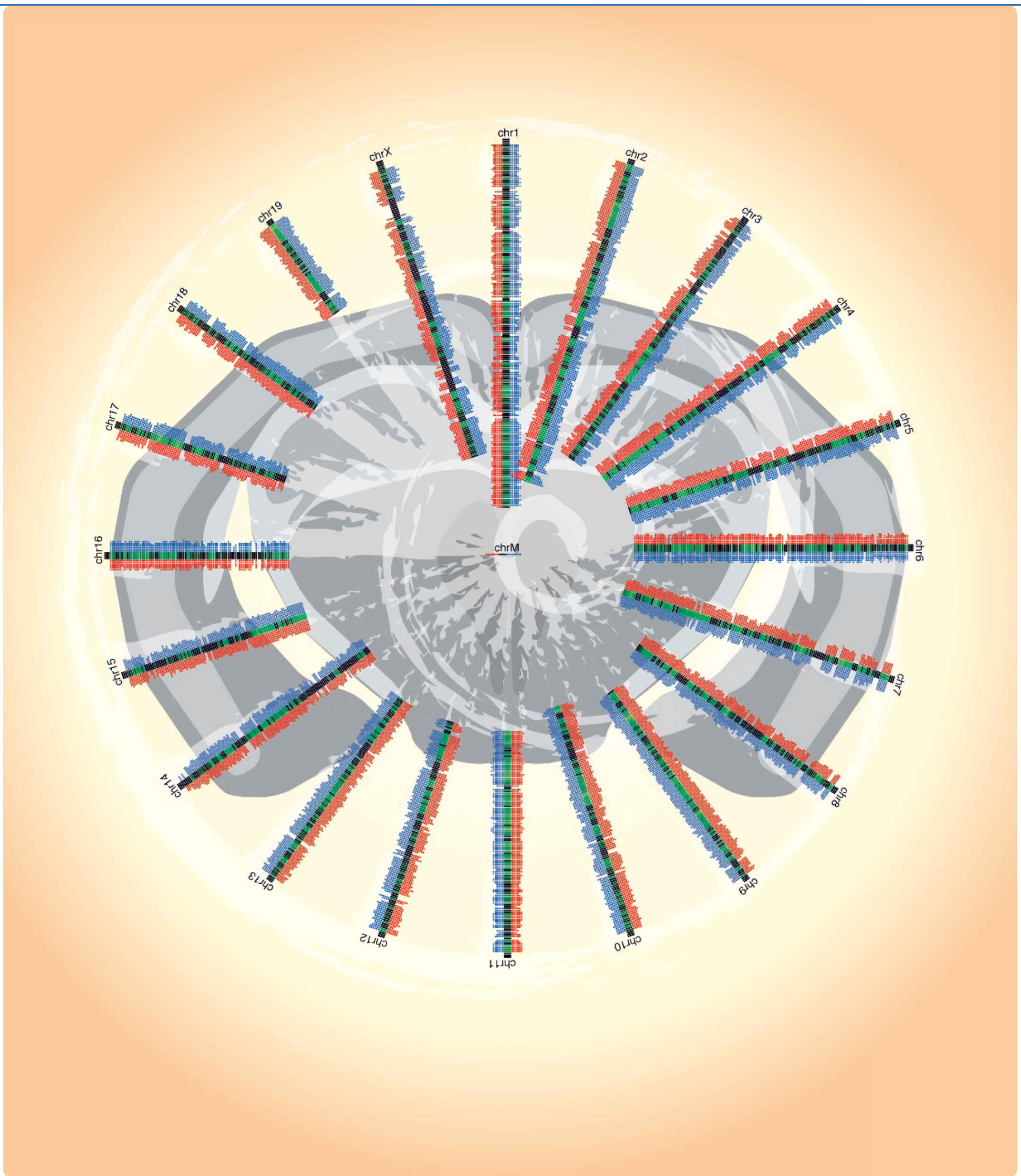
BMC Genomics



Comparative analysis of neural transcriptomes and functional implication of unannotated intronic expression

Sun *et al.*

BioMed Central

Sun *et al. BMC Genomics* 2011, **12**:494
http://www.biomedcentral.com/1471-2164/12/494 (10 October 2011)

BMC Genomics

RESEARCH ARTICLE Open Access

Comparative analysis of neural transcriptomes and functional implication of unannotated intronic expression

Yazhou Sun[1,3,4], Yaqiong Wang[2,5], Yi Hu[1,3], Gong Chen[1,3,4] and Hong Ma[1,2,3,4,5*]

Abstract

Background: The transcriptome and its regulation bridge the genome and the phenome. Recent RNA-seq studies unveiled complex transcriptomes with previously unknown transcripts and functions. To investigate the characteristics of neural transcriptomes and possible functions of previously unknown transcripts, we analyzed and compared nine recent RNA-seq datasets corresponding to tissues/organs ranging from stem cell, embryonic brain cortex to adult whole brain.

Results: We found that the neural and stem cell transcriptomes share global similarity in both gene and chromosomal expression, but are quite different from those of liver or muscle. We also found an unusually high level of unannotated expression in mouse embryonic brains. The intronic unannotated expression was found to be strongly associated with genes annotated for neurogenesis, axon guidance, negative regulation of transcription, and neural transmission. These functions are the hallmarks of the late embryonic stage cortex, and crucial for synaptogenesis and neural circuit formation.

Conclusions: Our results revealed unique global and local landscapes of neural transcriptomes. It also suggested potential functional roles for previously unknown transcripts actively expressed in the developing brain cortex. Our findings provide new insights into potentially novel genes, gene functions and regulatory mechanisms in early brain development.

Keywords: neural transcriptomes, stem cell, intronic expression, embryonic brain cortex, neonatal brain cortex

Background

It is well known that total gene numbers are similar among multicellular eukaryotes, and genome size does not correlate with organism complexity, which differs greatly in terms of development, physiology and behavior among eukaryotes [1]. The transcriptome and its regulation contribute significantly to eukaryotic diversity in the aforementioned complexity. The Functional Annotation of the Transcriptome of Mammalian Genome (FANTOM) projects (FANTOM 1-4) have demonstrated the complexity of transcriptomes in several aspects, including non-coding RNAs [2], antisense transcription [2,3], regulated retrotransposon expression [4], and alternative promoter usage, splicing and polyadenylation [5].

Recent high-throughput RNA-seq [6] technologies have provided unprecedented capability to analyze cellular, tissue-specific, or organismal gene activities across a broad spectrum. It also revealed the transcriptomic complexity during cell differentiation [7,8] and organ development [9]. Furthermore, individuals of the same species have transcriptomic differences such as expression variation among humans [10]. Another level of transcriptomic complexities has been revealed by extensive analysis of novel splicing variants from known exons [7-11]. In addition, thousands of transcripts from previously unannotated (non-exonic) genomic regions have been reported [7,8,10-13]; they are either named TUF (Transcripts of Unknown Function) [14] or unannotated TAR (Transcriptionally Active Region) [15].

* Correspondence: hxm16@psu.edu
[1]Department of Biology, Pennsylvania State University, University Park, PA 16802, USA
Full list of author information is available at the end of the article

Some of the unannotated TARs are large intergenic noncoding RNAs that function in embryonic stem cell pluripotency and cell proliferation [16,17], while most unannotated TARs have no known function.

It has been reported that undifferentiated human stem cells have elevated expression of unannotated TARs compared with differentiated neural progenitor cells [7]. Our recent study has also detected additional transcripts from intergenic regions and introns in mouse embryonic and neonatal brain cortices [9]. Mammalian neural development is a complex process involving cell division, cell differentiation, cell migration, axon guidance, synaptogenesis, and synaptic plasticity. The characterization of stage specific unannotated TARs during early brain development could provide clues regarding the roles these unannotated TARs might play in determining neural fate and in regulating neuronal functions.

To further investigate the transcriptome dynamics and to better understand the possible roles of unannotated TARs in early neural development, we have analyzed the RNA-seq datasets from embryonic and postnatal mouse brain cortices that we generated recently [9], as well as seven additional RNA-seq datasets covering both neural and non-neural tissues [7,18]. These nine transcriptome datasets include data from human embryonic stem cell (hESC) and its subsequently differentiated forms (N1, early initiation; N2, neural progenitor; and N3, early glial-like cell) [7], embryonic day 18 (E18) and postnatal day 7 (P7) mouse brain cortices [9], and adult mouse brain (AMB), liver (AML), and muscle (AMM) [18].

Through a systematic analysis of these nine datasets, we found several unique characteristics of the transcriptomes in early neural development. We found that, although the genome was not as pervasively transcribed as previously reported [19], most of the genomic regions at 1 Mb resolution had detectable RNA-seq signals. We also found that the transcriptomes from neural tissues possessed several genome-wide characteristics resembling those of stem cells. Interestingly, the E18 cortex shows the highest level of unannotated transcript expression compared to P7 and adult brains. Furthermore, the intronic unannotated transcripts are associated with GO terms for neurogenesis, neural signaling and negative regulation. Importantly, few of the unannotated TARs in E18 and P7 cortices are connected with known transcripts, suggesting potential novel functions of these TARs during brain development.

Results and discussion

Mapping RNA-seq data from mouse developing brains and other organs

To examine the genomic distribution of transcriptomic reads, we mapped all RNA-seq data by the TopHat software [20], which was designed to map RNA-seq data with moderate IT (information technology) infrastructure. Embryonic, neonatal and adult mouse data were mapped onto the mouse reference genome (UCSC mm9, NCBI build 37) as described in the Methods section. For comparison, adult mouse liver and muscle RNA-seq data were analyzed in the same manner. Human ESC and its differentiation data were mapped onto the human reference genome (UCSC hg19, NCBI build 37). Only uniquely matched reads were further analyzed. Because the amounts of data available for the downstream analysis varied for different tissues (Table 1), the read count data were first normalized against the available dataset size, measured in base pairs, for each tissue. To accommodate differences in the details of library preparation and sequencing procedures, we adjusted for the sequencing quality in all data sets according to the quality computation method of the Illumina 1.3 pipeline. E18 and P7 RNA-seq data had the largest percentage of mappable reads in this group, approximately 60%, while other data sets had about 30% mappable data. The AMB RNA-seq data had the largest quantity, with more than 1,000 million mapped base pairs.

Most 1-Mb genomic regions have transcriptional activity with uneven distribution at a finer scale

Despite a dataset size up to 1,000 Mbp (AMB), only about 2% of the mouse genome had been mapped with RNA-seq reads, unlike a previous report that suggested a more pervasively transcribed genome from detailed analyses of ~1% of the human genome [19]. A recent study using single- and paired-end RNA-seq and tiling arrays also concluded that the genome is not as pervasively transcribed as previously reported [12]. Nevertheless, when the mouse genome was divided into 1 million base-pair long intervals, we observed that 85% of the intervals had 100 or more detected RNA-seq reads (Figure 1). However, the majority of chromosome regions of the 100 kb size had no detected transcription, as illustrated for X-chromosome (Additional file 1, Fig. S1). Whereas many regions showed transcription at both the E18 and P7 stages, some regions were specifically active in one of these stages, suggesting that a selective set of genes are turned on and off from embryonic to neonatal brain stages, consistent with the previous finding of preferential expression of several thousand genes at either one of these stages [9]. Besides heterochromatin regions (centromeres and telomeres), some large genomic regions with very low annotated gene content had no detectable reads, including several regions in chromosome 7 (Figure 1) and the region in chromosome × from 26,000,000 to 32,000,000 bp (Additional file 1, Fig. S1).

Table 1 RNA-seq mapping result using TopHat

	hESC[1] human embryonic stem cell	**hN1[1] early hESC initiation**	**hN2[1] human neural progenitor**	**hN3[1] human early glial-like**	**E18 mouse brain cortex at E18**	**P7 mouse brain cortex at P7**	**Adult brain[2] mouse**	**Adult liver[2] mouse**	**Adult muscle[2] mouse**
Sequencing Platform	Illumina								
Accession Number	SRR037165 to SRR037170	SRR037193 to SRR037198	SRR037199 to SRR037205	SRR037220 to SRR037226	SRP007262	SRP007262	SRR001356 SRR001357 SRR006488 and SRR006489	SRR001358 SRR001359 SRR006490 and SRR006491	SRR001361 SRR001362 and SRR006492
Read Type	35 bp PE and 35 bp SE				36 bp PE and 36 bp SE		33 bp SE		
Original Read Count	14.4 M PE and 4.4 M SE	15.8 M PE and 7.0 M SE	19.6 M PE and 11.6 M SE	22.4 M PE and 3.0 M PE	10.0 M PE and 3.0 M SE	10.4 M PE and 3.6 M SE	89.0 M SE	75.9 M SE	59.9 M SE
Original RNA-seq data size (Mbp)	1,162	1,352	1,774	1,819	853	857	2,936	2,503	1,977
Mapped by Tophat* (Mbp)	401	458	628	594	491	525	1,086	683	571
Percentage Mapped (%)	35	34	35	33	58	61	37	27	29

* Allowing 1 hit by TopHat

1. Data from Wu *et al.* [7].

2. Data from Mortazavi *et al.* [18].

After further dividing the genome into 10 kb intervals, most of the 10 kb intervals had no detected transcription (Figure 2A). Most of the reads in highly expressed intervals (> 1,000 RNA-seq reads) were mapped to known exons. Among intervals with 1,000 to 10,000 detected reads, more intervals had intergenic transcripts than intronic transcripts, with very few intervals having all three types (exonic, intronic and intergenic) of transcripts. For intervals with 100 or more RNA-seq reads, there were 3 times more intervals expressing intronic signal in E18 stage than in P7 stage. The E18 stage also had slightly more intervals with intergenic transcripts than the P7 stage, although the numbers of intervals with exonic transcripts were similar between the two stages (Figure 2B).

We found that expression level was positively correlated with the exon contents of a given interval. The higher exon percentage an interval had, the higher the number of detected RNA-seq reads in the interval. At 1 Mb interval size, the Pearson's Correlation Coefficient between exonic read number and exonic content percentage was 0.60; in contrast, the Pearson's Correlation Coefficient between intronic read number and exonic content percentage was only 0.18. The Pearson's Correlation Coefficient between intergenic read number and exonic content percentage was also quite low, at 0.26 (Figure 3A and 3B). At 100 kb interval size, the correlation decreased but the general trend was maintained (Figure 3C). While the exon-rich regions had more reads for exons than introns or intergenic regions, the exon-sparse regions had similar numbers of intronic and intergenic reads as the exon-rich region. In addition, at 1 Mb resolution, the exonic expression level had modestly positive correlation with both intronic (R = 0.43) and intergenic expression level (R = 0.37) in the same region (Additional file 1, Fig. S2 **A** and **B**). However, at 100 kb resolution, the aforementioned correlation became very weak (Additional file 1, Fig. S2 **C** and **D**).

Transcriptomic comparison between neural tissues and other tissues

To understand the neural transcriptome characteristics, we compared mouse cortical RNA-seq data at E18 and P7 stages with available adult mouse brain, liver, and muscle RNA-seq data [18], as well as RNA-seq data from human embryonic stem cells (hESC) and neural cells (N1, N2, N3) immediately differentiated from hESC [7]. We first analyzed transcriptome properties at the chromosome level, using a method slightly modified from Mortazavi *et al.* [18] as detailed in the Methods section, and labeled as RPKM* (similar to RPKM; formula (1)).

In addition to the above-mentioned mapping of the 5 mouse datasets (E18 and P7 cortices, and adult brain, liver and muscle), we also mapped the 4 human RNA-seq datasets (hESC, N1, N2 and N3) onto the mouse

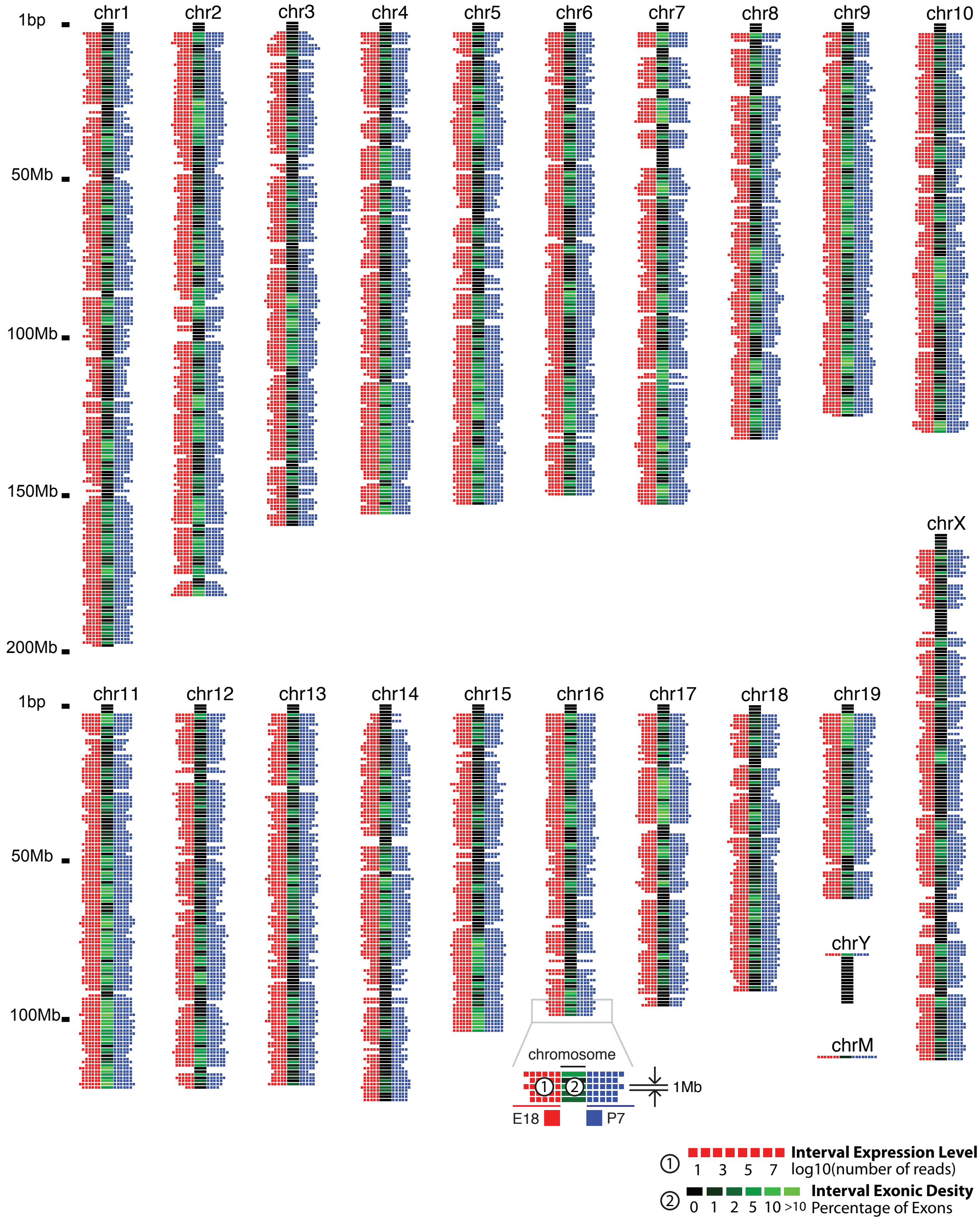


Figure 1 Genome-wide expression map at 1 Mb resolution for E18 and P7. Each 1 Mb interval is represented by a horizontal box, with color ranging from black to green. A box with black color means that there was no exon in this given interval. The brighter the green color, the higher the exon content (percentage) this given interval possesses. The scale at the bottom of the map illustrates color representation of the exon percentage of the region. Red square at the left side of the chromosome represents the detected number of RNA-seq reads for the given region in E18 stage, while the blue square at the right side represents the detected number of RNA-seq reads in P7 stage. Expression level is data size normalized. Scale for expression level is at the bottom of the map.

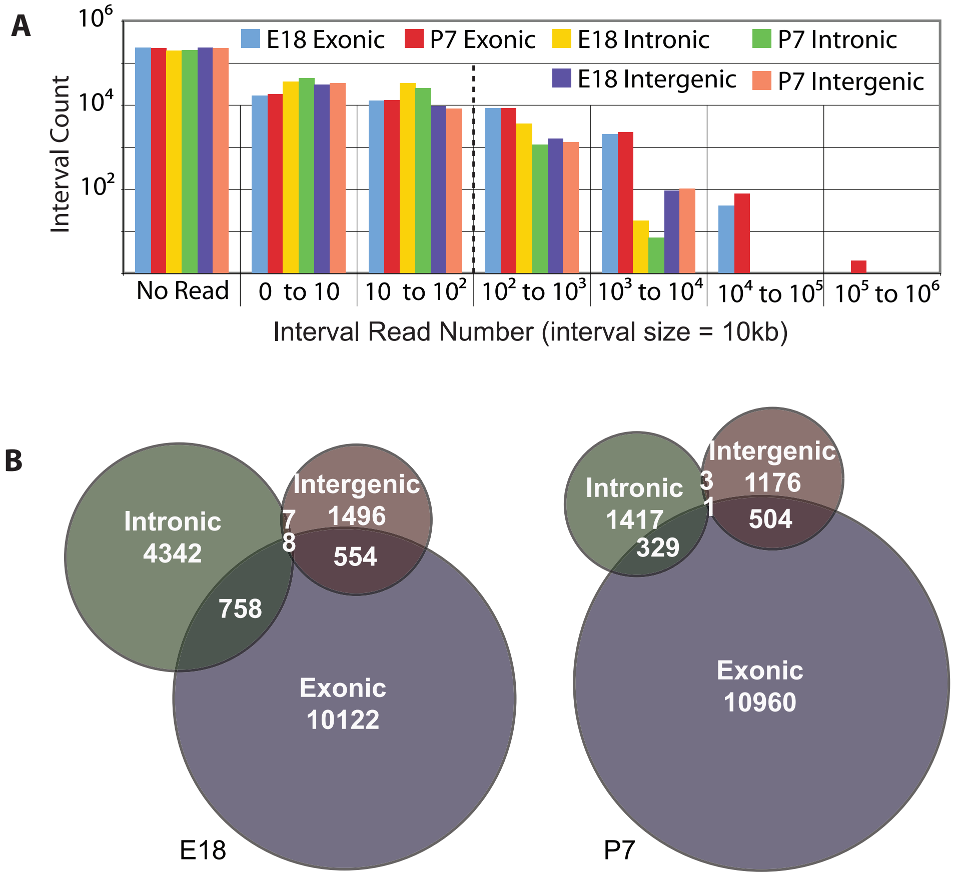


Figure 2 Genomic expression analysis in 10 Kb interval. A. Number of intervals in different categories of RNA-seq read coverage. RNA-seq reads were categorized as: *exonic*, which were reads mapped to known exon, *intronic*, which were reads mapped to known intron, and *intergenic*, which were reads mapped to known intergenic region. There were more highly expressed intervals with exonic expression than intronic or intergenic ones. B. Venn diagram illustration of the above result. Only intervals containing more than 100 reads were considered, as indicated by the dotted line in A.

reference genome (mm9 in the UCSC database; [21]). Based on the 85% identity calculated from coding regions between mouse and human genome previously [21], there would be on average 5 mismatches per 35 bp RNA-seq read length. We found that the threshold of 2 mismatches per 35 bp read achieved the best balance between specificity and sensitivity for this cross-species mapping. Increasing the number of allowed mismatches resulted in fewer uniquely and correctly mapped reads, while decreasing this number resulted in fewer total mappable reads. With the threshold of allowing maximum 2 mismatches for RNA-seq mapping, this would mean little cross-species mappable reads if the differences between coding regions of human and mouse were distributed evenly. Surprisingly, on average 6% of the total human RNA-seq reads could be mapped to the mouse genome, or 11.5% of the reads mappable to the human genome. The majority (80%) of the reads mapped to the mouse genome were also mapped to the human genome.

We then compared the chromosomal expression profile between mouse (E18, P7, AMB, AMM and AML) and human (hESC, N1, N2 and N3) samples. Despite the fact that smaller fractions of the human reads were mapped onto the mouse reference genome, largely due to DNA sequence similarity between these two species, both mouse and human neural data sets were highly similar in terms of expression level relative to total mapped reads at the scale of individual chromosomes (Additional file 1, Fig. S3A). Furthermore, all neural tissues had similar profiles to that of stem cells, but quite different from those of adult liver and muscle, particularly for some chromosomes, such as 5 and 7 (Additional file 1, Fig. S3B). To summarize the above information, we applied Correspondence Analysis on the chromosomal expression profile. The chromosomal expression distribution among all stages, using mouse genome as reference, was first measured in read counts (Additional file 2, Table S1). Correspondence Analysis

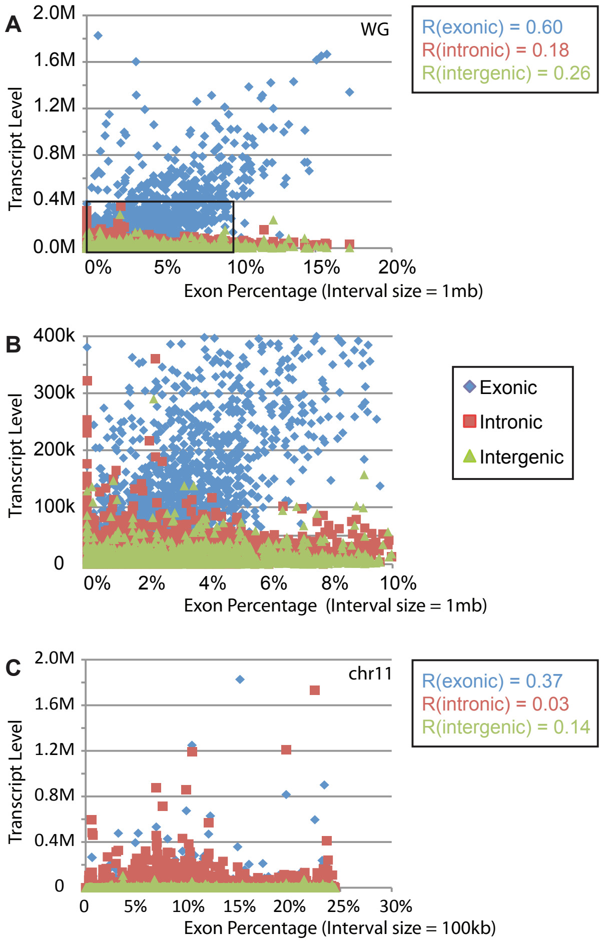


Figure 3 Correlation between exon density (measured by percentage of exon regions) and expression level. A. Interval size = 1 Mb. Regional percentage of exon vs. exonic expression (blue diamond), intronic expression (red square), intergenic expression (green triangle). B. Zoomed in view of the boxed region in A. C. Interval size = 100 kb. Only data from chromosome 11 were shown.

(CA) was performed using the ca package in R [22]. The first two dimensions resulted from the CA could explain 96% of the differences in the original 9 × 21 dimensions, which indicated that the first two dimensions were representative (Additional file 2, Table S2). The result was plotted using these two dimensions (Figure 4). E18 and P7 were clustered together with hESC, N1, N2 and N3. AMB was also very close to the aforementioned cluster. AML and AMM were significantly further away than AMB.

As an alternative way to compare the mouse and human data, we also mapped the human data onto the human reference genome (hg19; [23]), for comparative analysis using previously identified syntenic/orthologous genomic regions between mouse and human [21,23]. These studies defined 217 conserved syntenic blocks between the human and mouse genomes. Chromosomal expression profiles in early-differentiated human neural cells were very similar to that of human embryonic stem cells (Additional file 1, Fig. S3C). Even with different chromosome numbers and organizations, neural chromosomal expression profiles were also very similar between human neural cells and mouse neural tissue samples between syntenic/orthologous genomic regions (Additional file 1, Fig. S3B, C). For example, the most highly expressed chromosome in mouse was chromosome 11, whose human counterpart is chromosome 17, which was the second most highly expressed chromosome in human. The most highly expressed chromosome in human was chromosome 19, whose mouse counterparts are distributed on chromosomes 7, 8 and 19, among which chromosomes 7 and 19 were also highly expressed in neural tissues.

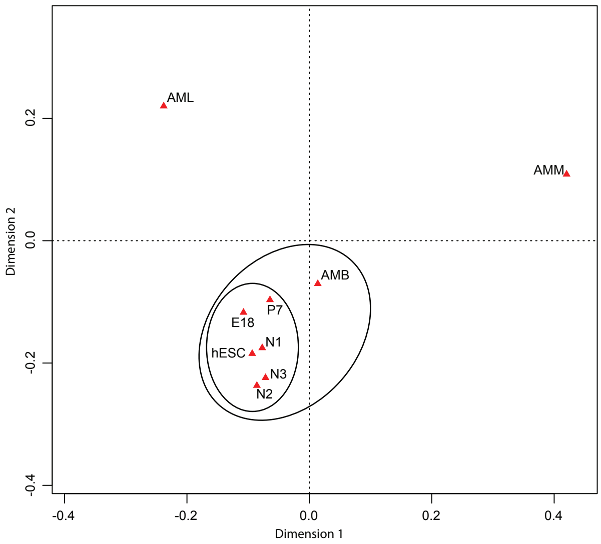


Figure 4 Correspondence Analysis result on stage similarity based on comparison between chromosomal expression profiles. Similar stages were closer to each other than dissimilar ones. Analysis done using ca package in R [22].

To assess the variation in expression levels between chromosomes for different tissues/organs, we calculated the standard deviation for the distribution of individual chromosome expression level, or RPKM* values, for each mapped transcriptome dataset (Additional file 1, Fig. S3D). We found that the standard deviation for E18 was the lowest among mouse samples, while the standard deviation for the stem cells was the lowest in human samples (Additional file 1, Fig. S3D). These results indicated that the mouse E18 brain cortex and human embryonic stem cells use chromosomes more evenly than other organs/tissues.

It is well known that the brain has a very high metabolic rate, consuming a significant amount of energy while lacking substantial energy reserve tissues. Thus normal brain functions depend on mitochondria as the crucial energy provider. To examine the mitochondrial genome expression-level changes across different developmental stages, we plotted the normalized mitochondrial expression level, measured in RPKM*, across all nine datasets and normalized against the dataset size. We found that in human datasets, compared with stem cells, differentiated neural cells had a higher level of mitochondrial expression (Additional file 1, Fig. S3E), increasing from the N1 to N2 stages, then maintaining a similarly high level at the N3 stage. Similarly in the mouse brain, mitochondrial expression progressively increased from the E18 embryonic stage to the P7 neonatal stage, then to the adult stage. The adult mouse brain had a similar level of mitochondrial expression to that of the adult liver, while the adult mouse muscle had the highest level of mitochondrial expression among all analyzed organs/tissues, consistent with high energy demand for muscle contraction.

To assess the similarity between neural and stem cell transcriptomes further, we compared the transcriptomes between human and mouse using only 1-to-1 orthologous gene pairs between these species. A total of 12168 orthologous gene pairs were identified using the MGI (http://www.informatics.jax.org) orthology database. The expression level of each gene was measured in RPKM, with the modification that detected base pairs from exons were used instead of read number in the RPKM formula to accommodate read length differences between datasets. Again the expression level for a given gene at a given stage was normalized against the RNA-seq dataset size. We then added 1 to the calculated expression level value for each gene, to ensure valid logarithm transformation. The calculated value was then $\log_2$ transformed. We first calculated the gene expression correlation between hESC and the rest of the samples. As expected, the correlation was high between hESC and the cells derived from hESC. Among the rest, E18 had the highest correlation with hESC (R = 0.61), and P7 had the second highest (R = 0.57). The data were then analyzed in MeV (MultiExpriment Viewer) from the TM4 suite [24]. After using different clustering methods (Hierarchical clustering and K-means) and different distance calculation methods (Pearson's Correlation and Euclidean Distance), we found that neural datasets from human and mouse were consistently grouped together with the stem cell dataset, separate from the liver and muscle datasets (Figure 5A). This suggested that the neural and stem cell transcriptomes were globally more similar in terms of orthologous gene expression than they were to liver and muscle transcriptomes.

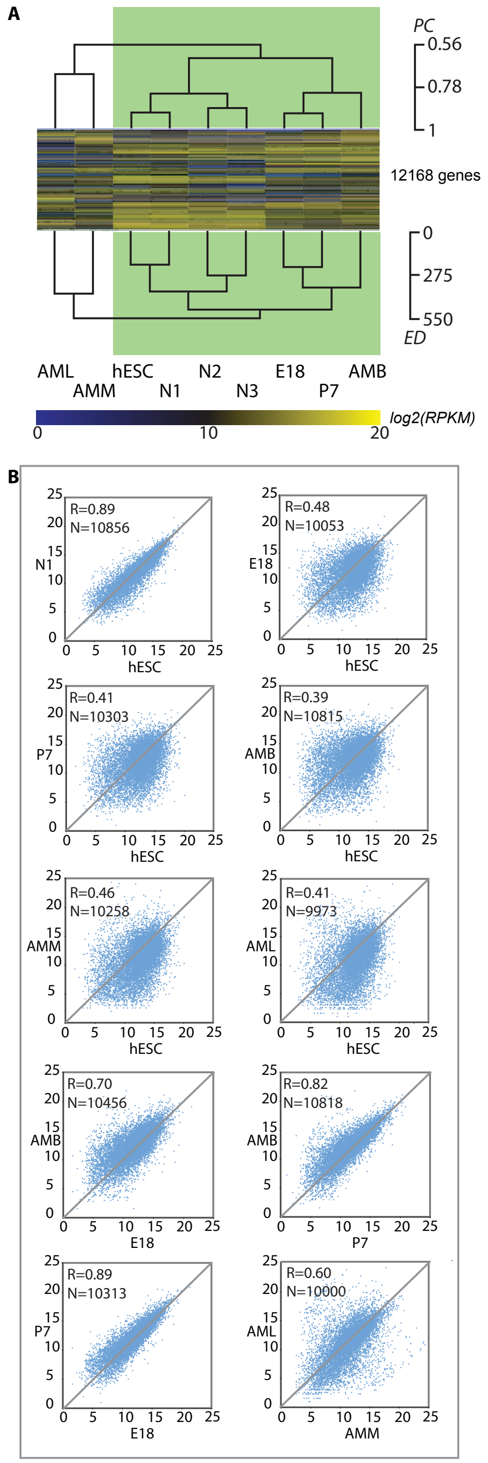


Figure 5 Comparisons of human and mouse expression data. **A**. Clustering of nine RNA-seq datasets using orthologous genes between human and mouse. Center panel: expression level heat map overview of 12168 orthologous genes. Expression level is measured in RPKM [18]. The calculated number is increased by 1 to ensure valid logarithm transformation, and then $\log_2$ transformed. The sample tree on top is derived from Pearson's Correlation (*PC*) distance matrix, while the one at the bottom is derived from

Euclidean Distance (*ED*) matrix. Both trees support the same topology, which grouped all neural stages with stem cell stage, while left muscle and liver in a separate group. Scale of color representation of the expression level is at the bottom of the figure. **B**. Scatter plot of orthologous gene expression level between selected stages. Genes without detectable expression were not included.

We then further analyzed the correlation of the expression in two different tissues/stages among co-expressed genes between the tissues/stages (Figure 5B and Additional file 1, Fig. S6). Among all mouse samples, although E18 was the one with the highest correlation with hESC, E18 was still more similar to mouse neural transcriptomes in terms of expression level correlation. In particular, E18 and P7 transcriptomes were much more correlated with each other than with hESC, suggesting that the similarity between E18 or P7 cortex and hESC is relatively limited.

We also analyzed the genes associated with the pluripotency of stem cells. *Sox2*, *Myc*, *Oct4* and *Klf4*, which are four genes that have been found in multiple studies to convert human and mouse somatic cells to induced pluripotent stem (iPS) cells [25-29]. Similarly, *Oct4* and *Sox2*, plus two other factors, *Nanog* and *LIN28A*, were also able to induce iPS cells from human somatic cells [30]. We found that *Sox2*, *Myc* and *Klf4* were detectable in all neural samples we analyzed (Figure 5A). Specifically, the E18 stage had the highest *Sox2* and *Myc* expression among all mouse samples. However, *Nanog* and *LIN28A* were not detected at either the E18 or P7 stage. RT-PCR experiment was also carried out with primers specifically targeting *Sox2*, *Myc* and *Klf4*. The results supported the expression of *Sox2*, *Myc* and *Klf4* in E18 and P7 cortices (Figure 6B). Due to the presence of multiple homologues of *Oct4* (also called *Pou5f1*) in the mouse genome, neither RNA-seq nor RT-PCR could identify specific expression for *Oct4/Pou5f1*. Interestingly, *Sox11*, which encodes a transcription factor and was previously reported to be expressed in glial cells [31], was highly expressed in the E18 cortex and significantly down-regulated in the P7 cortex (Figure 6A &6B). Because cortical neurons are mainly generated at the late embryonic stage whereas glial cells are mainly generated in the postnatal stage, the high level expression of *Sox11* in the E18 cortex suggests an additional role during early brain development besides its proposed function in glial cells.

Expression characteristics of genes associated with neurodevelopmental disorders

We further analyzed the expression of genes potentially associated with neurodevelopmental disorders in both neural tissues and stem cells. Autism spectrum disorders (ASD), together with schizophrenia and mental retardation, are typically characterized as neurodevelopmental disorders. Genome-wide association studies (GWAS) have identified many genes related to ASD. Among these genes, many have been found to relate to the GABAergic neurotransmission system. Here, we analyzed the expression of 20 genes encoding different GABAA receptor subunits and 25 genes that have been proposed to be associated with ASD [32,33] (Figure 7A and 7B). Although GABA receptor genes in general showed low expression levels in non-neural tissue and stem cells, the gene for GABA$_A$R alpha5 (GABRA5) subunit showed a very high level of expression in hESCs, which lacks a GABAergic system, suggesting a novel function of this gene in embryonic stem cells. Its expression was not detectable after initial differentiation (N1) and then observed again after further differentiating into neural cells (N2). It also showed an increase from E18 to P7, but a reduction from P7 to adult brain, consistent with more restricted localization in the adult brain [34]. In addition, GABRQ and GABRP also showed modest expression in hESCs. Furthermore, the gene encoding the GABA$_A$R delta (GABRD) subunit showed the highest expression level among all GABA$_A$R subunit genes in the adult brain. Since delta is specifically localized at extrasynaptic sites and mediates tonic inhibition rather than normal fast inhibition, this result emphasizes the importance of tonic inhibition in regulating adult brain activity. Finally, genes for GABA$_B$ receptor (GABBR1 and GABBR2) subunits showed significant differential expression during brain development, with GABBR1 dominant from E18 through adult brain while GABBR2 only expressed highly in the adult brain.

Among genes associated with ASD, *RGS4*, *DTNBP1*, *NLGN2*, *STX7*, *MECP2*, *ARVCF*, and *PPP3CC* all showed high-level expression from the embryonic to adult brain. One important finding was that while both *NLGN1* and *NLGN2* showed high-level expression at the E18 to P7 stages, consisting with their synaptogenic functions, *NLGN1* expression was significantly reduced in the adult brain, suggesting that the relevant function might be fulfilled by other cell adhesion molecules. This is also consistent with the current understanding that many cell adhesion molecules can trigger glutamatergic synapse formation as NLGN1 does, but only NLGN2 is capable of inducing GABAergic synaptogenesis [35,36]. One surprising finding is that *DISC1*, a well-studied gene associated with schizophrenia [37], showed very low expression at the E18/P7 stages and still low in adult brain. However, *DISC1* was modestly expressed in hESCs and the expression decreased after neural differentiation, suggesting that *DISC1* might play an important role in stem cell functions.

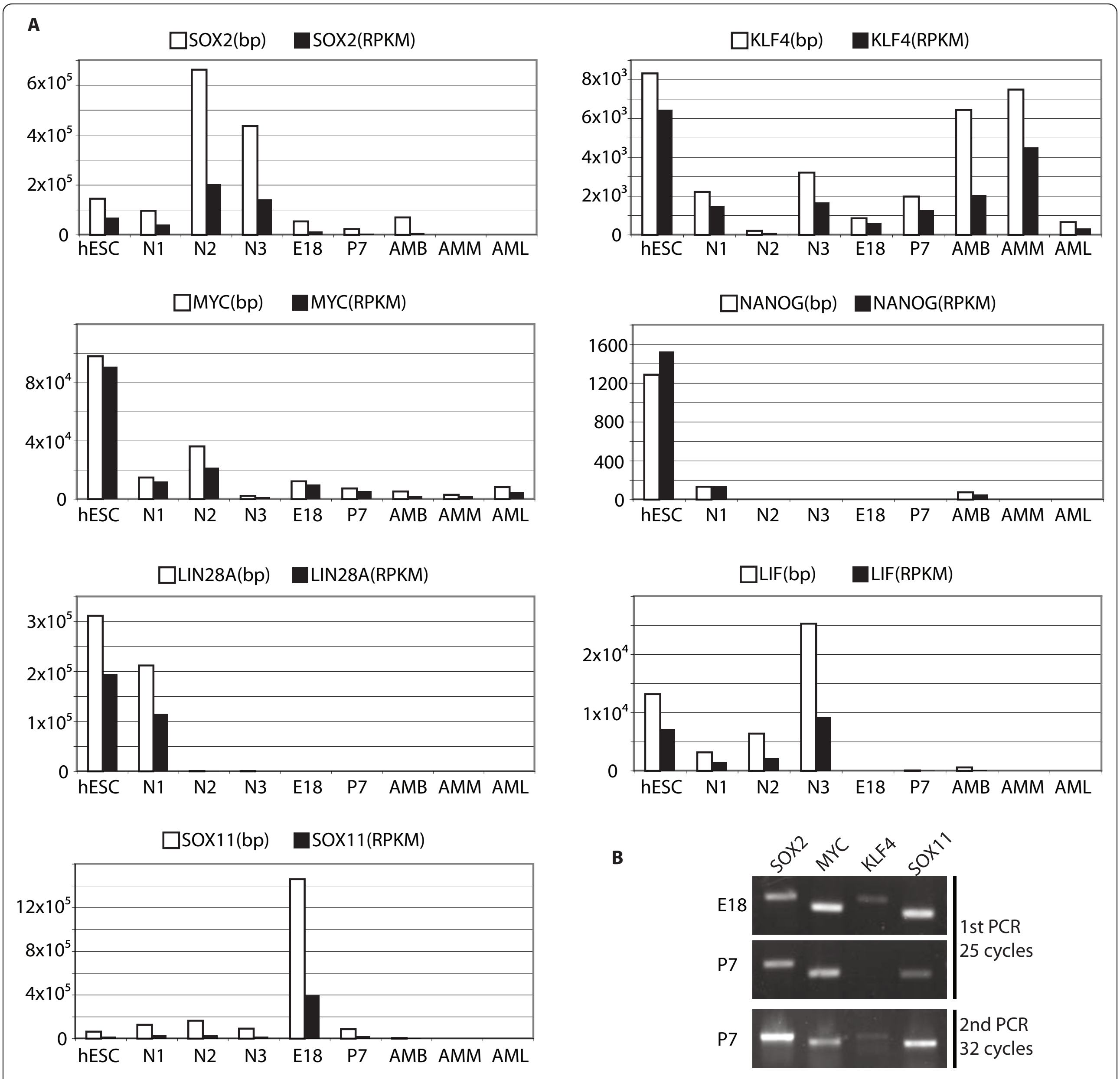


Figure 6 RNA-seq levels and RT-PCR analyses of pluripotency-related genes. A. Expression level of pluripotency-related genes across stages. Expression level was measured in both raw RNA-seq data size (bp) and RPKM methods. B. RT-PCR results showing the expression of *Sox2*, *Myc* and *Klf4*, with *Sox11* as the positive control.

Detection of unusually high levels of unannotated transcript expression level at E18 and P7 stages

To obtain an accurate set of unannotated TARs, we first subtracted the TARs overlapping with annotated exons, tRNAs or rRNAs. We then removed TARs overlapping with repeats. We also excluded TARs within 2 kb vicinity of the first and last exon to avoid promoter, TSS and TTS associated transcription activity, which have been previously studied [38-40]. We call the set of unannotated TARs generated from the aforementioned procedure the filtered unannotated TARs, and they include both intronic TARs and intergenic TARs. We found that the E18 stage had the highest percentages of both intronic and intergenic reads, at 5% and 3% of the E18 total data, respectively (Figure 8A). P7 stage had the second highest percentages, with about 1.1% and 2.6% for introns and intergenic regions, respectively. In comparison, almost all the other datasets had less than 1% of the data matching intronic or intergenic regions. Among human samples, surprisingly, the stem cell stage had the lowest percentage

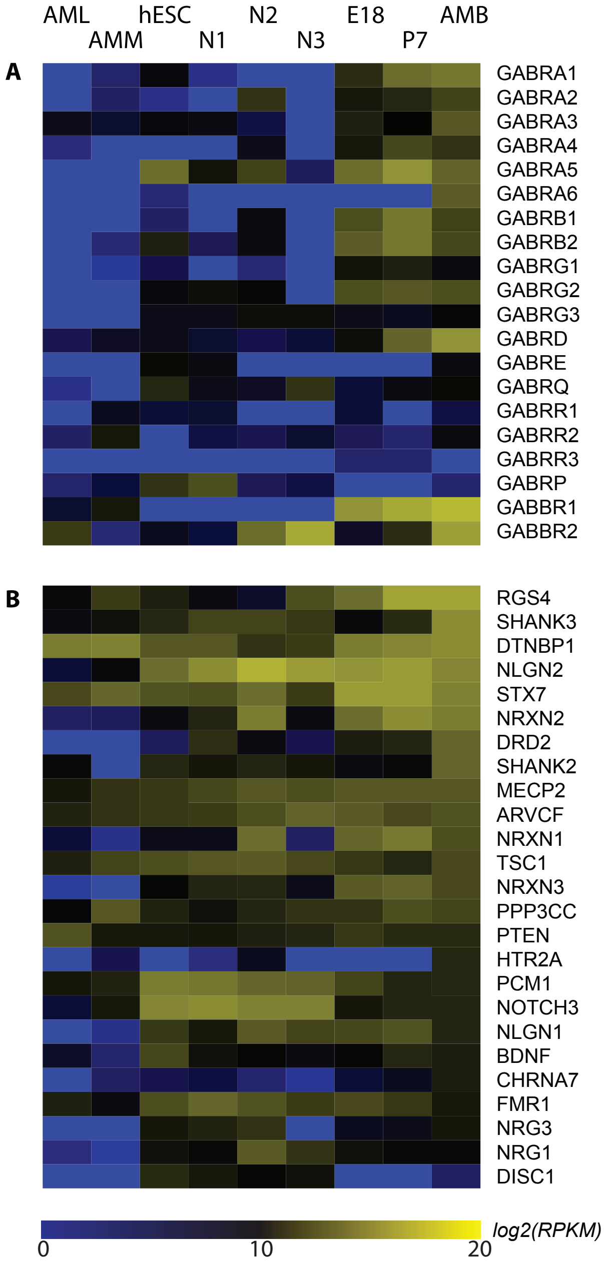


Figure 7 Expression of genes associated with neurodevelopmental disorders. A. Expression of selected GABAA receptor genes. B. Expression of selected Autism spectrum disorders associated genes.

of unannotated TARs, while the neural progenitor cell N2 stage had the highest. This result is different from a previous report [7], which did not use methods that rigorously removed repeats as in this study.

Concordant changes in expression levels between intronic TARs and flanking exons

To investigate the relationship between an intronic TAR and its flanking exons, we examined their respective expression levels (as measured in RPKM; [18]). There were 488 genes with intronic TARs at either the E18 or P7 stage. We found a strong positive correlation between the intronic expression and the flanking exonic expression. From E18 to P7, we found 436 genes with decreased intronic expression levels; among these 436 genes, 242 (56%) also had decreased exonic expression (Table 2). Even more strikingly, 52 genes had increased intronic expression level from E18 to P7 (Table 2), with 43 (83%) having increased exonic expression levels. The correlation was highly significant, with a p-value of 0.0001 from Fisher's Exact Test for the association of the exonic and intronic expression levels.

Few unannotated TARs were connected with known exons

The strong concordant correlation between the previously unannotated intronic TARs and flanking exons suggested that the intronic TAR and its flanking exon might be parts of the same RNA transcript. To test this hypothesis, we focused on the E18 and P7 datasets, which had the largest percentage of filtered unannotated TARs. A paired-end read with one end located in the unannotated TAR and the other in a known exon would be strong evidence that this intronic TAR and the known exon are parts of the same mRNA. However, it is in principle possible that the mapping positions could be erroneous. In addition, the existing mathematical and statistical models for determining the connection between TARs [8] are designed for RNA-seq data from cDNAs generated with random primers. They are not applicable to poly-dT primed data, which have a 3' bias. So we first devised a model suitable for both priming techniques (Methods; formula (2)), which reports the presence of the physical connection between expressed TARs and known exons. Using known adjacent exons and single exon genes (SGEs) with detected reads as positive and negative controls in a simulation, the formula had success rates of 93% and 100%, respectively.

Using formula 2, we found that only a very small percentage of the unannotated intronic TARs were connected with known exons (Table 3) as supported by the RNA-seq reads. Although a large fraction (70% for E18, 60% for P7) of the unannotated TARs was found to be connected to other regions mapped elsewhere, these mapped regions were very short, fragmented and unannotated regions with low read coverage. Surprisingly, some unannotated TARs were connected to read ends that mapped to multiple chromosomes, possibly due to mapping to repetitive sequences, erroneous mapping or possible cross-chromosome splicing (*trans*-splicing), a rare phenomenon that was previously observed [41].

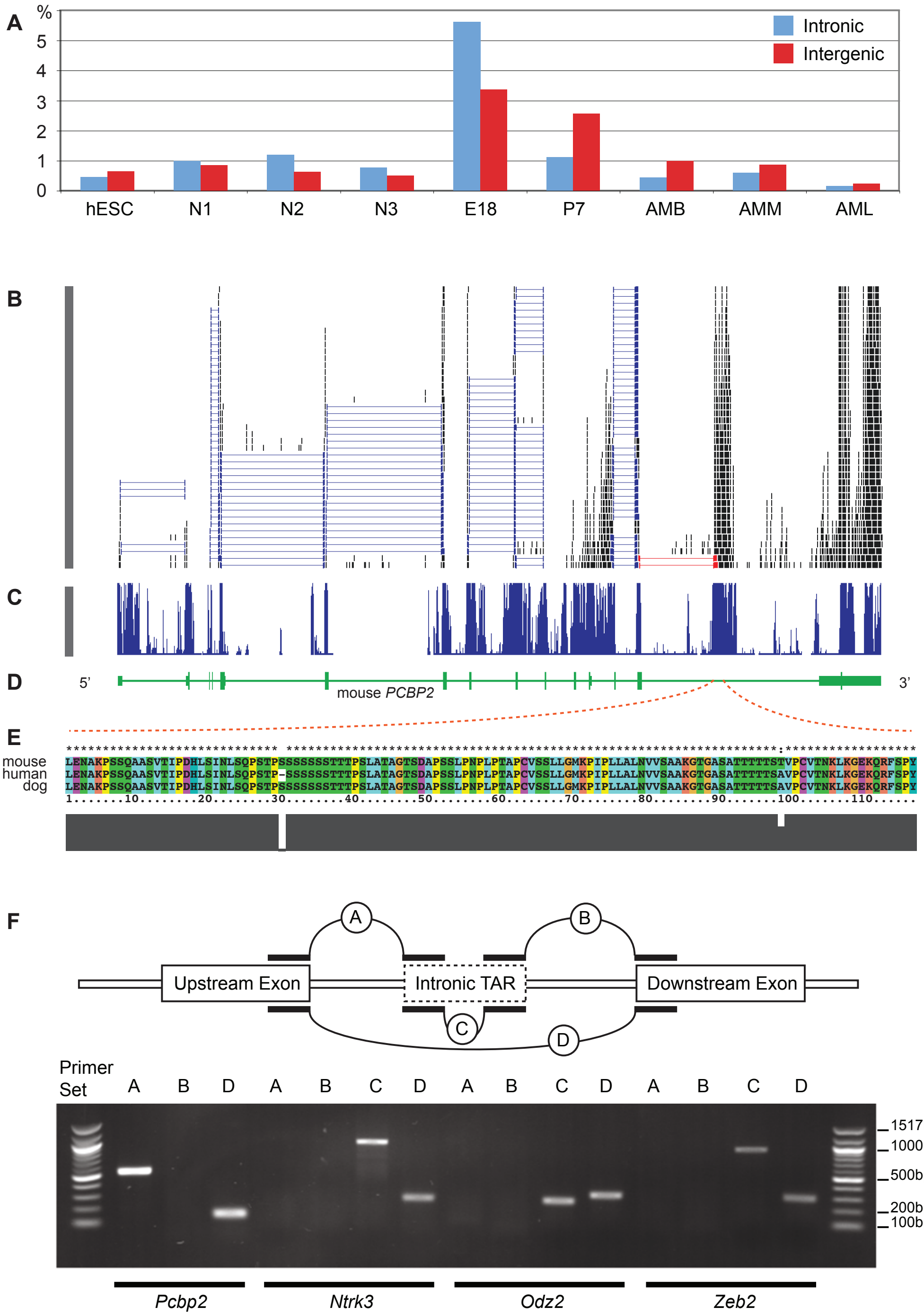


Figure 8 Detection of unannotated transcriptionally active regions (TARs). A. Percentage of TAR data of each dataset. Column in red color represents intronic TAR, and column in blue color represents intergenic TAR. X-axis represents different datasets, while Y-axis represents the percentage. B to E: An intronic TAR found located inside mouse *PCBP2*. B: RNA-seq reads mapped by TopHat to this location. Each black bar represents a RNA-seq read. Because the coverage on the 3' end (right side) of the gene is too high, only part of the reads mapped at this region is shown. Each blue horizontal line connecting two bars indicates a splice-junction-spanning read (read split by splicing activity). Red lines highlighted such reads spanning the intronic TAR and the exon on its left. C: conservation score from 30-Way multi-species genome alignment. Notice the region with high conservation score co-localized with the intronic TAR. D: current *PCBP2* gene model, with green box representing known exons. E: amino acid alignment of the ORF of intronic TAR and its human and dog homologues. Gray column plot below: each column represents the similarity of the amino acids in that column. F: RT-PCR validation for intronic TARs. For each intronic TAR inferred from RNA-seq mapping result, primers were picked to validate the connection between the intronic TAR and the flanking exon (upstream/downstream), between the flanking exons, as well as the expression of the intronic TAR.

Table 2 Intronic and exonic expression level changes for genes with intronic TARs from E18 to P7

	Also With Increased Intronic Expression	Also With Decreased Intronic Expression
Number of Genes With Increased Exonic Expression	43	194
Number of Genes With Decreased Exonic Expression	9	242

Comparing the intronic TARs with known mRNA and EST in NCBI databases

To test whether there is other evidence for the intronic TARs, we searched the data in NCBI's cDNA/mRNA and EST databases. Among 554 intronic TARs detected at E18 stage, 176 (32%) had no matches in NCBI databases. Similarly, among 168 intronic TARs at P7 stage, 49 (29%) had no matches in NCBI databases. Therefore, our results provide the first evidence for these TARs being expressed. Among the matching NCBI database entries, 11 (2%) of the 378 for the E18 stage and 7 (4%) of the 119 for the P7 stage were from the same stages, but none of them was from the brain cortex.

We then examined the splicing pattern of the mRNA and EST records matched to our detected intronic TARs and found two classes of intronic TARs: (1) with records suggesting that the TARs were standalone, without connection to known exons; (2) with some records suggesting that the TARs were standalone while other records suggesting that they were connected to known exons. 304 out of 378 (80%) intronic TARs at E18 and 75 out of 119 (63%) intronic TARs at P7 belonged to the first class. For the second class of intronic TARs, on average, the ratios for records supporting standalone transcripts to those for connections to known exons were 4.2 and 2.8 for the E18 and P7 stages, respectively. Taking together, the comparison with NCBI's cDNA/mRNA and EST databases strongly suggested that most of our detected intronic TARs were not connected with known exons and thus were novel transcripts.

Comparing the intronic TARs with known records in miRbase and lncRNAdb

We then compared our intronic TARs in miRNA database miRbase [42] and long non-coding RNA database lncRNAdb [43]. Although we found no significant hits in these two databases for any intronic TARs observed at P7 stage, we did find 12 and 6 hits for intronic TARs at E18 stage in miRbase and lncRNAdb, respectively (Additional file 2, Table S5 and S6). However, all 6 intronic TARs with hits in lncRNAdb were mapped to the same lncRNA, B2 SINE RNA, which was from a SINE repeat element. In addition, 11 of the 12 intronic TARs having hits in miRbase mapped to the same miR-1935 miRNA, and the remaining one mapped to miR-153-2. Otherwise, we did not detect significant hits for other types of RNAs.

Sequence conservation and coding potential of intronic TARs

To obtain clues about possible function of the intronic TARs using sequence similarity to other mammalian genes, we investigated whether unannotated TARs corresponded to any highly conserved region using the 30-Way Multiz Alignment & Conservation track in UCSC Genome Browser [44]. We found that there were 554 and 168 unannotated TARs at E18 and P7 stages, respectively; among these, 67 in E18 and 21 in P7 matched regions of highly conserved sequences. For example, a TAR on chromosome 15 (102324092-102324772) was localized to an intron of the mouse *PCBP2* gene encoding the major cellular poly(rC)-binding protein [45]. In addition, there were RNA-seq reads spanning this intronic TAR and the upstream exon (Figure 8B, red reads), indicating that this previously unannoted TAR was spliced with a known exon. Moreover, this TAR had a significant overlap with a highly conserved region located in the 3' most intron, which was identified by mammalian conservation study using 30-Way Multiz Alignment & Conservation track data (Figure 8C and 8D) [44]. An Open Reading Frame (ORF) was also predicted inside this TAR and was conserved among the *PCBP2* genes of human and dog (99% similar in amino acid sequences; Figure 7E), but not opossum. PCR and ABI 3730 resequencing results further verified that this TAR is indeed part of an mRNA (Figure 8F)

Table 3 Physical connection between unannotated TARs and known exons

		E18 Brain Cortex		P7 Brain Cortex	
		Intronic TAR	Intergenic TAR	Intronic TAR	Intergenic TAR
Connected with Known Exons		19	2	13	2
Not Connected with Known Exons	**Standalone**	117	45	28	41
	Non-Standalone (Same Chromosome)	404	71	123	46
	Other (Non-Standalone Multi-Chromosome)	14	8	4	5
	Total	**554**	**126**	**168**	**94**

with a connection between this TAR and the upstream exon, consistent with RNA-seq results. However, PCR product between this intronic TAR and the downstream exon was not detected, in agreement with the RNA-seq results. This TAR was very likely to represent an alternative 3' UTR with a potential coding region.

In addition, an intronic TAR with an ORF inside the *ATP2B1* gene located on chromosome 10 (98481907-98482067) shares 99.3% identity to the 20th exon of human *ATP2B1* isoform a (*ATP2B1a*) (Additional file 1, Fig. S4). Human *ATP2B1* has two splicing variants: *ATB2B1a* and *ATB2B1b*, which differ in the usage of the 20th exon. Previous studies showed that *ATP2B1a* has a specific expression at synapses whereas *ATB2B1b* is expressed in most tissues [45,46]. Thus this TAR is likely to encode a neuron-specific exon of the mouse *ATP2B1* gene. We also found another expressed region on chromosome 7 (112781296-112781396) that shares 87.5% identity with a part of the second 3' UTR exon of the human *Trim3* gene (Additional file 1, Fig. S4). *Trim3* (or *BERP*) is expressed in the brain and encodes a RING finger protein that regulates *GABAR* cell surface expression [47]. Another intronic TAR located in the *NRXN1* gene on chromosome 17 (90854147-90854636) has the potential for coding Neurexin 1, a neuronal cell adhesion molecule interacting with neuroligins to promote synapse formation and maturation [48]. The ORFs in these intronic TARs were highly similar to parts of human *ATP2B1*, *BERP* and *NRXN1* genes, respectively. A number of other intronic TARs, such as those in *CHD3*, *TSC22*, and *SRCAP*, were either similar to known human exons or supported by mouse gene predictions and mRNA and/or EST data in the NCBI database.

Three other intronic TARs were located, respectively, in the *Zeb2* gene on chromosome 2 (44953049-44955802), the *Ntrk3* gene on chromosome 7 (85484006-85485464), and the *Odz2* gene on chromosome 11 (36491704-36492013), within introns that are more than 10 kb long. These TARs did not match mRNA or EST records in the NCBI database, nor were they similar to protein sequences in the NCBI database. Nevertheless, these three TARs were conserved in rat, human, dog and opossum genomes, matching annotated introns in the orthologous genes in human and rat (Additional file 1, Fig. S5). Our RNA-seq data did not detect physical connection between the TARs and known exons; the lack of connection between the TARs and the flanking exons were further supported by the observations that PCR was successful when both primers were located inside a particular intronic TAR, but not able to generate products when a primer in the intronic TAR region was combined with another primer in one of the flanking exons (Figure 8F). As a control, the correct PCR product was obtained using primers matching the two flanking exons of the given intronic TAR (Figure 8F). Therefore, these three intronic TARs were most likely standalone transcripts that were not linked with the flanking exons.

Genes with intronic TARs were over-represented in GO terms closely associated with neural development

Although we have found that few of these intronic TARs were physically connected with the exons of the corresponding genes in the same mRNA, the fact the same genomic regions encode both the transcripts with exons and the intronic TARs suggests some association between these intronic TARs and the exonic genes. To further examine the functional implication of the intronic TARs, it is informative to study the corresponding genes. To study the nature of proteins encoded by the genes with intronic TARs, we analyzed their enrichment in Biological Process Gene Ontology (GO) using agriGO [49]. The GO annotations for all expressed genes at either stage were used as a reference for comparison to determine possible enrichment of specific GO categories. A total of 316 unique genes contained filtered intronic TARs among a total of 10657 genes with detected reads at E18, while only 119 genes contained filtered intronic TARs among 10901 genes expressed at P7. E18 data had 59 statistically overrepresented GO terms, but P7 data only had 4 statistically overrepresented GO terms, 3 of which are shared between these stages. The 60 overrepresented GO terms could be mapped onto only 3 major branches: neural signaling (Figure 9), neurogenesis (Figure 10) and regulation (Figure 11). The GO terms of these branches are closely related with neural developmental events occurring at E18 stages.

For neural signaling related GO terms at E18, two subgroups form largely parallel interactions: the first subgroup mainly functions in regulating neural system process, while the second subgroup carries out signal transmission (Figure 9). The terminal node was the regulation of synaptic transmission, which combines the aforementioned two functions. Among the nodes with strong statistical support were "transmission of nerve impulse", "neurological system process", and "synaptic transmission". Another interesting aspect is that the regulatory relations in the first subgroup were positive, while they were negative in the second subgroup. For neural signaling related GO terms at P7, however, only one function group, similar to the aforementioned 2nd subgroup, was identified. It had 3 nodes and the regulatory relations between the nodes were negative (Figure 9).

Nearly half (28, all at E18) of the enriched GO terms were for neurogenesis, with extensive interconnections

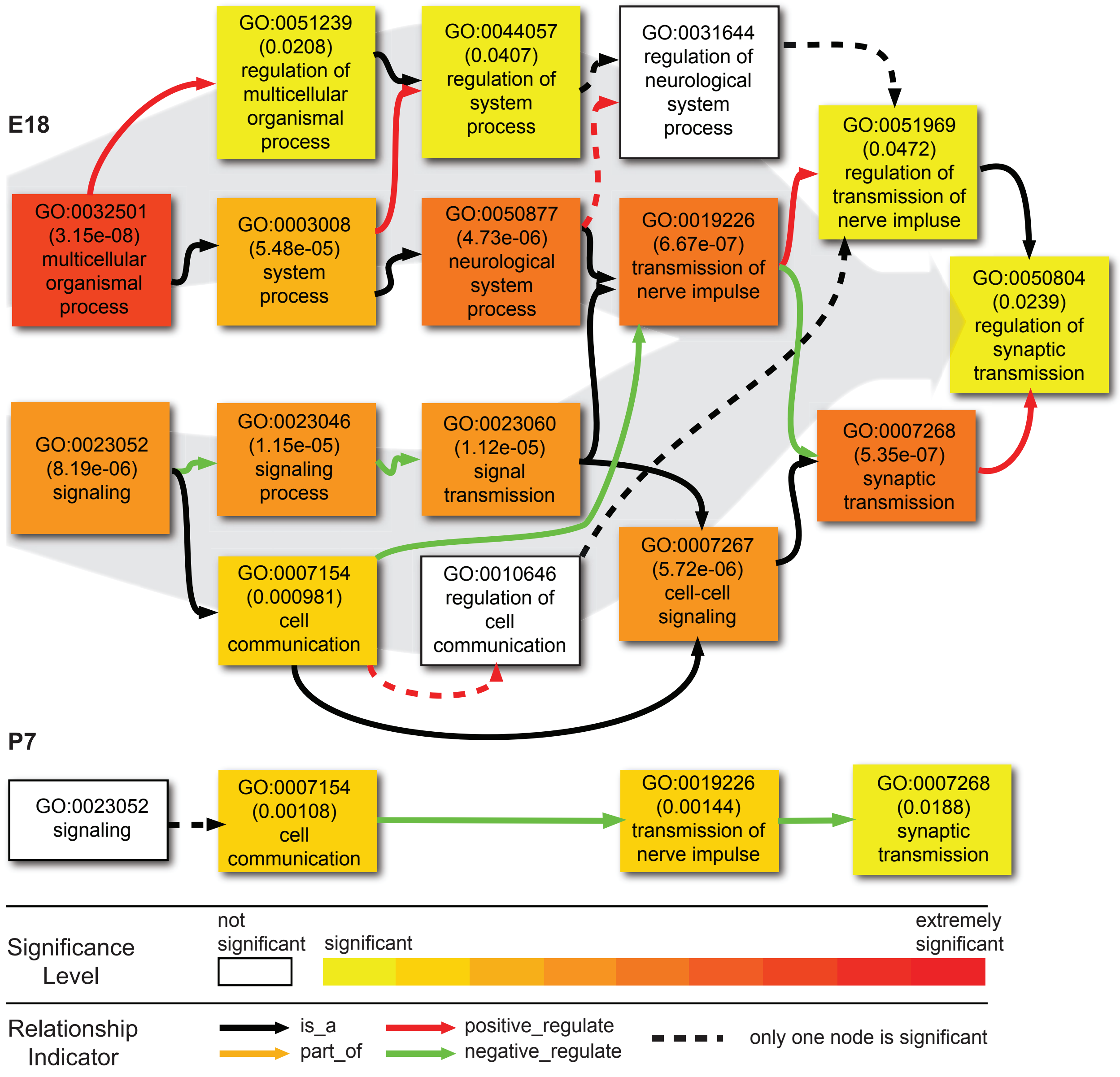

E18
GO:0051239 (0.0208) regulation of multicellular organismal process
GO:0044057 (0.0407) regulation of system process
GO:0031644 regulation of neurological system process
GO:0051969 (0.0472) regulation of transmission of nerve impluse
GO:0032501 (3.15e-08) multicellular organismal process
GO:0003008 (5.48e-05) system process
GO:0050877 (4.73e-06) neurological system process
GO:0019226 (6.67e-07) transmission of nerve impulse
GO:0050804 (0.0239) regulation of synaptic transmission
GO:0023052 (8.19e-06) signaling
GO:0023046 (1.15e-05) signaling process
GO:0023060 (1.12e-05) signal transmission
GO:0007268 (5.35e-07) synaptic transmission
GO:0007154 (0.000981) cell communication
GO:0010646 regulation of cell communication
GO:0007267 (5.72e-06) cell-cell signaling
P7
GO:0023052 signaling
GO:0007154 (0.00108) cell communication
GO:0019226 (0.00144) transmission of nerve impulse
GO:0007268 (0.0188) synaptic transmission
Significance Level
not significant
significant
extremely significant
Relationship Indicator
is_a
part_of
positive_regulate
negative_regulate
only one node is significant

Figure 9 Neural signaling GO terms significantly overrepresented by genes with intronic TARs. The coloring of the node box indicates the statistical support for the overrepresentation. The color scale at the bottom of the figure illustrates the different levels of statistical support. The node box with white color means that that node was not statistically overrepresented. However these are included to illustrate the relationship between statistically significant nodes.

between nodes and no obvious functional subgroups (Figure 10). The functions of these nodes include many aspects of neural development, such as cell morphogenesis, neurogenesis and neuron differentiation, eventually diverging into two termini: (1) axongenesis and axon guidance and (2) dendrite morphogenesis. It is also striking that all regulatory relationships were negative.

As shown in Figure 10, the third major group of overrepresented GO terms at E18 was for regulation. They were mainly about negative regulations, consistent with the negative regulatory relations identified between the majority of the nodes for neural signaling and neurogenesis groups. Specifically, strong statistical support was found for negative regulation of “metabolic process”, of “gene expression” and of “biosynthetic process”, ending in that of RNA polymerase II-dependent transcription. For P7, only one node received strong statistical support: regulation of molecular function (Figure 11).

Because axon guidance is critical for correct formation of the neural circuit during neural development, we

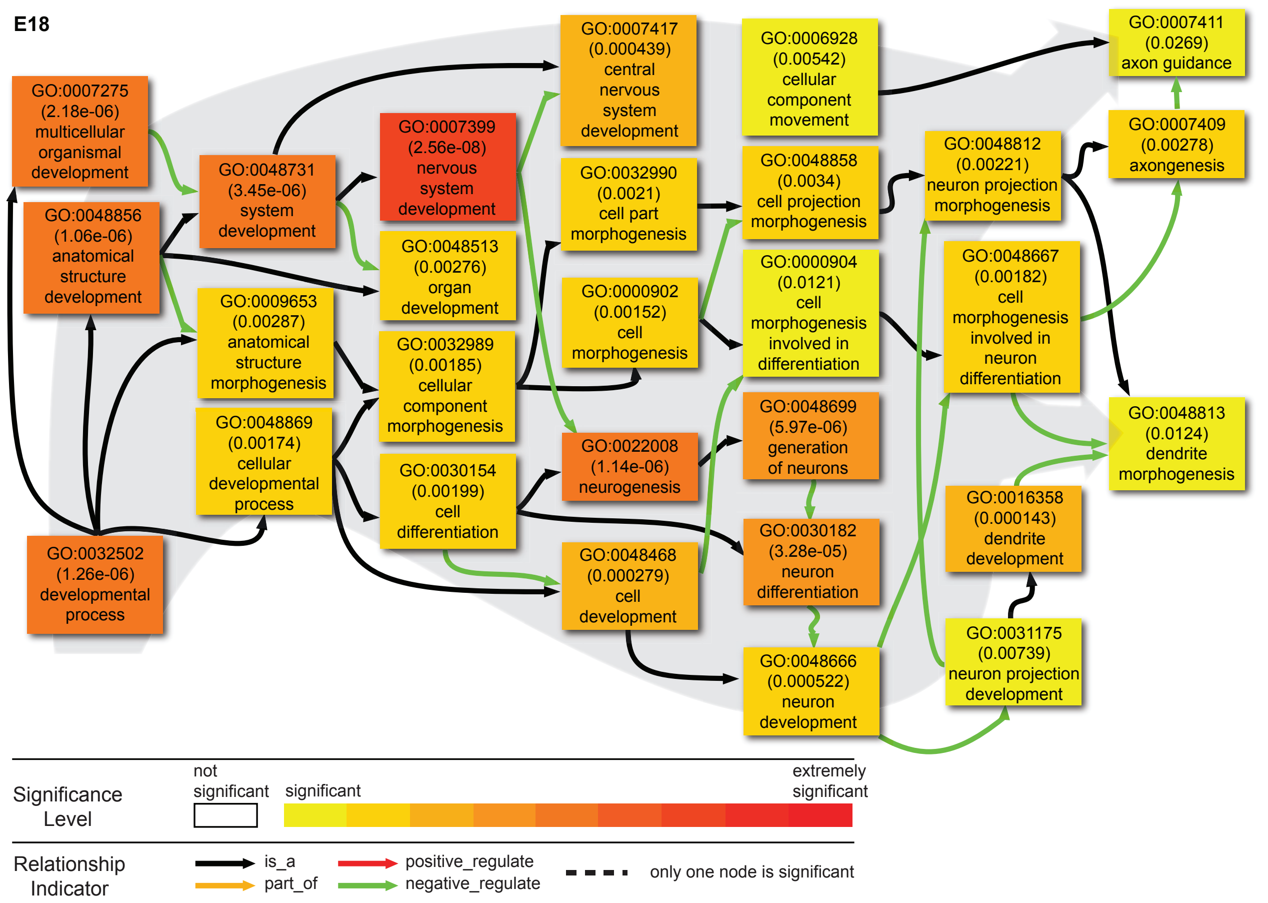

E18
GO:0007275 (2.18e-06) multicellular organismal development
GO:0048856 (1.06e-06) anatomical structure development
GO:0032502 (1.26e-06) developmental process
GO:0048731 (3.45e-06) system development
GO:0009653 (0.00287) anatomical structure morphogenesis
GO:0048869 (0.00174) cellular developmental process
GO:0007399 (2.56e-08) nervous system development
GO:0048513 (0.00276) organ development
GO:0032989 (0.00185) cellular component morphogenesis
GO:0030154 (0.00199) cell differentiation
GO:0007417 (0.000439) central nervous system development
GO:0032990 (0.0021) cell part morphogenesis
GO:0000902 (0.00152) cell morphogenesis
GO:0022008 (1.14e-06) neurogenesis
GO:0048468 (0.000279) cell development
GO:0006928 (0.00542) cellular component movement
GO:0048858 (0.0034) cell projection morphogenesis
GO:0000904 (0.0121) cell morphogenesis involved in differentiation
GO:0048699 (5.97e-06) generation of neurons
GO:0030182 (3.28e-05) neuron differentiation
GO:0048666 (0.000522) neuron development
GO:0048812 (0.00221) neuron projection morphogenesis
GO:0048667 (0.00182) cell morphogenesis involved in neuron differentiation
GO:0016358 (0.000143) dendrite development
GO:0031175 (0.00739) neuron projection development
GO:0007411 (0.0269) axon guidance
GO:0007409 (0.00278) axongenesis
GO:0048813 (0.0124) dendrite morphogenesis
Significance Level
not significant
significant
extremely significant
Relationship Indicator
is_a
part_of
positive_regulate
negative_regulate
only one node is significant

Figure 10 Neurogenesis GO terms significantly overrepresented by genes with intronic TARs. The coloring of the node box indicates the statistical support for the overrepresentation. The color scale at the bottom of the figure illustrates the different levels of statistical support. The node box with white color means that that node was not statistically overrepresented. However these are included to illustrate the relationship between statistically significant nodes.

then further analyzed the axon guidance node at E18. A total of 10 genes with filtered intronic TARs were assigned to this node by GO. According to the KEGG Pathway database [50], among the 10 genes, *Robo1*, *Robo2*, *Nrp1*, *Dcc*, *Ephb1* and *Ephb2* were all receptors involved in axon guidance pathway. They were all involved in the regulation of the cytoskeleton dynamics and axon repulsion activity. The average ratio of intronic read number to exonic read number for these 10 genes was 3.9 at E18, but was 1.8 at P7 (Table 4). As a comparison, we also examined three well-known genes: *Myc*, β-actin (*Actb*) and tubulin (*Tuba1a*), and found that their expression levels (exonic reads) were very similar between E18 and P7. However, the average ratios of intronic read number to exonic read number for these three genes were 0.023 and 0.051 at E18 and P7, respectively. These ratios were less than 1/100 (E18) or 1/30 (P7) of those for the 10 axon guidance associated genes. From a different perspective, for the 10 axon guidance associated genes, the average exonic reads ratio (E18/P7) was 1.61, suggesting a slight reduction in expression. In contrast, while the average intronic reads ratio (E18/P7) was 3.67, representing a much bigger reduction of the intronic transcripts. As a reference, the 3 house keeping genes had an average ratio of 1.09 for E18/P7 exonic reads, whereas the intronic read numbers for the 3 genes were too low to compare accurately. Therefore, the 10 axon guidance-related genes had significantly more intronic reads than exonic reads (P-value = 0.001, Chi-square with Yates correction) and more so at E18 than P7 (P-value = 0.001), suggesting a possible role of these intronic transcripts in modulating axon guidance at E18 cortex. **Conclusions**

Neural and stem cell transcriptome

In this study, we have investigated global characteristics of embryonic and neonatal neural transcriptomes, and compared with transcriptomes of the adult brain and

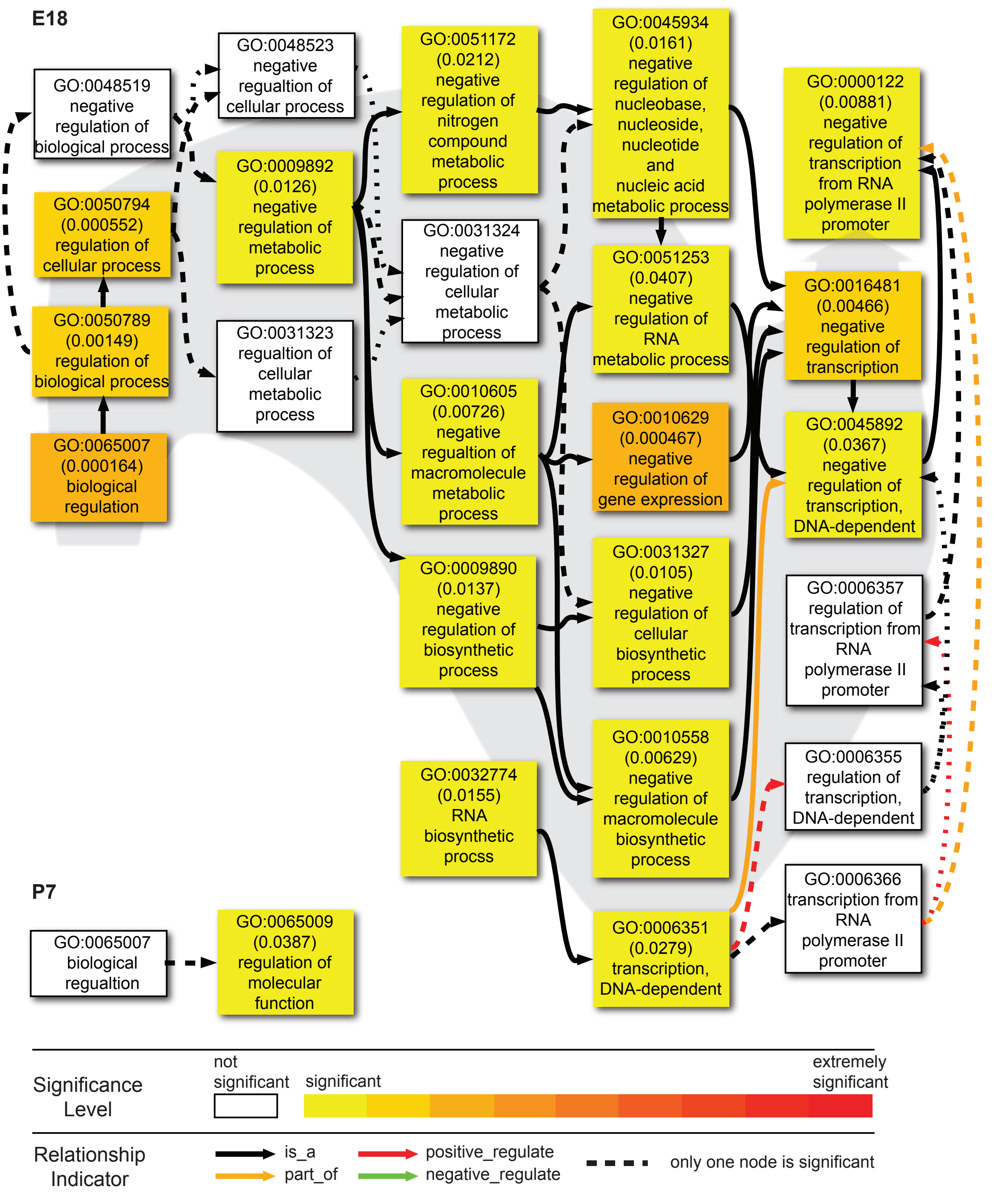

E18
GO:0048519 negative regulation of biological process
GO:0048523 negative regualtion of cellular process
GO:0051172 (0.0212) negative regulation of nitrogen compound metabolic process
GO:0045934 (0.0161) negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
GO:0000122 (0.00881) negative regulation of transcription from RNA polymerase II promoter
GO:0050794 (0.000552) regulation of cellular process
GO:0009892 (0.0126) negative regulation of metabolic process
GO:0031324 negative regulation of cellular metabolic process
GO:0051253 (0.0407) negative regulation of RNA metabolic process
GO:0016481 (0.00466) negative regulation of transcription
GO:0050789 (0.00149) regulation of biological process
GO:0031323 regualtion of cellular metabolic process
GO:0010605 (0.00726) negative regualtion of macromolecule metabolic process
GO:0010629 (0.000467) negative regulation of gene expression
GO:0045892 (0.0367) negative regulation of transcription, DNA-dependent
GO:0065007 (0.000164) biological regulation
GO:0009890 (0.0137) negative regulation of biosynthetic process
GO:0031327 (0.0105) negative regulation of cellular biosynthetic process
GO:0006357 regulation of transcription from RNA polymerase II promoter
GO:0032774 (0.0155) RNA biosynthetic procss
GO:0010558 (0.00629) negative regulation of macromolecule biosynthetic process
GO:0006355 regulation of transcription, DNA-dependent
GO:0006351 (0.0279) transcription, DNA-dependent
GO:0006366 transcription from RNA polymerase II promoter
P7
GO:0065007 biological regualtion
GO:0065009 (0.0387) regulation of molecular function
Significance Level
not significant
significant
extremely significant
Relationship Indicator
is_a
part_of
positive_regulate
negative_regulate
only one node is significant

Figure 11 Regulation GO terms significantly overrepresented by genes with intronic TARs. The coloring of the node box indicates the statistical support for the overrepresentation. The color scale at the bottom of the figure illustrates the different levels of statistical support. The node box with white color means that that node was not statistically overrepresented. However these are included to illustrate the relationship between statistically significant nodes.

Table 4 Exonic and intronic RNA-seq read number comparison between E18 and P7 for exon guidance related genes.

Gene Name	Number of detected reads in E18		Number of detected reads in P7	
	From Exon	From Intron	From Exon	From Intron
Slit3	43	1012	96	827
Robo2	2123	4900	1415	796
Robo1	1569	1872	425	362
Nrp1	991	1688	533	416
Nrcam	775	1150	1434	730
Klf7	1032	826	1088	328
Gli3	295	133	83	51
Ephb2	697	845	548	456
Ephb1	2006	1773	834	435
Dcc	1294	6509	281	1246
Read Ratio (Intron/Exon)	**3.860**		**1.802**	
Intronic Read Ratio (E18/P7)	**3.67**			
Exonic Read Ratio (E18/P7)	**1.61**			
Myc	358	13	229	8
Actb	8847	144	8236	503
Tuba1a	4029	63	3724	214
Read Ratio (Intron/Exon)	**0.023**		**0.051**	
Intronic Read Ratio (E18/P7)	**0.30**			
Exonic Read Ratio (E18/P7)	**1.09**			

embryonic stem cells. We found that embryonic and neonatal brain cortex transcriptomes correspond to most genomic regions at large scale of megabase intervals, but are unevenly distributed with positive correlation to exon density. In addition, neural transcriptomes are similar to that of embryonic stem cells, more than those of liver and muscle, in several features including chromosome level expression (Additional file 1, Fig. S3A and B), and expression pattern of orthologous genes (Figure 5). Also, the E18 brain cortex transcriptome and hESC transcriptome showed relatively even chromosomal distribution and had lower mitochondrial expression.

Other than these global similarities, we noted another shared characteristic between neural expressed genes and genes important for pluripotent stem cells. Specifically, three genes, *Sox2*, *Myc* and *Klf4* were detectable in all six neural samples (Figure 6), with high levels in E18 or P7 transcriptomes. The expression of these genes suggests that neural cells might need fewer factors to be converted to stem cells. Indeed Kim *et al* found that only two factors (*Oct4* with either *Klf4* or c-*Myc*), instead of four, were needed to revert neural stem cell to iPS cells [51]. Therefore, the similarity in transcriptome, including the expression of specific genes, such as *Sox2*, *Myc* and *Klf4*, between neural cells and stem cells suggests that neural cells might retain certain stem cell properties and have greater potential to be reprogrammed to be pluripotent.

Intronic TARs as standalone RNA regulators in early brain development

The mapping results of our transcriptome datasets revealed significant levels of intronic reads. We found that only a small portion of the intronic transcripts that we detected was on the same RNAs with any known exons. Recently, Klevebring *et al.* have reported that about 50% of the intronic expression was actually from the antisense strand [11], different from the sense exon-containing mRNAs of the same gene. Thus the intronic TARs detected here share some characteristic with the antisense transcripts, although our data lacked the strand information. Our finding that the level of intronic transcript is positively correlated with that of flanking exons is consistent with previous studies that antisense transcription may have both concordant and discordant regulation relative to the adjacent exons [3]. Furthermore, Faghihi *et al.* also reported regulation involving antisense RNA not mediated by the conventional RNA interference pathway [52], indicating additional mechanisms are important. Our results that the E18 brain cortex has significantly higher levels of intronic transcripts than other tissues/organs strongly suggest that such non-coding transcripts play important roles in regulating gene expression during embryonic brain development.

We also found that the mouse E18 embryonic brain had a concordant relation between intronic transcript and flanking exonic expression. This is unlike previous studies showing preferential localization of antisense transcripts in the upstream and downstream regions of the gene [53-55]. Our data have further indicated that the E18 embryonic brain showed enrichment of genes with intronic TARs in GO categories that are closely associated with neural functions. The E18 cortical neurons are actively engaged in neurogenesis, including axonogenesis and synaptogenesis. For the significant GO terms associated with neurogenesis, all the regulatory relations between nodes were negative (Figure 10). Moreover, an entire group of significant nodes was about negative regulation (Figure 11). However, at P7, intronic TARs were no longer associated with either neurogenesis or negative regulation. These findings suggest the involvement of intronic TARs in stage-specific regulation of neural developmental.

Recently, a subset of long ncRNAs was found to have an enhancer-like function [56]. Our data also indicated a correlation between the change in intronic transcript expression level and the change in the expression level of the corresponding gene. For example, the increased intronic expression is correlated with increased exonic expression for 10 axon guidance associated genes, whereas such correlation was not found for a control set of 3 housekeeping genes. The positive correlation in expression change between intronic TARs and the flanking exons further supports the idea that they have regulatory interactions, although it is formally possible that the intronic transcripts have functions unrelated to the genes represented by the flanking exons.

Our transcriptome analyses have revealed possible important mechanisms for gene function and its regulation in the developing brain, and uncovered a strong similarity to stem cells. These results provide a number of novel insights regarding neural developmental gene functions that can be further investigated using molecular genetic, biochemical and electrophysiological experiments.

Methods

RNA-seq mapping

RNA-seq data for hESC, N1, N2 and N3 were obtained from NCBI Sequence Read Archive SRP002079. RNA-seq data for adult mouse brain, liver and muscle tissues were obtained from NCBI Sequence Read Archive SRA001030. RNA-seq data for mouse embryonic day 18 and postnatal day 7 brain cortices were the same as described previously [9]. Its NCBI Sequence Read Archive accession number is SRP007262. The protocol for dissection of the mouse cortex was approved by IACUC committee of Pennsylvania State University and in accordance with the US Federal guidelines. All quality scores were then transformed into FASTQ ASCII code by original quality score plus 64. TopHat was selected for mapping these RNA-seq data. SRP002079 data were mapped onto human genome (UCSC hg19, NCBI Build 37), and the rest were mapped onto mouse genome (UCSC mm9, NCBI Build 37), both with the following parameters: –solexa-qual, -g 1. The same parameters were used when we mapped all nine datasets onto mouse reference genome. Although TopHat was instructed to report only unique hit (-g 1), it sometimes could not fully suppress multiple hits (personal communication with Cole Trapnell, TopHat author, on Feb 22nd 2011). Results were then further screened against RepeatMasker [57] database of the corresponding species to further eliminate possible ambiguous hits.

Normalizing against data size and chromosome size

Normalization was done according to previously published RPKM method [18] with the following adjustment: length normalization was done against chromosome size L_C when we were studying the chromosomal expression level distribution. Also, unlike the original RPKM concept, detected base pair size C_{bp} was used instead of read numbers to accommodate different read-lengths from different RNA-seq datasets (33 bp, 35 bp and 36 bp). For the same reason, total mapped base pairs N_{bp} was used in normalization against data size. And thus the RPKM* label was used here to distinguish these differences:

$$RPKM^* = \frac{C_{bp} \times 10^{12}}{N_{bp} \times L_C} \tag{1}$$

Data size normalization was done against mappable data size instead of original data size generated from sequencer. This was to accommodate systematic sequence quality and mapping percentage differences from different datasets.

Unannotated transcriptionally active region (TAR) calling

After the RNA-seq data were mapped to the target genome, regions with continuous read coverage that were within close proximity to each other were then chained together, thus forming the transcriptionally active regions (TARs). Only TARs longer than 100 bp and with more than 5X coverage were considered. These TARs were then compared with UCSC Known Gene [58]. TARs that did not overlap with UCSC Known Gene annotation features were then compared with known tRNA annotation [59] and custom-complied rRNA annotation. To further eliminate possible false-positives from repeat, TARs that were not included in any of the above annotations were then mapped back to genome with BLAST [60]. All regions with significant hits elsewhere in the genome were discarded. The remaining unannotated TARs were then filtered by their distance to their nearest exons. All unannotated TARs that were too close to known exons or genes were discarded as these may originated from previously reported small exon variations [61].

RT-PCR validation for intronic TARs and the connection between intronic TARs and flanking exons

Sample preparation and RNA extraction were done according to the procedures described previously [9]. For the RT-PCR experiment, total RNA was isolated from mice E18 and P7 cortical tissues by using Ambion RNAqueous-Midi Total RNA Isolation Kit (Catalog#1911). One microgram of RNA was reverse transcribed into cDNA by using Biolabs DyNAmo cDNA Synthesis Kit (Catalog# F-470L).

To validate the expression of several pluripotency-related genes: approximately 1/20 of the first strand cDNAs was used as a template for PCR with gene-specific primers. PCR was carried out for 25 cycles of 94°C for 20s, 54°C for 30s, and 74°C for 40s. 10 ul of PCR products was separated on 0.8% (w/v) agarose gels containing ethidium bromide and visualized by UV light. A secondary PCR was performed for P7 with same primers by using 1 ul first round PCR products, for 32 cycles.

To validate the expression of specific intronic TARs, primer sequences were chosen within the intronic TAR, between the intronic TAR and the upstream exon, between the intronic TAR and the downstream exon, and between the upstream exon and the down stream exon. RT-PCR was carried out using the cDNAs as template with Taq polymerase for 22 cycles (add the temperature info). PCR product was sequenced at the Genomics Core Facility at Penn State using an ABI 3730 machine.

Models to determine the physical connection of unannotated TARs with known exons/transcript for both poly-dT and random primed RNA-seq data

The general question can be abstracted to how to determine whether a given detected transciptionally active region (TAR) was on the same RNA with other exons/transcripts using paired-end information, i.e., there was a physical connection between the given TAR and another exon. If a given unannotated TAR is long and has many internal RNA-seq reads, its number of paired-end reads with one end located at a known exon should mean differently if a given unannotated TAR is short and with comparably less internal RNA-seq reads. We thus propose that the support for the aforementioned physical connection between the unannotated TAR and a known exon should be evaluated as a function of the length of the given unannotated TAR, the coverage (number of internal RNA-seq reads) of the unannoated TAR and the number of paried-end RNA-seq reads linking this unannotated TAR and another known exon. RNA-seq mappers also tend to have a lower mapping capability if they need to map a partial read at the end of an exon or TAR.

We first assumed that the paired-end read distribution inside a given TAR were either in a uniform distribution (in the case of using random primer), or in a skewed distribution (in the case of using ploy-T primer). Under this assumption, if the given TAR was part of a larger transcript, the number of paired-end reads at its end(s) should be similar to the average number of paired-end reads over the entire TAR. The following formula was used to calculate the estimated number of paired-end reads (Ne) at only one end of the given TAR, on the assumption that this TAR was part of a larger transcript:

$$Ne = \frac{Ni \times Lc}{2 \times Lr} - \frac{T \times Ni \times (Ls - M)}{Lr} \qquad (2)$$

Ni, two times the total number of paired-end reads with both ends located inside the given region (to reflect each end in a pair).

Ls, read length for one end of a paired-end read.

Lc, clone length of a paired-end read, which is 2×Ls plus the insert size.

Lr, length of the given TAR.

M, maximum number of allowed mismatches of the mapping algorithm.

T, a correction factor. Splice junction spanning reads will have two partial matches to two discrete genomic regions. This value represents the success rate of the algorithm in mapping partial reads to the end of a given region, normally between 0 and 1. 0 means the algorithm cannot map partial reads to the end of a given region. 1 means the algorithm can map 100% of the partial reads to the end of a given region. Given the fact TopHat is designed to do RNA-seq mapping, the T value we picked was 0.99.

If data indicated that there was a significant amount of links from the given TAR to both upstream and downstream exons, the Ne should be doubled since there should be reads covering both ends of the given TAR. If the size of the given region Lr was smaller than that of the clone length Lc, then by theory all paired-end reads from this TAR should be reads linking this given TAR with other region(s).

If the actual number of paired-end reads connecting a given TAR with other regions was significantly less than (in this case we used 20%) *Ne*, the given TAR is thought to be a standalone transcript. Otherwise, this region was inferred as non-standalone, which means some level of splicing activity. More specifically, if a significant portion of the aforementioned paired-end reads had the other ends located in annotated exon(s), this given TAR was thought to be part of a known transcript. However, if the aforementioned reads were connecting more than one chromosome, then this TAR was thought to be multi-chromosome linked.

Testing model effectiveness

To test the effectiveness of the proposed formula (formula (2)) in determining the connection between a given unannotated TAR and known exon(s), we must have positive controls that are known to be detectable in our dataset and are also known to be on the same RNAs with known exons. An expressed exon from a multi-exon gene would meet this requirement and should be able to serve as our positive control. To ensure that these exons were truly connected to known exons by RNA-seq reads, these exons were selected by hand through manual inspection of the RNA-seq mapping results using Integrative Genomics Viewer (IGV, http://www.broadinstitute.org/igv). For each selected exon, we made sure that there were multiple reads spanning the selected exon with at least one other known exon. To ensure a true representation of the genome wide situation from our test data, these selected exons were picked from different chromosomes, with different RNA-seq read coverage, different locations within a given gene and different relative distances to the 3' end (Additional file 2, Table S3). The proposed formula (formula (2)) was able to identify such exons as being physically connected with known exons with a success rate of 93%.

We also performed a negative control test to determine the effectiveness of the proposed formula in determining whether a given transcript has a physical connection with any known exon(s). Single exon genes (SEGs) in mouse genome were selected as the negative controls since they are known to be a standalone transcript. We first identified a list of SEGs which had RNA-seq reads in our dataset (Additional file 2, Table S4). The proposed formula was able to identify these selected SEGs as not being physically connected with any known exon(s) at a success rate of 100% (14% of the selected SEGs were determined, however, as multi-chromosome linked).

Gene Ontology (GO) analysis

Reference mouse GO annotation was obtained from the Jackson Laboratory's MGI site (http://www.informatics.jax.org). Expressed genes were inferred from RNA-seq mapping results mapped to UCSC Known Gene. Expressed genes were then compared with reference mouse GO annotation. Identifier conversion between the UCSC Known Gene and the GO annotation was done using in-house script. Among all GO terms, only Biological Process GO terms were analyzed. We first calculated the number of genes mapped to a given GO term. For a gene with multiple GO terms, all terms were considered because one gene may be involved in multiple biological processes. If one GO term node was counted, all its parental nodes were excluded. Four sets of GO annotation were produced using the aforementioned procedure: all expressed genes in E18, all expressed genes in P7, genes with intronic TAR(s) in E18 and genes with intronic TAR(s) in P7.

For a given stage, GO annotation for the entire transcriptome and GO annotation for only genes containing intronic TARs were compared using agriGO server [49]. The statistical significance was determined by Fisher's Exact Test, with Bonferroni Correction. The p-value threshold was preset at 0.05 and only GO terms with more than 5 hits were reported.

Accession Numbers

Human samples: [SRA: SRP002079], adult mouse samples: [SRA: SRA001030], brain cortices samples: [SRA: SRP007262].

Additional material

Additional file 1: this file includes figure S1 to S7.

Additional file 2: this file includes supplemental data and analysis results relevant to this study, table S1 to S6, and figure legends for figure S1 to S7.

List of abbreviations

TAR: transcriptionally active region; hESC: human embryonic stem cell; N1: early initiation stage of hESC; N2: neural progenitor derived from hESC; N3: early glial-like cell derived from hESC; E18: embryonic day 18 brain cortex; P7: postnatal day 7 brain cortex; AMB: adult mouse brain; AMM: adult mouse muscle; AML: adult mouse liver; RPKM: reads per kilobase of exon(s) per million mapped reads; GO: gene ontology;

Acknowledgements

We thank X. Han and X. Zhou for helpful discussion and comments on this manuscript. We specially thank Prof. Naomi Altman for her input on statistical analysis on comparison between chromosome expression profiles. This work was supported by the Biology Department, Eberly College of Sciences, and the Huck Institutes of the Life Sciences, the Pennsylvania State University. G.C. was supported by an NIH grant (MH083911). Y.S. was supported by the Intercollege Graduate Program in Genetics and the Biology Department, the Pennsylvania State University. H.M. was also supported by Fudan University.

Author details

[1]Department of Biology, Pennsylvania State University, University Park, PA 16802, USA. [2]State Key Laboratory of Genetic Engineering and Institute of Plant Biology, School of Life Sciences, Fudan University, Shanghai 200433, P.

R. China. [3]The Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802, USA. [4]Intercollege Graduate Program in Genetics, Pennsylvania State University, University Park, PA 16802, USA. [5]Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, P. R. China.

Authors' contributions

HM designed research; YS and YH performed research; YS, YW and GC analyzed data; YS, YW, GC and HM wrote the paper. All authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 13 April 2011 Accepted: 10 October 2011
Published: 10 October 2011

References

1. Thomas CA Jr: **The genetic organization of chromosomes.** *Annu Rev Genet* 1971, **5**:237-256.
2. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, *et al*: **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.** *Nature* 2002, **420**:563-573.
3. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, *et al*: **Antisense transcription in the mammalian transcriptome.** *Science* 2005, **309**:1564-1566.
4. Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, *et al*: **The regulated retrotransposon transcriptome of mammalian cells.** *Nat Genet* 2009, **41**:563-571.
5. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, *et al*: **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309**:1559-1563.
6. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The transcriptional landscape of the yeast genome defined by RNA sequencing.** *Science* 2008, **320**:1344-1349.
7. Wu JQ, Habegger L, Noisa P, Szekely A, Qiu C, Hutchison S, Raha D, Egholm M, Lin H, Weissman S, *et al*: **Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing.** *Proc Natl Acad Sci USA* 2010, **107**:5254-5259.
8. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol* 2010, **28**:511-515.
9. Han X, Wu X, Chung WY, Li T, Nekrutenko A, Altman NS, Chen G, Ma H: **Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA sequencing.** *Proc Natl Acad Sci USA* 2009, **106**:12741-12746.
10. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature* 2010, **464**:768-772.
11. Klevebring D, Bjursell M, Emanuelsson O, Lundeberg J: **In-depth transcriptome analysis reveals novel TARs and prevalent antisense transcription in human cell lines.** *PLoS One* 2010, **5**:e9762.
12. van Bakel H, Nislow C, Blencowe BJ, Hughes TR: **Most "dark matter" transcripts are associated with known genes.** *PLoS Biol* 2010, **8**:e1000371.
13. Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak S, Mis E, Zegar C, Gutwein MR, Khivansara V, *et al*: **The Landscape of *C. elegans* 3'UTRs.** *Science* 2010.
14. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, *et al*: **Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution.** *Science* 2005, **308**:1149-1154.
15. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, *et al*: **Global identification of human transcribed sequences with genome tiling arrays.** *Science* 2004, **306**:2242-2246.
16. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, *et al*: **Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.** *Nature* 2009, **458**:223-227.
17. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, *et al*: **Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression.** *Proc Natl Acad Sci USA* 2009, **106**:11667-11672.
18. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**:621-628.
19. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, *et al*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447**:799-816.
20. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25**:1105-1111.
21. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al*: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420**:520-562.
22. Nenadic O, Greenacre M: **Correspondence analysis in R, with two- and three-dimensional graphics: The ca package.** *Journal of Statistical Software* 2007, **20.**
23. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860-921.
24. Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J: **TM4 microarray software suite.** *Methods Enzymol* 2006, **411**:134-193.
25. Takahashi K, Yamanaka S: **Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.** *Cell* 2006, **126**:663-676.
26. Okita K, Ichisaka T, Yamanaka S: **Generation of germline-competent induced pluripotent stem cells.** *Nature* 2007, **448**:313-317.
27. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S: **Induction of pluripotent stem cells from adult human fibroblasts by defined factors.** *Cell* 2007, **131**:861-872.
28. Aoi T, Yae K, Nakagawa M, Ichisaka T, Okita K, Takahashi K, Chiba T, Yamanaka S: **Generation of pluripotent stem cells from adult mouse liver and stomach cells.** *Science* 2008, **321**:699-702.
29. Marchetto MC, Carromeu C, Acab A, Yu D, Yeo GW, Mu Y, Chen G, Gage FH, Muotri AR: **A model for neural development and treatment of Rett syndrome using human induced pluripotent stem cells.** *Cell* 2010, **143**:527-539.
30. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, Nie J, Jonsdottir GA, Ruotti V, Stewart R, *et al*: **Induced pluripotent stem cell lines derived from human somatic cells.** *Science* 2007, **318**:1917-1920.
31. Kuhlbrodt K, Herbarth B, Sock E, Enderich J, Hermans-Borgmeyer I, Wegner M: **Cooperative function of POU proteins and SOX proteins in glial cells.** *J Biol Chem* 1998, **273**:16050-16057.
32. Geschwind DH, Levitt P: **Autism spectrum disorders: developmental disconnection syndromes.** *Curr Opin Neurobiol* 2007, **17**:103-111.
33. Polleux F, Lauder JM: **Toward a developmental neurobiology of autism.** *Ment Retard Dev Disabil Res Rev* 2004, **10**:303-317.
34. Paysan J, Fritschy JM: **GABAA-receptor subtypes in developing brain. Actors or spectators?** *Perspect Dev Neurobiol* 1998, **5**:179-192.
35. Graf ER, Zhang X, Jin SX, Linhoff MW, Craig AM: **Neurexins induce differentiation of GABA and glutamate postsynaptic specializations via neuroligins.** *Cell* 2004, **119**:1013-1026.
36. Dong N, Qi J, Chen G: **Molecular reconstitution of functional GABAergic synapses with expression of neuroligin-2 and GABAA receptors.** *Mol Cell Neurosci* 2007, **35**:14-23.
37. Jaaro-Peled H, Hayashi-Takagi A, Seshadri S, Kamiya A, Brandon NJ, Sawa A: **Neurodevelopmental mechanisms of schizophrenia: understanding disturbed postnatal brain maturation through neuregulin-1-ErbB4 and DISC1.** *Trends Neurosci* 2009, **32**:485-495.
38. Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, Faulkner GJ, Lassmann T, Forrest AR, Grimmond SM, Schroder K, *et al*: **Tiny RNAs associated with transcription start sites in animals.** *Nat Genet* 2009, **41**:572-578.

39. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, *et al*: **RNA maps reveal new RNA classes and a possible function for pervasive transcription.** *Science* 2007, **316**:1484-1488.
40. Fejes-Toth K, Sotirova V, Sachidanandam R, Assaf G, Hannon GJ, Kapranov P, Foissac S, Willingham AT, Duttagupta R, Dumais E, *et al*: **Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs.** *Nature* 2009, **457**:1028-1032.
41. Li H, Wang J, Mor G, Sklar J: **A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells.** *Science* 2008, **321**:1357-1361.
42. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics.** *Nucleic Acids Res* 2008, **36**:D154-158.
43. Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS: **lncRNAdb: a reference database for long noncoding RNAs.** *Nucleic Acids Res* 2011, **39**: D146-151.
44. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, *et al*: **The UCSC Genome Browser database: update 2010.** *Nucleic Acids Res* 2010, **38**:D613-619.
45. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, *et al*: **GeneCards Version 3: the human gene integrator.** *Database (Oxford)* 2010, **2010**:baq020.
46. Kenyon KA, Bushong EA, Mauer AS, Strehler EE, Weinberg RJ, Burette AC: **Cellular and subcellular localization of the neuron-specific plasma membrane calcium ATPase PMCA1a in the rat brain.** *J Comp Neurol* 2010, **518**:spc1.
47. Cheung CC, Yang C, Berger T, Zaugg K, Reilly P, Elia AJ, Wakeham A, You-Ten A, Chang N, Li L, *et al*: **Identification of BERP (brain-expressed RING finger protein) as a p53 target gene that modulates seizure susceptibility through interacting with GABA(A) receptors.** *Proc Natl Acad Sci USA* 2010, **107**:11883-11888.
48. Missler M, Sudhof TC: **Neurexins: three genes and 1001 products.** *Trends Genet* 1998, **14**:20-26.
49. Du Z, Zhou X, Ling Y, Zhang Z, Su Z: **agriGO: a GO analysis toolkit for the agricultural community.** *Nucleic Acids Res* 2010, **38(Suppl)**:W64-70.
50. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs.** *Nucleic Acids Res* 2010, **38**:D355-360.
51. Kim JB, Zaehres H, Wu G, Gentile L, Ko K, Sebastiano V, Arauzo-Bravo MJ, Ruau D, Han DW, Zenke M, *et al*: **Pluripotent stem cells induced from adult neural stem cells by reprogramming with two factors.** *Nature* 2008, **454**:646-650.
52. Faghihi MA, Wahlestedt C: **RNA interference is not involved in natural antisense mediated regulation of gene expression in mammals.** *Genome Biol* 2006, **7**:R38.
53. Sun M, Hurst LD, Carmichael GG, Chen J: **Evidence for a preferential targeting of 3'-UTRs by cis-encoded natural antisense transcripts.** *Nucleic Acids Res* 2005, **33**:5533-5543.
54. Core LJ, Waterfall JJ, Lis JT: **Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters.** *Science* 2008, **322**:1845-1848.
55. Faghihi MA, Wahlestedt C: **Regulatory roles of natural antisense transcripts.** *Nat Rev Mol Cell Biol* 2009, **10**:637-643.
56. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q, *et al*: **Long noncoding RNAs with enhancer-like function in human cells.** *Cell* 2010, **143**:46-58.
57. Smit A, Hubley R, Green P: **RepeatMasker Open-3.0.** 2010 [http://www.repeatmasker.org].
58. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D: **The UCSC Known Genes.** *Bioinformatics* 2006, **22**:1036-1046.
59. Chan PP, Lowe TM: **GtRNAdb: a database of transfer RNA genes detected in genomic sequence.** *Nucleic Acids Res* 2009, **37**:D93-97.
60. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
61. Chern TM, van Nimwegen E, Kai C, Kawai J, Carninci P, Hayashizaki Y, Zavolan M: **A simple physical model predicts small exon length variations.** *PLoS Genet* 2006, **2**:e45.

doi:10.1186/1471-2164-12-494
Cite this article as: Sun *et al.*: Comparative analysis of neural transcriptomes and functional implication of unannotated intronic expression. *BMC Genomics* 2011 **12**:494.