Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development

Sharmistha Pal, Ravi Gupta, Hyunsoo Kim, et al.

Genome Res. 2011 21: 1260-1272 originally published online June 28, 2011
Access the most recent version at doi:10.1101/gr.120535.111

Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development

Sharmistha Pal,[1,2,4] Ravi Gupta,[1,2,4] Hyunsoo Kim,[1] Priyankara Wickramasinghe,[1] Valérie Baubet,[2] Louise C. Showe,[1,2,3] Nadia Dahmane,[2] and Ramana V. Davuluri[1,2,5]

[1]*Center for Systems and Computational Biology, The Wistar Institute, Philadelphia, Pennsylvania 19019, USA;* [2]*Molecular and Cellular Oncogenesis Program, The Wistar Institute, Philadelphia, Pennsylvania 19019, USA;* [3]*Immunology Program, The Wistar Institute, Philadelphia, Pennsylvania 19019, USA*

Despite our growing knowledge that many mammalian genes generate multiple transcript variants that may encode functionally distinct protein isoforms, the transcriptomes of various tissues and their developmental stages are poorly defined. Identifying the transcriptome and its regulation in a cell/tissue is the key to deciphering the cell/tissue-specific functions of a gene. We built a genome-wide inventory of noncoding and protein-coding transcripts (transcriptomes), their promoters (promoteromes) and histone modification states (epigenomes) for developing, and adult cerebella using integrative massive-parallel sequencing and bioinformatics approach. The data consists of 61,525 (12,796 novel) distinct mRNAs transcribed by 29,589 (4792 novel) promoters corresponding to 15,669 protein-coding and 7624 noncoding genes. Importantly, our results show that the transcript variants from a gene are predominantly generated using alternative transcriptional rather than splicing mechanisms, highlighting alternative promoters and transcriptional terminations as major sources of transcriptome diversity. Moreover, H3K4me3, and not H3K27me3, defined the use of alternative promoters, and we identified a combinatorial role of H3K4me3 and H3K27me3 in regulating the expression of transcripts, including transcript variants of a gene during development. We observed a strong bias of both H3K4me3 and H3K27me3 for CpG-rich promoters and an exponential relationship between their enrichment and corresponding transcript expression. Furthermore, the majority of genes associated with neurological diseases expressed multiple transcripts through alternative promoters, and we demonstrated aberrant use of alternative promoters in medulloblastoma, cancer arising in the cerebellum. The transcriptomes of developing and adult cerebella presented in this study emphasize the importance of analyzing gene regulation and function at the isoform level.

[Supplemental material is available for this article.]

The flow of genomic information, from DNA to RNA to protein, is a highly complex process in mammalian cells (Moore and Proudfoot 2009). An important aspect of this complexity is the generation of alternative gene products from a single gene locus (e.g., *TP53, DSCAM, GPHN*), which can occur through transcriptional or post-transcriptional (splicing) mechanisms (Schmucker 2007; Hollstein and Hainaut 2010). The use of alternative transcriptional initiation and/or termination (transcriptional events) can give rise to different pre-mRNAs, which can further undergo alternative splicing (splicing events), leading to multiple mRNA/transcript variants from the same gene. Therefore, a gene can potentially yield an extensive array of gene products—alternative transcript (transcriptome) and alternative protein (proteome) isoforms—thereby expanding the repertoire of the gene products in the mammalian genomes. Although the functional consequence of differential expression of alternative isoforms is known for some genes, the magnitude of alternative isoform expression at the genome scale is still not well understood. Indeed, recent evidence suggests that almost all multi-exon human genes generate multiple mRNA variants that differ either in protein-coding regions and/or regulatory untranslated regions (UTR) (Davuluri et al. 2008; Pan et al. 2008; Wang et al. 2008; Trapnell et al. 2010). For example, the alternative promoter usage of *TP73* results in two protein isoforms that perform opposing biological functions (Muller et al. 2006), and their balanced expression is a crucial factor in normal development and disease (Tomasini et al. 2008). In contrast, nine distinct mRNAs are produced from the *BDNF* gene through the use of alternative promoters, which differ in their 5′UTR but translate the same protein. The distinct 5′UTRs function as the regulatory region responsible for the differential expression and localization of *BDNF* transcripts (Pruunsild et al. 2007). Interestingly, a subset of differentially spliced transcript variants was recently found to be associated with poor prognosis in a large clinical cohort of patients with breast cancer (Dutertre et al. 2010). However, after nearly a decade since the completion of the human genome draft sequence, we still consider the "gene" as the basic functional unit in the genome (Check Hayden 2010). Indeed, we need a paradigm shift from the current "gene centric approach" to a "gene isoform centric approach," making the study of gene isoforms an important aspect of biological networks. Therefore, acquiring the knowledge of all possible gene isoforms and their in vivo expression patterns in specific cell populations and tissues, as well as their developmental

[4] These authors contributed equally to this work.
[5] Corresponding author.
E-mail rdavuluri@wistar.org.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.120535.111.

21:1260–1272 © 2011 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/11; www.genome.org

stages, is a necessary first step for understanding the isoform-specific functions of a gene and identification of disease-relevant gene isoforms from bystander alternative forms of a gene.

The role of alternative promoter is particularly critical in transcriptional regulation, since their precise utilization allows the balanced expression of corresponding transcript variants in different cell and/or developmental contexts. However, the molecular mechanisms of how these multiple promoters are selectively used under different cellular conditions are still unclear. The possible mechanisms include diverse core-promoter structure at alternative promoters (D'Alessio et al. 2009), variable concentration of *cis*-regulatory elements in the upstream promoter region, and regional epigenetic mechanisms such as DNA methylation, histone modifications, and chromatin remodeling (Davuluri et al. 2008). Although the role of epigenetic modifications in gene regulation is well established (Hatchwell and Greally 2007), the influence of such mechanisms on the choice of alternative promoter usage has been demonstrated for only a handful of genes (Archey et al. 1999; Dammann et al. 2000; Ventura et al. 2002).

To understand the function of individual transcript variants and their epigenetic regulation in both normal and disease conditions, high-resolution maps of in vivo mRNA isoform expression (*transcriptome*), along with the epigenetic modification states (*epigenome*) of the promoters (*promoterome*) that transcribe the pre-mRNAs, are necessary. Here, we performed massive parallel sequencing-based methods (mRNA-seq and ChIP-seq) and integrative bioinformatics analysis to (1) identify the active and repressed promoters, their expressed transcript variants, and their epigenetic profiles; (2) generate a digital inventory of transcripts including the transcript variants of a gene; and (3) evaluate the role of histone modifications in the selection of promoters during brain development using the mouse cerebellum as a model. The relatively simple structure and cell composition of the cerebellum makes it an excellent model system to investigate developmental questions such as cell differentiation or migration (Zervas et al. 2005). A critical aspect of its development (proliferation and differentiation of the major cell type, the granule neurons) occurs during postnatal stages, and some genes that are important for normal development are known to express multiple transcripts and/or protein isoforms in the brain (e.g., *Trp73*, *Gsk3b*, *Nrg1*) (Flames et al. 2004; Tomasini et al. 2008; Hur and Zhou 2010; Wilhelm et al. 2010). It is well established that disruption of tightly regulated and coordinated gene expression, which occurs in a temporal and spatial manner during normal development, is observed in the cerebella of patients with psychiatric disorders like schizophrenia, autism, anxiety, attention-deficit hyperactivity disorders, and in medulloblastoma (Grimmer and Weiss 2006; Ten Donkelaar and Lammens 2009). Here, we show, using medulloblastoma as an example, that some of the developmentally regulated mRNA variants are aberrantly expressed, emphasizing the importance of studying altered isoform expression rather than only gene expression in developmental diseases. This integrated resource is provided as a database called the Mammalian Development Transcriptome Database (MDevTrDb) (http://mdevtrdb.wistar.upenn.edu/).

Results

mRNA-seq reveals transcriptome diversity and numerous novel transcripts in developing mouse cerebella

To determine the diversity and abundance of expressed transcripts in cerebellar development, we performed massive parallel sequencing of poly(A)-tailed mRNAs. We focused our study on postnatal development when most of the cerebellum expansion, patterning, and maturation occurs, and selected the development stages corresponding to (1) early increase of granule neuron precursor cells (GNPs) (postnatal day 0, P0); (2) peak of GNP proliferation (P5); (3) end of GNP proliferation and beginning of GNP differentiation (P15); and (4) the fully mature cerebellum comprised of mostly differentiated cells (Adult-P56). Deep sequencing of cDNAs from each of these stages yielded a total of 149 million reads, and 95% of the filtered reads mapped to the mouse transcriptome and genome (Supplemental Table S7). Furthermore, the high quality of our mRNA-seq data was revealed by the mapping of 80%–87% of the reads in all stages to either the exon or the exon–exon junctions of combined known gene models (RefSeq, Vega, UCSC, Ensembl, MGI) and by a lack of 3′end alignment bias (Supplemental Fig. S1A).

Having established the quality of the mRNA-seq data, we sought to identify the poly(A)-tailed transcriptome across the four cerebellar developmental stages. Our analysis found 21,301 novel splice junctions and 12,565 novel exons with significant sequence enrichment ($P \leq 10^{-5}$) relative to known gene model transcripts, and we observed that the 5′ and 3′ UTR regions for 545 and 1460 transcripts, respectively, were longer than the combined known gene model transcripts. These longer UTR regions were supported by one or more of the novel gene model transcripts (Aceview, SGP, TROMER, XenoRef, Genscan, Geneid) (Supplemental Fig. S1B; Supplemental Tables S1, S2, S8). In addition, we found 30,475 genomic regions that have significant expression ($P \leq 10^{-5}$) in one or more developmental stages. These expressed contigs (hereafter referred to as "novel expressed contigs"), with a mean length of 740 bp, lack any gene/transcript annotations, computational predictions, and are localized in both intergenic (mostly at 5′ or 3′ end of the genes) and intronic regions (Supplemental Fig. S1C; Supplemental Tables S3, S8). On the aligned mRNA-seq reads, we applied the IsoformEx algorithm (Kim et al. 2010) to identify and estimate the expression of transcripts and transcript variants of a gene (known and novel), and found 92,990 transcripts ($P \leq 0.01$) that correspond to 20,055 protein-coding (based on RefSeq, Enterz, and Vega gene) and 17,197 noncoding genes (Supplemental Table S4). Next, we compared the expression of both noncoding and protein-coding transcripts and their variants across the four postnatal stages of the cerebellum, which is visualized as a heat map generated by hierarchical clustering (Fig. 1A). We observed variation in expression of both protein-coding and noncoding gene transcripts between development stages. Interestingly, a large number of noncoding gene transcripts are highly expressed during early (P0–P5) development. Our transcript/isoform-level analysis captures the stage-specific regulation of isoform expression, as demonstrated for *Dclk1*, a phenomenon lost in gene-centric approaches (Fig. 1B). *Dclk1*, a member of the doublecortin gene family involved in neurogenesis and neuronal migration, is expressed from two distinct promoters and generates four transcript variants/protein isoforms (Fig. 1B, bottom) (Dijkmans et al. 2010). The isoforms (1 and 2) driven by the upstream promoter (promoter 1) are highly expressed in early P0–P5 stages, and the downstream promoter (promoter 2) derived transcripts (isoforms 3 and 4) are P15 and adult specific. It is worth noting that the distinct isoforms of DCLK1 differ in their functional domains, with isoform 2 possessing both the doublecortin domain (responsible for microtubule binding) and the catalytic ser/thr protein kinase domain, and isoform1 lacking the kinase domain. Moreover, isoforms 3 and 4, which arise from alternative

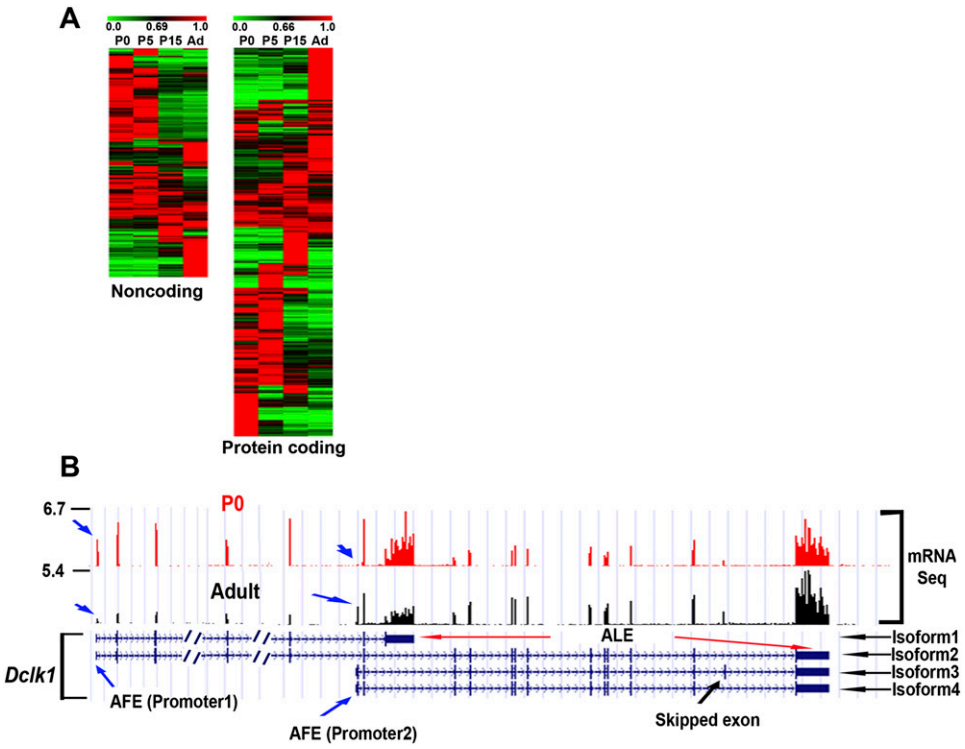

Figure 1. Expression of alternative transcripts during cerebellum development determined by mRNA-seq. (*A*) Heat map shows hierarchical clustering based on relative, normalized expression for protein-coding and noncoding transcripts and transcript variants in P0, P5, P15, and adult cerebellum. (*B*) Wiggle profile of mRNA-seq data reveals a developmental, stage-specific pattern of expression for the four isoforms of *Dclk1* in the cerebellum. While promoter 1–driven transcripts are mostly expressed in early development, P0 and P5 (only P0 shown) promoter 2–driven transcripts are mostly expressed in adult cerebella.

splicing of promoter 2-driven pre-mRNA and which differ in their C-terminal regions, are missing the doublecortin domains.

Following our finding that there are ~2.5 transcripts per gene and that transcript variants are differentially expressed during postnatal cerebellar development, we wanted to identify the mechanism(s) that generate the diverse cerebellar transcriptome. For this purpose we analyzed the occurrence of both transcriptional and splicing events using an approach independent of IsoformEx-based transcript expression (Table1). In order to distinguish the transcript variants that could arise due to either alternative transcription or alternative splicing, we categorized alternative first exons (AFE) and alternative last exons (ALE) events as "transcriptional events" and exon skipping (ES), intron retention (IR), alternative 5′ splice site (A5SS), and alternative 3′ splice site (A3SS) events as "splicing event." Only those exons that do not overlap are considered for AFE and ALE events. If any two exons, including the first and last exons, overlap, but differ in their 5′ or 3′ splice sites, those are counted in A5SS and A3SS events, respectively. Therefore, the presence of nonoverlapping AFE and ALE represent the alternative promoters and alternative transcription termination, respectively. Using the above criteria, we generated a library of all known transcriptional and splicing events from UCSC, RefSeq, Ensembl, MGI, and Vega gene models (see Methods for details). For example, we found that *Gad1* generates different pre-mRNAs at transcription through the usage of AFE and ALE events, and that *Tnc* undergoes ES on the single pre-mRNA to produce transcript variants during development that translate distinct proteins (Supplemental Fig. S2). Additionally, there are genes such as *Dclk1* that demonstrate the use of AFE, ALE, and ES events (Fig. 1B). By using the library of known transcriptional and splicing events, the mRNA-seq data revealed that a total of 6764 genes (e.g., *Dclk1*, *Hdgf*, *Tpm1*, *Gad1*) used one or both of the transcriptional events (AFE and ALE), which is significantly higher than the number of genes (3077 genes; e.g., *Tnc*, *Gli2*) that used at

Table 1. Identification of known alternative events in the mouse cerebellum tissue

	Alt event type	Reference set	Both isoforms expressed				
			P0	P5	P15	Adult	Overall
Transcriptional events [No. of events (No. of genes)]							
	AFE	26147 (9334)	8873 (3606)	8508 (3482)	8786 (3554)	8001 (3194)	11880 (4649)
	ALE	22469 (8519)	8176 (3447)	8266 (3476)	8153 (3418)	8090 (3372)	10486 (4297)
Alternative splicing events [No. of events (No. of genes)]							
	Exon skipping	20569 (7547)	1579 (1192)	1613 (1214)	1585 (1140)	1224 (903)	2592 (1810)
	Intron retention	2389 (2144)	1038 (845)	1025 (835)	1001 (808)	945 (764)	1230 (980)
	A5SS	1659 (1425)	251 (236)	238 (219)	234 (216)	194 (182)	389 (356)
	A3SS	3151 (2393)	423 (383)	421 (384)	419 (382)	358 (324)	643 (580)

Using the combined known gene models, we generated a library of all known alternative events in the mouse genome (reference set) and identified the events occurring in each stage, as well as over the course of cerebellum development (overall). If the isoforms are coexpressed in a stage, then it is counted as an alt event in that stage, and in case the alternative isoforms are expressed during different developmental stages, the alt event will be included in the overall category.

least one of the splicing events (Table 1). While ES is the most prominent of all of the identified splicing events in cerebellar development, both the AFE and ALE events are used in more than double the number of genes showing ES events, demonstrating that the use of alternative transcriptional events is more widespread than alternative splicing to diversify the transcriptome during development.

We therefore generated a comprehensive library of expressed transcripts, including those with modified 5′ and/or 3′ UTR regions, by integrating mRNA-seq with the combined knowledgebase of transcripts from known and novel gene models. The developmental transcriptome (inventory of transcript variants/isoforms) and their expression status in each developmental stage will be a key resource for isoform-driven studies in the normal development of the mammalian brain and its diseased states.

Pol II promoters and corresponding transcript variants in postnatal cerebellar development

To identify the promoters driving the cerebellar transcriptome diversity observed above, we used the integrative approach outlined in Figure 2A. Briefly, we generated the RNA Pol II (Pol II)-binding and H3K4me3 enrichment profiles in each developmental stage to identify the Pol II promoters of the postnatal cerebellum, and categorized the promoters as "active" or "inactive" in each stage by integrating with mRNA-seq data. We choose Pol II binding and H3K4me3 enrichment to identify promoters because they have been shown to occupy sites of transcription initiation, including paused TSS and promoters with abortive transcription initiation (Guenther et al. 2007). Deep sequencing of ChIP-enriched DNA from P0, P5, P15, and adult cerebella using antibodies against Pol II

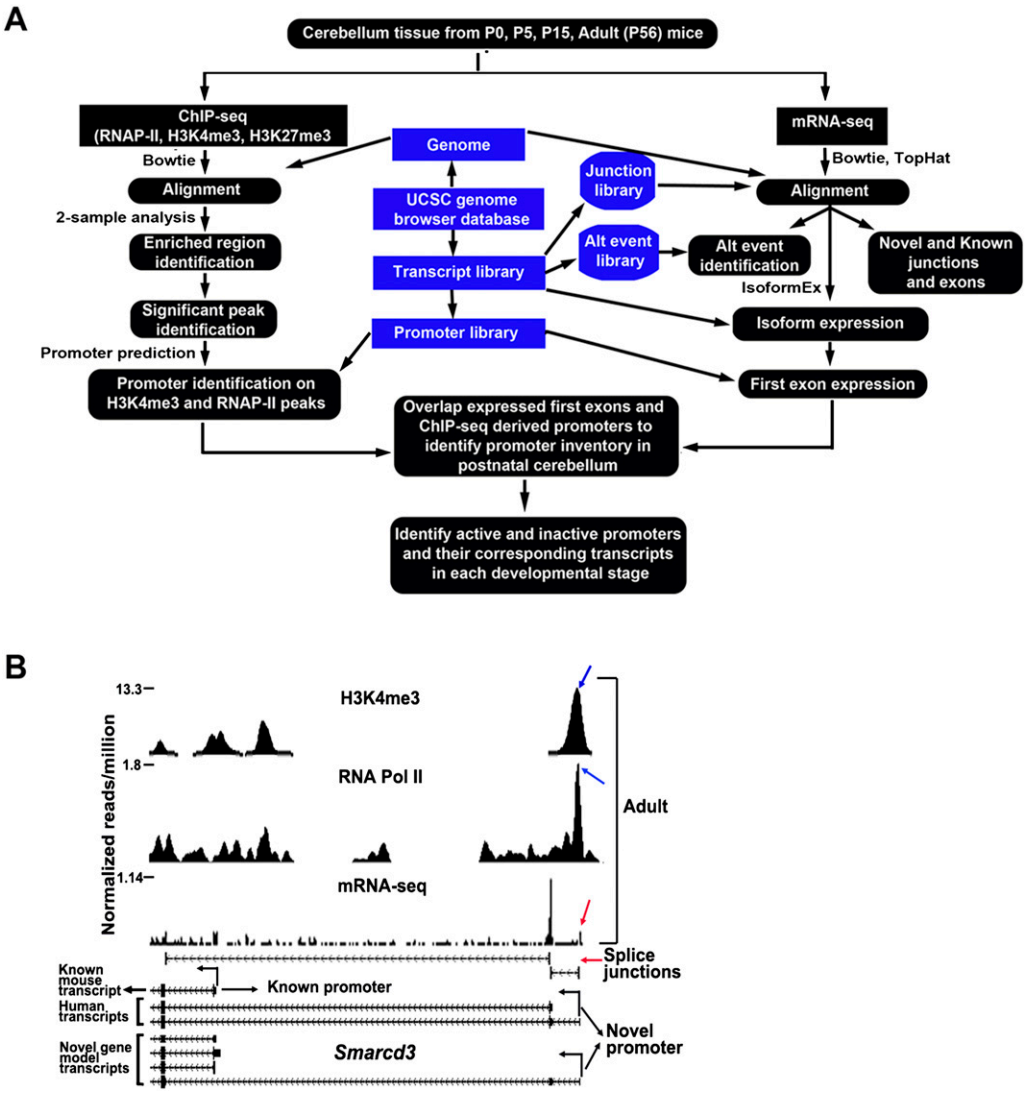

Figure 2. Identification of known and novel promoters and transcripts expressed in the cerebellum. (*A*) Integrative approach for the identification of promoter and corresponding transcript inventory used in mouse cerebella after birth. We performed mRNA-seq and ChIP-seq experiments with antibodies against Pol II, H3K4me3, and H3K27me3. The flowchart indicates the multiple steps we followed to determine the promoterome and transcriptome of each postnatal stage during cerebellum development. (*B*) An example of a novel promoter and corresponding transcript identified through the integrative approach. *Smarcd3* is also expressed from a novel upstream promoter that was identified by promoter prediction on both the Pol II and H3K4me3 enrichment profiles (blue arrows), and the expression of the corresponding transcript was determined from mRNA-seq profiles (red arrows). It is worth noting that the novel promoter and transcript is known in humans and is also supported by novel gene models of mouse.

and H3K4me3, yielded 8.5–10 million and 10.4–11.5 million uniquely aligned reads, respectively, with a high enrichment of reads around known TSSs (Supplemental Fig. S3A; Supplemental Table S9). Next, we applied our recent promoter identification algorithm on the Pol II and H3K4me3 enrichment profiles at each stage and predicted a total of 37,130 promoters that included 11,733 novel promoters (Supplemental Table S5) (Gupta et al. 2010; Sun et al. 2011). We have tested the ability of five predicted novel promoters to drive luciferase expression and found promoter activity in all, indicating successful prediction of novel promoters (Supplemental Fig. S3B).

We then prepared an inventory of Pol II promoterome and their respective transcripts expressed in postnatal cerebella. We first integrated the Pol II bound and/or H3K4me3 marked promoters with corresponding transcripts expressed in each developmental stage by overlapping the predicted Pol II promoters from ChIP-seq data with the first exons of mRNA-seq transcripts (both known and novel) and novel expressed contigs (for novel promoters/transcripts). A promoter is designated as "active" if it has an annotated transcript with significant expression in that developmental stage; otherwise, it is considered as "inactive." Overall, we found that 15,669 protein-coding, 7624 noncoding genes, and 750 novel expressed contigs were expressed during cerebellar development through the activity of 29,589 promoters (85.5% overlap with CAGE promoter clusters) that generated a total of 61,525 mRNAs, including transcript variants (Table 2; Supplemental Table S6). Of these, 16% of the promoters and 21% of the transcripts are novel and are not annotated in any of the current known mouse gene models, for example, the upstream promoter for *Smarcd3* (Fig. 2B). We predicted a Pol II promoter ~24 Kb upstream of the known mouse *Smarcd3* promoter based on both Pol II binding and H3K4me3 enrichment, and this prediction was supported by the expression of the corresponding novel transcripts that contain two novel exons and novel splice junctions. In addition, we identified 7541 promoters (92% are novel) that lacked supporting transcript expression from our mRNA-seq data in all stages analyzed and these were not considered for further analysis. Moving forward, our analyses will focus on the integrated set of active and inactive promoters and the resulting transcriptome.

Alternative promoter usage and differential expression of transcript variants in cerebellar development

Our goal was to provide a digital inventory of all promoters and their mRNAs, including the transcript variants, along with their estimated expression values in each cerebellar developmental stage. Having identified both known and novel Pol II promoters and corresponding transcript variants, expressed at each stage, we analyzed the use of alternative promoters in the postnatal development model of a cerebellum. We found that 50.3% of genes expressed in the cerebellum, which includes 11,173 (71%) of the protein-coding genes (e.g., *Pax6*) and 532 (7%) of the noncoding genes (e.g., *Meg3*), generate multiple transcripts (multitranscript genes) and express a total of 50,122 transcripts through the use of alternative transcriptional and/or splicing events (Fig. 3A,B; data not shown). The remaining genes (4496 protein coding and 7092 noncoding) express a single transcript, thus excluding the occurrence of any alternative event in these genes in a postnatal cerebellum. Moreover, we found that while 52% of multitranscript genes use alternative promoters and almost 81% use one of the alternative transcriptional events (AFE/alternative promoter and/or ALE/alternative transcriptional termination), 68% use alternative splicing events (for distributions of genes in different categories, refer to Fig. 3A). Interestingly, 47% of genes exhibit both alternative transcriptional and splicing events. Further inspection revealed that both alternative splicing (AS) and alternative transcriptional termination events are significantly ($P = 2.2 \times 10^{-16}$ χ^2 test) more prevalent in the multipromoter genes than in single promoter genes (Supplemental Table S10A). These findings point to a central role for alternative promoters in generating transcriptome diversity.

Following the identification of genes exhibiting only transcriptional events or only splicing events or both transcriptional and splicing events to generate transcript variants, we investigated whether distinguishable nervous-system development functions could be associated with the genes in these three classes. Using ingenuity pathway analysis (IPA), we observed that multitranscript genes exhibiting only splicing mechanisms participate in the branch termination of axons (arborization), the number and orientation of neuritis, proliferation, and the differentiation of neural precursors, while genes with alternative transcriptional events are involved in regeneration, distribution, neurotransmission of neurons, neurogenesis, and the formation of cerebellar folia and vermis. The genes with both transcriptional and splicing mechanisms were contributing to the morphogenesis of neurites, dendrites, the migration of neurons, and the transport of synaptic vessels and synaptic plasticity (Fig. 3C).

Next, we studied the distribution of promoter usage and transcript expression in the four cerebellar stages and observed that 83% (24,450) of the promoters were active and 68% (41,850) of the transcripts were expressed in all stages, while a small percentage of promoters (4%) and transcripts (9%) were exclusively expressed in one stage (Supplemental Table S10B). Furthermore, 8% of the promoters, which transcribe 17% of the transcripts, are active only during early postnatal development (P0–P15), while only 1.7% of the promoters corresponding to 2.8% of the transcripts are adult specific (Supplemental Table S10B). For example,

Table 2. The integrative approach (Fig. 2A) was used to identify the active promoters and the corresponding expressed transcripts at each of the four postnatal development stages

	Active promoters			Inactive promoters			Expressed transcripts			
Stage	Known	Novel	Total	Known	Novel	Total	Known	Novel	Total	Genes
P0	23,131	4014	27,145	1666	778	2444	41,402	10,964	52,366	22,505
P5	23,099	4030	27,129	1698	762	2460	41,140	10,975	52,115	22,360
P15	23,148	4295	27,443	1649	497	2146	41,592	11,212	52,804	22,526
Adult	22,915	4425	27,340	1882	367	2249	40,328	11,023	51,351	22,311
Overall	24,797	4792	29,589				48,729	12,796	61,525	24,370

"Overall" represents the nonredundant number of promoters/transcripts/genes in the four stages.

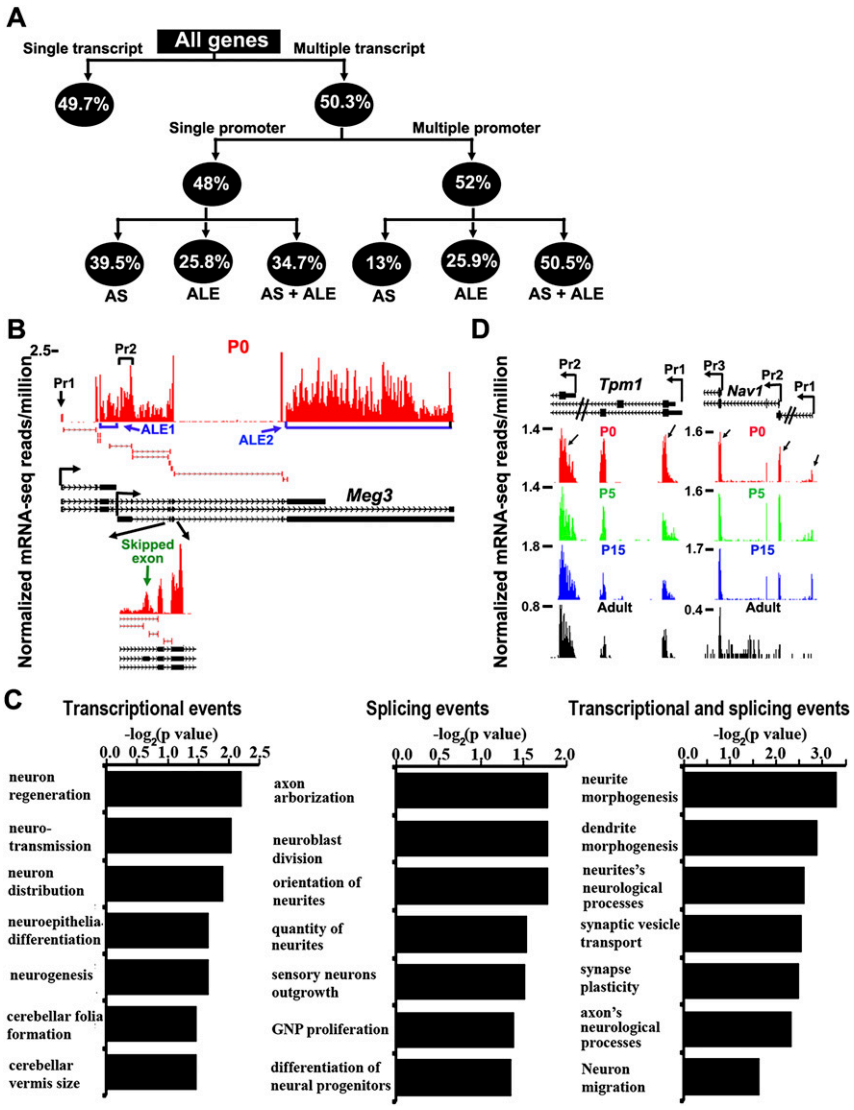

Figure 3. Alternative promoter usage and expression of promoters and transcripts in the four stages of postnatal cerebella. (*A*) The occurrence of alternative splicing (AS) and alternative last exon (ALE) events in both single-promoter and multipromoter-driven genes that express multiple transcripts in postnatal cerebellum. (*B*) Wiggle profile of mRNA-seq data for *Meg3* shows the expression of multiple transcript variants through the use of alternative promoters/alternative first exon (AFE), ALE, and AS in P0. The black and blue arrows point to the coexpressed AFE and ALE, respectively. (*Bottom*) The zoomed in regions of exon skipping (ES) (green arrow) and supporting splice junctions in P0. (*C*) Ingenuity pathway analysis reveals the enriched functions ($P < 0.05$) among the genes that exhibit only transcriptional events, only splicing events, and both transcriptional and splicing events. (*D*) Wiggle profile for *Tpm1* (*left*) and *Nav1* (*right*) from mRNA-seq shows the expression of multiple promoters and transcripts in all stages or in a stage-restrictive manner, respectively. While both promoters are active for *Tpm1*, only promoter 3 for *Nav1* is active in all four stages. Promoters 1 and 2 of *Nav1* are development specific and not active in adult cerebella. The promoter-specific AFE expression is indicated by the black arrows. (*Top*) The pre-mRNAs for each gene and the corresponding promoter and TSS.

both promoters of *Tpm1* are active in all stages, in contrast to *Nav1*, which possesses alternative promoters that are specific to the P0–P15 cerebellum (Fig. 3D). Analysis of the transcription factor (TF) binding sites for the stage-specific transcripts versus ubiquitous transcripts revealed that few TF motifs (for example, AP3M1 in P0, HLF in P5, AP4E1 and NFAT5 in P15, and TCF3, YY1, LMO2, JUN, and ACTR1A in adult) were significantly enriched (Fisher's test $P < 0.01$) in stage-specific promoters.

Although 83% of promoters are active in all of the stages, the expression of corresponding mRNAs and mRNA variants varies significantly (more than twofold) across the stages. By performing a pairwise comparison of the transcript expression, from mRNA-seq estimates in all four developmental stages, we observed that the transcript expression changes (up or down) by a minimum of 6910 (23%) promoters from P0 to P5, and a maximum of 15,791 (53.3%) promoters from P0 to adult (Supplemental Table S10C). This differential expression from alternative promoters during development has been confirmed by quantitative RT–PCR for 10 genes (*Fgf9*, *Sox17*, *Pax6*, *Olfm1*, *Ptch1*, *Tpm1*, *Hdgf*, *Axin2*, *Rassf1*, and *Gad1*) (Supplemental Table S11A,B). The differential expression of alternative transcripts driven by alternative promoters could be attributed partly to the presence of some distinguishable TF motifs. We found that the binding sites of SP1, E2F, KLF9, TCFAP2A, TCFAP2C, ZFP161, CHURC1, AP2A1, and GTF2IRD1 transcription factors were significantly enriched in promoters, driving higher expression (major promoters). Similarly, the binding sites of YY1, GRLF1, POU2F1, ZBTB16, ZEB1, and CEBPG were significantly present in minor promoters of transcripts with lower expression. Interestingly, for some multipromoter genes, altered gene expression is reflected by a similar pattern of changes for all of the transcripts transcribed by alternative promoters (e.g., *Pax6*, *Amz2*, *Casp7*, *Nudcd2*, *Nav1*, *Grm1*, *Rfx4*), while for others (e.g., *Rassf1*, *Gad1*, *Etv1*, *Dclk1*, *Lims1*, *Olfm1*, *Dnm1l*), switch-like behavior is observed where one transcript variant shows increased expression and the other shows decreased expression across P0 to adult. This switch-like behavior information is lost in gene centric expression analyses, which can be biologically very significant in light of the fact that transcript variants can produce proteins with distinct biological roles. For example, the two protein isoforms of *Dnm1l*, whose deletion results in abnormal cerebellum development, with purkinje cells carrying fewer and large mitochondria, which differ in their N-termini, are differentially expressed during cerebellum development. The smaller DNM1L isoform carries a truncated dynamin-N domain that contains the GTPase activity (Wakabayashi et al. 2009). Similarly, the transcription factor *Rfx4*, which is known to modulate SHH signaling (critical for normal cerebellum development) through the regulation of ciliogenesis, generates two protein isoforms that are similarly regulated during development, but differ by the presence of the DNA-binding

domain (Ashique et al. 2009). Taken together, our results suggest that the developing cerebellum's transcriptome is highly diverse and that this diversity is partly mediated through alternative promoter use that can change the proteome during development.

Relationship of H3K4me3 and H3K27me3 epigenetic marks at the promoters including alternative promoters with transcript expression

To conduct correlative analysis of active and inactive chromatin modifications with mRNA expression profiles during development, we chose to study H3K4me3 and H3K27me3. H3K4me3 was selected as the active chromatin mark, since it has been shown to occur near TSS (±1 Kb) that are either active or maintained in a paused state (Guenther et al. 2007). To study the repressive chromatin, H3K27me3 was analyzed because of its role in development. It has been observed that the PRC2 complex, the H3K27 methyltransferase, is highly expressed in the neural progenitor cells and is critical for neurogenic to gliogenic transition during brain development (Hirabayashi et al. 2009). We generated repressive chromatin profiles of H3K27me3 by performing ChIP-seq in the four developmental stages of the cerebellum (Supplemental Fig. S3A; Supplemental Table S9). Surprisingly, the average enrichment of H3K27me3 around TSS is significantly higher in Adult than in P0, a phenomenon less prominent for H3K4me3 (cf. Supplemental Fig. S3A, center and right).

We first analyzed the influence of H3K4me3 and H3K27me3 modifications at active promoters on the expression levels of the corresponding transcripts. We divided the expressed transcripts into three categories as low, medium, and high based on mRNA expression, and then plotted the average enrichment of either lysine modification around the TSS (Fig. 4A; Supplemental Fig. S4A). Globally, we observed a direct correlation of promoter activity (corresponding transcript expression) with the H3K4me3 mark and an inverse relationship with H3K27me3. Next, we fine-tuned the analysis and monitored the relationship of each mark with the average expression from clusters of 50 promoters (Fig. 4B; Supplemental Fig. S4B). We found that there is an exponential rather than a linear relationship of mRNA expression, and the enrichment of the histone modification marks, which fits well with the use of a single exponential model with a high percentage of explained variation (R^2) in each developmental stage (Fig. 4B; Supplemental Fig. S4B).

Since CpG-rich and Non-CpG promoters are structurally different (Davuluri et al. 2001), and H3K4me3 and H3K27me3 marks are known to be highly enriched near CpG rich promoters (Ku et al. 2008; Thomson et al. 2010), we repeated the correlation analysis by dividing the promoters into two classes—CpG and Non-CpG promoters. We found that the exponential models fitted the relationship between mRNA expression and H3K4me3 (Fig. 4C; Supplemental Fig. S4C) or H3K27me3 (Fig. 4D; Supplemental Fig. S4D) enrichment with better R^2 for CpG promoters than the overall data. However, for Non-CpG promoters, the exponential relationship of mRNA expression with promoter enrichment of H3K4me3 was good, although with lower R^2, but not with promoter enrichment of H3K27me3. Our results demonstrate that both H3K4me3 and H3K27me3 are highly biased toward CpG-rich promoters.

Role of H3K4me3 and H3K27me3 modifications in alternative promoter selection during development

Since promoter enrichment profiles of H3K4me3 and H3K27me3 are highly correlated with the transcript expression profiles, we checked the percentage of active and inactive promoters that were marked by these two marks (Table 3; Fig. 5B). We found that on average 69% of the active promoters were marked by H3K4me3, while only 38% of the inactive promoters were marked by H3K27me3 during development, suggesting that H3K4me3 plays a more prominent role in gene activation than H3K27me3 plays in transcriptional silencing (Table 3). Furthermore, the inactive promoters that were only enriched with H3K4me3 might be either paused or generating transcripts with short half-life such that no RNA is detected by mRNA-seq. Similarly, ~8% of active promoters were marked exclusively by H3K27me3, and we speculate that

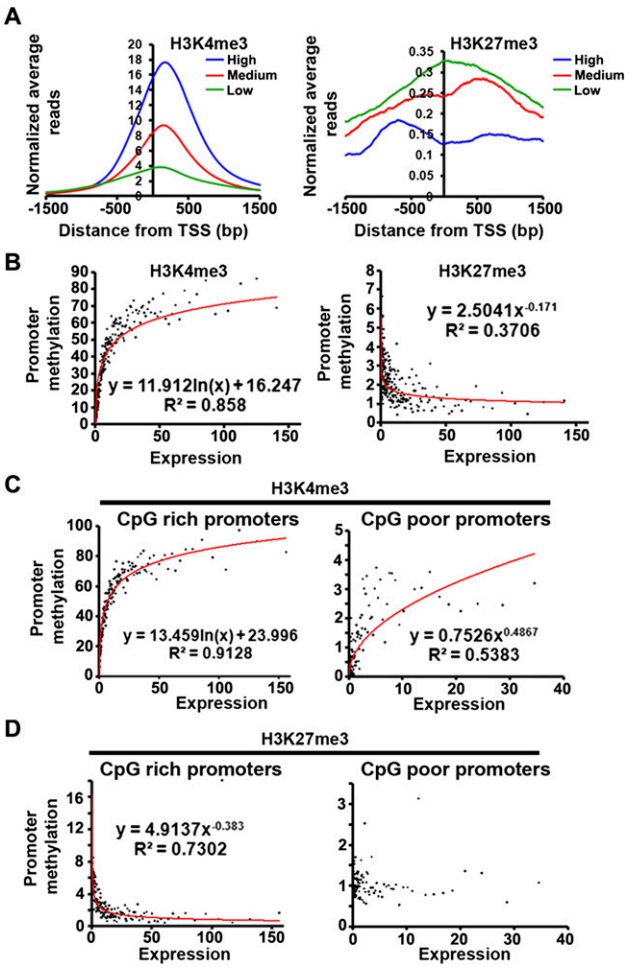

Figure 4. Correlation of H3K4me3 and H3K27me3 enrichment at the promoter with the expression of corresponding transcripts. (*A*) The active promoters of P0 were divided into three groups based on the expression of their transcripts into low, medium, and high expressed promoters, and the global profile of H3K4me3 and H3K27me3 were plotted for each group around the TSS. (*B*) In order to find the mathematical relation between H3K4me3 or H3K27me3 at the promoters with transcript expression, promoters were clustered into groups of 50 promoters based on their transcript expression. The figure shows the scatter plot and the best fitted curve for the average level of H3K4 or K27 trimethylation as a function of average cluster expression. (*C*,*D*) Role of H3K4me3 and H3K27me3 modification on transcript expression driven by CpG-rich and CpG-poor promoters. The promoters analyzed in *B* were first divided into CpG-rich and CpG-poor categories, and clusters of 25 promoters were formed in each category based on the expression of the corresponding transcripts. Next, the average levels of H3K4me3 (*C*) and H3K27me3 (*D*) in each cluster were plotted as a function of average cluster expression as in *B*.

Table 3. The occupancy of active and inactive promoters with H3K4 or/and H3K27 trimethylated histones in P0, P5, P15, and adult cerebellum

	Promoters are marked by (in %)							
	H3K4me3		H3K27me3		Both		None	
Stage	Active	Inactive	Active	Inactive	Active	Inactive	Active	Inactive
P0	42	22	12	24	28	20	18	34
P5	46	25	6	14	23	19	25	42
P15	52	29	2	12	17	19	27	40
Adult	49	26	9	25	18	19	24	30

the lack of PRC1 binding maintains these promoters as active. Although a significant number of promoters in both active and inactive promoter groups are marked by both chromatin modifiers, the active promoters show a higher H3K4me3/H3K27me3 ratio compared with the inactive promoters (Fig. 5B; Supplemental Fig. S5), a relationship clearly illustrated on the alternative promoters of *Dclk1* and *Ndrg4* (Fig. 5A).

Since the highly expressed transcript variants of *Dclk1* and *Ndrg4* genes have significantly higher H3K4me3 and lower H3K27me3 enrichment at respective promoters than those of lowly expressed counterparts (Fig. 5A), we performed a genome-wide analysis to determine the influence of H3K4me3 and H3K27me3 on the differential expression of transcript variants from alternative promoters. We selected the alternative promoters of two-promoter genes and found that the promoter with higher levels of H3K4me3 exhibited higher expression of the corresponding transcript, while H3K27me3 had no significant role in the choice of predominantly transcribed promoters (Fig. 5C).

H3K4me3 and H3K27me3 display combinatorial regulation of promoter-specific transcript expression during development

Having analyzed the role of H3K4 and H3K27 trimethylation in regulating alternative promoters, we investigated their role in refining the expression of transcripts or their variants from a promoter across different developmental stages. We performed this analysis on those transcripts/transcript variants whose expression was up- or down-regulated by at least twofold between any two developmental stages, and whose respective promoters carried both of the marks. Multivariate analysis showed that the up-regulation of transcript expression was associated with increased H3K4me3 and reduced H3K27me3 promoter enrichment, while the opposite was true for down-regulated promoters (Fig. 6; Supplemental Fig. S6).

Our analysis revealed that the rise and fall of H3K4me3 enrichment is relatively more important for the up- and down-regulation of expression compared with changes in H3K27me3 (clearly visualized in Supplemental Fig. S6). Taken together, our results suggest that the expression of the transcripts, including transcript variants of the same gene, is regulated during development by the enrichment of H3K4me3 and H3K27me3 in the promoter regions.

Multitranscript genes are associated with neurological disorders and transcript variants are aberrantly expressed in medulloblastoma

Developmental gene expression signatures that are disrupted in diseases such as schizophrenia, have been identified by various studies (Torkamani et al. 2010). Similarly, certain phenomena that occur during development were also observed in diseases, for example, epithelial mesenchymal transition, which is crucial for normal development and tissue repair and also participates in tumor metastasis (Thiery et al. 2009; Hatten and Roussel 2011). Since we observed extensive changes in the expression of transcript variants during normal development, we wanted to analyze whether genes that have been associated with specific diseases tend to generate multiple transcripts during development. Moreover, aberrant

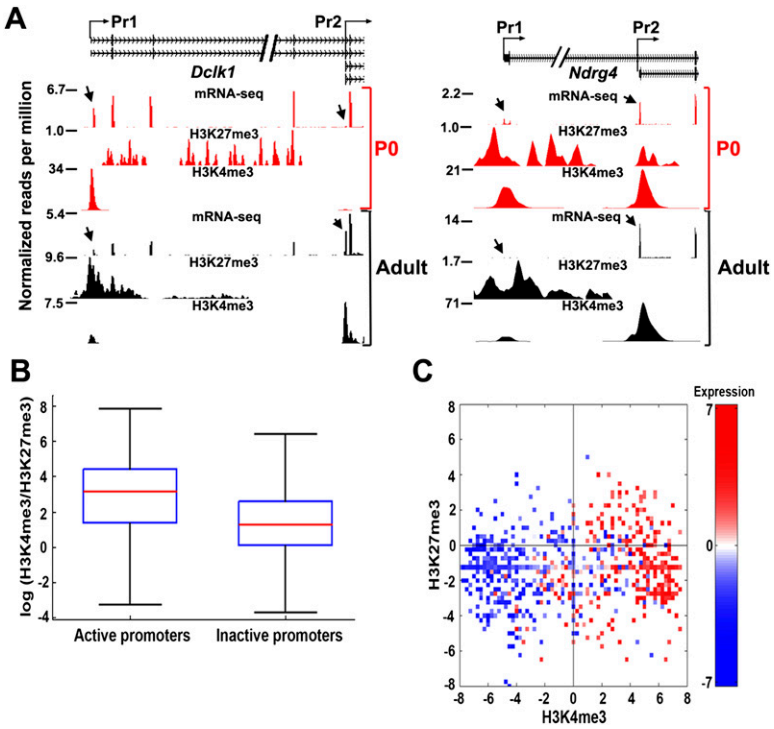

Figure 5. Distinct histone modification profile of active and inactive promoters and their role in alternative promoter selection during development. (*A*) Wiggle profiles of mRNA expression and H3K4me3, H3K27me3 modifications on the alternative promoters of *Dclk1* and *Ndrg4* in P0 (red) and adult (black) cerebella. The arrows point to either the expression of AFE or the enrichment of H3K4me3 and H3K27me3 at the alternative promoters. The loss of transcript expression from promoter (Pr) 1 for *Dclk1*and *Ndgr4* in adult cerebellum is marked by reduced/loss of H3K4me3 and highly enriched H3K27me3. In contrast, the increase in expression from Pr2 parallels high H3K4me3 and loss of H3K27me3. (*B*) Box plot shows the distribution of relative enrichment of H3K4me3 over H3K27me3 on the active and inactive promoters marked by both marks in P0. (*C*) Heat map shows the distribution of relative H3K4me3 and H3K27me3 enrichment in $\log_2$ scale on the alternative promoters of two promoter genes, where the expression of alternative promoters differs at least twofold. Expression is represented as $\log_2$ fold change between alternative promoters of a gene. Red means up-regulation and blue represents down-regulation of expression.

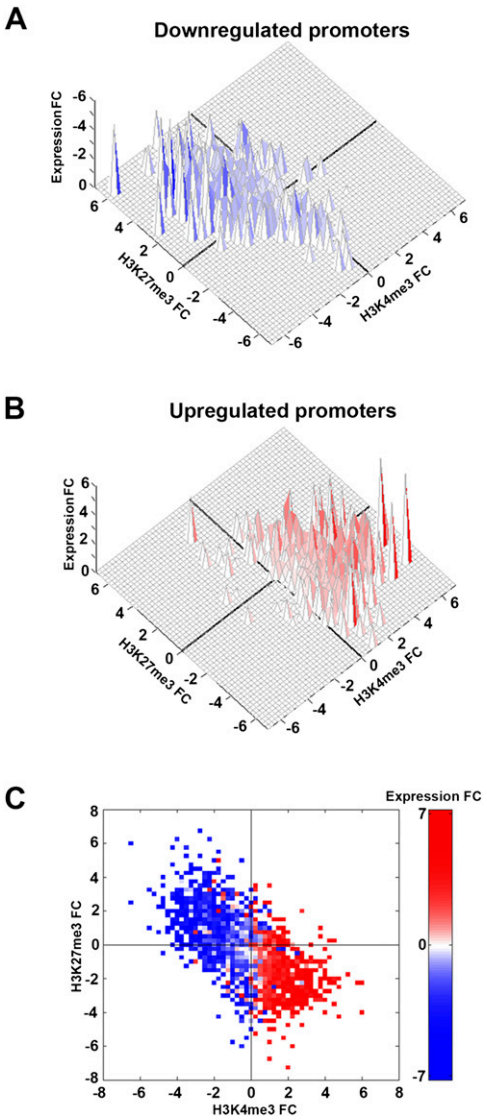

Figure 6. Three-dimensional graphical visualization of H3K4me3 and H3K27me3 enrichments near promoters and expression of corresponding transcripts. For this analysis, we selected the promoters that correspond to transcripts whose expression were either up- or down-regulated during development by at least twofold between any two developmental stages and were also simultaneously marked by H3K4me3 and H3K27me3. The 3D plots show the relationship between transcript expression and relative enrichment (fold change: FC) of both marks on the down-regulated (in blue, *A*) and up-regulated (in red, *B*) promoters of genes that are differentially expressed during cerebellar development. The *x*, *y*, and *z*-axis represent the $\log_2$ fold change (FC) in H3K4me3, H3K27me3, and expression, respectively, between stages, with scale being negative for down-regulated promoters and positive for up-regulated promoters. (*C*) Heatmap shows the relationship of changes in transcript expression with the trimethylation of H3K4 and H3K27 on respective promoters of developmentally regulated transcripts shown *above* in *A* and *B*.

expression of one transcript variant (or alternative promoter) over another has been observed for many genes like *MYC*, *RASSF1*, and *TP73* in cancer (Davuluri et al. 2008). We investigated the prevalence of multitranscript genes among the genes involved in cerebellar development, the neurological disorders—autism and schizophrenia—and cancer of the cerebellum medulloblastoma (MB). Genes related to cerebellum development were downloaded from the MGI website (http://www.informatics.jax.org/phenotypes.shtml) and represent genes whose genetic alteration results in abnormal cerebellum development in mouse. An autism-related gene list was downloaded from the AutDB database, which collects genes from various published genetics studies on human autism spectrum disorders (ASD) (Basu et al. 2009). Human schizophrenia and MB gene lists were obtained from published gene expression studies (Kho et al. 2004; Torkamani et al. 2010). We found that ~85% of the genes that are connected to abnormal cerebellar development in mouse and the diseases—autism, schizophrenia, and MB in humans—belong to multitranscript genes, with ~50% of the genes being expressed from alternative promoters (Table 4). Overall, the genes associated with abnormal development and disease show a higher enrichment for alternative transcriptional events when compared with all of the genes expressed in cerebellum (Abnormal development: $P = 4.5 \times 10^{-14}$; Autism: $P = 4.9 \times 10^{-8}$; Schizophrenia: $P = 2.2 \times 10^{-16}$; MB: $P = 8.07 \times 10^{-5}$). Our analysis suggests the involvement of specific transcript variants in these diseases and emphasizes the importance of isoform centric studies.

To validate our hypothesis, we measured the expression of transcript variants from alternative promoters in MB primary tumors and cell lines. MB is the most common childhood brain cancer that arises in the developing cerebellum. We examined the expression of 22 transcripts (10 genes) that are differentially expressed during development, and the genes have been implicated in cancer. We performed quantitative RT–PCR on RNA isolated from primary mouse MBs and tumor cell lines obtained from *Ptch1*$^{+/-}$; *Trp53*$^{-/-}$ mice (Wetmore et al. 2001). We found that the promoters of *Fgf9*, *Sox17*, *Pax6*, *Olfm1*, *Gad1*, and *Axin2* are either silenced or highly repressed in the primary MB and cell lines (Table 5; Supplemental Table S11C). In contrast, *Hdgf*, *Tpm1*, *Ptch1*, and *Rassf1* exhibit promoter-specific expression in MB, thus attributing the expression of these genes to specific transcript variants. We observed specific up-regulation of promoter 2 and down-regulation of promoter 1 in MB for both *Hdgf* and *Tpm1*. Moreover, in the case of *Ptch1*, though promoter 1 is repressed varying from 2.5 to fivefold, promoter 2 is highly repressed or silenced in MB samples, while for *Rassf1*, promoter 1 is activated in a manner similar to the developmental expression (P0–P5) in MB tumor and cell lines (Table 5). It is worth noting that the huge changes in expression from specific promoters of *Rassf1* and *Tpm1* are not that striking at the gene level due to the low level of expression for these transcript variants. The use of alternative promoters generates distinct protein isoforms for each of these four genes. For example, in the case of *Hdgf*, the use of promoter 2 generates a protein that has a truncated PWWP domain, a domain that functions as a methyl-lysine recognition motif and binds to methylated H4K20. All together, our results show aberrant expression of specific transcript variants of a gene in medulloblastoma, further demonstrating the importance of isoform centric studies in identifying disease-specific novel biomarkers and therapeutic targets.

Discussion

By utilizing massive parallel sequencing technology and integrative bioinformatics methods (Hawkins et al. 2010), our experiments provide the first genome-wide inventory of the transcriptome and promoterome, and maps of active (H3K4me3) and repressed (H3K27me3) epigenetic marks in postnatal developing and adult mouse cerebella. These integrated data, the first of their

Table 4. Most disease and MB cancer-related genes generate multiple transcripts through the use of alternative transcriptional events–alternative promoters and alternative transcription termination/alternative last exon (ALE)

	Single transcript genes	Multiple transcript genes		
Disease/disorder	Single promoter and single last exon	Single promoter (ALE)	Multiple promoter (ALE)	Total genes
Abnormal cerebellum development	19	41 (22)	59 (44)	119
Autism	29	62 (36)	104 (77)	195
Schizophrenia	507	1152 (641)	1461 (1061)	3120
Medulloblastoma	46	113 (57)	150 (110)	309

The numbers within parentheses indicate the number of genes with alternative transcriptional termination/ALE that belong to either single-promoter and multipromoter genes.

kind not only for cerebellar development, but also for any developing organ in human or mouse, provide a critical resource for further studies of the isoform-level gene regulation in cerebellar function, development, and disease. The resource includes a database of 61,525 transcripts, including alternative mRNA variants along with their promoters and epigenetic marks, and a digital inventory of gene and mRNA isoform expression measurements across the three early developmental stages and adult mouse cerebella (Supplemental Table S6).

Almost the entire body of literature on gene regulation credits transcriptome diversity to alternative splicing and rarely to alternative transcription in mammalian cells (Moore and Proudfoot 2009). Surprisingly, the integrative analysis of the data here highlights the widespread use of both transcriptional events (AFEs and ALEs), followed by an ES splicing event in generating alternative transcript variants during development. The implication is that the transcriptome diversity in development seems to arise during transcription, in the form of multiple pre-mRNAs with different transcriptional starts and/or transcriptional ends. These pre-mRNAs further undergo alternative splicing (AS), mostly using the "ES" mechanism. Similar usage of alternative events was observed in a recent study that analyzed 15 diverse human tissue and cell line transcriptomes, although ES was reported as the top-ranking event, followed by AFE and ALE (Wang et al. 2008). A survey of human embryonic kidney cell line and a B cell line also showed that ES was the most prevalent form of alternative splicing (Sultan et al. 2008). Both of these studies report "ES" as the major driver of transcriptome diversity, and we speculate that this discrepancy could stem from the use of combined gene models and improved 5′ and 3′ annotations over the last few years (from sources like GENCODE). To address this issue, we performed our alternative event analysis on individual gene models (UCSC, RefSeq, Vega, or Ensembl) and arrived at the same conclusion, irrespective of the gene model used in the analysis—alternative transcription is the major source of transcriptome diversity in developing cerebella (Supplemental Table S12).

In addition to detecting many novel splice junctions and exons for numerous genes, we redefined the 5′ and 3′ UTR regions for over 500 genes that are expressed in the cerebellum. This is essential in light of our current knowledge of the regulatory role of UTRs in gene expression, such as through the binding of miRNAs (Bartel 2009). Importantly, we found more than 30,000 novel transcribed regions (with significant expression in one or more developmental stages) that are not annotated in any of the existing genomic annotations. Although we predicted Pol II promoters for only 3.1% of these novel expressed contigs, we observed simultaneous occupancy of Pol II and H3K4me3 on 67% of these genomic regions. Many of these intergenic contigs lie in the 5′ or 3′ end of known genes, and the enrichment of Pol II and H3K4me3 in these expressed contigs suggests that some may be noncoding RNAs or sORF/short peptide RNAs that possess noncanonical Pol II promoters, or Pol III promoters that interact with Pol II, thus explaining their presence in these regions (Barski et al. 2010; Oler et al. 2010).

We found that ~50% of the multitranscript genes (25% of all expressed genes) use multiple promoters, which is quite substantial given the fact that the ChIP-seq and mRNA-seq experiments were performed using only four different libraries (P0, P5, P15, adult). The previous estimates of genes utilizing alternative promoters, ranging from 52% to 58%, were based on 5′ end identification from 164 and 145 different libraries (Carninci et al. 2006; Kimura et al. 2006). Also, we observed that genes with multiple promoters exhibit increased alternative splicing as well as alternative transcriptional termination (last exons) usage. This is not surprising, because splicing is cotranscriptional, and a causal link between promoter usage and splicing has been established (Cramer et al. 1997; Kornblihtt 2005). Similarly, a link between promoter selection and transcriptional termination has been ob-

Table 5. Transcript variants' expression from distinct promoters for five genes in MB primary tumor and tumor-derived cell lines from *Ptch*$^{+/-}$; *p53*$^{-/-}$ mice relative to normal adult cerebella

		Primary MB tumor	MB tumor derived cell lines	
Gene	Expression at	*Ptch*$^{+/-}$; *Trp53*$^{-/-}$	*Ptch*$^{+/-}$; *Trp53*$^{-/-}$ CL1	*Ptch*$^{+/-}$; *Trp53*$^{-/-}$ CL2
Tpm1	Pr1	−0.73697	−0.32193	−1.73697
	Pr2	4.916477	4.554589	3.364572
	Gene level	3.925999	3.217231	2.765535
Hdgf	Pr1	−1.32193	−1.73697	−2.32193
	Pr2	1.584963	1.678072	0.847997
	Gene level	2.035624	1.536053	0.847997
Rassf1	Pr1	16.33816	18.49413	14.85788
	Pr2	1.722466	2.432959	0.378512
	Gene level	2.392317	2.292782	1.070389
Ptch1	Pr1	−1.73697	−1.73697	−3.32193
	Pr2	ND[a]	ND[a]	ND[a]
	Gene level	−4.64386	−5.64386	−6.64386
Axin2	Pr1	−2.55639	−3.32193	−3.8365
	Pr2	−4.32193	−1	−4.32193
	Gene level	−6.64386	−4.05889	−5.05889

The transcript annotation corresponding to each promoter is presented in Supplemental Table S11A. The expression values represent the $\log_2$ of relative expression of MB versus normal, averaged over three PCR reactions.
[a]Not detected (ND).

served for *Mid1* (Winter et al. 2007). These findings are crucial to our understanding of mechanisms generating the transcript diversity, especially in light of our current knowledge on the expression and functions of distinct isoforms for some genes. For example, MAPT isoforms, which form the neuronal microtubule networks, are expressed through a combination of alternative promoters and splicing, and are developmentally regulated and involved in neurological disorders (Buee et al. 2000). Similarly, the expression of multiple isoforms for the alpha and beta subunit of Na+/K+ transporting ATPase in a tissue and development-regulated fashion confer different biochemical properties for the Na+/K+ ATPase ion channels (Lingrel 1992). Our study has identified the expression of transcript variants for 11,955 genes in the cerebellum during postnatal development, and as an example, we found four genes (*Dab1*, *Fyn*, *Reln*, and *Cnr1*) belonging to the reelin-signaling pathway that generate multiple protein isoforms; it would be interesting to study how each protein isoform modulates the pathway. Although the role of epigenetic histone methylation in regulating gene expression is well established (Kurdistani and Grunstein 2003; Hatchwell and Greally 2007), there are no reports showing that such mechanisms influence the choice of alternative promoters. By profiling the chromatin states in pluripotent and lineage-committed cells, it was shown that the promoter state reflected the lineage commitment of mouse embryonic stem cells and that alternative promoters have multiple and distinctive chromatin states (Mikkelsen et al. 2007), suggesting that an active state at one of the promoters is sufficient to drive the expression of the corresponding transcript (Davuluri et al. 2008). Here, our results show that the levels of H3K4me3 on alternative promoters identify the highly transcribed promoter for multipromoter genes, and that both H3K4me3 and H3K27me3 work in concert to developmentally regulate promoters that dictate the expression of corresponding mRNA variants. Further, our results indicate more dominant roles for H3K4me3 than H3K27me3 in fine-tuning the expression during development, which is supported by previous studies reporting an interaction of RNA Pol II with H3K4me3. As previously reported, we also observed a strong bias of H3K4 and H3K27 trimethylation toward CpG islands (Ku et al. 2008; Thomson et al. 2010), and the epigenetic modifications regulating and fine-tuning the expression of Non-CpG promoters is an open question. Moreover, our detailed analysis establishes a nonlinear relationship between H3K4 and H3K27 trimethylation and promoter-specific isoform expression, and points to a lack of correlation between H3K27me3 on Non-CpG promoters and their transcript expression.

We have observed that the multipromoter genes are also enriched in various diseases like cancer and neurological disorders. Deregulated expression of certain gene isoforms like *TP73*, *LEF1*, and *MYC* have been studied and reported in various cancers. Our analysis on the expression of alternative promoter-driven transcripts from 10 genes in mouse medulloblastoma reveals that four genes (*Rassf1*, *Ptch1*, *Tpm1*, and *Hdgf*) exhibit isoform-specific alteration of gene expression in MB, a phenomenon not captured by measuring changes in gene expression, thus suggesting a role for specific gene isoforms in cancer. The mRNA variants can either alter the UTRs, which impacts mRNA stability and protein translation or/and protein coding regions. In either case, identifying the transcriptome at the isoform level could be more critical than only using gene-level information to understand the molecular aberrations associated with the initiation, progression, and maintenance of cancer.

We have profiled the transcriptome and identified the active promoters responsible for their expression in postnatal developing and adult cerebella. We provide these results as a searchable database (MDevTrDb), which is a critical resource to the scientific community to study molecular abnormalities in various diseases and disorders of the cerebellum and their relationship to the normal events occurring during development.

Methods

ChIP and mRNA sequencing, quantitative RT–PCR analysis

About 0.5 g of mouse cerebellum tissues collected from CD1 mice at postnatal days 0, 5, 15, or 56 were used for performing ChIP-seq, as described in the Supplemental Methods. To identify the expressed transcripts, mRNA-seq was performed on 10 μg of total RNA from each stage (details in the Supplemental methods). To measure the expression of alternative promoter-driven transcripts during normal development and in medulloblastoma, quantitative PCR was performed on cDNA from P0, P5, P15, adult cerebellum, and medulloblastoma cell lines and tumors (Supplemental Methods).

Bioinformatics analysis of mRNA-seq data

The mRNA-seq data analysis, which involves generating the reference set of alternative events, alignment of mRNA-seq data and analysis, identification of alternative events, and estimation of transcript expression, is described in the Supplemental Methods.

Clustering analysis for protein-coding and noncoding transcripts

For the clustering, we selected transcripts that had an estimated expression of a minimum of 1 RPKM in at least one of the developmental stages. To generate the protein-coding transcripts list we selected only those protein-coding transcripts that belong to protein-coding genes based on RefSeq/Enterz/Vega definitions. For the noncoding transcript list, transcripts are selected based on the following criteria: (1) The transcript does not overlap with pseudogenes (Vega annotation definition used), (2) the transcript is present at least 1 Kb away from known protein-coding gene boundary, and (3) if the transcript lies within a protein-coding gene, then it should be on the opposite strand of the protein-coding gene.

We performed hierarchical clustering of transcript expression using the MeV package (Saeed et al. 2006). An average linkage clustering method was selected for clustering, and Euclidean distance was used for distance metric calculation. Each row in the heatmap represents a transcript/expressed locus. A default value of 10^{-6} was added to each data in the heatmap in order to avoid zeros in the expression value. Each row was then normalized based on the highest expression of the corresponding transcript/locus among P0, P5, P15, and adult stages. This was done to capture the pattern of relative expression across four stages, irrespective of expression values.

Bioinformatics analysis of ChIP-seq data

The ChIP-seq data (RNAP-II, H3K4me3, H3K27me3, IgG-Control) analysis is comprised of three major stages: (1) alignment, (2) peak identification, and (3) promoter prediction, and is presented in the Supplemental Methods.

Promoter identification and annotation

To identify promoters, we applied our recently developed promoter prediction program (Gupta et al. 2010) to each significant

peak obtained from Pol II and H3K4me3 ChIP-seq data as detailed in the Supplemental Methods.

Correlation studies of promoter H3K4me3 and H3K27me3 with expression of corresponding transcripts

We performed studies to address the role of H3K4me3 and H3K27me3 in the regulation of transcript expression at three different levels: (1) the individual impact of H3K4 and H3K27 trimethylation at CpG rich and poor promoters on corresponding transcript expression, (2) the role of H3K4me3 and H3K27me3 in the choice of alternative promoters, and (3) the combinatorial role of H3K4 and K27 trimethylation in regulating expression during development. The detail of criteria for each analysis is presented in the Supplemental Methods.

Transcription factor (TF) motif analysis

Motif analysis was performed on the set of two promoter genes. The MATCH program was used to identify the TFBS in each promoter (−1 Kb to +1 Kb around TSS), and only those motifs that were conserved (conservation score cutoff = 0.4) using the euarchontoglires phastcon scoring system were considered for further analysis. To identify motifs associated with major versus minor promoters, we first defined the alternative promoters of a gene as major and minor promoters at each stage. Next, we determined the frequency of occurrence for each TF in the minor versus major promoter lists and performed Fisher's exact test to determine whether a given TF was significantly ($P < 0.01$) associated with one of the promoter classes.

To identify the motifs associated with stage-specific promoters, we first identified the stage-specific promoters of P0, P5, P15, and adult stage, and the set of ubiquitous promoters in the cerebellum. We defined the promoters that have a minimum expression of 2 RPKM and are expressed at a level at least fourfold higher than the stage with the second highest expression as stage-specific promoters. In contrast, the promoters with a minimum 1 RPKM expression and no greater than a 1.5-fold difference between the expressions from the most active versus least active stages are considered "ubiquitous promoters." Next, using MATCH, the TF-binding sites were identified, conserved sites were selected, TF motif frequency was calculated, and significant association with stage-specific promoters versus ubiquitous promoters was determined using Fisher's exact test, as above.

Data access

The mRNA-seq and ChIP-seq data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under series accession no. GSE23525.

Acknowledgments

This work was supported by an NHGRI/NIH grant (# R01HG003362) and an American Cancer Society Research Scholar grant (# RSG-07-097-01) to R.V.D. R.V.D. holds a Philadelphia Healthcare Trust Endowed Chair Position; research in his laboratory is partially supported by the Philadelphia Healthcare Trust. We thank V. Tatard for providing some of the tissues used in this study and H. Riethman for critical reading of the manuscript. Work in the N.D. laboratory was supported by the American Cancer Society and the Brain Tumor Society. The use of resources in the Genomics and Bioinformatics Shared Facilities of The Wistar Cancer Center (grant # P30 CA010815) are gratefully acknowledged.

References

Archey WB, Sweet MP, Alig GC, Arrick BA. 1999. Methylation of CpGs as a determinant of transcriptional activation at alternative promoters for transforming growth factor-beta3. *Cancer Res* **59:** 2292–2296.

Ashique AM, Choe Y, Karlen M, May SR, Phamluong K, Solloway MJ, Ericson J, Peterson AS. 2009. The Rfx4 transcription factor modulates Shh signaling by regional control of ciliogenesis. *Sci Signal* **2:** ra70. doi: 10.1126/scisignal.2000602.

Barski A, Chepelev I, Liko D, Cuddapah S, Fleming AB, Birch J, Cui K, White RJ, Zhao K. 2010. Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. *Nat Struct Mol Biol* **17:** 629–634.

Bartel DP. 2009. MicroRNAs: target recognition and regulatory functions. *Cell* **136:** 215–233.

Basu SN, Kollu R, Banerjee-Basu S. 2009. AutDB: a gene reference resource for autism research. *Nucleic Acids Res* **37:** D832–D836.

Buee L, Bussiere T, Buee-Scherrer V, Delacourte A, Hof PR. 2000. Tau protein isoforms, phosphorylation and role in neurodegenerative disorders. *Brain Res Brain Res Rev* **33:** 95–130.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38:** 626–635.

Check Hayden E. 2010. Human genome at ten: Life is complicated. *Nature* **464:** 664–667.

Cramer P, Pesce CG, Baralle FE, Kornblihtt AR. 1997. Functional association between promoter structure and transcript alternative splicing. *Proc Natl Acad Sci* **94:** 11456–11460.

D'Alessio JA, Wright KJ, Tjian R. 2009. Shifting players and paradigms in cell-specific transcription. *Mol Cell* **36:** 924–931.

Dammann R, Li C, Yoon JH, Chin PL, Bates S, Pfeifer GP. 2000. Epigenetic inactivation of a RAS association domain family protein from the lung tumour suppressor locus 3p21.3. *Nat Genet* **25:** 315–319.

Davuluri RV, Grosse I, Zhang MQ. 2001. Computational identification of promoters and first exons in the human genome. *Nat Genet* **29:** 412–417.

Davuluri RV, Suzuki Y, Sugano S, Plass C, Huang TH. 2008. The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet* **24:** 167–177.

Dijkmans TF, van Hooijdonk LW, Fitzsimons CP, Vreugdenhil E. 2010. The doublecortin gene family and disorders of neuronal structure. *Cent Nerv Syst Agents Med Chem* **10:** 32–46.

Dutertre M, Lacroix-Triki M, Driouch K, de la Grange P, Gratadou L, Beck S, Millevoi S, Tazi J, Lidereau R, Vagner S, et al. 2010. Exon-based clustering of murine breast tumor transcriptomes reveals alternative exons whose expression is associated with metastasis. *Cancer Res* **70:** 896–905.

Flames N, Long JE, Garratt AN, Fischer TM, Gassmann M, Birchmeier C, Lai C, Rubenstein JL, Marin O. 2004. Short- and long-range attraction of cortical GABAergic interneurons by neuregulin-1. *Neuron* **44:** 251–261.

Grimmer MR, Weiss WA. 2006. Childhood tumors of the nervous system as disorders of normal development. *Curr Opin Pediatr* **18:** 634–638.

Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130:** 77–88.

Gupta R, Wikramasinghe P, Bhattacharyya A, Perez FA, Pal S, Davuluri RV. 2010. Annotation of gene promoters by integrative data-mining of ChIP-seq Pol-II enrichment data. *BMC Bioinformatics* **11:** S65. doi: 10.1186/1471-2105-11-S1–S65.

Hatchwell E, Greally JM. 2007. The potential role of epigenomic dysregulation in complex human disease. *Trends Genet* **23:** 588–595.

Hatten ME, Roussel MF. 2011. Development and cancer of the cerebellum. *Trends Neurosci* **34:** 134–142.

Hawkins RD, Hon GC, Ren B. 2010. Next-generation genomics: an integrative approach. *Nat Rev Genet* **11:** 476–486.

Hirabayashi Y, Suzki N, Tsuboi M, Endo TA, Toyoda T, Shinga J, Koseki H, Vidal M, Gotoh Y. 2009. Polycomb limits the neurogenic competence of neural precursor cells to promote astrogenic fate transition. *Neuron* **63:** 600–613.

Hollstein M, Hainaut P. 2010. Massively regulated genes: the example of TP53. *J Pathol* **220:** 164–173.

Hur EM, Zhou FQ. 2010. GSK3 signalling in neural development. *Nat Rev Neurosci* **11:** 539–551.

Kho AT, Zhao Q, Cai Z, Butte AJ, Kim JY, Pomeroy SL, Rowitch DH, Kohane IS. 2004. Conserved mechanisms across development and tumorigenesis revealed by a mouse development perspective of human cancers. *Genes Dev* **18:** 629–640.

Kim H, Bi Y, Davuluri RV. 2010. Estimating the expression of transcript isoforms from mRNA-Seq via nonnegative least squares. In *Proceedings of the 10th IEEE International Conference on Bioinformatics and Bioengineering (BIBE-2010)*, pp. 296–297, Philadelphia, PA.

Kimura K, Wakamatsu A, Suzuki Y, Ota T, Nishikawa T, Yamashita R, Yamamoto J, Sekine M, Tsuritani K, Wakaguri H, et al. 2006. Diversification of transcriptional modulation: Large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res* **16:** 55–65.
Kornblihtt AR. 2005. Promoter usage and alternative splicing. *Curr Opin Cell Biol* **17:** 262–268.
Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, Mikkelsen TS, Presser A, Nusbaum C, Xie X, Chi AS, et al. 2008. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet* **4:** e1000242. doi: 10.1371/journal.pgen.1000242.
Kurdistani SK, Grunstein M. 2003. Histone acetylation and deacetylation in yeast. *Nat Rev Mol Cell Biol* **4:** 276–284.
Lingrel JB. 1992. Na,K-ATPase: isoform structure, function, and expression. *J Bioenerg Biomembr* **24:** 263–270.
Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448:** 553–560.
Moore MJ, Proudfoot NJ. 2009. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* **136:** 688–700.
Muller M, Schleithoff ES, Stremmel W, Melino G, Krammer PH, Schilling T. 2006. One, two, three–p53, p63, p73 and chemosensitivity. *Drug Resist Updat* **9:** 288–306.
Oler AJ, Alla RK, Roberts DN, Wong A, Hollenhorst PC, Chandler KJ, Cassiday PA, Nelson CA, Hagedorn CH, Graves BJ, et al. 2010. Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. *Nat Struct Mol Biol* **17:** 620–628.
Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40:** 1413–1415.
Pruunsild P, Kazantseva A, Aid T, Palm K, Timmusk T. 2007. Dissecting the human BDNF locus: bidirectional transcription, complex splicing, and multiple promoters. *Genomics* **90:** 397–406.
Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J. 2006. TM4 microarray software suite. *Methods Enzymol* **411:** 134–193.
Schmucker D. 2007. Molecular diversity of Dscam: recognition of molecular identity in neuronal wiring. *Nat Rev Neurosci* **8:** 915–920.
Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, et al. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321:** 956–960.
Sun H, Wu J, Wickramasinghe P, Pal S, Gupta R, Bhattacharyya A, Agosto-Perez FJ, Showe LC, Huang TH, Davuluri RV. 2011. Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq. *Nucleic Acids Res* **39:** 190–201.
Ten Donkelaar HJ, Lammens M. 2009. Development of the human cerebellum and its disorders. *Clin Perinatol* **36:** 513–530.
Thiery JP, Acloque H, Huang RY, Nieto MA. 2009. Epithelial-mesenchymal transitions in development and disease. *Cell* **139:** 871–890.
Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr AR, Deaton A, Andrews R, James KD, et al. 2010. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464:** 1082–1086.
Tomasini R, Tsuchihara K, Wilhelm M, Fujitani M, Rufini A, Cheung CC, Khan F, Itie-Youten A, Wakeham A, Tsao MS, et al. 2008. TAp73 knockout shows genomic instability with infertility and tumor suppressor functions. *Genes Dev* **22:** 2677–2691.
Torkamani A, Dean B, Schork NJ, Thomas EA. 2010. Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia. *Genome Res* **20:** 403–412.
Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28:** 511–515.
Ventura A, Luzi L, Pacini S, Baldari CT, Pelicci PG. 2002. The p66Shc longevity gene is silenced through epigenetic modifications of an alternative promoter. *J Biol Chem* **277:** 22370–22376.
Wakabayashi J, Zhang Z, Wakabayashi N, Tamura Y, Fukaya M, Kensler TW, Iijima M, Sesaki H. 2009. The dynamin-related GTPase Drp1 is required for embryonic and brain development in mice. *J Cell Biol* **186:** 805–816.
Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470–476.
Wetmore C, Eberhart DE, Curran T. 2001. Loss of p53 but not ARF accelerates medulloblastoma in mice heterozygous for patched. *Cancer Res* **61:** 513–516.
Wilhelm MT, Rufini A, Wetzel MK, Tsuchihara K, Inoue S, Tomasini R, Itie-Youten A, Wakeham A, Arsenian-Henriksson M, Melino G, et al. 2010. Isoform-specific p73 knockout mice reveal a novel role for ΔNp73 in the DNA damage response pathway. *Genes Dev* **24:** 549–560.
Winter J, Kunath M, Roepcke S, Krause S, Schneider R, Schweiger S. 2007. Alternative polyadenylation signals and promoters act in concert to control tissue-specific expression of the Opitz Syndrome gene MID1. *BMC Mol Biol* **8:** 105. doi: 10.1186/1471-2199-8-105.
Zervas M, Blaess S, Joyner AL. 2005. Classical embryological studies and modern genetic analysis of midbrain and cerebellum development. *Curr Top Dev Biol* **69:** 101–138.

Received January 7, 2011; accepted in revised form May 23, 2011.