Supplementary Figure legends:

Figure S1 (A) mRNA-seq data generated from P0, P5, P15 and adult cerebellum is of high quality. Bar chart shows the alignment of mRNA-seq reads (in %) to exons, exon-exon splice junctions, intron and interegene regions (left) and the distribution of reads in the first (center) and last exon (right) . We observe very low alignment of mRNA-seq reads to intergenic locations and a lack of 3' end bias due to polyA purification in our datasets. (B) Extended 5' and 3' UTR for some genes based on mRNA-seq data. An example is shown for 3'UTR (up panel) and 5'UTR (down panel), which are incomplete in the UCSC database and have been extended based on our mRNA-seq data. Our data is in agreement with the UTR presented in the novel gene models- Aceview gene models and Tromer transcriptome database in the examples shown in panel B. (C) An example of expressed genomic contigs identified through mRNA-seq analysis. Wiggle profile of mRNA-seq read distribution on a chromosome1 location (1146bp) that lacks any known or predicted transcript information. This region is highly conserved in vertebrates and seems similarly expressed in all postnatal stages of cerebellum.

Figure S2: Expression of alternative events during cerebellum development. (A and B) Wiggle profile of mRNA-seq reads enrichment on *Gad1* (A) and *Tnc* (B) in P0 (red) and adult (black) cerebellum demonstrates the expression of AFE and ALE in *Gad1* and alternative splicing event in *Tnc*. In (A), the expression of second transcript's AFE and ALE is low and hence a zoomed in profile of the corresponding region is shown for P0 and adult stage. The blue arrows point to the events and the black arrows point to the specific exons and their expression. The detected splice junctions are shown below each wiggle profile.

Figure S3: (A) Distribution of reads obtained from the sequencing of anti-RNA Pol-II (left), anti-H3K4me3 (center), and anti-H3K27me3 (right) enriched DNA around TSS. The read distribution profiles peak near TSS and hence suggest that Pol-II, H3K4me3 and H3K27me3 enrichment is observed mostly in and nearpromoter regions. (B) Novel promoters are identified by our

analysis. Luciferase assays were performed in Daoy (human medulloblastoma) and NIH3T3 (mouse fibroblast) cells to test the activity of five dnovel promoters that are not known in the UCSC, RefSeq, Ensembl, and Vega databases. These novel promoters were identified through our integrative approach detailed in the main manuscript. The coordinates of the cloned novel promoters and their gene annotation is as follows: NP1-Chr11:58052738-58053305 for *Pgbd2*; NP2-Chr12:12949469-12950388 *for N-Myc*; NP3- Chr4:153455349-153456410 for *Trp73*; NP4-Chr12:3833906-3834893 for *Dnmt3A*; NP5-Chr2:105520689-105521709 for *Pax6*. The ctrl1, 2 represent control regions without any known promoter or promoter prediction. *Dll1* promoter was used as a positive control in the assay. Briefly, 1.8 µg of pGL3 basic or pGL3basic-novel promoter and 0.2 µg of pGL4-renilla luciferase vector were transfected using Lipofectamine2000 in the cells for 48hr prior to the assay of luciferase and renilla luciferase activity using the dual luciferase assay kit. All transfections were normalized based on renilla luciferase expression.

Figure S4: (A) Plot shows the relative enrichment of H3K4 and K27 trimethylation on the expressed promoters (-1500 to +1500 bp) that were subdivided into three groups as “low”, “medium”, and “high” based on the expression of the corresponding transcripts. (B) The relationship of H3K4me3 enrichment at the promoter and the expression from the promoters is not linear (left). The active promoters of P5, P15 and Adult were individually divided into clusters of 50 promoters and the average expression of the cluster was plotted against the average enrichment for H3K4 trimethylation. The correlation of H3K27 trimethylation at the promoter and expression was analyzed as above for H3K4me3 (right). (C) Association of H3K4 trimethylation with expression from CpG and Non-CpG promoters. Active promoters of P5, P15 and adult cerebellum were divided into CpG rich and CpG poor promoters and then subdivided into clusters of 25 promoters based on the expression of their transcripts. The average H3K4me3 enrichment was plotted against the average expression for individual clusters belonging to CpG

rich and poor class. (D) H3K27me3 does not correlate with the expression from CpG poor promoters. Similar analysis as above for figure S4C was performed to look at the link of H3K27trimethylation with CpG richness of promoters.

Figure S5: Active promoters marked by both H3K4me3 and H3K27me3 show a higher H3K4me3/H3K27me3 ratio and inverse is observed on inactive promoters. (A) Box plot showing the distribution of log(H3K4me3/H3K27me3) on active and inactive promoters of P5, P15, and adult stages. (B) Box plot shows the distribution of ChIP-seq reads for H3K4me3 and H3K27me3 around TSS among the active and inactive promoters of each stage (P0, P5, P15, and adult).

Figure S6: H3K4me3 and H3K27me3 fine tune transcript expression. Contour maps show that promoters that are regulated through a combination of H3K4me3 and H3K27me3 during development show a higher H3K4me3 and lower H3K27me3 for upregulated promoters (A) and vice versa is observed for down-regulated promoters (B). X and Y axes show the relative enrichment of H3K4me3 and H3K27me3 in $\log_2$ scale.

Supplementary tables:

Table S7: Alignment of mRNA-seq reads from the four cerebellum development stages. Raw reads (column 1) represent the output from the genome analyzer II and filtered reads (column 2) correspond to the reads that are retained after the removal of raw mRNA-seq reads that align to the contamination library (mitochondrial genome, ribosomal RNA, and adapters).

Stage	Raw reads ($x10^6$)	Filtered reads ($x10^6$)	Read alignment ($x10^6$) to				Aligned reads ($x10^6$)	Unaligned reads ($x10^6$)
			Exon	Splice junction	Intron	Intergenic region		
P0	38.6	36.57	29.2	2.73	2.72	.11	34.76	1.8
P5	37.6	36.21	28.7	2.74	2.77	.10	34.31	1.89

P15	39.3	37.31	29.4	2.61	3.33	.12	35.46	1.79
ADULT	33.3	30.84	22.9	1.94	4.09	.12	29.05	1.68
Total	**148.8**	**140.96**	**110.2**	**10.04**	**12.91**	**.45**	**133.6**	**7.16**

Table S8: Expression of genomic contigs, modified 5' and 3' UTR, and novel transcripts in the mouse cerebellum tissue.

	P0	**P5**	**P15**	**Adult**	**Overall**
Genomic contigs	10503	11187	14623	25444	30475
Changed 5'UTR	377	338	406	364	545
Changed 3' UTR	1128	1099	1152	1190	1460

Table S9: Initial bioinformatic analysis of ChIP-seq data obtained using antibodies against Pol-II, H3K4me3, and H3K27me3 from each of the four stages of cerebellum development .

	Number of			
	reads obtained	**uniquely aligned reads**	**reads in enriched region**	**significant peaks**
Pol-II antibody				
P0	15785020	10269224	3940977	159641
P5	15587053	10051468	2626651	62459
P15	14811287	9247969	1666264	121842
Adult	14471100	8570495	2878246	39457
Total	**60654460**	**38139156**	**11112138**	**383399**
H3K4me3 antibody				
P0	13839721	11549150	9179177	22429
P5	12675451	11532148	10403531	21339
P15	13229230	11458293	10538784	23286

Adult	11731508	10472017	9810110	21432
Total	**51475910**	**45011608**	**39931602**	**88486**
H3K27me3 antibody				
P0	15748989	7746529	4399639	264767
P5	16022467	6582334	2801693	216106
P15	16670029	10896378	4432090	91668
Adult	16425590	11342030	7515760	107799
Total	**64867075**	**36567271**	**19149182**	**680340**
Non specific IgG antibody				
P0	14287598	6704310	N/A	N/A
P5	15834449	9053711	N/A	N/A
P15	17287429	11181763	N/A	N/A
Adult	16225380	5907454	N/A	N/A
Total	**63634856**	**32847238**	**N/A**	**N/A**

Table S10A: Distribution of single and multi-promoter genes to alternatively spliced (AS) and not alternatively spliced (No AS) groups (left) and to a group that have alternative last exon(s) (ALE) and another lacking ALE (No ALE) (right) for performing Chi-square test .

	AS	**No AS**	**Total**			**ALE**	**No ALE**	**Total**
Single promoter genes	4279	12880	17129		**Single promoter genes**	3464	13665	17129
Multi-promoter genes	3909	2249	6158		**Multi-promoter genes**	4706	1452	6158
Total	8158	15129	23287		**Total**	8170	14117	23287

Table S10B: Distribution of active promoters and expressed transcripts during cerebellum development. The table shows the numbers of promoters and transcripts that are specifically expressed in each stage or in various combinations of stages.

Stage(s)	**Active promoters**	**Expressed transcripts**
P0	295	1417
P5	240	1250

P15	244	1487
Adult	494	1739
P0 and P5	340	1318
P0 and P15	203	1007
P0 and Adult	155	710
P5 and P15	171	809
P5 and Adult	197	776
P15 and Adult	548	1383
P0, P5 and P15	756	2886
P0, P5 and Adult	425	1511
P0,P15 and Adult	521	1667
P5, P15 and Adult	550	1715
P0, P5, P15 and Adult	24450	41850
Overall	**29589**	**61525**

Table S10C: Two stage comparisons show the changes in promoter activity during postnatal cerebellum development.

	Promoters whose transcript expression		
	Increases	**Decreases**	**Does not change**
P0-P5	3287	3623	24627
P0-P15	5880	6206	19451
P0-Adult	7414	8377	15746
P5-P15	5914	5893	19730
P5-Adult	7429	8014	16094
P15-Adult	4782	5190	21565

Table S11A: The transcripts corresponding to each promoter whose expression has been measured by quantitative real time PCR.

Gene	Promoter	Corresponding transcripts
Tpm1	Pr1	OTTMUST00000048462;ENSMUST00000113695;ENSMUST00000113690;ENSMUST00000113684
	Pr2	ENSMUST00000113697;ENSMUST00000030185;OTTMUST00000048446;ENSMUST00000113705;ENSMUST00000113701;OTTMUST00000048464;ENSMUST00000034928;MTR000652.9.1507.12;ENSMUST00000113687;NM_001164248;OTTMUST00000048463
Hdgf	Pr1	OTTMUST00000067281
	Pr2	MTR000535.3.1137.2;NM_008231
Rassf1	Pr1	OTTMUST00000075668;OTTMUST00000075660;OTTMUST00000075658;ENSMUST00000122225
	Pr2	OTTMUST00000075661
Ptch1	Pr1	ENSMUST00000021921;NM_204960;UC007QXY.1;PTCH1.ASEP07
	Pr2	PTCH1.BSEP07
Axin2	Pr1	OTTMUST00000006621;ENSMUST00000106712
	Pr2	OTTMUST00000006624
Fgf9	Pr1	FGF9.CSEP07-UNSPLICED;NM_013518
	Pr2	UC007UDY.1
Sox17	Pr1	UC007AEY.1
	Pr2	NM_011441;UC007AFC.1;UC007AFB.1

Gad1	Pr1	UC008JZM.1;OTTMUST00000083437
	Pr2	OTTMUST00000083433
Olfm1	Pr1	OTTMUST00000027196;OLFM1.ISEP07;MTR001015.2.574.5;ENSMUST00000113920
	Pr2	MTR001015.2.574.1;OTTMUST00000027193
Pax6	Pr1	OTTMUST00000035860
	Pr2	ENSMUST00000111083
	Pr3	OTTMUST00000084075;ENSMUST00000111087;OTTMUST00000035642;ENSMUST00000090397;OTTMUST00000035652;ENSMUST00000037848
	Pr4	ENSMUST00000111088;OTTMUST00000035653;ENSMUST00000111086;ENSMUST00000111084

Table S11B: Expression of the indicated genes at the gene level and at individual alternative promoters (Pr) (in log2 scale) in normal cerebellum tissue from P0, P5, P15 mice relative to normal adult mice cerebellum. The expression values represent the averages of three PCR reactions.

Gene	Expression at	P0	P5	P15
Tpm1	**Pr1**	0	0.454176	0.641546
	Pr2	-1.39593	-0.21759	0.948601
	Gene level	0.443607	0.565597	0.594549
Hdgf	**Pr1**	-1.73697	-1.47393	-0.76121
	Pr2	1.014355	1.475085	0.411426
	Gene level	2.235727	0.214125	0.545968
Rassf1	**Pr1**	1.765535	1.678072	0.137504
	Pr2	0.678072	0.432959	0.137504
	Gene level	0.9855	-0.59946	0.137504
Ptch1	**Pr1**	-0.62149	-1.39593	1.378512
	Pr2	0.333424	0.831877	2.084064
	Gene level	0.263034	0.650765	1.321928
Axin2	**Pr1**	1.280956	1.130931	0.925999
	Pr2	3.173127	2.575312	1.38405
	Gene level	2.419539	0.739848	1.070389
Fgf9	**Pr1**	-0.32193	0.454176	0.367371
	Pr2	-1.18442	-1.08927	-0.30401
	Gene level	-0.25154	0.321928	0.226509
Sox17	**Pr1**	0.757023	1.195348	1.31034
	Pr2	0.669027	1.713696	1.232661
	Gene level	0.62293	1.85997	1.464668
Gad1	**Pr1**	-2	-2	-0.41504
	Pr2	10.39874	11.29864	9.236014
	Gene level	-0.71312	-0.94342	0.903038
Olfm1	**Pr1**	1.195348	1.589763	1.15056
	Pr2	-0.91594	-1.25154	0.214125
	Gene level	0.650765	-1.39593	1.077243
Pax6	**Pr1**	0.970854	1.280956	0.799087
	Pr2	0.070389	0.748461	0.189034
	Pr3	3.193772	3.590961	2.157044

	Pr4	0.956057	1.367371	0.505891
	Gene level	2.944858	0.910733	1.084064

Table S11C: Expression (in log2 scale) estimates at the gene level and alternative promoters for five genes in primary medulloblastoma tumor and tumor derived cell lines from *Ptch+/-;p53-/-* mice relative to normal adult cerebellum. The expression values represent the averages of three PCR reactions.

Gene	**Expression At**	**Primary MB tumor**	**MB tumor derived cell lines**	
		Ptch+/; p53-/-	**Ptch+/; p53-/- CL1**	**Ptch+/; p53-/- CL2**
Fgf9	**Pr1**	-8.70275	ND*	ND*
	Pr2	ND*	ND*	ND*
	Gene level	-9.96578	ND*	-10.9658
Sox17	**Pr1**	ND*	ND*	ND*
	Pr2	-8.96578	ND	-8.96578
	Gene level	ND*	ND*	ND*
Gad1	**Pr1**	ND*	ND*	ND*
	Pr2	ND*	ND*	ND*
	Gene level	-12.2877	-15.6096	-14.6096
Olfm1	**Pr1**	ND*	ND*	ND*
	Pr2	-4.32193	-8.96578	-6.44222
	Gene level	-7.38082	-8.38082	-9.96578
Pax6	**Pr1**	-4.32193	-3.32193	ND*
	Pr2	-3.47393	-3.8365	ND*
	Pr3	-2.7369656	-2.7369656	ND*
	Pr4	-3.6438562	-3.6438562	-13.287712
	Gene level	-2.7369656	-3.1844246	-12.287712

*Not detected (ND)

Table S12: Distribution of alt events using single gene model.

Only UCSC gene model						
	Reference Set	**Observed alternative presented as # of events (# of genes)**				
	# of events (# of genes)	P0	P5	P15	Adult	Overall
Transcriptional events						
AFE	9696 (4177)	3034 (1411)	2881 (1354)	3069 (1421)	2583 (1200)	4069 (1864)

ALE	10927 (4621)	2820 (1307)	2863 (1331)	2903 (1341)	2942 (1357)	3834 (1762)
Splicing events						
Exon Skipping	14052 (3616)	1087 (883)	1093 (869)	1100 (836)	835 (657)	1688 (1272)
Intron Retention	368 (335)	132 (123)	132 (122)	124 (115)	123 (113)	178 (165)
A5SS	748 (687)	187(176)	172(159)	163(153)	131(126)	272(239)
A3SS	1500 (1279)	293(276)	292(276)	295(278)	243(232)	442(402)

Only RefSeq gene model						
	Reference Set	**Observed alternative presented as # of events (# of genes)**				
	# of events (# of genes)	P0	P5	P15	Adult	Overall
Transcriptional events						
AFE	2892 (1235)	838 (392)	823 (380)	842 (397)	693 (324)	1149 (524)
ALE	1797 (749)	737 (341)	719 (332)	722 (331)	701 (321)	851 (392)
Splicing events						
Exon Skipping	2099 (1262)	437 (377)	457 (387)	431 (352)	334 (281)	661 (531)
Intron Retention	37 (37)	12 (12)	14 (14)	10 (10)	10 (10)	17 (17)
A5SS	152 (137)	58 (54)	52 (46)	45 (42)	45 (43)	86 (85)
A3SS	397 (339)	97 (88)	93 (88)	93 (86)	78 (72)	144 (121)

Only Ensembl gene model		
	Reference	**Observed alternative presented as # of events (# of genes)**

	Set					
	# of events (# of genes)	P0	P5	P15	Adult	Overall
Transcriptional events						
AFE	7981 (3442)	2281 (1066)	2146 (1007)	2347 (1088)	2042 (947)	3202 (1460)
ALE	6452 (2793)	2192 (1019)	2213 (1030)	2222 (1024)	2227 (1018)	2943 (1341)
Splicing events						
Exon Skipping	9448 (4072)	1002 (785)	1004 (787)	1024 (763)	793 (605)	1574 (1146)
Intron Retention	767 (673)	73 (70)	63 (61)	67 (65)	66 (64)	99 (96)
A5SS	421 (388)	95 (91)	89 (81)	88 (83)	73 (68)	151 (128)
A3SS	1002 (880)	174 (163)	160 (151)	178 (166)	144 (134)	259 (231)

Only VEGA gene model						
	Reference Set	**Observed alternative presented as # of events (# of genes)**				
	# of events (# of genes)	P0	P5	P15	Adult	Overall
Transcriptional events						
AFE	18714 (6363)	6392 (2577)	6211 (2507)	6503 (2592)	6166 (2428)	8830 (3367)
ALE	15546 (5733)	5400 (2325)	5518 (2369)	5544 (2361)	5691 (2407)	7364 (3033)
Splicing events						
Exon Skipping	13872 (5389)	1031 (778)	1035 (794)	1031 (758)	804 (599)	1732 (1222)

Intron Retention	1835 (1368)	827 (668)	832 (687)	811 (662)	819 (648)	1073 (838)
A5SS	1127 (960)	137 (131)	132 (124)	132 (124)	115 (107)	250 (217)
A3SS	2070 (1537)	248 (221)	230 (203)	233 (208)	198 (173)	391 (328)

Experimental procedures

ChIP and mRNA sequencing

About 0.5 gram of mouse cerebellum tissues collected from CD1 mice at postnatal days 0, 5, 15, or 56 were used to prepare solubilized chromatin. For each stage multiple mice cerebellum tissues were pooled together for each ChIP experiment (for P0 about 16-18, for P5 about 10-12, for P15 about 5-8 and for P56 3-4 cerebellum tissues were pooled). Tissues were minced finely and cross linked with 1% formaldehyde for 10 min at room temperature. To stop cross-linking, glycine was added to a final concentration of 0.125 M. Next, the tissue samples were treated to isolate l nuclei and cross-linked chromatin was fragmented to a size range of 0.2-0.6Kb as described in Lee et al., 2006). Chromatin immunoprecipitation was performed, using 10µg of anti-Pol-II or H3K4me3 or H3K27me3 or IgG antibody that had been immobilized on Dynal magnetic beads. The antibodies against Pol-II, H3K4me3, H3K27me3 were purchased Abcam Inc. (ab5408, ab1012, ab6002 respectively). Following immunoprecipitation, the bound nucleoprotein complexes were extensively washed and the ChIP enriched DNA was eluted and purified using the Qiagen PCR purification kit as per manufacturer's instructions (Qiagen Inc.). This purified DNA was quantitated by picogreen assay and 10ng of the enriched DNA was further processed according to the Illumina Inc. instructions to sequence ChIP enriched DNA.

Two whole freshly dissected mouse cerebellum tissues from postnatal CD1 mice (P0, P5, P15, P56) was finely minced and resuspended in five volumes of Tri reagent (w/v). Total RNA was isolated from the Tri reagent-cell suspension according to the manufacturer's instructions (Sigma Inc.). The final RNA pellet was resuspended in DEPC-treated water and

concentration was measured by using nanodrop. To check the quality and integrity of total RNA, samples were analyzed on the Bioanalyzer (Illumina Inc). For mRNA sequencing, 10µg of total RNA was processed according to the instructions from Illumina Inc. to sequence mRNA, which include poly A tailed RNA purification, RNA fragmentation, cDNA synthesis, end repair and adapter ligation, purification of 175-225bp cDNAs. The purified cDNAs was enriched by 15 cycles of PCR as suggested by the Illumina's mRNA-seq protocol before proceeding to sequencing on the GAII (Illumina Inc). The PCR amplification step might render the mRNA-seq methodology semi-quantitative, though the limited amplification is believed to be within linear range.

Quantitative RT-PCR

Approximately 0.25µg of total RNA was reverse transcribed (RT) to generate cDNA using SuperscriptII following DNAseI treatment according to manufacturer's instructions (Invitrogen Inc.). We designed primers that would uniquely amplify a single transcript isoform to perform quantitative PCR. Using the specific primers for 22 distinct alternative promoter driven mRNA variants corresponding to 10 genes, we performed SYBR green based PCR on the reverse transcribed cDNA from each stage of cerebellum tissue (P0, P5, P15 and P56) as well as *Ptch+/-;p53-/-* medulloblastoma cell lines and primary tumors.

Bioinformatics analysis of mRNA-seq data

Generating the reference set for the study of alternative events and their occurrence in cerebellum from mRNA-seq data: We used 13 gene models and a phylogeny-based exon model for our study. All of these reference models were downloaded from UCSC genome browser database (Hsu et al., 2006). We divided the downloaded gene models into two sets: known and predicted gene models. The known gene models include five well annotated gene models: RefSeq (Pruitt et al., 2005), Vega (Wilming et al., 2008), Ensembl, UCSC (Hsu et al.,

2006) and MGI (Mouse genome informatics) (Eppig et al., 2007) . We consider these models as known because they are the most refered mouse gene models in the literature till date. The predicted gene models include eight tracks: Aceview (Thierry-Mieg and Thierry-Mieg, 2006), XenoRef (Kent, 2002) (non-mouse Ref Genes), TROMER (Lottaz et al., 2003), SGP (http://big.crg.cat/bioinformatics_and_genomics), MGC (http://mgc.nci.nih.gov/), N-SCAN (Gross and Brent, 2006), Genscan (Burge and Karlin, 1997), and Geneid (http://big.crg.cat/bioinformatics_and_genomics). The predicted models are generated either using purely computational approach or applying computational technique on experimental evidences such as ESTs, cDNAs. Additionally, the exon track based on phylogenetic information used in this study is called exoniphy (Kent, 2002). Finally, a non-redundant set of 112,537 transcripts were generated by combining the gene models and we defined a given gene as protein-coding, if there exists atleast one protein-coding transcript for the coprresponding gene in RefSeq, Entrez, and/or Vega gene models.

Alignment of mRNA-Seq data and analysis:

We follow a 4-step procedure to align mRNA-Seq sequences. These steps include: (1) filtering the raw reads, (2) alignment to junction library (3) alignment of genome (4) de novo assembly for novel junction discovery. Bowtie (Langmead et al., 2009) and TopHat (Trapnell et al., 2009) programs were used for alignment using thefirst 32-bases of each sequence read,.

(1) Filtering: In this step, mRNA-seq reads are compared with contamination sources (non PolyA tailed genes). The contamination sources include genes encoded by the mitochondrial genome, rRNA, and adapter sequences. The reads matching contamination library sequences were removed from further analysis.

(2) Alignment to junction library: This step involve mapping of filtered reads to non-redundant junction libraries. In order to generate a comprehensive list of splice junctions we first defined

gene boundaries based on known gene models and then extracted exon coordinates falling within the gene boundary from all of the 13 gene models. We obtained 1,806,576 non-redundant exon coordinate (419,575 known; 1,387,001 putative), and all possible combinations of splice junctions were generated for the exons belonging to a given gene. Overalll, we obtained 45,514,016 splicing junctions that includ 260,505 known (supported by a transcript from known gene models) and 45,253,511 putative junctions. For each of the splice junctions, 56 base sequences (last 28 bases of upstream exon and first 28 bases of downstream exon) were extracted and a bowtie index file was generated. Since we align 32-bases, the selection of 28 bases from both side of junctions ensures the mapping of atleast 4 bases to both exons. A similar approach to generate splice junction library was followed in previous mRNA-seq studies (Pan et al., 2008; Wang et al., 2008). For high confidence alignment, we consider only uniquely mapped reads and do not allow any mismatches for alignment.

(3) Alignment to Genome: The unaligned reads from step 2 were mapped to the reference genome. We allowed upto 20 multiple mapping with upto 2 mismatches for alignment. Multiple mapping was allowed in order to take care of genes present in multiple copies in the genome.

(4) De novo assembly for novel junction discovery: We applied TopHat program on the reads, which did not align to the reference genome in step 3. This step helps in discovering novel junctions arising from novel exons.

Identification of alternative events

We generated a library of the following alternative events: alternative first exon (AFE), alternative last exon (ALE), skipped exons, retained introns, alternative 5' splice sites (A5SS), and alternative 3' splice sites (A3SS). In our present study we considered non-overlapping first exons and last exons of genes as alternative first and last exon, respectively. We only included the transcripts present in combined known gene models for generating alternative event library.

In order to identify the expressed alternative events, we look for the presence of mRNA-seq reads in the corresponding exons (significant only if >= 1 read per base) and/or splice junctions (significant only if >= 2 distinct reads) participating in the event. If both isoforms that form an alternative event are found to be significantly expressed, then the event is considered to occur in postnatal cerebellum.

Estimation of transcript expression from mRNA-Seq data by IsoformEx

The observed number of reads from mRNA-Seq is usually summation of expression level of multiple transcripts. This occurs due to presence of several constitutive and overlapping exons from multiple transcripts. Thus, it is important to identify the expression levels of each transcript variants from observed number of reads inside exons. The expression of each transcript was estimated using our recently developed program – IsoformEx. By using combined transcript model having known transcripts (RefSeq, Vega, UCSC known gene, MGI, Ensembl) and predicted transcripts (Aceview, etc), we identified non-overlapping exon slices, splice junctions, and distinct isoform clusters (p-value < 0.01). The RPKM values of exon slices were computed by reads mapped to corresponding genomic regions, and the RPKM values of splice junctions were determined by reads mapped to splice junction database that was built from the combined transcript model. Multiple mapped reads were appropriately weighted for computing the RPKM of each exon slice. For each cluster, we built a combined transcript structure matrix having both an exon structure matrix and a splice junction structure matrix, of which element indicates whether a transcript uses the specific exon slice or the splice junction. Under the assumption that the observed RPKM values came from the summation of the expression of transcripts in an isoform cluster, we constructed an optimization problem with nonnegativity constraints to obtain proportions of contributions of the transcripts. The nonnegative constraints can be explained by the following nonnegative features of the given problem: (1) The RPKM value represents the molecular concentration of the fragments of mRNA, which should be nonnegative, (2) The

expression of each transcript in an isoform cluster should be nonnegative, i.e. zero expression (zero contribution to RPKM) or nonzero expression (nonzero contribution to the amount of tags inside exon slices or splice junctions). The optimization problem was solved by nonnegative least squares algorithm (Hanson, 1974) that is a well established optimization methodology, so the whole estimation of transcript variant abundance was computationally fast and numerically stable. More detailed description of the estimation approach based on nonnegative least squares can be found in (Kim et al., 2010).

Bioinformatics analysis of ChIP-seq data

ChIP-Seq data analysis: The ChIP-Seq data (anti-RNA Pol-II, anti-H3K4me3, anti-H3K27me3, and IgG-Control) analysis comprised of 3 major stages: (i) alignment, (ii) peak identification and (iii) promoter prediction. We used Bowtie program for the alignment of ChIP-Seq data and only. uniquely mapped reads with up to 2-mismatches were considered for analysis. The significant peak identification was performed using a two step procedure. In the first step statistically significant enriched genomic regions (of length = 1Kb) were identified. A region is defined as statistically significant, if the difference in number of reads between experiment (RNA Pol-II, H3K4me3, H3K27me3) and control IgG samples, within the region is higher than a given cutoff read count calculated using a *p*-value ≤ 0.05. Each ChIP-Seq read distribution in genome can be considered as Poisson distribution and difference of two Poisson distribution is given by a Skellam distribution (Skellam et al. 1946). Skellam probability mass function is given by,

$$P(n,\lambda_1,\lambda_2) = e^{-(\lambda_1+\lambda_2)}\left(\frac{\lambda_1}{\lambda_2}\right)^{n/2} I_n\left(2\sqrt{\lambda_1\lambda_2}\right)$$

where *n* represents number of reads, λ_1 , λ_2 are mean number of reads for two different samples, and $I_n\left(2\sqrt{\lambda_1\lambda_2}\right)$ is a modified Bessel function. After identification of enriched regions

the significantly enriched peaks in the experimental data were identified based on threshold read count obtained from a second cut-off *p*-value <= 0.01.

Promoter identification and annotation

To identify promoters, we apply our recently developed promoter prediction program (Gupta et al.2010) on each significant peak obtained from anti-RNA Pol-II and anti-H3K4me3 ChIP-seq data. The program divides the significant peaks into promoter and non-promoter peak classes. The predicted promoter peaks present within +/-1Kb of the first exon from known gene model transcripts are defined as known promoters and the remaining promoter peaks are refered as novel promoters. If a novel promoter peak is found to be within +/-1 Kb of the first exon corresponding to the novel gene model transcripts, then it is termed as assigned novel promoter otherwise the promoter peaks are left unassigned. Further, the unassigned promoter peaks are combined if they are within 500 bases. We also included the non-promoter peaks that are present within +/- 1 Kb of known transcript's first exon in the known promoter category.

H3K4me3 and H3K27me3 marks at the promoter and expression of corresponding transcripts.

Individual relationship of H3K4me3 or H3K27me3 enrichment with expression from promoters

We performed this analysis on the active promoters at three levels-(i) globally, (ii) clusters of 50 promoters, (iii) CpG rich vs CpG poor promoter clusters. (i) For the global analysis all promoters were divided based on the corresponding transcript expression into high, medium and low expression groups, and the enrichment of either H3K4me3 or H3K27me3 around the TSS (-1.5Kb to + 1.5Kb) was plotted for each group. (ii) Next, we performed a detailed analysis, where the promoters were grouped in clusters of 50 based on the expression of their transcripts. Only

promoters that are atleast 1.5 Kb apart were considered for this analysis to avoid the effect of H3K4me3 and H3K27me3 spreading to neighboring promoters. Finally, a scatter plot was generated between the average expression of the cluster and the average methylation of the cluster for either H3K4me3 or H3K27me3 computed from the enrichment of the marks in the region (-1.0Kb to +1.0Kb) around TSS for each promoter in the cluster. The best possible curve based on highest R^2 value was fitted for each scatter plot. (iii) To determine the association of either mark with CpG, we first divided all active promoters into CpG rich and CpG poor category. Both classes of promoters were then subgrouped based on the expression of their transcripts into clusters of 25 promoters and analysis was performed as described above for (ii).

Role of H3K4 and K27 trimethylation in the choice of alternative promoters

To address this issue we selected the alternative promoters belonging to two-promoter genes, where the alternative promoter driven transcript expression differ by atleast two fold in the same stage, such that one promoter is upregulated while the other is downregulated with respect to one another. We combined all these alternative promoters identified from each of the four stages and computed the log[H3K4me3(promoter1)/ H3K4me3 (promoter2)] and log[H3K27me3(promoter1)/ H3K27me3 (promoter2)] on each set of alternative promoters . We plotted a heatmap for the expression change of the alternative promoters and the corresponding log ratios of both methylation marks to visualize the impact of either mark on the selection of the upregulated/downregulated alternative promoter.

Role of H3K4me3 and H3K27me3 in regulating expression during development

To analyze the regulation of transcript expression during development and the contribution of H3K4me3 and H3K27me3 in the process, we performed the analysis on the set of promoters that are developmentally regulated. First we performed all possible two stage comparisons e.g P0-P5, P0-adult and so on to identify the promoters whose expression is either up or down

regulated between the two stages and combined the different sets to create the list of developmentally regulated promoters. Next, we selected only those promoters that were marked by both methylation marks in each of the two stages from where they were initially identified as being developmentally regulated. These are the promoters that are being regulated during development through the enrichment of both H3K4 and H3K27 trimethylation. Then we computed the log[H3K4me3(stage 1)/ H3K4me3 (stage 2)] and log[H3K27me3(stage 1)/ H3K27me3 (stage 2)] for each upregulated and downregulated promoter. We generated a 3-D plot with log[expression (stage 1)/ expression (stage 2)] on the Z-axis and Log ratios of H3K4me3 and H3K27me3 on X and Y axes respectively.

Supplemental References

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J Mol Biol *268*, 78-94.

Eppig, J.T., Blake, J.A., Bult, C.J., Kadin, J.A., and Richardson, J.E. (2007). The mouse genome database (MGD): new features facilitating a model system. Nucleic Acids Res *35*, D630-637.

Gross, S.S., and Brent, M.R. (2006). Using multiple alignments to improve gene prediction. J Comput Biol *13*, 379-393.

Gupta, R., Wikramasinghe, P., Bhattacharyya, A., Perez, F.A., Pal, S., and Davuluri, R.V. Annotation of gene promoters by integrative data-mining of ChIP-seq Pol-II enrichment data. BMC Bioinformatics *11 Suppl 1*, S65.

Hanson, C.L.L.a.R.J. (1974). Solving Least Squares Problems (Prentice-Hall).

Hsu, F., Kent, W.J., Clawson, H., Kuhn, R.M., Diekhans, M., and Haussler, D. (2006). The UCSC Known Genes. Bioinformatics *22*, 1036-1046.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-664.

Kim, H., Bi, Y., and Davuluri, R.V. (2010). Estimating the Expression of Transcript Isoforms from mRNA-Seq via Nonnegative Least Squares. Paper presented at: Proceedings of the the 10th IEEE International Conference on Bioinformatics and Bioengineering (BIBE-2010) (Philadelphia, PA, USA).

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Lee, T.I., Johnstone, S.E., and Young, R.A. (2006). Chromatin immunoprecipitation and microarray-based analysis of protein location. Nat Protoc *1*, 729-748.

Lottaz, C., Iseli, C., Jongeneel, C.V., and Bucher, P. (2003). Modeling sequencing errors by combining Hidden Markov models. Bioinformatics *19 Suppl 2*, ii103-112.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet *40*, 1413-1415.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res *33*, D501-504.

Thierry-Mieg, D., and Thierry-Mieg, J. (2006). AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome Biol *7 Suppl 1*, S12 11-14.

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

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476.

Wilming, L.G., Gilbert, J.G., Howe, K., Trevanion, S., Hubbard, T., and Harrow, J.L. (2008). The vertebrate genome annotation (Vega) database. Nucleic Acids Res *36*, D753-760.