Specific expression of long noncoding RNAs in the mouse brain

Tim R. Mercer*, Marcel E. Dinger*, Susan M. Sunkin†, Mark F. Mehler‡, and John S. Mattick*§

*Australian Research Council (ARC) Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, St. Lucia, QLD 4072, Australia; †Allen Institute for Brain Science, Seattle, WA 98103; and ‡Institute for Brain Disorders and Neural Regeneration, Departments of Neurology, Neuroscience and Psychiatry, and Behavioral Sciences, Einstein Cancer Center and Rose F. Kennedy Center for Research in Mental Retardation and Developmental Disabilities, Albert Einstein College of Medicine, Bronx, New York, NY 10461

Edited by Huda Y. Zoghbi, Baylor College of Medicine, Houston, TX, and approved November 21, 2007 (received for review July 17, 2007)

A major proportion of the mammalian transcriptome comprises long RNAs that have little or no protein-coding capacity (ncRNAs). Only a handful of such transcripts have been examined in detail, and it is unknown whether this class of transcript is generally functional or merely artifact. Using *in situ* hybridization data from the Allen Brain Atlas, we identified 849 ncRNAs (of 1,328 examined) that are expressed in the adult mouse brain and found that the majority were associated with specific neuroanatomical regions, cell types, or subcellular compartments. Examination of their genomic context revealed that the ncRNAs were expressed from diverse places including intergenic, intronic, and imprinted loci and that many overlap with, or are transcribed antisense to, protein-coding genes of neurological importance. Comparisons between the expression profiles of ncRNAs and their associated protein-coding genes revealed complex relationships that, in combination with the specific expression profiles exhibited at both regional and subcellular levels, are inconsistent with the notion that they are transcriptional noise or artifacts of chromatin remodeling. Our results show that the majority of ncRNAs are expressed in the brain and provide strong evidence that the majority of processed transcripts with no protein-coding capacity function intrinsically as RNAs.

genomics | neuroscience | transcriptomics | imprinting | subcellular

Although only 1.2% of the mammalian genome encodes proteins, it is now evident that most of the genome is transcribed to yield complex patterns of interlaced and overlapping transcripts that include tens of thousands of long (>200 nt) noncoding RNAs (ncRNAs) (1, 2). Although a small number of long ncRNAs have been functionally characterized (3), it remains a matter of debate whether the majority are biologically meaningful or merely transcriptional "noise" (4–7). The few long ncRNAs that have been characterized to date exhibit a diverse range of functions (3, 8) and expression in specific cell types and/or localization to specific subcellular compartments (9–12). The determination of whether many more long ncRNAs are functional may considerably impact our understanding of various fundamental biological processes and significantly influence the approaches used to investigate them.

If this class of long ncRNAs is indeed functional, one would expect that they would, in the main, show developmentally regulated and cell-specific expression patterns. The Allen Brain Atlas (ABA) is a large-scale study of the adult mouse brain that comprehensively catalogues and maps the patterns of gene expression that underlie brain development and function on a genome-wide scale (13). The ABA used high-throughput RNA *in situ* hybridization (ISH) to visualize the expression of over 20,000 mainly protein-coding transcripts from the mouse transcriptome at cellular resolution. We discovered that the ABA (13) also contained ISH data for many long ncRNAs. Our analysis of these data provides a landscape perspective of ncRNA expression and provides compelling evidence that this class of RNA is intrinsically functional. Furthermore, by comparing the expression profiles of protein-coding genes with ncRNAs that can be associated via their genomic context, we reveal intriguing functional insights for many individual previously uncharacterized ncRNAs.

Results

Identification and Expression of ncRNA Transcripts. To identify ncRNAs that were targeted within the ABA, we implemented a stringent filtering approach. First, we mapped all ABA probes to the mouse genome and then excluded all probes that could not be uniquely associated with a full-length transcript. To omit probes that targeted protein-coding transcripts, we filtered out any probes that (*i*) overlapped with protein-coding mRNAs in mouse [including unannotated 3′ untranslated regions (UTRs)] (14) as defined by RefSeq (15), Mammalian Gene Collection (16) and UCSC Known Genes (17); (*ii*) matched homologous protein-coding genes in other vertebrate genomes; and (*iii*) targeted transcripts that contained open reading frames (ORFs) predicted by CRITICA (18) and an independent ORF detection algorithm (see *Methods*). We calculated that the final step alone was able to correctly classify 97.2% of RefSeq genes as protein-coding (see *Methods*). Nevertheless, we cannot rule out the possibility that a fraction of the ncRNA subset encodes very small proteins (19) or distant misannotated or unrecognized untranslated exons associated with protein-coding transcripts. In total, we found that 1,328 of the ≈20,000 probes in the ABA (13) targeted transcripts that lacked significant protein-coding potential, including many previously characterized functional ncRNAs [supporting information (SI) Table 1].

The ABA mapped the ISH images to a common anatomical framework that enables the localization and relative quantification of transcript expression in major neuroanatomical brain regions. We used this relative quantified expression data to associate expression of ncRNAs to specific neuroanatomical structures. In summary, we identified 849 ncRNAs that exhibited cellular expression above background (see *Methods* and SI Table 1), and this subset of ncRNAs displayed a wide range of expression profiles, similar to that observed for protein-coding genes in the mouse brain. A comparison of mRNA and ncRNA expression levels reveals that, on average, mRNAs exhibit a higher expression level in the brain. However, we also found that the level of expression of ncRNAs is more variable among the 12 regions of the brain, suggesting that at least some ncRNAs are

Author contributions: T.R.M. and M.E.D. contributed equally to this work; T.R.M. and M.E.D. designed research; T.R.M. and M.E.D. performed research; M.E.D. and S.M.S. contributed new reagents/analytic tools; T.R.M., M.E.D., S.M.S., and M.F.M. analyzed data; and T.R.M., M.E.D., and J.S.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

§To whom correspondence should be addressed. E-mail: j.mattick@imb.uq.edu.au.

This article contains supporting information online at www.pnas.org/cgi/content/full/0706729105/DC1.

more specifically expressed than mRNAs (SI Fig. 5). To validate the expression data in the ABA, we intersected the noncoding probe targets with the GNF Atlas (20), SAGE (21), and CAGE (1) libraries (SI Fig. 6). In total, 606 (71%) expressed ABA probes targeted common transcripts to the GNF Atlas or had CAGE or SAGE evidence for transcription. In addition, 494 (58%) expressed ABA probe target transcripts exhibit evidence of expression in tissues other than brain. A comparison of the lengths and genomic spans of the ncRNAs and mRNAs targeted in the ABA indicated that the ncRNAs were shorter in terms of both transcript length and genomic span than mRNAs (SI Fig. 7). Furthermore, although the primary sequence of long ncRNAs generally is less conserved than protein-coding exons (22), 329 (39%) of the expressed ncRNAs analyzed in the ABA were predicted to contain conserved secondary structures (see *Methods*, SI Table 2, and SI Fig. 7).

The biological relevance of the observed ncRNA expression patterns was supported by examining previously described long ncRNAs, such as *Evf* (23), *Gtl2* (24), *Gomafu* (25), and *Sox2ot* (26), that were targeted in the ABA. Each example was consistent with, and in some cases elaborated on, previous findings. *Evf2* interacts *in trans* with the homeobox transcription factor Dlx2 to regulate the expression of the *Dlx6* gene that it encompasses (23). Consistent with this function, *Evf* exhibits a coincident expression profile with *Dlx2* (SI Fig. 8), which supports the proposed role for *Evf2* in neuronal differentiation (23). *Gtl2*, an ncRNA initially identified by a gene trap insertion that causes a dwarfism phenotype (24), is strongly expressed throughout the adult brain. The ABA also shows a highly similar expression profile for the ncRNAs *Rian* and *Mirg*, which are thought to be transcribed as a single polycistronic transcript that includes *Gtl2* (27) (SI Fig. 9). *Gomafu* is expressed in a distinct set of neurons in the mouse nervous system where it is thought to constitute a cell-type-specific component of the nuclear matrix, which controls gene expression or DNA metabolism (25). Accordingly, the ABA shows a specific expression profile for *Gomafu* within in the brain, with apparently nuclear-restrained expression in distinct cells in the hippocampus, cerebral cortex, and olfactory bulb (SI Fig. 8). The *Sox2* gene, an important regulator of neurogenesis, lies within the intron of a long, alternatively spliced ncRNA, *Sox2ot*, which is a highly conserved transcript present from zebrafish to human (28). One isoform of *Sox2ot* originates from a distal ultraconserved element that has been shown to have enhancer function in the developing forebrain. Furthermore, dynamic bivalent domains in embryonic stem cells and differentiated neurons (26) overlap the transcription start site of *Sox2ot*. A role for *Sox2ot* expression in development also is suggested by its exclusive expression in regions associated with neurogenesis in the ABA (SI Fig. 8). Although microRNAs fall outside the scope of the ISH platform because of their small size, we identified an ncRNA, *AK021368*, which encompasses the microRNA *mir-101a*. *mir-101a* recently was shown to regulate the translation of *Cox2* during embryo implantation (29). It is interesting to note that *AK021368* and *Cox2* exhibit overlapping expression profiles in the cerebral cortex, raising the possibility that a similar regulatory mechanism occurs in the brain.

A systematic analysis of the expressed ncRNAs in the ABA revealed 60 ncRNAs that were expressed at a high level throughout the brain in an apparently ubiquitous manner (SI Table 3 and SI Fig. 10) and 513 ncRNAs that exhibited distinct regionally enriched expression profiles (SI Table 4). The remainder showed no strong expression or detectable enrichment in any one region. Many of the ncRNAs that showed regionally enriched expression profiles were associated with discrete functional subregions of the brain. This finding was particularly apparent in the hippocampus, cerebral cortex, olfactory bulb, and cerebellum. The hippocampus can be divided into four functionally distinct subregions—the dentate gyrus, CA1, CA2, and CA3 (Fig. 1*a*)—that may be delineated by anatomical and gene markers. We identified a number of ncRNAs that exhibited expression profiles specific to these subregions (e.g., Fig. 1*b*; SI Fig. 11). Furthermore, the hippocampal subregions also act cooperatively, and we accordingly identified a number of examples of ncRNAs that showed striking combinations of regionally enriched expression of ncRNAs in each of these subfields (e.g., Fig. 1 *c* and *d*; SI Fig. 11). Similarly, the cerebral cortex can be divided into six distinct layers (Fig. 1*e*), each with specialized functions that are subject to specific efferent and afferent signals and comprising unique neuronal cell types. We identified ncRNAs with expression profiles that delineate cortical laminae (Fig. 1 *f*–*h*; SI Fig. 12). The identification of laminar-specific ncRNAs may lead to further insights into the molecular mechanisms that orchestrate cortical laminar specialization and functional connectivity, both of which currently are incompletely understood. In the three distinct layers of the cerebellar cortex (Fig. 1*i*), we also were able to identify ncRNAs that were enriched in each of these subregions (Fig. 1 *j*–*l*; SI Fig. 13). Similar observations of specific ncRNA expression also were identified in particular subregions of the olfactory bulb (SI Fig. 14).

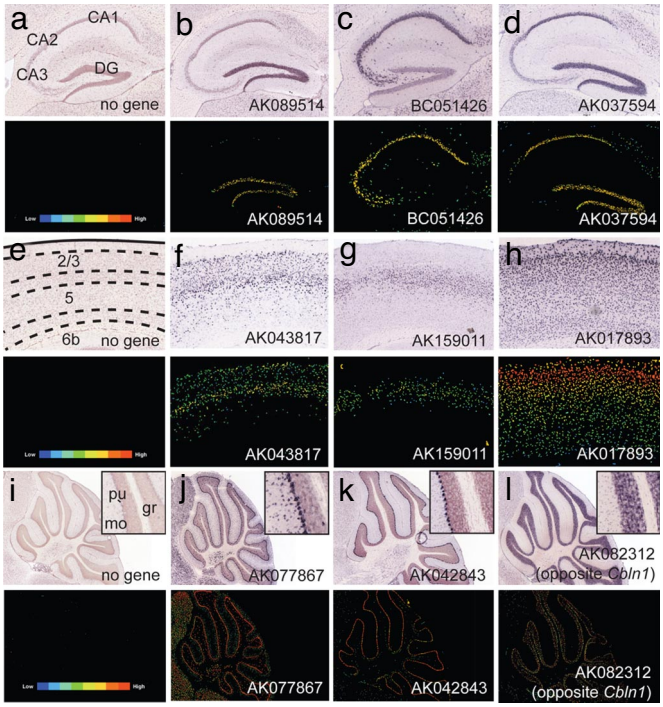

Fig. 1. Regionally enriched expression of ncRNAs in the hippocampus, cerebral cortex, and cerebellum. ISH images of ncRNA expression (accession nos. indicated) in sagittal plane accompanied by false-color heat map below. (*a*) No probe control in hippocampus, with the functionally distinct CA1, CA2, CA3, and dentate gyrus (DG) subfields indicated. (*b*–*d*) Enriched ncRNA expression in DG (*b*) and CA1–CA3 (*c*) and DG and CA1 in combination (*d*). (*e*) No probe control with labeled cortical layer boundaries. (*f*–*h*) Enriched ncRNA expression that correlates with specific cortical laminae. (*i*) No probe control in cerebellum, with the molecular (MO), Purkinje (PU), and granular (GR) layers indicated on detailed *Inset*. (*j*–*l*) Enriched ncRNA expression associated with cerebellar subregions. ncRNA *AK082312* (*l*) is transcribed opposite *Cbln1*, a gene crucial in maintaining the synapse integrity and plasticity in Purkinje cells (62). Further examples are illustrated in SI Figs. 11–14.

Genomic Characterization of ncRNA Transcripts and Comparative Expression with Associated Protein-Coding Genes. The functional characterization of the hundreds of expressed ncRNAs shown here presents a formidable task. However, because a number of previously characterized long ncRNAs regulate the expression of

NEUROSCIENCE

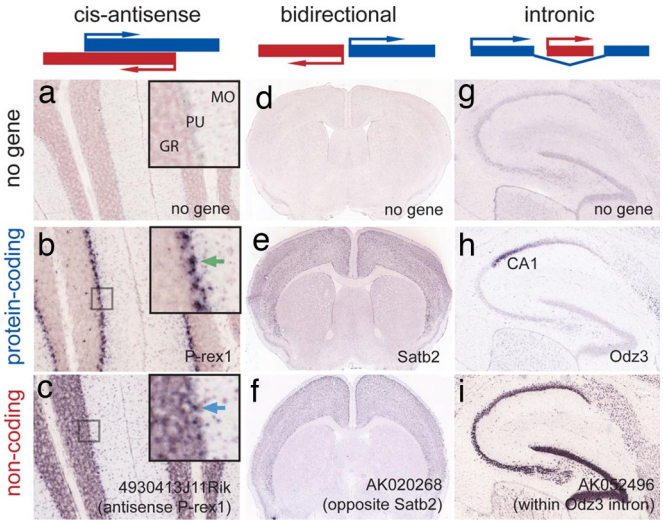

Fig. 2. Expression of ncRNAs associated with protein-coding genes. (*a–c*) *cis*-antisense. (*a*) No probe control showing cerebellum in sagittal plane. (*b*) *P-rex1* is specifically expressed in the Purkinje cell layer in the cerebellum. (*c*) In contrast, an ncRNA that is transcribed antisense to the *P-rex1* 3′ UTR is expressed throughout the granular layer and in a restricted subcellular manner within Purkinje cells. (*d–f*) Bidirectional pairs. (*d*) No probe control of cerebral cortex in coronal plane (CX). *Satb2* is expressed in the cerebral cortex (*e*) similar to an ncRNA transcribed opposite the *Satb2* gene (*f*). (*g–i*) Intronic. (*g*) No probe control showing labeled hippocampus in sagittal plane. *Odz3* is expressed in a gradient in the CA1 hippocampal subfield (*h*) in contrast to an ncRNA located in a *Odz3* intron that is strongly expressed throughout the hippocampus proper (*i*). Further examples are illustrated in SI Fig. 15.

adjacent or overlapping protein-coding genes (23, 30), the examination of the genomic context of ncRNAs may provide preliminary insight into their function, at least at the biological if not the mechanistic level. Therefore, we identified all brain-expressed ncRNAs originating from complex transcriptional loci that encompass protein-coding genes and classified them according to their genomic relationship with protein-coding genes as *cis*-antisense, intronic, or bidirectional (SI Table 5 and SI Fig. 15). Inspection of these genes revealed ≈24% (66 of 278) had associated gene ontology terms relating to neurological function (SI Table 6).

If the relationship between an ncRNA and its associated protein-coding gene is functionally significant, we would expect to observe an equivalent genomic organization conserved in other species. Therefore, we identified ncRNA transcripts that were positionally conserved in the human genome. In total, 39% (109 of 278) of ncRNA transcripts had positional equivalents in the human transcriptome, including 36% (67 of 182) intronic, 39% (20 of 51) bidirectional, and 27% (12 of 44) *cis*-antisense examples. The prevalence of positional equivalents is similar to previous studies (31, 32) and supports the significance of the association of these ncRNAs with protein-coding genes.

***cis*-Antisense Transcripts.** Transcriptional profiling has shown that antisense transcription is prevalent in the mammalian genome (33), and several studies indicate its importance in regulating diverse biological functions (34–36). We identified 44 ncRNAs in the ABA that are antisense to the exons of protein-coding genes (SI Table 5). These antisense ncRNAs often share varied and complex expression relationships with their sense protein-coding transcripts. For example, *P-rex1*, a gene involved in neuronal migration, and its antisense ncRNA partner are both expressed in the cerebral cortex (Fig. 2*b*). However, in the cerebellum, *P-rex1* is specifically expressed in the Purkinje cell layer, whereas the associated antisense ncRNA is expressed within the granular and molecular layer (Fig. 2*c*).

Bidirectional Transcript Pairs. A major organizational theme within the mammalian transcriptome is the prevalence of bidirectional transcript pairs (31, 32), where expression of two transcripts is initiated in close proximity but in opposite directions. We identified 51 ncRNAs that form bidirectional pairs with protein-coding genes (SI Table 5). Many of the ncRNA transcripts exhibit similar expression profiles to their protein-coding partners, suggesting they may be subject to shared regulatory pressures and involved in related neurological processes. For example, *Satb2*, a chromatin remodeling gene expressed by specific cortical neurons (37), and its ncRNA partner show concordant cortical expression profiles (Fig. 2 *e* and *f*).

Similar to previous investigations (31, 32), we also find other complex and often discordant expression relationships between bidirectional ncRNA and protein-coding gene pairs. For instance, one highly conserved ncRNA expressed in the mouse brain is homologous to a human ncRNA that intersects with a recently described human accelerated region (38) and, like the human homolog, is transcribed opposite to the *Klhl14* gene. Although both *Klhl14* and the associated ncRNA are similarly expressed in the anterior olfactory nucleus, the ncRNA also is expressed in the cerebellum and hippocampus (SI Fig. 16). Similarly, an ncRNA bidirectional to *Camkk1*, a gene shown to be involved in male-specific memory formation, exhibits a discordant expression profile with *Camkk1* in the hippocampus, olfactory bulb, and cerebellum (SI Fig. 15). These examples that exhibit discordant expression of ncRNAs relative to their associated protein-coding partners challenges the assertion that long ncRNA transcription occurs solely to "open" chromatin to promote the expression of neighboring protein-coding genes (5, 6), in which we would expect to see concordant expression between the ncRNA and the associated protein-coding genes.

Intronic Transcripts. Instances of regulatory ncRNAs located within introns of protein-coding genes have been previously ascribed neurological functions (39, 40). Among the expressed ncRNAs, we identified 182 ncRNAs that map within the introns of protein-coding genes (SI Table 5). The possibility that these intronic ncRNAs are solely a consequence of persevering nonfunctional lariats is unlikely because we also observe (*i*) intronic ncRNAs expressed in the cytoplasm, (*ii*) nonexpressed intronic ncRNAs of highly expressed host protein-coding genes, and (*iii*) intronic ncRNAs that exhibit a discordant expression profile to their host protein-coding gene (SI Fig. 17). For example, an ncRNA encompassed within the intron of *Odz3* is strongly expressed throughout all hippocampal fields (Fig. 2*i*), whereas expression of the *Odz3* gene itself is restricted to the CA1 subfield (Fig. 2*h*). This also is illustrated by an ncRNA within the intron of *Rorα*, where the intronic ncRNA exhibits a markedly different expression profile in Purkinje cells compared with the *Rorα* gene itself (SI Fig. 15). It has been proposed that RNA signals reside within introns to regulate processes related to that of their host genes (41), and the specific and regulated expression of intronic ncRNAs has been recently reported in humans (42). Similarly, we report here that intronic-derived ncRNAs can exhibit specific and independent expression profiles relative to their host protein-coding gene, supporting the notion that they are biologically significant.

ncRNAs and Imprinting in the Brain. Imprinted loci often give rise to long antisense ncRNAs such as *Air* (43) that trigger the imprinting of neighboring genes. Imprinted genes fulfill essential roles in the development and functioning of the brain (44), and several imprinted genes also undergo imprinting (or lack of) in a neuron-specific manner (34, 45). From a previous survey of

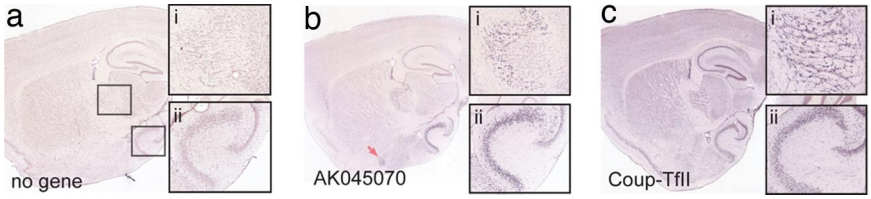

Fig. 3. Expression of *Coup-TfII* and imprinted antisense ncRNA. (*a*) No probe control in sagittal plane showing detail of thalamus (*Inset i*) and ventral hippocampus (*Inset ii*). (*b*) Imprinted antisense ncRNA *AK045070* is expressed in the cortical amygdala area (red arrow), reticular nucleus of the thalamus (*Inset i*), and the ventral hippocampus (*Inset ii*). (*c*) *Coup-TfII* is similarly expressed in the reticular nucleus of the thalamus (*Inset i*) and in the piriform cortex with additional expression in the granular and Purkinje cell layers of the cerebellum, the ventral (*Inset ii*) and dorsal hippocampus.

candidate imprinted mouse transcripts (46), 34 corresponded to ncRNAs expressed in the ABA (SI Table 7 and SI Fig. 9), which included a number of previously characterized imprinted ncRNAs, including *Copg2as* (47), *Gtl2* (24), *Rian* (48), and *Mirg* (49).

The restricted expression of imprinting antisense ncRNA can result in region-specific silencing of neighboring genes. For example, neuron-specific lack of *Air* expression results in neuron-specific lack of silencing at the *Ig2fr* locus (34). Therefore, the heterogeneous expression of another imprinting antisense ncRNA, *Kcnq1ot1* (50), within the brain may similarly result in region- or cell-specific imprinting of neighboring genes (SI Fig. 9). We also have identified a candidate imprinted long antisense ncRNA, *AK045070*, that encompasses >45 kb and shows expression within the thalamus, pyramidal cells in the ventral hippocampus, and cortical amygdalar area (Fig. 3*b*). The adjacent candidate imprinted expressed gene, *Coup-TfII*, an important nuclear receptor involved in regulating various hormonal and brain functions, also exhibits a similar expression profile within these regions (Fig. 3*c*). We suggest that *AK045070* is a favorable antisense ncRNA candidate to trigger the silencing of neighboring genes in a manner similar to *Air* and *Kcnq1ot1*.

Subcellular Expression of ncRNAs. The specific expression profiles of many of these ncRNAs, many of which show exquisite patterns, are indicative of biological meaning and function. However, despite the fact that a significant number of ncRNAs display differential splicing in different cells (1, 11), it has been argued that that the apparently specific expression of noncoding transcripts may be simply cell-type-specific transcriptional noise (4–6) (i.e., an artifact of the differential open chromatin conformation between cell lineages) or are products of transcription that occurs simply to modify chromatin architecture in different cells (5, 7). However, in the few cases that have been studied in more detail, the ncRNAs have been shown to be trafficked to specific subcellular (nonchromosomal) locations (9–12), which is inconsistent with these transcripts' being artifactual.

Within this study, we observed ncRNA expression that appeared to be associated with neuronal extensions or nuclei or appeared as distinct foci within the cell body (Fig. 4). Comparisons between previously examined ncRNAs (25, 48) that are represented in the ABA revealed that the ISH data were indeed informative of subcellular localization, particularly in the cerebellar Purkinje cells whose large size enables resolution of cellular substructures. Therefore, to obtain further insight into ncRNA subcellular localization, we exhaustively examined ncRNAs that were expressed in Purkinje cells (13) (SI Fig. 18). We used previously characterized ncRNAs (27, 51) and mRNAs with known subcellular expression profiles as crude markers of subcellular compartments. From the 88 (10% of expressed ncRNAs) that were expressed in Purkinje cells, 25 (29%) showed a nuclear restricted expression profile, a further 54 (61%) appeared as foci or speckles, and the remaining 9 (10%) were diffusely expressed throughout the soma (SI Fig. 18). This diversity in subcellular localization would be unexpected of artifactual transcription and rather is suggestive of regulated expression. Although we cannot extrapolate the subcellular localization apparent within Purkinje cells to ncRNA expression in the brain, similarly diverse and high levels of mRNA localization have been reported recently in a global analysis of *Drosophila* embryogenesis where RNA localization was thought to have major roles in organizing cellular architecture and function (52).

Discussion

Among the ≈20,000 genes catalogued in the ABA, we identified 849 long transcripts with little or no protein-coding potential that were expressed in the adult mouse brain. Many of these ncRNAs showed regionally enriched expression profiles similar to that observed for protein-coding mRNAs. In addition, viewing the ncRNAs in their genomic context revealed potential functional implications of their expression profiles, particularly with respect to ncRNAs associated with well characterized neurological genes. From a more general perspective, the majority of protein-coding genes can be associated with ncRNAs (1, 31). Given that most of the ncRNAs that have been functionally characterized to date are in some way involved in the regulation of nearby protein-coding genes (3, 8), it appears to be increasingly likely that these mechanisms may indeed be representative of a genome-wide phenomena. However, in this study we were not able to detect any consistent relationship between the expression pattern of an ncRNA and its protein-coding partner for any of the genomic contexts examined. Although in some cases the

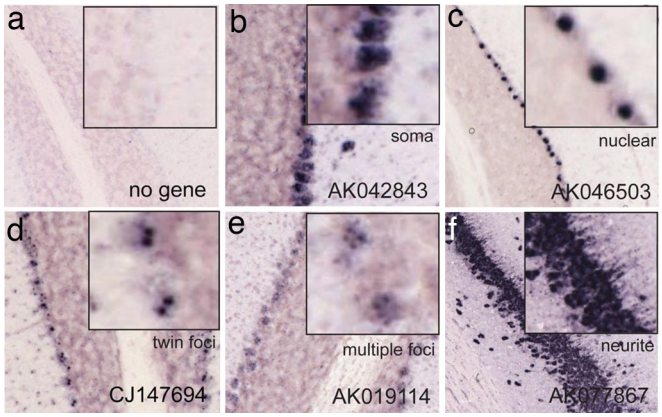

Fig. 4. Subcellular localization of ncRNAs. (*a*) No probe control showing Purkinje cell layer (PU) in cerebellum in sagittal plane. (*b–e*) ncRNAs exhibiting a range of subcellular expression profiles within Purkinje cells: (*b*) expressed throughout the soma; (*c*) expressed throughout the nucleus (ncRNAs known to be exclusively retained in the nucleus were used as indicators for nuclear staining; see SI Fig. 18); (*d*) expressed as twin nuclear foci; (*e*) expressed as multiple foci; and (*f*) expressed in proximal neurite extensions in the hippocampus. Further examples are illustrated in SI Fig. 18.

functions may simply be independent, recent studies (23, 53) suggest that the mechanisms and pathways that underlie the regulatory functions of ncRNAs associated with protein-coding genes maybe diverse, indirect, and complex, and thereby confound the identification of any simple relationship.

The observation that the majority of long ncRNAs examined here exhibit highly specific expression and are posttranscriptionally processed (the majority are spliced and/or polyadenylated), coupled with the recent finding that long ncRNA sequences, splice sites, and promoters are subject to purifying selection (54), strongly suggests that these transcripts are functional and as such considerably expands the repertoire of transcripts that likely play a role in mammalian biology. Because there was no obvious bias in the selection of the noncoding transcripts included in the ABA study, the subset analyzed here may be considered broadly representative of such transcripts generally, of which 34,000 currently are annotated (1). Therefore, because ≈64% of the noncoding subset examined here showed expression above background, we can infer the likely existence of ≈20,000 brain-expressed long ncRNAs, which supports the suggestion that ncRNA, at least in part, underlies the complexity of the brain (55, 56) and, given their presence in other tissue types (1), also are likely to be involved in other aspects of developmental and cellular biology in mammals. If this is the case, it will be a monumental challenge to understand the role of this new and expanding class of ncRNA transcripts, analogous to the continuing efforts to dissect the biochemical and biological functions of the proteome. This study highlights the importance of considering ncRNA in genome-scale experiments, which are commonly restricted to protein-coding genes.

Although the data from the ABA targets only ≈4% of the known noncoding transcriptome, this study nevertheless represents an important early step in appreciating the significance of ncRNA in brain biology and not only provides compelling evidence that many of these transcripts are intrinsically functional but also identifies many for future study. We have compiled the analysis performed here into a searchable database to facilitate the further investigation of ncRNAs in the ABA (http://jsm-research.imb.uq.edu.au/abancrna).

Methods

Probe Mapping. Sequences for 20,098 probes were obtained from the Allen Institute for Brain Science and mapped to the February 2006 (NCBI Build 36) assembly of the mouse genome using BLAT (57) (parameters: minScore = 50, minIdentity = 99, stepSize = 5, tileSize = 11, and ooc = 11.ooc). Probes that could not be reliably mapped were excluded from the study. Mapping data are available on request.

Classification of Probes as Protein-Coding or Nonprotein Coding. Because the probes used by the ABA were derived from a range of sources (including RefSeq, Mammalian Gene Collection, Celera, The Institute for Genomic Research, RIKEN, and Unigene) there is no existing standard classification that encompasses the total set of transcripts targeted by the probes. Therefore, we developed an informatic pipeline to classify probes based on a combination of current gene annotations and protein-prediction software. To associate probes with transcripts, the mapped positions of the probes were intersected with all 3′ UTR, CDS, and 5′ UTR annotations [19,803 RefSeq genes (15), 31,863 UCSC Known Genes (17), 20,407 Mammalian Gene Collection genes (16)], full-length cDNA transcripts (220,902 transcripts from the UCSC Genome Browser ''All mRNA'' track), and all orthologous coding regions of RefSeq sequences from other organisms [113,785 regions from the UCSC Genome Browser ''Other RefSeq'' track (58)] as of March 2007. Probes that could not be associated with full-length transcripts were omitted from the study. To account for unannotated 3′ UTRs, annotated 3′ UTRs were extended with ESTs and mRNAs that formed continuous transcribed fragments (transfrags). The sequences of the targeted transcripts as well as the probe sequences themselves then were analyzed for their protein-coding capacity using CRITICA (18) as previously described (19). CRITICA was used on the basis that previous comparisons of protein-prediction algorithms show it to be the most effective individual tool for discriminating between coding and noncoding transcripts (19). Additionally, CRITICA is able to detect statistically significant regions that encode proteins as small as 50 aa (59) that typically are excluded when using other approaches. Although CRITICA alone correctly identifies 94.4% of RefSeq genes as protein-coding, we also added a further filtering step to remove any transcripts that contain ORFs >120 codons that comprise at least a third of the transcript length. Using these parameters, we retain a large proportion of transcripts that CRITICA does not consider to contain statistically significant ORFs, while still further decreasing our potential false-detection rate. The combination of these two filters correctly predicts 97.2% of RefSeq genes as protein-coding. In summary, probes were classified as coding if (*i*) any targeted transcript (including extended 3′ UTRs) had annotated-protein coding potential, (*ii*) any targeted transcript or the probe sequence itself had significant protein-coding potential as predicted by CRITICA or contained an ORF >120 codons that comprised at least one third of the transcript length, or (*iii*) any targeted transcript intersected with any orthologous region that is annotated as protein-coding in another organism. Probes that did not match any of these criteria were classified as noncoding.

Classification and Comparison of Expression Patterns. Expression level and density data for the ABA were obtained from the Allen Institute for Brain Science. The basis for these measurements has been described previously (13). The expression of transcripts was considered to be regionally enriched if both their expression level was >10 and they ranked in the ''High Expression Level'' class in the Anatomic Search for a particular region in the ABA (www.brain-map.org). Transcripts were considered broadly expressed if their expression level was greater than 10 in 11 neuroanatomical regions (cerebellum, cortex, pons, medulla, midbrain, striatum, olfactory bulb, hippocampus, thalamus, hypothalamus, and pallidum) as defined in the ABA. To compare the expression levels of probes targeting ncRNAs and mRNAs, the regional maximum and coefficient of variation were calculated from the ABA expression level data and plotted on a c-kernel density plot with R. A two-tailed Mann–Whitney test was used to determine the significance of the difference in coefficient of variance between mRNA and ncRNA expression levels.

Genomic Context of Probes. We determined the genomic context of noncoding probes in relation to protein-coding genes. Noncoding transcripts were defined into three categories as follows: (*i*) *cis*-antisense probes were defined where at least 50% of the probe mapped to the opposite strand of a 5′ UTR, CDS, or 3′ UTR; (*ii*) intronic probes were defined where at least 50% of the probe mapped within the intron of a protein-coding gene; and (*iii*) bidirectional probes were defined as noncoding probes that targeted transcripts that had been identified previously as belonging to a bidirectional pair (31). The ABA probe data are searchable by genomic context (and other various criteria) online at http://jsm-research.imb.uq.edu.au/abancrna.

Positional Conservation of ncRNAs. Positional conservation of ncRNAs associated with protein coding genes was determined by examining syntenic genomic regions spanned by the mouse ncRNA transcripts using the Human Chained Alignments tool in the UCSC Genome Browser (58). We individually examined the human syntenic region for evidence of noncoding transcription with equivalent genomic organization (i.e., intronic, bidirectional, or *cis*-antisense). In cases where positionally conserved transcripts were identified, we checked for sequence homology using the BLAT tool (57).

Secondary Structure Predictions of ncRNAs. The mouse-centric genome-wide alignment of vertebrates (''multiz17way'') was downloaded from the UCSC Genome Browser (58). The alignments included sequences of 17 species: mouse (mm8), rat (rn4), rabbit (oryCun1), human (hg18), chimp (panTro2), macaque (rheMac2), dog (canFam2), cow (bosTau2), armadillo (dasNov1), elephant (loxAfr1), tenrec (echTel1), opossum (modDom4), chicken (galGal2), frog (xenTro2), zebrafish (danRer4), Tetraodon (tetNig1), and Fugu (fr1). The alignments were preprocessed by using rnazWindow.pl with default parameters. RNAz was used to predict regions with conserved secondary structure as described previously (60). The secondary structural composition of the noncoding transcripts targeted in the ABA was determined by intersecting the chromosomal positions of the RNAz structural predictions (using confidence threshold levels of $P > 0.5$ and $P > 0.9$) with the chromosomal positions of the noncoding transcripts. Specific secondary structures were determined and visualized with mFOLD (61).

ACKNOWLEDGMENTS. We thank the Allen Institute for Brain Science for providing raw data, Cas Simons for assistance in the preparation of charts, Michael Pheasant and anonymous reviewers for constructive comments on the manuscript, and our laboratory colleagues for stimulating discussions.

T.R.M. is supported by an Australian Postgraduate Award. M.E.D. is funded by a Foundation for Research, Science, and Technology, New Zealand Fellowship. S.M.S. is supported by the Allen Institute for Brain Science, founded by Paul G. Allen and Jody Patton. M.F.M. is supported by grants from the National Institutes of Health, as well as by the F. M. Kirby, the Rosanne H. Silbermann, the Alpern Family, the Lipid, and the Roslyn and Leslie Goldstein Foundations. J.S.M. is supported by an Australian Research Council Federation Fellowship, the University of Queensland, and the Queensland state government.

1. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, *et al*. (2005) *Science* 309:1559–1563.
2. Furuno M, Pang KC, Ninomiya N, Fukuda S, Frith MC, Bult C, Kai C, Kawai J, Carninci P, Hayashizaki Y, *et al*. (2006) *PLoS Genet* 2:e37.
3. Prasanth KV, Spector DL (2007) *Genes Dev* 21:11–42.
4. Brosius J (2005) *Trends Genet* 21:287–288.
5. Chakalova L, Debrand E, Mitchell JA, Osborne CS, Fraser P (2005) *Nat Rev Genet* 6:669–677.
6. Struhl K (2007) *Nat Struct Mol Biol* 14:103–105.
7. Pauler FM, Koerner MV, Barlow DP (2007) *Trends Genet* 23:284–292.
8. Shamovsky I, Nudler E (2006) *Sci STKE* 2006:pe40.
9. Prasanth KV, Prasanth SG, Xuan Z, Hearn S, Freier SM, Bennett CF, Zhang MQ, Spector DL (2005) *Cell* 123:249–263.
10. Willingham AT, Orth AP, Batalov S, Peters EC, Wen BG, Aza-Blanc P, Hogenesch JB, Schultz PG (2005) *Science* 309:1570–1573.
11. Ginger MR, Shore AN, Contreras A, Rijnkels M, Miller J, Gonzalez-Rimbau MF, Rosen JM (2006) *Proc Natl Acad Sci USA* 103:5781–5786.
12. Clemson CM, Chow JC, Brown CJ, Lawrence JB (1998) *J Cell Biol* 142:13–23.
13. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ, *et al*. (2007) *Nature* 445:168–176.
14. Moucadel V, Lopez F, Ara T, Benech P, Gautheret D (2007) *Nucleic Acids Res* 35:1947–1957.
15. Pruitt KD, Tatusova T, Maglott DR (2005) *Nucleic Acids Res* 33:D501–D504.
16. Gerhard DS, Wagner L, Feingold EA, Shenmen CM, Grouse LH, Schuler G, Klein SL, Old S, Rasooly R, Good P, *et al*. (2004) *Genome Res* 14:2121–2127.
17. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D (2006) *Bioinformatics* 22:1036–1046.
18. Badger JH, Olsen GJ (1999) *Mol Biol Evol* 16:512–524.
19. Frith MC, Bailey TL, Kasukawa T, Mignone F, Kummerfeld SK, Madera M, Sunkara S, Furuno M, Bult CJ, Quackenbush J, *et al*. (2006) *RNA Biol* 3:40–48.
20. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, *et al*. (2004) *Proc Natl Acad Sci USA* 101:6062–6067.
21. Khattra J, Delaney AD, Zhao Y, Siddiqui A, Asano J, McDonald H, Pandoh P, Dhalla N, Prabhu AL, Ma K, *et al*. (2007) *Genome Res* 17:108–116.
22. Pang KC, Frith MC, Mattick JS (2006) *Trends Genet* 22:1–5.
23. Feng J, Bi C, Clark BS, Mady R, Shah P, Kohtz JD (2006) *Genes Dev* 20:1470–1484.
24. Schuster-Gossler K, Bilinski P, Sado T, Ferguson-Smith A, Gossler A (1998) *Dev Dyn* 212:214–228.
25. Sone M, Hayashi T, Tarui H, Agata K, Takeichi M, Nakagawa S (2007) *J Cell Sci* 120:2498–2506.
26. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T-K, Koche RP, *et al*. (2007) *Nature* 448:553–560.
27. Tierling S, Dalbert S, Schoppenhorst S, Tsai CE, Oliger S, Ferguson-Smith AC, Paulsen M, Walter J (2006) *Genomics* 87:225–235.
28. Fantes J, Ragge NK, Lynch SA, McGill NI, Collin JR, Howard-Peebles PN, Hayward C, Vivian AJ, Williamson K, van Heyningen V, *et al*. (2003) *Nat Genet* 33:461–463.
29. Chakrabarty A, Tranguch S, Daikoku T, Jensen K, Furneaux H, Dey SK (2007) *Proc Natl Acad Sci USA* 104:15144–15149.
30. Martianov I, Ramadass A, Serra Barros A, Chow N, Akoulitchev A (2007) *Nature* 445:666–670.
31. Engstrom PG, Suzuki H, Ninomiya N, Akalin A, Sessa L, Lavorgna G, Brozzi A, Luzi L, Tan SL, Yang L, *et al*. (2006) *PLoS Genet* 2:e47.
32. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM (2004) *Genome Res* 14:62–66.
33. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, *et al*. (2005) *Science* 309:1564–1566.
34. Yamasaki Y, Kayashima T, Soejima H, Kinoshita A, Yoshiura K, Matsumoto N, Ohta T, Urano T, Masuzaki H, Ishimaru T, *et al*. (2005) *Hum Mol Genet* 14:2511–2520.
35. Alfano G, Vitiello C, Caccioppoli C, Caramico T, Carola A, Szego MJ, McInnes RR, Auricchio A, Banfi S (2005) *Hum Mol Genet* 14:913–923.
36. Lapidot M, Pilpel Y (2006) *EMBO Rep* 7:1216–1222.
37. Britanova O, Akopov S, Lukyanov S, Gruss P, Tarabykin V (2005) *Eur J Neurosci* 21:658–668.
38. Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R, *et al*. (2006) *PLoS Genet* 2:e168.
39. Tarabykin V, Stoykova A, Usman N, Gruss P (2001) *Development* 128:1983–1993.
40. Taft RJ, Pheasant M, Mattick JS (2007) *BioEssays* 29:288–299.
41. Mattick JS (1994) *Curr Opin Genet Dev* 4:823–831.
42. Nakaya HI, Amaral PP, Louro R, Lopes A, Fachel AA, Moreira YB, El-Jundi TA, da Silva AM, Reis EM, Verjovski-Almeida S (2007) *Genome Biol* 8:R43.
43. Sleutels F, Zwart R, Barlow DP (2002) *Nature* 415:810–813.
44. Davies W, Isles AR, Wilkinson LS (2005) *Neurosci Biobehav Rev* 29:421–430.
45. Yamasaki K, Joh K, Ohta T, Masuzaki H, Ishimaru T, Mukai T, Niikawa N, Ogawa M, Wagstaff J, Kishino T (2003) *Hum Mol Genet* 12:837–847.
46. Nikaido I, Saito C, Mizuno Y, Meguro M, Bono H, Kadomura M, Kono T, Morris GA, Lyons PA, Oshimura M, *et al*. (2003) *Genome Res* 13:1402–1409.
47. Lee YJ, Park CW, Hahn Y, Park J, Lee J, Yun JH, Hyun B, Chung JH (2000) *FEBS Lett* 472:230–234.
48. Hatada I, Morita S, Obata Y, Sotomaru Y, Shimoda M, Kono T (2001) *J Biochem (Tokyo)* 130:187–190.
49. Seitz H, Youngson N, Lin SP, Dalbert S, Paulsen M, Bachellerie JP, Ferguson-Smith AC, Cavaille J (2003) *Nat Genet* 34:261–262.
50. Thakur N, Tiwari VK, Thomassin H, Pandey RR, Kanduri M, Gondor A, Grange T, Ohlsson R, Kanduri C (2004) *Mol Cell Biol* 24:7855–7862.
51. Schuster-Gossler K, Simon-Chazottes D, Guenet JL, Zachgo J, Gossler A (1996) *Mamm Genome* 7:20–24.
52. Lecuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, Cerovina T, Hughes TR, Tomancak P, Krause HM (2007) *Cell* 131:174–187.
53. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, *et al*. (2007) *Cell* 129:1311–1323.
54. Ponjavic J, Ponting CP, Lunter G (2007) *Genome Res* 17:556–565.
55. Muotri AR, Gage FH (2006) *Nature* 441:1087–1093.
56. Mehler MF, Mattick JS (2006) *J Physiol* 575:333–341.
57. Kent WJ (2002) *Genome Res* 12:656–664.
58. Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A, *et al*. (2007) *Nucleic Acids Res* 35:D668–D673.
59. Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, Kawai J, Carninci P, Hayashizaki Y, Bailey TL, Grimmond SM (2006) *PLoS Genet* 2:e52.
60. Washietl S, Hofacker IL, Lukasser M, Huttenhofer A, Stadler PF (2005) *Nat Biotechnol* 23:1383–1390.
61. Zuker M (2003) *Nucleic Acids Res* 31:3406–3415.
62. Hirai H, Pang Z, Bao D, Miyazaki T, Li L, Miura E, Parris J, Rong Y, Watanabe M, Yuzaki M, *et al*. (2005) *Nat Neurosci* 8:1534–1541.