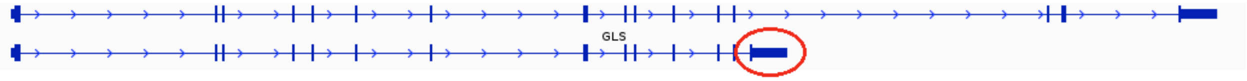

Figure 26 The isoforms of *GLS*

GLS is expressed as 2 isoforms in the human brain. Here the longer isoform (KGA) is depicted as per its layout on the hg19 genome assembly above the shorter isoform (GAC). The developmentally regulated exon predicted by our methods is circled in red.

Our inquiry into in vivo quantifications of the *GLS* gene product identified a substantial presence of the 15th exon in the very early stages of brain development. Exon 15 was classed as developmental by our methods due to a temporal trajectory that is unlike any of the other informative, robust exon trajectories associated with the *GLS* gene. The 15th exon peaks at conception and rapidly tails off suggesting that the shortened isoform is present for a brief window before being down-regulated by in accordance with the maturation of the brain. Figure 27 shows the clustering of the trajectories for the *GLS* gene.

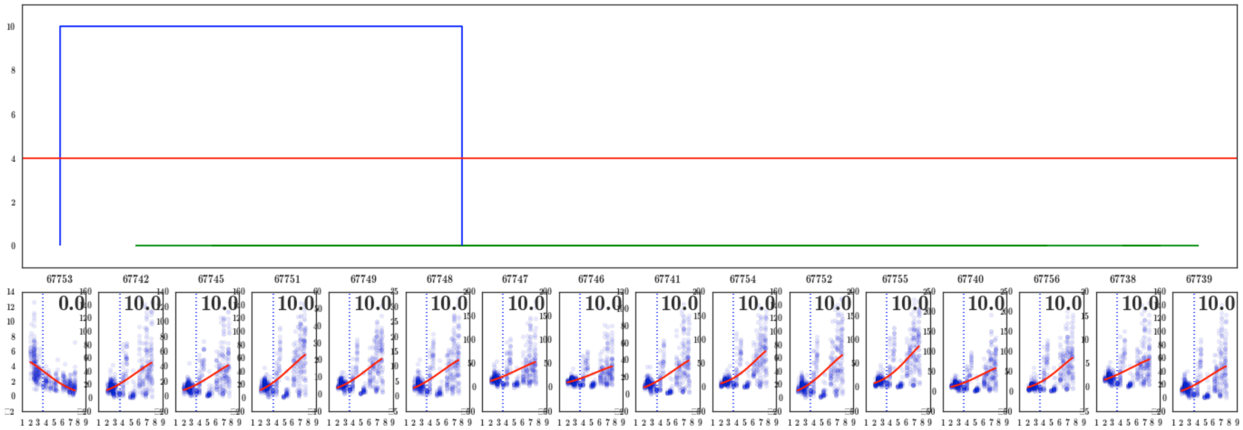

Figure 27 Clustering of *GLS* exon trajectories.

All informative and robust trajectories save the rightmost have a gradual increase in expression with a peak situated at the end of lifespan. The exception is the terminator schematized above which peaks early in development and tapers off immediately indicating the short-term presence of the CAG isoform in human fetal brains at the start of development.

Figure 28 offers a close-up of the validation expression data corroborating the Brainspan expression data for the terminating exon (right-most trajectory in Figure 25).

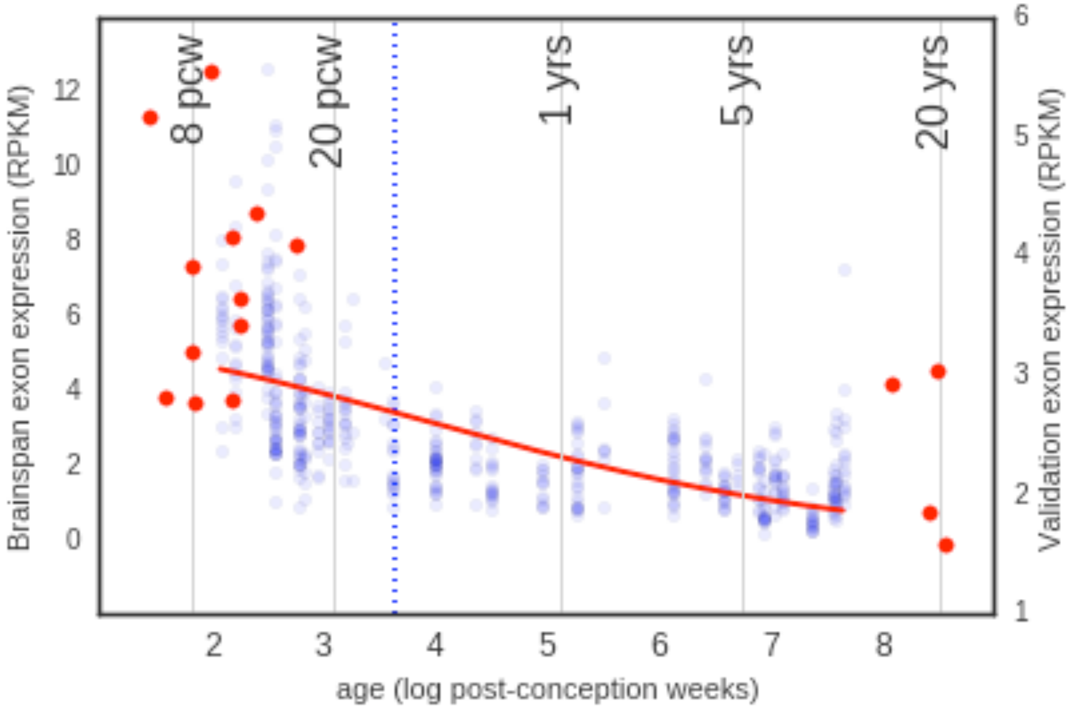

Figure 28 *GLS* differentially expressed terminator exons

Our models, abstracted as the red curves predict developmental trajectories that peak in at conception. Brainspan samples is plotted in blue according to the left y-axis, greater saturation indicates a greater density of samples. Validation data is in red, plotted according to the right y-axis. Expression in validation samples declines from conception to adulthood.

This *GLS* terminator is corroborated by three different datasets: predicted by the Brainspan dataset and validated our aggregated validation dataset (provided by 2 different researchers). A *GLS* orthologue in mice was also validated (Figure 20).

5.4 Concluding remarks

The results of this study are subject to some limitations. We first acknowledge the difficulty in exon expression elucidation given the heterogeneity of the brain's cellular composition. Therefore, exon expression is confounded, in at least some cases, with the maturation of the tissue. Despite the complicated composition of the brain, this tissue was chosen in part because the stringency of its regulation would favor robust trajectories. Given the success of the method in this complex tissue, an important avenue for further work lies in the application of this model-based framework in other tissue types, such as blood or the liver.

A natural expansion on this vein of research would be the inclusion of bio-informatics tools to estimate cell types and their counts among the tissue heterogeneity used to construct the Brainspan transcriptomes. Such counts would inform our process to exclude candidates that show marked temporal differences in exon expression likely due to changes in tissue heterogeneity coincident with aging. Computational methods exist have begun to be developed[114,115] and assessed.[116] Given the relatively new onset of such technique, they would require some non-trivial analysis prior to their integration to our methods.

A more proximal, yet still only partial, solution to this concern of cellularity lies in the availability of substructure data for the brain. Brainspan data exists for substructures of the brain on which our approach can be applied although the number of samples is quite low and precludes cross-validation straetgies. Validation datasets specific to these regions (required to assess candidates) are not yet publicly available. Should further transcriptomic datasets of these same brain regions become available for use in candidate validation, our methods could be applied to these more specific brain regions as delineated in the Brainspan dataset.

A further limitation includes the precision of the trajectories reviewed. Models were built on RNAseq data available from a 2010 assay of a limited set of donor transcriptomes. As a result, differences between exon peak expression patterns were only validated for quite high values of dissimilarities. While a finer resolution would not have been immediately useful given the sparsity of our limited validation dataset, it is anticipated that in further application of these methods more samples will aid in the detection of trajectories whose peaks differ only slightly thus inviting comparisons between more proximal developmental windows (rather than only fetal and adults, for example).

In light of these limitations the high number of developmentally regulated exons predicted and validated using our methods is instructive in developing an understanding of the frequency of developmentally regulated exon expression in the brain. Studies show that the occurrence of alternative splicing events varies by tissue[117] and that the liver and the testis in particular, experience a high number of AS events (as does the brain).[118] These tissues are natural targets of further studies to elucidate the role of AS events in their development.

These findings suggest a non-trivial frequency of developmental exon presence in transcripts in the brain. We anticipate further applications of these techniques to other datasets will reveal a comparable number of events of the developmental regulation of exons. This suggests the immediate applicability of this approach and its results in prioritizing candidate SNPs discovered in genome wide association studies (GWAS). For example one might hypothesize that the inclusion of temporally sensitive exons in a transcript is more likely to bring about pathology either by its untimely inclusion, for instance in the development of brain cancer[6] or simply by the presence of a time-sensitive active domain such as the increased phosphorylation of an adult-specific exon in MAPT in cases of dementia.[109] Application of our methods on a larger scale would greatly expand the use of such tools for biologists conducting sequence-specific protocols.

In addition, the results provided by our approach to identify developmentally regulated exons within a tissue, when referenced with data sets that associate pathology and genome segments, may provide interpretability as to the physiological cause underlying disease and traction for further research. Current annotation sets duly lay out the various isoforms present in the various tissues yet this representation is misleading as a tissue's population changes over the course of development. The application of our methods may shed more light on the developmental course of alternative splicing events in the hopes of rounding out this body of knowledge.

6. References

1. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40,** 1413–1415 (2008).
2. Andreadis, A., Gallego, M. E. & Nadal-Ginard, B. Generation of protein isoform diversity by alternative splicing: mechanistic and biological implications. *Annu. Rev. Cell Biol.* **3,** 207–242 (1987).
3. Blencowe, B. J. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci.* **25,** 106–110 (2000).
4. Faustino, N. A. & Cooper, T. A. Pre-mRNA splicing and human disease. *Genes Dev.* **17,** 419–437 (2003).
5. Cáceres, J. F. & Kornblihtt, A. R. Alternative splicing: multiple control mechanisms and involvement in human disease. *TRENDS Genet.* **18,** 186–193 (2002).
6. Kleinman, C. L. *et al.* Fusion of TTYH1 with the C19MC microRNA cluster drives expression of a brain-specific DNMT3B isoform in the embryonal brain tumor ETMR. *Nat. Genet.* **46,** 39–44 (2014).
7. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
8. Venter, J. C. *et al.* The sequence of the human genome. *Science (80-.).* **291,** 1304–1351 (2001).
9. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* **22,** 1760–1774 (2012).
10. de Klerk, E. & 't Hoen, P. A. C. Alternative mRNA transcription, processing, and translation: Insights from RNA sequencing. *Trends Genet.* **31,** 128–139 (2015).
11. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463,** (2010).
12. Barbosa-Morais, N. L. *et al.* The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science (80-.).* **338,** 1587–1593 (2012).
13. Lee, Y., Rio, D. C., Biology, S. & Biology, C. Mechanisms and Regulation of

Alternative Pre-mRNA splicing. *Annu Rev Biochem* 291–323 (2015). doi:10.1146/annurev-biochem-060614-034316.Mechanisms

14. Calarco, J. A. *et al.* Global analysis of alternative splicing differences between humans and chimpanzees. *Genes Dev.* **21,** 2963–2975 (2007).
15. Lipovich, L. *et al.* High-throughput RNA sequencing reveals structural differences of orthologous brain-expressed genes between western lowland gorillas and humans. *J. Comp. Neurol.* **524,** 288–308 (2016).
16. Rakic, P. Evolution of the neocortex: a perspective from developmental biology. *Nat. Rev. Neurosci.* **10,** 724–735 (2009).
17. Rubenstein, J. L. R. Annual research review: development of the cerebral cortex: implications for neurodevelopmental disorders. *J. Child Psychol. Psychiatry* **52,** 339–355 (2011).
18. Levine, A. & Durbin, R. A computational scan for U12-dependent introns in the human genome sequence. *Nucleic Acids Res.* **29,** 4006–4013 (2001).
19. Burge, C. B., Padgett, R. a & Sharp, P. a. Evolutionary fates and origins of U12-type introns. *Mol. Cell* **2,** 773–785 (1998).
20. Will, C. L. & Lührmann, R. Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.* **3,** 1–2 (2011).
21. Moore, M. J. & Sharp, P. A. Evidence for two active sites in the spliceosome provided by stereochemistry of pre-mRNA splicing. *Nature* **365,** 364–368 (1993).
22. Toor, N., Keating, K. S., Taylor, S. D. & Pyle, A. M. Crystal structure of a self-spliced group II intron. *Science (80-.).* **320,** 77–82 (2008).
23. Wachtel, C. & Manley, J. L. Splicing of mRNA precursors: the role of RNAs and proteins in catalysis. *Mol. Biosyst.* **5,** 311–6 (2009).
24. Reed, R. Mechanisms of fidelity in pre-mRNA splicing. *Curr. Opin. Cell Biol.* **12,** 340–345 (2000).
25. Shepard, P. J. & Hertel, K. J. The SR protein family. *Genome Biol.* **10,** 242 (2009).
26. Sugnet, C. W., Kent, W. J., Ares, M. & Haussler, D. Transcriptome and genome conservation of alternative splicing events in humans and mice. in *Pacific Symposium on Biocomputing* **9,** 66–77 (2004).

27. Ichtchenko, K., Nguyen, T. & Südhof, T. C. Structures, Alternative Splicing, and Neurexin Binding of Multiple Neuroligins. *J. Biol. Chem.* **271,** 2676–2682 (1996).

28. Birzele, F., Csaba, G. & Zimmer, R. Alternative splicing and protein structure evolution. *Nucleic Acids Res.* **36,** 550–558 (2008).

29. Pohl, M., Bortfeldt, R. H., Grützmann, K. & Schuster, S. Alternative splicing of mutually exclusive exons—A review. *Biosystems* **114,** 31–38 (2013).

30. Hutton, M. *et al.* Association of missense and 5[prime]-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature* **393,** 702–705 (1998).

31. Sposito, T. *et al.* Developmental regulation of tau splicing is disrupted in stem cell-derived neurons from frontotemporal dementia patients with the 10 + 16 splice-site mutation in MAPT. *Hum. Mol. Genet.* **24,** 5260–5269 (2015).

32. Proudfoot, N. J., Furger, A. & Dye, M. J. Integrating mRNA Processing with Transcription. *Cell* **108,** 501–512 (2016).

33. Zhao, J., Hyman, L. & Moore, C. Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol. Mol. Biol. Rev.* **63,** 405–445 (1999).

34. Braunschweig, U. *et al.* Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res.* **24,** 1774–1786 (2014).

35. Marquez, Y., Brown, J. W. S., Simpson, C., Barta, A. & Kalyna, M. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res.* **22,** 1184–1195 (2012).

36. Sebé-Pedrós, A. *et al.* Regulated aggregative multicellularity in a close unicellular relative of metazoa. *Elife* **2,** e01287 (2013).

37. GALANTE, P. A. F., SAKABE, N. J. O., KIRSCHBAUM-SLAGER, N. & DE SOUZA, S. J. Detection and evaluation of intron retention events in the human transcriptome. *RNA* **10,** 757–765 (2004).

38. Ge, Y. & Porse, B. T. The functional consequences of intron retention: alternative splicing coupled to NMD as a regulator of gene expression. *Bioessays* **36,** 236–243 (2014).

39. Yap, K., Lim, Z. Q., Khandelia, P., Friedman, B. & Makeyev, E. V. Coordinated regulation of neuronal mRNA steady-state levels through developmentally

controlled intron retention. *Genes Dev.* **26,** 1209–1223 (2012).

40. Jung, H. *et al.* Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat Genet* **47,** 1242–1248 (2015).

41. Shandilya, J. & Roberts, S. G. E. The transcription cycle in eukaryotes: from productive initiation to RNA polymerase II recycling. *Biochim. Biophys. Acta* **1819,** 391–400 (2012).

42. Schaefer, S. & Nadeau, J. H. The ggenetics of epigenetic inheritance: modes, molecules and mechanisms. *Q. Rev. Biol.* **90,** 381–415 (2015).

43. Almouzni, G. & Cedar, H. Maintenance of Epigenetic Information. *Cold Spring Harb. Perspect. Biol.* **8,** (2016).

44. Breitbart, R. E., Andreadis, A. & Nadal-Ginard, B. Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. *Annu. Rev. Biochem.* **56,** 467–495 (1987).

45. Kornblihtt, A. R. Promoter usage and alternative splicing. *Curr. Opin. Cell Biol.* **17,** 262–268 (2005).

46. Chih, B., Gollan, L. & Scheiffele, P. Alternative Splicing Controls Selective Trans-Synaptic Interactions of the Neuroligin-Neurexin Complex. *Neuron* **51,** 171–178 (2006).

47. Vuong, C. K., Black, D. L. & Zheng, S. The neurogenetics of alternative splicing. *Nat. Rev. Neurosci.* **17,** 265–81 (2016).

48. Li, Q., Lee, J.-A. & Black, D. L. Neuronal regulation of alternative pre-mRNA splicing. *Nat. Rev. Neurosci.* **8,** 819–831 (2007).

49. Zheng, S. & Black, D. L. Alternative Pre-mRNA Splicing in Neurons, Growing Up and Extending Its Reach. *Trends Genet.* **29,** 442–448 (2013).

50. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45,** 580–585 (2013).

51. Xue, Y. *et al.* Direct conversion of fibroblasts to neurons by reprogramming PTB-regulated microRNA circuits. *Cell* **152,** 82–96 (2013).

52. Spellman, R., Llorian, M. & Smith, C. W. J. Crossregulation and functional redundancy between the splicing regulator PTB and its paralogs nPTB and ROD1. *Mol. Cell* **27,** 420–434 (2007).

53. Makeyev, E. V, Zhang, J., Carrasco, M. A. & Maniatis, T. The MicroRNA miR-

124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Mol. Cell* **27,** 435–448 (2007).

54. Linares, A. J. *et al.* The splicing regulator PTBP1 controls the activity of the transcription factor Pbx1 during neuronal differentiation. *Elife* e09268 (2015).

55. Calarco, J. A. *et al.* Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. *Cell* **138,** 898–910 (2009).

56. Raj, B. *et al.* A global regulatory mechanism for activating an exon network required for neurogenesis. *Mol. Cell* **56,** 90–103 (2014).

57. Palm, K., Metsis, M. & Timmusk, T. Neuron-specific splicing of zinc finger transcription factor REST/NRSF/XBR is frequent in neuroblastomas and conserved in human, mouse and rat. *Mol. brain Res.* **72,** 30–39 (1999).

58. Tabuchi, A. *et al.* REST4-mediated modulation of REST/NRSF-silencing function during BDNF gene promoter activation. *Biochem. Biophys. Res. Commun.* **290,** 415–420 (2002).

59. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol* **17,** 1–19 (2016).

60. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotech* **31,** 46–53 (2013).

61. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7,** 562–578 (2012).

62. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech* **28,** 511–515 (2010).

63. Shi, Y. & Jiang, H. rSeqDiff: Detecting Differential Isoform Expression from RNA-Seq Data Using Hierarchical Likelihood Ratio Test. *PLoS One* **8,** e79448 (2013).

64. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* **10,** 1185–1191 (2013).

65. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22,** 2008–2017 (2012).

66. Wang, W., Qin, Z., Feng, Z., Wang, X. & Zhang, X. Identifying differentially

spliced genes from two groups of RNA-seq samples. *Gene* **518,** 164–170 (2013).

67. Shen, S. *et al.* rMATS : Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci.* **111,** 5593–5601 (2014).

68. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).

69. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

70. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7,** 1009–1015 (2010).

71. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).

72. Roberts, A. & Pachter, L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Meth* **10,** 71–73 (2013).

73. Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotech* **32,** 462–464 (2014).

74. Angelini, C., Canditiis, D. De & Feis, I. De. Computational approaches for isoform detection and estimation: good and bad news. *BMC Bioinformatics* **15,** 135 (2014).

75. Lu, J., Tomfohr, J. K. & Kepler, T. B. Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics* **6,** 165 (2005).

76. Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26,** 493–500 (2010).

77. Lu, J., Tomfohr, J. K. & Kepler, T. B. Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics* **6,** 165 (2005).

78. Robinson, M. D. & Smyth, G. K. Moderated statistical tests for assessing

differences in tag abundance. *Bioinforma.* **23,** 2881–2887 (2007).

79. Ramoni, M. F., Sebastiani, P. & Kohane, I. S. Cluster analysis of gene expression dynamics. *Proc. Natl. Acad. Sci.* **99,** 9121–9126 (2002).

80. Schliep, A., Schönhuth, A. & Steinhoff, C. Using hidden Markov models to analyze gene expression time course data. *Bioinformatics* **19,** i255–i263 (2003).

81. Kim, S. Y., Imoto, S. & Miyano, S. Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Brief. Bioinform.* **4,** 228–235 (2003).

82. Holter, N. S., Maritan, A., Cieplak, M., Fedoroff, N. V & Banavar, J. R. Dynamic modeling of gene expression data. *Proc. Natl. Acad. Sci.* **98,** 1693–1698 (2001).

83. Chen, T. Modeling gene expression with differential equations. *Pacific Symp. Biocomput.* **4,** 4 (1999).

84. Brainspan consortium. BrainSpan: Atlas of the Developing Human Brain. (2011). at <http://developinghumanbrain.org>

85. Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res.* **40,** 84–90 (2012).

86. Habegger, L. *et al.* RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* **27,** 281–283 (2011).

87. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

88. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).

89. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930 (2014).

90. Stiles, J. & Jernigan, T. L. The Basics of Brain Development. *Neuropsychol. Rev.* **20,** 327–348 (2010).

91. Tau, G. Z. & Peterson, B. S. Normal Development of Brain Circuits. *Neuropsychopharmacology* **35,** 147–168 (2010).

92. Oliphant, T. E. Python for Scientific Computing. *Comput. Sci. Eng.* **9,** 10–20

(2007).

93. van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **13,** 22–30 (2011).

94. McKinney, W. Data Structures for Statistical Computing in Python. in *Proceedings of the 9th Python in Science Conference* (eds. van der Walt, S. & Millman, J.) 51–56 (2010).

95. Jones, E., Oliphant, T., Peterson, P. & others. SciPy: Open source scientific tools for Python. at <http://www.scipy.org/>

96. Coleman, T. F. & Li, Y. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optim.* **6,** 418–445 (1996).

97. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43,** D662–D669 (2015).

98. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinforma.* **27,** 3423–3424 (2011).

99. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma.* **26,** 841–842 (2010).

100. Kaufman, L. & Rousseeuw, P. J. *Finding groups in data: an introduction to cluster analysis*. **344,** (John Wiley & Sons, 2009).

101. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12,** 656–664 (2002).

102. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).

103. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).

104. Liu, B. P., Fournier, A., GrandPre, T. & Strittmatter, S. M. Myelin-associated glycoprotein as a functional ligand for the Nogo-66 receptor. *Science* **297,** 1190–1193 (2002).

105. Wang, T., Xiong, J.-Q., Ren, X.-B. & Sun, W. The role of Nogo-A in neuroregeneration: a review. *Brain Res. Bull.* **87,** 499–503 (2012).

106. Wang, X. *et al.* Localization of Nogo-A and Nogo-66 receptor proteins at sites

of axon--myelin and synaptic contact. *J. Neurosci.* **22,** 5505–5515 (2002).

107. Zhou, X. *et al.* Exploring long-range genome interaction data using the WashU Epigenome Browser. *Nat. Methods* **10,** 10.1038/nmeth.2440 (2013).

108. Lindwall, G. & Cole, R. D. Phosphorylation affects the ability of tau protein to promote microtubule assembly. *J. Biol. Chem.* **259,** 5301–5305 (1984).

109. Lovestone, S. *et al.* Alzheimer's disease-like phosphorylation of the microtubule-associated protein tau by glycogen synthase kinase-3 in transfected mammalian cells. *Curr. Biol.* **4,** 1077–1086 (1994).

110. Brosnan, J. T. Interorgan amino acid transport and its regulation. *J. Nutr.* **133,** 2068S--2072S (2003).

111. Shen, J. *Modeling the glutamate--glutamine neurotransmitter cycle*. *Transcellular Cycles underlying Neurotransmission* (Frontiers Media SA, 2015).

112. Cardona, C. *et al.* Expression of Gls and Gls2 glutaminase isoforms in astrocytes. *Glia* **63,** 365–382 (2015).

113. Wang, Y., Huang, Y., Zhao, L., Li, Y. & Zheng, J. Glutaminase 1 is essential for the differentiation, proliferation, and survival of human neural progenitor cells. *Stem Cells Dev.* **23,** 2782–2790 (2014).

114. Li, Y. & Xie, X. A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues. *BMC Bioinformatics* **14,** S11 (2013).

115. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12,** 453–457 (2015).

116. Mohammadi, S., Zuckerman, N., Goldsmith, A. & Grama, A. A Critical Survey of Deconvolution Methods for Separating cell-types in Complex Tissues. *arXiv Prepr. arXiv1510.04583* (2015).

117. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476 (2008).

118. Yeo, G., Holste, D., Kreiman, G. & Burge, C. B. Variation in alternative splicing across human tissues. *Genome Biol* **5,** R74 (2004).

7. Appendix

Table A1: Developmentally regulated exons predicted by Dmeans, Dmedian, Dmax, Zscores and Dual-Peak

Exons predicted and validated using the Dmeans, Dmedian, Dmax, Zscores and the dual-peak queries methods as detailed in Sections 4.4.2 through 4.5.2. Locations are given with regards to the hg19 genome assembly coordinates.

Table A1

Gene	Exon Location	A	mu	sigma	NRMSE	Cassette	Delta PSI	RHO	Simple	Dmax	Dmean	Dmedian	Zscore
PPHLN1	chr12:42778741-42778798	7.46	6.78	3.35	0.61	Y	0.35	2.13	Y	Y	Y	Y	Y
PTPRD	chr9:8437197-8437239	5.54	7.02	3.19	0.93	Y	0.27	2.81	Y	Y	Y	Y	Y
PRPF18	chr10:13639644-13639671	2.60	6.95	3.01	0.93	Y	0.17	1.11	Y	Y	Y	Y	Y
PHLDB1	chr11:118492077-118492253	4.98	0.00	2.93	0.71	N	0.52	1.59	Y	Y	Y	Y	Y
ADAM23	chr2:207470513-207470604	2.69	2.78	1.41	0.78	Y	0.25	2.67	Y	Y	Y	Y	Y
EPB41L1	chr20:34680601-34680735	5.95	2.77	0.91	0.89	N	0.07	2.52	Y	Y	Y	Y	Y
PFKP	chr10:3167288-3167441	5.50	3.19	0.87	0.95	Y	0.15	9.01	Y	Y	Y	Y	Y
KCNMA1	chr10:78761164-78761338	2.78	2.78	0.79	0.88	Y	0.63	3.11	Y	Y	Y	Y	Y
SCN2A	chr2:166165674-166165766	8.45	2.73	0.67	0.88	Y	0.43	1.50	Y	Y	Y	Y	Y
DNAJC6	chr1:65775227-65775621	10.05	2.68	0.66	0.65	N	0.37	2.77	Y	Y	Y	Y	Y
NRCAM	chr7:107866088-107866145	49.49	2.45	0.38	0.81	Y	0.19	0.50	Y	Y	Y	Y	Y
APBB2	chr4:40859001-40859209	128.73	0.00	1.27	0.98	N	0.01	1.92	N	Y	Y	Y	Y
DLG4	chr17:7108264-7108487	7.92	2.66	0.84	0.68	N	0.13	2.12	N	Y	Y	Y	Y
MADD	chr11:47290926-47291195	2.53	2.61	0.72	0.76	N	0.09	1.53	N	Y	Y	Y	Y
GAS7	chr17:10101524-10101868	5.36	2.56	0.40	0.87	N	0.00	1.11	N	Y	Y	Y	Y
SIPA1L1	chr14:72052997-72053645	15.44	2.52	0.03	0.90	N	0.11	1.07	N	Y	Y	Y	Y
DMD	chrX:31144758-31144820	2.55	7.12	3.83	0.74	Y	0.43	1.82	Y	N	Y	Y	N
WDR37	chr10:1102797-1102908	7.33	0.00	3.09	0.78	N	0.42	1.61	N	Y	Y	N	Y
CDK14	chr7:90225810-90226032	13.49	0.00	2.66	0.67	N	0.24	1.57	N	N	Y	N	Y
SEPT5	chr22:19701986-19702154	32.23	2.77	1.50	0.68	N	0.00	1.64	N	N	Y	Y	N
CRTC2	chr1:153921309-153921376	4.96	3.02	0.72	0.99	Y	0.56	3.62	N	N	Y	Y	N
FXR1	chr3:180693100-180693192	23.50	0.00	8.98	0.40	Y	0.08	0.27	Y	N	N	N	N
GRIA1	chr5:153174180-153174295	20.07	10.00	8.70	0.70	Y	0.40	1.83	Y	N	N	N	N
PPHLN1	chr12:42726461-42726504	2.93	0.00	8.65	0.76	Y	0.05	0.19	Y	N	N	N	N
RAPGEF1	chr9:134494437-134494596	17.68	0.00	8.61	0.61	Y	0.16	0.52	Y	N	N	N	N
NARF	chr17:80441060-80441180	1.53	10.00	8.52	0.81	Y	0.08	1.68	Y	N	N	N	N
NBR1	chr17:41322510-41322960	1.80	10.00	8.49	0.53	N	0.00	1.39	Y	N	N	N	N
ZNF138	chr7:64291317-64291454	2.15	6.81	8.44	0.68	Y	0.26	1.25	Y	N	N	N	N
WNK1	chr12:988738-989197	7.96	0.00	8.26	0.68	Y	0.38	1.48	Y	N	N	N	N
XPA	chr9:100438052-100438237	1.88	10.00	8.23	0.92	N	0.01	1.19	Y	N	N	N	N
CCDC65	chr12:49312467-49312686	7.68	10.00	8.23	0.61	Y	0.09	0.04	Y	N	N	N	N
NFASC	chr1:204946808-204946853	10.68	0.00	8.18	0.79	Y	0.25	0.86	Y	N	N	N	N
ZC4H2	chrX:64254508-64254593	2.01	10.00	8.12	0.94	N	0.00	2.61	Y	N	N	N	N
CRTC1	chr19:18885709-18885796	10.00	10.00	7.99	0.56	Y	0.37	0.90	Y	N	N	N	N
G3BP2	chr4:76579166-76579265	66.32	10.00	7.92	0.66	Y	0.33	0.85	Y	N	N	N	N
FAM49B	chr8:130891634-130891717	43.18	0.00	7.87	0.51	Y	0.17	0.49	Y	N	N	N	N
KALRN	chr3:123881555-123883533	1.72	10.00	7.84	0.53	N	0.64	1.44	Y	N	N	N	N
PTPRG	chr3:62216898-62216985	1.89	6.88	7.70	0.84	Y	0.26	0.82	Y	N	N	N	N
TSC2	chr16:2132436-2132505	4.08	10.00	7.69	0.75	Y	0.19	1.36	Y	N	N	N	N
CCDC136	chr7:128455667-128455985	5.83	0.00	7.60	0.73	Y	0.37	2.36	Y	N	N	N	N
NFASC	chr1:204970297-204970414	9.98	10.00	7.59	0.84	Y	0.19	1.54	Y	N	N	N	N
FAM49B	chr8:130892621-130892704	14.87	9.20	7.48	0.71	Y	0.05	0.37	Y	N	N	N	N
WNK1	chr12:980430-980514	16.69	10.00	7.44	0.57	Y	0.17	0.78	Y	N	N	N	N
TTLL5	chr14:76203908-76203950	1.77	10.00	7.43	0.95	Y	0.27	1.79	Y	N	N	N	N
MTMR1	chrX:149882950-149883004	3.50	10.00	7.35	0.80	Y	0.41	1.51	Y	N	N	N	N
NFASC	chr1:204971723-204971876	13.05	10.00	7.25	0.83	Y	0.17	1.53	Y	N	N	N	N
SLC6A9	chr1:44463526-44463697	7.85	7.14	7.23	0.71	Y	0.06	0.10	Y	N	N	N	N
SORBS1	chr10:97131740-97131806	3.14	10.00	7.19	0.89	Y	0.10	0.55	Y	N	N	N	N
NRXN1	chr2:50573828-50574892	10.67	0.00	7.13	0.89	N	0.22	2.33	Y	N	N	N	N
MTRR	chr5:7873485-7873639	6.35	10.00	7.09	0.52	Y	0.22	0.11	Y	N	N	N	N
PTPRD	chr9:8499646-8499840	4.03	1.88	7.07	0.88	Y	0.11	0.61	Y	N	N	N	N
PTPRD	chr9:9734532-9734571	7.00	0.00	7.03	0.94	Y	0.11	0.10	Y	N	N	N	N
ZNF827	chr4:146684241-146684274	2.13	6.82	6.98	0.90	Y	0.49	0.56	Y	N	N	N	N

Table A1

Gene	Exon Location	A	mu	sigma	NRMSE	Cassette	Delta PSI	RHO	Simple	Dmax	Dmean	Dmedian	Zscore
LIN7A	chr12:81241877-81241954	2.19	10.00	6.94	0.94	Y	0.10	0.52	Y	N	N	N	N
SHC4	chr15:49143384-49143445	2.87	10.00	6.87	0.83	Y	0.07	0.11	Y	N	N	N	N
ATP8A1	chr4:42592827-42592901	12.24	10.00	6.77	0.90	Y	0.31	0.89	Y	N	N	N	N
MTRR	chr5:7856899-7857005	1.88	0.00	6.76	0.88	Y	0.09	0.91	Y	N	N	N	N
CCDC136	chr7:128452761-128452983	6.68	0.00	6.71	0.77	Y	0.49	2.46	Y	N	N	N	N
PFKP	chr10:3162083-3162236	137.49	10.00	6.70	0.57	Y	0.11	0.20	Y	N	N	N	N
GPHN	chr14:67452398-67452455	5.81	10.00	6.57	0.76	Y	0.30	1.98	Y	N	N	N	N
SOS1	chr2:39216410-39216455	3.36	10.00	6.56	0.79	Y	0.06	0.44	Y	N	N	N	N
MYO18A	chr17:27406744-27406794	4.50	10.00	6.50	0.78	N	0.06	3.00	Y	N	N	N	N
KIAA1841	chr2:61300592-61300686	6.05	0.00	6.50	0.63	Y	0.07	0.34	Y	N	N	N	N
SEC16A	chr9:139339503-139339563	6.78	10.00	6.47	0.66	Y	0.54	1.21	Y	N	N	N	N
PRDM2	chr1:14095533-14095668	2.02	10.00	6.45	0.84	Y	0.15	0.02	Y	N	N	N	N
RP11-48B14.2	chr17:3567488-3567566	48.85	0.00	6.28	0.44	Y	0.06	0.22	Y	N	N	N	N
RAPGEF1	chr9:134479347-134479440	3.12	10.00	6.24	0.72	Y	0.11	1.38	Y	N	N	N	N
ATP8A1	chr4:42596305-42596379	12.42	0.00	6.22	0.82	Y	0.38	1.15	Y	N	N	N	N
CCDC136	chr7:128451853-128452366	5.83	0.00	6.19	0.92	Y	0.48	2.38	Y	N	N	N	N
RGS7	chr1:240964754-240964808	4.38	0.00	6.16	0.85	Y	0.30	4.70	Y	N	N	N	N
CADPS	chr3:62499312-62499381	13.77	0.00	6.08	0.85	Y	0.06	0.60	Y	N	N	N	N
RP11-48B14.2	chr17:3594249-3594321	5.50	10.00	6.04	0.90	Y	0.06	2.86	Y	N	N	N	N
MACF1	chr1:39925489-39925504	4.24	10.00	5.98	0.87	Y	0.07	1.93	Y	N	N	N	N
MKL1	chr22:40929593-40929815	2.29	10.00	5.95	0.94	N	0.14	4.68	Y	N	N	N	N
KIAA0226	chr3:197417944-197418019	4.06	10.00	5.85	0.81	Y	0.19	1.63	Y	N	N	N	N
KIAA1841	chr2:61386452-61386574	2.82	10.00	5.76	0.79	Y	0.06	1.52	Y	N	N	N	N
SORBS1	chr10:97131082-97131184	9.67	10.00	5.72	0.93	Y	0.14	0.92	Y	N	N	N	N
NRCAM	chr7:107808721-107808847	22.93	10.00	5.63	0.74	Y	0.06	1.06	Y	N	N	N	N
KIAA1841	chr2:61384996-61385147	9.40	10.00	5.62	0.61	Y	0.13	2.12	Y	N	N	N	N
NRCAM	chr7:107807365-107807518	20.16	10.00	5.54	0.73	Y	0.08	1.36	Y	N	N	N	N
SORBS1	chr10:97096277-97097051	6.90	10.00	5.41	0.83	Y	0.06	3.62	Y	N	N	N	N
CDC42BPA	chr1:227406963-227407121	2.28	0.00	5.40	0.85	Y	0.16	0.88	Y	N	N	N	N
VLDLR	chr9:2651414-2651498	3.65	8.28	5.32	0.82	Y	0.21	1.26	Y	N	N	N	N
KIDINS220	chr2:8865407-8865729	5.65	10.00	5.29	0.73	N	0.00	5.47	Y	N	N	N	N
CADPS	chr3:62498425-62498443	31.99	10.00	5.27	0.89	Y	0.58	2.07	Y	N	N	N	N
SORBS1	chr10:97110965-97111133	21.68	8.67	5.26	0.58	Y	0.57	2.09	Y	N	N	N	N
KCNQ2	chr20:62043127-62043235	5.06	10.00	5.06	0.78	Y	0.07	3.15	Y	N	N	N	N
ZFR2	chr19:3851811-3852612	6.56	2.97	4.97	0.70	N	0.15	1.63	Y	N	N	N	N
ADAM23	chr2:207474633-207474724	42.63	10.00	4.92	0.69	Y	0.45	1.39	Y	N	N	N	N
LRRFIP1	chr2:238628165-238628210	9.00	9.61	4.90	0.76	Y	0.58	2.80	Y	N	N	N	N
PACS2	chr14:105852021-105852054	29.31	10.00	4.90	0.54	Y	0.60	4.12	Y	N	N	N	N
KIDINS220	chr2:8867016-8867073	10.76	10.00	4.84	0.74	Y	0.06	5.50	Y	N	N	N	N
KCNQ2	chr20:62062692-62062722	25.90	8.24	4.74	0.69	Y	0.78	2.82	Y	N	N	N	N
DCLK2	chr4:151120179-151120230	23.21	10.00	4.74	0.70	Y	0.40	4.68	Y	N	N	N	N
SORBS1	chr10:97135729-97135813	10.59	7.64	4.62	0.69	Y	0.56	2.04	Y	N	N	N	N
DCLK2	chr4:151174625-151174708	12.70	0.00	4.51	0.69	Y	0.12	8.18	Y	N	N	N	N
PALM	chr19:740351-740483	39.33	9.13	4.49	0.46	Y	0.17	1.86	Y	N	N	N	N
LRRFIP1	chr2:238626402-238626452	7.89	9.24	4.40	0.76	Y	0.65	3.68	Y	N	N	N	N
SLC22A23	chr6:3285169-3285193	9.49	10.00	4.31	0.85	Y	0.42	4.69	Y	N	N	N	N
SH3KBP1	chrX:19705734-19705809	15.45	9.23	4.25	0.85	Y	0.69	4.67	Y	N	N	N	N
NRXN1	chr2:50201105-50201341	4.61	9.09	4.19	0.88	N	0.04	1.21	Y	N	N	N	N
LRRFIP1	chr2:238622901-238622919	5.29	8.89	4.14	0.83	Y	0.59	2.97	Y	N	N	N	N
EPB41L3	chr18:5630375-5630640	3.60	7.21	4.05	0.97	N	0.00	1.44	Y	N	N	N	N
R3HDM2	chr12:57682639-57682693	15.03	7.04	3.96	0.80	Y	0.63	3.99	Y	N	N	N	N
KCNMA1	chr10:78673814-78673895	22.70	10.00	3.95	0.83	Y	0.76	7.72	Y	N	N	N	N
DNAJC6	chr1:65730423-65730615	10.76	8.26	3.88	0.89	N	0.30	2.32	Y	N	N	N	N

Table A1

Gene	Exon Location	A	mu	sigma	NRMSE	Cassette	Delta PSI	RHO	Simple	Dmax	Dmean	Dmedian	Zscore
LRRFIP1	chr2:238636518-238636578	4.31	7.74	3.65	0.87	Y	0.43	2.38	Y	N	N	N	N
MYO18A	chr17:27412621-27412666	7.23	7.07	3.58	0.80	Y	0.13	0.74	Y	N	N	N	N
MAGI1	chr3:65344738-65344827	8.08	6.96	3.47	0.86	Y	0.48	1.93	Y	N	N	N	N
MYH10	chr17:8433940-8434003	18.09	10.00	3.36	0.81	Y	0.41	11.44	Y	N	N	N	N
BTBD10	chr11:13461430-13461826	2.95	6.74	3.32	0.94	N	0.04	3.07	Y	N	N	N	N
SH3KBP1	chrX:19713112-19713169	4.52	6.95	3.14	0.92	Y	0.27	2.97	Y	N	N	N	N
KALRN	chr3:124237246-124238697	45.50	6.86	3.10	0.66	N	0.63	5.14	Y	N	N	N	N
EPB41L1	chr20:34793770-34795702	8.61	10.00	3.09	0.92	N	0.04	7.92	Y	N	N	N	N
NFASC	chr1:204931235-204931286	4.47	7.15	2.74	0.99	Y	0.16	2.07	Y	N	N	N	N
NCAM1	chr11:113133567-113135919	48.69	6.90	2.72	0.64	N	0.25	5.09	Y	N	N	N	N
CDC42BPA	chr1:227198578-227198764	4.18	7.21	2.14	0.74	Y	0.23	9.50	Y	N	N	N	N
FAM107B	chr10:14646238-14646417	15.60	0.00	2.11	0.93	N	0.01	1.27	Y	N	N	N	N
FAM107B	chr10:14613761-14614188	4.20	6.75	1.65	0.80	N	0.97	11.52	Y	N	N	N	N
GRIA1	chr5:153175035-153175150	45.12	2.73	1.36	0.82	Y	0.32	0.67	Y	N	N	N	N
MAPT	chr17:44091608-44091690	247.69	2.93	1.27	0.71	Y	0.06	1.31	Y	N	N	N	N
MACF1	chr1:39930766-39930784	40.52	2.96	1.08	0.74	Y	0.29	0.54	Y	N	N	N	N
VLDLR	chr9:2641376-2641499	15.97	2.64	0.59	0.77	Y	0.05	0.45	Y	N	N	N	N
SORBS1	chr10:97082502-97082562	3.06	2.73	0.54	0.94	Y	0.15	10.28	Y	N	N	N	N
PTPRD	chr9:8454579-8454594	12.63	2.60	0.50	0.84	Y	0.21	0.74	Y	N	N	N	N
PTPRD	chr9:8523512-8523524	14.79	2.63	0.49	0.78	Y	0.21	0.54	Y	N	N	N	N
PTPRD	chr9:8526626-8526644	18.40	2.61	0.47	0.79	Y	0.28	0.59	Y	N	N	N	N
DLG2	chr11:83393200-83393468	15.28	2.56	0.38	0.76	N	0.50	3.82	Y	N	N	N	N
EPB41L3	chr18:5540338-5540555	89.38	2.52	0.03	0.84	N	0.28	2.34	Y	N	N	N	N

Table A2

Exons predicted and validated using the Hierarchical clustering method detailed in Sections 4.5.3. Locations are given with regards to the hg19 genome assembly coordinates.

Table A2

Gene	Location	A	mu	sigma	NRMSE	Cassette	Delta PSI	RHO fold change
EPB41L3	chr18:5540338-5540555	89.38	2.52	0.03	0.84	N	0.28	2.34
TIAM2	chr6:155411422-155411513	66.53	2.53	0.03	0.86	N	0.00	4.28
RBFOX1	chr16:6533179-6533545	24.22	2.52	0.03	0.83	N	0.15	1.92
CADPS	chr3:62516328-62516487	48.54	2.14	0.04	1.17	Y	0.10	1.17
CCDC136	chr7:128431463-128431605	32.76	4.24	0.17	1.52	N	0.50	1.17
PEX26	chr22:18606922-18607071	55.61	4.25	0.28	1.24	N	0.02	2.27
TIAM2	chr6:155450305-155451551	18.95	2.53	0.36	0.71	N	0.00	2.90
TIAM2	chr6:155532337-155532441	32.84	2.53	0.39	0.71	N	0.00	2.49
TIAM2	chr6:155504370-155504634	30.67	2.55	0.39	0.69	N	0.00	1.90
PDE4D	chr5:58481014-58481088	11.32	2.56	0.46	0.81	Y	0.06	0.26
MTUS1	chr8:17579220-17579730	4.60	2.46	0.52	1.10	N	0.61	2.40
ANK3	chr10:61871491-61871524	5.49	2.51	0.53	0.83	N	0.35	3.61
SORBS1	chr10:97082502-97082562	3.06	2.73	0.54	0.94	Y	0.15	10.28
TSC2	chr16:2127598-2127727	21.67	2.69	0.72	0.47	N	0.55	2.69
MAPT	chr17:43971747-43972052	52.14	2.80	0.74	0.74	N	0.00	3.65
HECW1	chr7:43152197-43152536	4.86	2.36	0.80	0.74	N	0.00	2.23
ANK2	chr4:114244914-114244950	9.03	7.23	0.84	0.86	Y	0.05	0.94
MCF2L	chr13:113656030-113656297	3.44	2.79	0.86	0.75	N	0.21	1.88
CADPS	chr3:62479319-62479340	11.52	7.50	0.88	1.10	Y	0.12	2.19
ABR	chr17:982053-982241	16.11	2.37	0.89	0.48	N	0.24	5.46
NRXN1	chr2:50693598-50693625	12.97	2.89	0.90	0.81	Y	0.21	2.78
MAPT	chr17:44055740-44055806	244.53	2.86	1.02	0.71	N	0.00	2.43
PSD3	chr8:18793543-18793652	5.73	2.22	1.04	0.85	N	0.00	2.86
MACF1	chr1:39930766-39930784	40.52	2.96	1.08	0.74	Y	0.29	0.54
MAPT	chr17:44039686-44039836	186.58	2.87	1.09	0.72	N	0.00	2.45
RERE	chr1:8684368-8684439	21.55	2.49	1.12	0.69	N	0.01	1.53
TIAM2	chr6:155282262-155282613	3.02	2.74	1.16	0.70	N	0.06	2.35
RERE	chr1:8674619-8674745	12.05	2.47	1.18	0.68	Y	0.06	0.60
AC073479.1	chr2:6122109-6122204	1.96	2.52	1.23	0.84	N	0.71	12.82
TRIM2	chr4:154074269-154074422	4.33	2.03	1.25	0.65	N	0.37	3.19
MAPT	chr17:44091608-44091690	247.69	2.93	1.27	0.71	Y	0.06	1.31
APBB2	chr4:40859001-40859209	128.73	0.00	1.27	0.98	N	0.01	1.92
MAPT	chr17:44064405-44064461	194.83	2.82	1.29	0.69	N	0.00	2.54
GRIA1	chr5:153175035-153175150	45.12	2.73	1.36	0.82	Y	0.32	0.67
ABAT	chr16:8814572-8814791	47.39	0.00	1.43	1.23	N	0.17	2.24
HNRNPL	chr19:39340339-39340617	400.33	0.00	1.60	1.08	N	0.30	1.39
AC073479.1	chr2:6122405-6125029	2.85	2.63	1.64	0.70	N	0.22	3.52
FAM107B	chr10:14613761-14614188	4.20	6.75	1.65	0.80	N	0.97	11.52
NASP	chr1:46072992-46074009	216.58	0.00	1.94	0.57	N	0.19	1.37
ARAP1	chr11:72463372-72463448	2.35	5.32	2.02	1.74	N	0.63	2.42
PHLDB1	chr11:118478305-118478414	2.67	5.67	2.03	1.12	N	0.00	1.91
WASF3	chr13:27254171-27254338	24.76	5.77	2.08	0.99	Y	0.44	3.10
FAM107B	chr10:14646238-14646417	15.60	0.00	2.11	0.93	N	0.01	1.27
CDC42BPA	chr1:227198578-227198764	4.18	7.21	2.14	0.74	Y	0.23	9.50
ARAP1	chr11:72443559-72443642	2.52	5.38	2.23	1.25	Y	0.13	0.81
CDH23	chr10:73574708-73575702	2.74	5.19	2.25	1.83	N	0.00	1.02
KIAA0930	chr22:45607214-45607313	6.56	6.38	2.35	1.09	N	0.07	9.20
DLGAP1	chr18:3656083-3656113	26.42	6.96	2.43	1.18	Y	0.59	11.47
MTUS1	chr8:17554765-17555179	11.30	6.38	2.56	0.71	N	0.34	1.66
MACF1	chr1:39802853-39803003	2.03	5.69	2.58	1.20	Y	0.22	1.45
CDK14	chr7:90225810-90226032	13.49	0.00	2.66	0.67	N	0.24	1.57
SMARCA2	chr9:2017438-2017502	1.70	4.86	2.66	1.29	N	0.04	1.75
PAK1	chr11:77122818-77123109	16.01	6.20	2.70	1.09	N	0.37	1.04
FEZ1	chr11:125321306-125321416	2.66	6.39	2.74	0.79	N	0.01	1.90
NFASC	chr1:204931235-204931286	4.47	7.15	2.74	0.99	Y	0.16	2.07
HSD17B4	chr5:118812236-118812419	2.49	6.92	2.76	1.11	N	0.03	7.87
ANK3	chr10:61926348-61926411	5.02	7.43	2.77	1.10	Y	0.13	1.47
SORBS1	chr10:97154757-97154832	6.91	6.27	2.86	1.01	Y	0.55	4.17
SORBS1	chr10:97154367-97154430	1.89	6.41	2.90	1.13	Y	0.22	3.51
MAPT	chr17:44067243-44067441	2.09	5.64	2.91	0.97	Y	0.19	1.34
PHLDB1	chr11:118492077-118492253	4.98	0.00	2.93	0.71	N	0.52	1.59
RTN4	chr2:55237222-55237585	95.78	7.44	2.94	0.75	N	0.63	5.26
SORBS1	chr10:97194378-97194474	4.73	5.61	2.95	0.99	Y	0.25	1.56
GPC1	chr2:241394254-241394437	3.83	4.93	2.96	1.22	N	0.02	5.98

Table A2

Gene	Location	A	mu	sigma	NRMSE	Cassette	Delta PSI	RHO fold change
PSD3	chr8:18541283-18541631	13.63	6.43	2.97	1.06	N	0.54	1.74
NFASC	chr1:204955037-204955242	5.06	7.45	3.03	1.01	Y	0.16	2.66
AGAP3	chr7:150817606-150818120	29.13	7.43	3.03	0.40	N	0.29	1.77
KALRN	chr3:124237246-124238697	45.50	6.86	3.10	0.66	N	0.63	5.14
MAPT	chr17:44051750-44051837	8.07	10.00	3.17	0.90	N	0.04	1.72
NFASC	chr1:204953154-204953457	2.80	7.59	3.19	0.85	N	0.23	2.85
R3HDM1	chr2:136374237-136374327	21.74	5.37	3.21	0.91	Y	0.34	0.95
CLTA	chr9:36210621-36210658	21.94	6.13	3.25	0.65	Y	0.43	2.92
PDE4DIP	chr1:145039589-145040002	25.99	4.84	3.26	0.47	N	0.05	1.51
PHLDB1	chr11:118526569-118526602	34.93	8.54	3.28	0.78	Y	0.33	11.40
PHLDB1	chr11:118485273-118485909	2.48	7.25	3.30	0.74	N	0.00	1.75
PDE4DIP	chr1:145015683-145016011	54.10	5.59	3.33	0.43	N	0.00	1.37
GAD1	chr2:171672623-171672766	3.55	0.00	3.36	1.51	N	1.00	4.18
MYH10	chr17:8433940-8434003	18.09	10.00	3.36	0.81	Y	0.41	11.44
R3HDM1	chr2:136373721-136373763	18.93	5.51	3.38	0.94	Y	0.42	1.12
SEMA6A	chr5:115808580-115808631	5.92	4.30	3.42	1.16	Y	0.27	1.65
DLGAP1	chr18:3874147-3874325	7.27	7.43	3.47	0.98	N	0.33	2.42
MAGI1	chr3:65344738-65344827	8.08	6.96	3.47	0.86	Y	0.48	1.93
MAP4	chr3:47894652-47894842	64.29	5.87	3.48	0.60	Y	0.07	0.36
ABR	chr17:1082960-1083131	11.13	5.70	3.55	0.86	N	0.05	1.38
ABR	chr17:1028517-1028702	33.34	5.81	3.56	0.67	N	0.00	1.55
MYO18A	chr17:27412621-27412666	7.23	7.07	3.58	0.80	Y	0.13	0.74
WHSC2	chr4:2043442-2043630	2.12	5.10	3.62	0.73	N	0.00	2.23
LRRFIP1	chr2:238636518-238636578	4.31	7.74	3.65	0.87	Y	0.43	2.38
OXR1	chr8:107696473-107696587	47.03	8.39	3.66	0.79	N	0.00	1.44
MKL2	chr16:14280573-14280892	6.64	7.36	3.69	1.13	N	0.46	1.31
KIAA0528	chr12:22655672-22655738	8.61	10.00	3.77	0.87	Y	0.48	6.84
PDE4DIP	chr1:144994590-144995082	19.89	5.93	3.77	0.46	N	0.01	1.41
RP11-48B14.2	chr17:3582883-3583078	9.55	5.94	3.81	0.75	N	0.00	1.80
MAP4	chr3:48057823-48057902	10.20	10.00	3.83	0.89	N	0.18	10.24
PDE4D	chr5:59817877-59817947	11.06	10.00	3.85	1.07	N		3.69
DLGAP1	chr18:3845231-3845359	8.96	8.38	3.88	1.11	N	0.28	4.83
RERE	chr1:8877218-8877702	5.09	0.00	4.03	0.64	N	0.00	3.00
GS1-124K5.12	chr7:66049331-66049453	1.65	4.11	4.03	0.83	Y	0.15	0.87
EPB41L3	chr18:5630375-5630640	3.60	7.21	4.05	0.97	N	0.00	1.44
PTK2	chr8:142011223-142011478	3.16	0.00	4.06	0.74	N	0.00	1.21
SH3PXD2A	chr10:105452785-105452930	2.05	6.79	4.06	0.71	N	0.00	1.27
MAP4	chr3:47917174-47917390	81.23	6.05	4.14	0.53	Y	0.05	0.34
LRRFIP1	chr2:238622901-238622919	5.29	8.89	4.14	0.83	Y	0.59	2.97
ACTR2	chr2:65469105-65469197	4.81	10.00	4.15	0.76	Y	0.13	2.59
NRXN1	chr2:50201105-50201341	4.61	9.09	4.19	0.88	N	0.04	1.21
NRXN1	chr2:50848342-50848387	6.05	5.60	4.20	0.92	Y	0.08	0.19
OXR1	chr8:107718608-107719918	28.37	9.06	4.21	0.74	N	0.00	1.06
CDK14	chr7:90338564-90339271	3.29	8.85	4.21	0.85	N	0.51	2.04
ARFGAP1	chr20:61915202-61915232	12.46	7.23	4.22	0.67	Y	0.57	2.33
SLC22A23	chr6:3285169-3285193	9.49	10.00	4.31	0.85	Y	0.42	4.69
NFE2L1	chr17:46134393-46134483	7.21	7.05	4.34	0.62	Y	0.12	0.30
GOLT1B	chr12:21668171-21668204	3.59	7.06	4.36	0.95	Y	0.13	1.51
PLCH1	chr3:155203935-155203995	2.97	8.71	4.39	1.14	Y	0.80	3.33
LRRFIP1	chr2:238626402-238626452	7.89	9.24	4.40	0.76	Y	0.65	3.68
RNPS1	chr16:2317175-2317238	14.09	10.00	4.43	0.62	N	0.34	4.77
OXR1	chr8:107371703-107371864	28.09	10.00	4.43	0.86	N	0.00	1.18
WHSC2	chr4:2019391-2019469	2.16	6.15	4.43	0.80	Y	0.08	1.62
AC073479.1	chr2:6141114-6141407	2.86	0.00	4.48	0.74	N	0.00	3.15
NUMA1	chr11:71723304-71723488	5.65	9.58	4.49	0.57	N	0.31	1.67
OXR1	chr8:107715133-107715318	34.51	9.44	4.49	0.82	N	0.01	1.07
MAP4	chr3:47908735-47908828	81.11	6.69	4.49	0.45	Y	0.13	0.20
DCLK2	chr4:151174625-151174708	12.70	0.00	4.51	0.69	Y	0.12	8.18
GAD1	chr2:171699078-171699269	2.62	0.00	4.51	1.36	N	0.17	4.51
PACRGL	chr4:20754187-20754530	1.81	5.11	4.52	0.74	N	0.00	2.05
SF1	chr11:64544970-64545233	1.62	5.48	4.53	0.70	N	0.04	1.91
PEX26	chr22:18609120-18609801	46.41	9.19	4.54	0.83	N	0.00	1.25
PRDM11	chr11:45230424-45230758	3.95	8.59	4.54	0.53	Y	0.31	1.61
RP11-48B14.2	chr17:3591305-3591399	10.18	7.36	4.55	0.73	N	0.00	1.79

Table A2

Gene	Location	A	mu	sigma	NRMSE	Cassette	Delta PSI	RHO fold change
PDE4DIP	chr1:144871695-144871881	19.81	10.00	4.57	0.78	Y	0.46	3.57
RERE	chr1:8716031-8716500	15.76	0.00	4.57	0.58	N	0.00	1.53
ATG16L1	chr2:234182636-234182687	7.26	10.00	4.58	0.91	Y	0.32	1.81
SORBS1	chr10:97135729-97135813	10.59	7.64	4.62	0.69	Y	0.56	2.04
AC024560.3	chr3:197338442-197338493	4.64	6.21	4.62	1.02	N	0.00	4.59
ARNT2	chr15:80735177-80735366	10.10	10.00	4.64	0.90	Y	0.35	3.26
ZNF273	chr7:64353789-64353834	3.97	10.00	4.67	1.21	Y	0.07	1.48
ANKS1B	chr12:99201621-99201696	35.52	10.00	4.71	0.79	Y	0.52	6.76
DCLK2	chr4:151120179-151120230	23.21	10.00	4.74	0.70	Y	0.40	4.68
KCNQ2	chr20:62062692-62062722	25.90	8.24	4.74	0.69	Y	0.78	2.82
SRRM3	chr7:75902370-75902913	7.00	10.00	4.75	1.36	N	0.06	1.23
CIZ1	chr9:130953558-130953868	3.19	0.00	4.75	0.83	N	0.36	1.99
SRRM3	chr7:75911931-75912190	2.20	4.76	4.75	1.22	N	0.00	1.05
ANK3	chr10:61867945-61868044	3.82	10.00	4.82	1.00	Y	0.16	2.28
XPNPEP3	chr22:41322272-41328819	1.38	6.29	4.82	0.47	N	0.01	1.23
TBC1D24	chr16:2547710-2547728	12.83	10.00	4.85	0.77	Y	0.56	2.18
EPB41L3	chr18:5415816-5416377	17.29	10.00	4.87	0.72	N	0.06	1.03
ATP1B3	chr3:141594965-141595236	11.51	0.00	4.87	0.60	N		3.01
POLR2F	chr22:38437074-38437113	11.34	10.00	4.88	0.65	N	0.00	4.26
SRPK2	chr7:105029690-105029838	2.14	5.30	4.89	0.90	N	0.00	1.42
PACS2	chr14:105852021-105852054	29.31	10.00	4.90	0.54	Y	0.60	4.12
LRRFIP1	chr2:238628165-238628210	9.00	9.61	4.90	0.76	Y	0.58	2.80
TAOK3	chr12:118677031-118677077	3.67	7.50	4.92	0.86	Y	0.06	0.72
KIAA1841	chr2:61390186-61390367	37.15	8.51	4.92	0.37	N	0.01	1.99
LRRFIP2	chr3:37132957-37133029	7.94	7.99	4.96	0.62	Y	0.69	2.53
SND1	chr7:127637524-127638129	3.02	10.00	4.97	0.57	N	0.01	1.74
GIT2	chr12:110383064-110383154	2.38	7.76	4.98	0.82	Y	0.44	1.44
ATP1B3	chr3:141620977-141621069	32.20	10.00	5.01	0.67	Y	0.35	2.40
GRIA3	chrX:122599524-122599639	11.18	10.00	5.05	0.66	Y	0.27	1.44
ANKMY1	chr2:241418838-241419072	1.44	6.62	5.06	0.84	N	0.00	1.19
KCNQ2	chr20:62043127-62043235	5.06	10.00	5.06	0.78	Y	0.07	3.15
INPP4A	chr2:99198038-99198284	5.46	10.00	5.07	0.72	N	0.31	2.08
PACRGL	chr4:20703763-20703844	2.17	6.60	5.09	0.81	N	0.32	1.47
ATP9B	chr18:77137246-77138278	6.72	7.30	5.19	0.39	N	0.00	1.46
AC073479.1	chr2:6125556-6125850	5.49	0.00	5.19	0.68	N	0.02	1.25
DLGAP1	chr18:3496029-3499392	52.95	8.38	5.20	0.60	N	0.00	1.37
PSD3	chr8:18662213-18662408	9.93	0.00	5.25	0.68	N	0.00	2.18
SORBS1	chr10:97110965-97111133	21.68	8.67	5.26	0.58	Y	0.57	2.09
CADPS	chr3:62498425-62498443	31.99	10.00	5.27	0.89	Y	0.58	2.07
PDE4DIP	chr1:144942179-144942690	1.74	6.33	5.30	0.83	N	0.00	1.71
PICALM	chr11:85689112-85689758	2.90	10.00	5.32	0.52	N	0.48	1.91
CDC42BPA	chr1:227406963-227407121	2.28	0.00	5.40	0.85	Y	0.16	0.88
SORBS1	chr10:97096277-97097051	6.90	10.00	5.41	0.83	Y	0.06	3.62
MKL2	chr16:14173144-14173211	4.18	0.00	5.42	1.10	N	0.00	1.77
AGAP3	chr7:150831486-150831655	44.36	0.00	5.43	0.47	N	0.00	2.09
MAP2	chr2:210555326-210555572	2.45	10.00	5.44	0.84	N	0.02	3.82
AGFG1	chr2:228414774-228414822	20.72	10.00	5.45	0.72	Y	0.37	2.87
DLGAP1	chr18:3879111-3880140	14.31	0.00	5.48	0.74	N	0.00	1.34
GLS	chr2:191797382-191800015	5.25	0.00	5.48	0.48	N	0.45	2.59
OXR1	chr8:107749747-107749828	62.41	10.00	5.49	0.63	Y	0.20	0.23
AGAP3	chr7:150835229-150835400	13.17	0.00	5.50	0.51	N	0.08	2.18
ARHGAP21	chr10:24879124-24879408	28.41	10.00	5.58	0.74	Y	0.38	2.25
RTN4	chr2:55276880-55277734	82.38	0.00	5.60	0.59	N	0.02	1.14
AGAP3	chr7:150825419-150825771	18.85	0.00	5.61	0.54	N	0.00	2.15
PDE4DIP	chr1:144930583-144932552	15.23	0.00	5.62	0.67	N	0.32	1.11
KIAA1841	chr2:61384996-61385147	9.40	10.00	5.62	0.61	Y	0.13	2.12
SORBS1	chr10:97131082-97131184	9.67	10.00	5.72	0.93	Y	0.14	0.92
TCOF1	chr5:149763300-149763816	3.77	10.00	5.73	0.52	N	0.25	2.15
RP11-539I5.1	chr10:118592511-118592884	1.93	0.00	5.74	0.67	N	0.00	1.14
KIAA1841	chr2:61386452-61386574	2.82	10.00	5.76	0.79	Y	0.06	1.52
N4BP2L2	chr13:33112754-33112970	9.86	0.00	5.78	0.51	N	0.00	2.33
C7orf50	chr7:1166892-1167024	19.68	0.00	5.80	0.44	N	0.00	1.19
TIAM2	chr6:155228863-155230002	4.89	0.00	5.83	0.38	N	0.07	2.80
KIAA0226	chr3:197417944-197418019	4.06	10.00	5.85	0.81	Y	0.19	1.63

Table A2

Gene	Location	A	mu	sigma	NRMSE	Cassette	Delta PSI	RHO fold change
ZYX	chr7:143084854-143084880	3.99	10.00	5.87	1.08	Y	0.24	3.82
PSD3	chr8:18622958-18623048	8.08	0.00	5.87	0.74	N	0.00	1.42
R3HDM1	chr2:136289024-136289203	4.61	0.00	5.91	0.67	N	0.00	1.86
AGAP3	chr7:150840823-150841523	59.79	1.82	5.92	0.40	N	0.00	1.54
MKL1	chr22:40929593-40929815	2.29	10.00	5.95	0.94	N	0.14	4.68
KALRN	chr3:124397036-124397484	2.35	0.00	5.97	0.80	N	0.01	1.07
MACF1	chr1:39925489-39925504	4.24	10.00	5.98	0.87	Y	0.07	1.93
ATP9B	chr18:77105427-77105572	5.68	6.43	5.99	0.53	N	0.00	1.16
RP11-48B14.2	chr17:3593898-3593974	6.53	9.03	6.00	0.83	N	0.00	2.38
DNM1L	chr12:32858758-32859036	2.44	10.00	6.00	0.73	N	0.37	2.17
ME3	chr11:86383315-86383678	3.30	0.57	6.04	0.66	N	0.00	1.35
RP11-48B14.2	chr17:3594249-3594321	5.50	10.00	6.04	0.90	Y	0.06	2.86
HECW1	chr7:43490426-43490528	14.09	0.00	6.05	0.70	Y	0.18	0.98
AGAP3	chr7:150836654-150839139	6.21	0.00	6.07	0.46	N	0.00	1.94
AMPD2	chr1:110162458-110162900	4.07	0.00	6.08	0.55	N	0.00	1.01
CADPS	chr3:62499312-62499381	13.77	0.00	6.08	0.85	Y	0.06	0.60
MARK2	chr11:63675731-63675776	10.98	7.34	6.11	0.47	Y	0.19	1.11
SMARCA2	chr9:2015341-2015404	6.94	0.00	6.12	0.87	N	0.00	1.62
RNH1	chr11:506608-506821	6.49	4.87	6.14	0.39	N	0.26	1.03
RGS7	chr1:240964754-240964808	4.38	0.00	6.16	0.85	Y	0.30	4.70
CCDC136	chr7:128454691-128454973	4.44	0.00	6.18	0.81	N	0.35	2.38
CCDC136	chr7:128451853-128452366	5.83	0.00	6.19	0.92	Y	0.48	2.38
CCDC136	chr7:128450192-128450420	6.18	0.00	6.20	0.78	N	0.44	2.65
PLXNB1	chr3:48470661-48470890	1.82	0.00	6.21	0.94	N	0.50	1.35
ATP8A1	chr4:42596305-42596379	12.42	0.00	6.22	0.82	Y	0.38	1.15
SCAF11	chr12:46354413-46355105	4.46	9.49	6.22	0.59	N	0.20	1.89
CHKA	chr11:67849957-67849987	3.11	10.00	6.24	1.17	Y	0.07	0.01
RAPGEF1	chr9:134479347-134479440	3.12	10.00	6.24	0.72	Y	0.11	1.38
SON	chr21:34944174-34944209	2.64	10.00	6.26	1.18	N	0.00	3.83
PAK1	chr11:77184596-77185107	4.19	0.00	6.27	0.73	N	0.03	1.19
RP11-48B14.2	chr17:3567488-3567566	48.85	0.00	6.28	0.44	Y	0.06	0.22
MORN1	chr1:2252691-2253018	3.10	10.00	6.32	1.38	N	0.00	1.92
CYTH1	chr17:76673076-76673132	5.04	10.00	6.32	0.78	N	0.09	1.35
OXR1	chr8:107738239-107738537	6.80	0.00	6.33	0.62	N	0.64	2.76
ANK3	chr10:61841907-61841934	8.50	0.00	6.36	1.12	Y	0.69	1.91
DDHD2	chr8:38130887-38130960	9.59	0.00	6.37	0.65	Y	0.11	7.61
MATR3	chr5:138611797-138611839	2.46	8.78	6.38	0.88	N	0.00	2.02
N4BP2L2	chr13:33109905-33111164	19.32	0.00	6.44	0.43	N	0.02	2.40
PRDM2	chr1:14095533-14095668	2.02	10.00	6.45	0.84	Y	0.15	0.02
MAPT	chr17:44095983-44096096	303.40	0.00	6.45	0.57	N	0.00	1.26
NCKAP1	chr2:183889705-183889723	27.01	9.21	6.45	0.74	Y	0.29	0.96
SEC16A	chr9:139339503-139339563	6.78	10.00	6.47	0.66	Y	0.54	1.21
FNIP1	chr5:131046270-131046354	3.58	0.00	6.48	0.78	Y	0.24	0.13
KALRN	chr3:124398304-124398596	1.75	0.00	6.48	0.84	N	0.00	1.07
NRXN1	chr2:51153075-51153093	14.51	0.00	6.49	0.71	Y	0.23	0.71
KIAA1841	chr2:61300592-61300686	6.05	0.00	6.50	0.63	Y	0.07	0.34
CHKA	chr11:67828959-67829297	2.82	10.00	6.50	0.85	N	0.16	1.38
MYO18A	chr17:27406744-27406794	4.50	10.00	6.50	0.78	N	0.06	3.00
PSD3	chr8:18661869-18662112	1.92	0.00	6.52	0.78	N	0.00	1.72
SOS1	chr2:39216410-39216455	3.36	10.00	6.56	0.79	Y	0.06	0.44
ATG13	chr11:46685546-46685699	10.26	10.00	6.56	0.52	N	0.23	1.28
GPHN	chr14:67452398-67452455	5.81	10.00	6.57	0.76	Y	0.30	1.98
MTMR2	chr11:95647405-95647476	2.68	0.00	6.60	0.74	Y	0.05	0.13
PPFIA2	chr12:82152178-82152580	5.71	0.00	6.65	0.62	N	0.00	1.53
PSD3	chr8:18658506-18658892	1.83	0.00	6.65	0.71	N	0.00	1.57
RP11-33B1.1	chr4:120464907-120465000	1.54	2.45	6.66	0.88	Y	0.06	0.66
MTRR	chr5:7862063-7862170	2.11	0.00	6.67	0.93	N	0.00	1.78
SHFM1	chr7:96279526-96279831	2.03	10.00	6.67	0.81	N	0.35	1.81