HHS Public Access

Author manuscript

Neuron. Author manuscript; available in PMC 2016 January 21.

Published in final edited form as:

Neuron. 2015 January 21; 85(2): 275–288. doi:10.1016/j.neuron.2014.12.024.

DeCoN: genome-wide analysis of in vivo transcriptional dynamics during pyramidal neuron fate selection in neocortex

Bradley J. Molyneaux[#1,5], **Loyal A. Goff**[#1,2,3,6], **Andrea C. Brettler**[1], **Hsu-Hsin Chen**[1], **Siniša Hrvatin**[1], **John L. Rinn**[1,2,4,†], and **Paola Arlotta**[1,2,†]

[1]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, 02138, United States

[2]Broad Institute of MIT and Harvard, Cambridge, MA, 02139, United States

[3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 02139, United States

[4]Department of Pathology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, 02115, United States

[#] These authors contributed equally to this work.

Abstract

Neuronal development requires a complex choreography of transcriptional decisions to obtain specific cellular identities. Realizing the ultimate goal of identifying genome-wide signatures that define and drive specific neuronal fates has been hampered by enormous complexity in both time and space during development. Here, we have paired high-throughput purification of pyramidal neuron subclasses with deep profiling of spatiotemporal transcriptional dynamics during corticogenesis to resolve lineage choice decisions. We identified numerous features ranging from spatial and temporal usage of alternative mRNA isoforms and promoters to a host of mRNA genes modulated during fate specification. Notably, we uncovered numerous long non-coding RNAs with restricted temporal and cell type specific expression. To facilitate future exploration, we provide an interactive online database to enable multidimensional data mining and dissemination. This multi-faceted study generates a powerful resource and informs understanding of the transcriptional regulation underlying pyramidal neuron diversity in the neocortex.

[†]Correspondence to paola_arlotta@harvard.edu and john_rinn@harvard.edu.

[5]Current address: Pittsburgh Institute for Neurodegenerative Diseases, Departments of Neurology and Critical Care Medicine, University of Pittsburgh, Pittsburgh, PA, 15213, United States

[6]Current address: Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, 21205, United States

Supplemental Information

Supplemental information includes three figures, six tables, the GTF file of the assembled reference transcriptome, and Supplemental Experimental Procedures and can be found with this article online.

Introduction

The myriad and complexity of neuronal networks present in the mammalian brain provide the basis for critical faculties such as sensory perception, motor behavior, and cognition. The neocortex in particular plays a critical role in computing higher-order brain functions, which are executed by an extreme diversity of cortical neuronal classes. Decoding the origin of this variety of neurons and defining the rules that shape and maintain neuronal diversity in the neocortex, and in the CNS more broadly, holds great potential but is still an unmet goal.

Addressing this challenge requires both a high-throughput neuronal subclass purification method and an integrative approach that considers dynamic, multilayered transcriptional regulation during the acquisition of distinct neuronal identities. Similarly, combinatorial profiling of multiple neuronal subtypes obtained from the same tissue may be required to understand cross-regulatory events that shape circuits.

A variety of genetic and surgical approaches have been used to attempt to resolve neocortical complexity and characterize distinct neuronal cell types (Arlotta et al., 2005; Ayoub et al., 2011; Belgard et al., 2011; Doyle et al., 2008; Fertuzinhos et al., 2014; Fishell and Heintz, 2013; Heiman et al., 2008; Lobo et al., 2006; Molyneaux et al., 2009; Sugino et al., 2005). However, it remains challenging to mark and purify defined neuronal subclasses, especially during developmental timelines, where the specificity of individual markers is dynamic, and in a scalable manner that enables high-throughput molecular profiling across many populations. We present here a broadly applicable approach leveraging a combination of experimental and systems-wide analyses to address this critical need. Specifically, we have incorporated the knowledge gained from other approaches to enable combinatorial immunodetection of nuclear markers to resolve neocortical pyramidal neuron populations and their temporal changes during cortical development.

We sought to investigate the transcriptome of three different subpopulations of cortical pyramidal neurons, selected for their diversity of targets, their importance for cortical function, and their direct clinical relevance. We profiled subcerebral projection neurons (ScPN), which include the clinically significant corticospinal motor neurons (CSMN), callosal projection neurons (CPN), and corticothalamic projection neurons (CThPN). To characterize these subpopulations throughout development, we used immunostaining against unique combinations of transcription factors for fluorescence-activated cell sorting (FACS) (Hrvatin et al., 2014). Using this approach we identified molecular signatures that distinguish between populations. In fact, we could discriminate among pyramidal neuron types, for the first time, at the height of key decisions in fate specification, migration, and axon targeting.

To comprehensively and systematically characterize the molecular signatures underlying neuronal diversity we performed whole transcriptome analyses by massively parallel RNA sequencing. In total, we identified 8864 genes with significant differential expression, 812 alternative promoter switches, 1068 changes in protein-coding sequences, and 1181 genes that demonstrate significant shifts in their relative isoform abundance during corticogenesis. Increasing evidence suggests that long noncoding RNAs (lncRNAs) play essential roles in

the specification and maintenance of cell identity (Dinger et al., 2008; Grote et al., 2013; Guttman et al., 2011; Mercer and Mattick, 2013; Mercer et al., 2008; Ponting et al., 2009; Ramos et al., 2013; Sauvageau et al., 2013). Accordingly, we assembled 5195 lncRNAs from these data, 806 of which exhibit differential expression across neuronal types over time.

Given the complexity and depth of this resource, we have developed an intuitive web-based utility to facilitate data dissemination (http://rinnlab.rc.fas.harvard.edu/pyramidal/), exploration, and future investigation, enabling researchers to navigate the full extent of the dataset. Specifically, this utility allows researchers to dynamically explore groups of genes meeting user defined expression criteria, examine upstream regulatory mechanisms, identify expression data at the isoform level, and examine processed RNA-seq reads to identify new exons and promoters for genes of interest. The data are integrated with other online resources such as the Allen Brain Atlas and the UCSC Genome Browser to facilitate cross-platform discovery.

Collectively, this work provides an approach to purify multiple classes of neurons from the same tissue without the need for genetic labeling. The labeling and purification procedure is compatible with high throughput RNA sequencing, and it enabled the generation of a deep resource of global transcriptional controls over the developmental divergence of individual classes of pyramidal neurons. The scalability of the methods and independence from genetic labels makes this platform universally applicable to transcriptional and epigenetic profiling and amenable to the screening of primary classes of neurons from the human brain.

Results

Scalable purification of molecularly defined neuronal populations

To investigate the transcriptional dynamics of neuronal fate decisions, we chose to focus on pyramidal neurons of the neocortex, a region of the brain with extraordinary neuronal diversity that remains underexplored at the molecular level. We purified and profiled three classes of pyramidal neurons based on the differential expression of a combination of three transcription factors: BCL11B (CTIP2), TLE4, and SATB2 (Figure 1A). Layer V subcerebral projection neurons, which include corticospinal motor neurons, are identified by their high BCL11B, low TLE4, and low SATB2 expression (Molyneaux et al., 2007). Corticothalamic projection neurons and subplate neurons (hereafter collectively referred to as CThPN) are identified by high TLE4 (Allen and Lobe, 1999), moderate BCL11B (Arlotta et al., 2005), and low SATB2 expression. Interhemispheric callosal projection neurons are identified by high SATB2, absent BCL11B, and absent TLE4 expression (Alcamo et al., 2008; Arlotta et al., 2005). Cortical tissue was dissociated to a single cell suspension, and cells were fixed with paraformaldehyde and concomitantly permeabilized by saponin prior to immunocytochemistry with chosen combinations of antibodies for FACS (Figure 1B-F; Experimental Methods).

Despite changes in marker gene expression at different stages of development, we were able to use our labeling and FACS strategy to reliably distinguish between the lineages of ScPN, CPN and CThPN as early as E15.5 (Figure 1B-E). For example, even though BCL11B

functions as an ideal discriminating marker later in development, BCL11B expression at E15.5 is roughly equivalent between subcerebral projection neurons and corticothalamic projection neurons, precluding their distinction. However, differing levels of TLE4 can distinguish the two cell types at this stage. As a result, over time, we can detect changes in levels of expression of each transcription factor and therefore distinguish these three neuronal classes as they diverge from each other during development (FACS plots; Figure 1B-1E). We can also detect temporal changes in the relative abundance of each cell type. For example, from E15.5 to P1 CPN increase from 5.4% of total cells to 26%, ScPN decrease from 6% to 1.1%, and CThPN remain constant at approximately 6%.

To understand gene-regulatory changes in these populations we performed systematic and comprehensive whole transcriptome analyses. Briefly, we generated RNA-seq libraries with two biological replicates for each neuronal type and across several developmental stages (E15.5, E16.5, E18.5, and P1). Despite fixation and reverse crosslinking, we obtained high-quality total RNA from all purified cell populations with RNA integrity numbers (RIN) ranging from 6.4 to 9 (median of 7.1). Approximately 100,000 cells were required to obtain 200 nanograms of total RNA, which served as input for the standard Illumina TruSeq RNA-seq library preparation. Libraries were sequenced to a mean of over 100 million mapped 100 base pair paired-end reads per replicate (Table S1; Figure S1A-C). Merged assemblies were generated as described in (Trapnell et al., 2012b) and detailed in Methods, and used as input for our previously-described long noncoding RNA (lncRNA) identification pipeline (Cabili et al., 2011). LncRNAs with a minimum of 3× coverage were appended to the list of known UCSC protein-coding genes to establish a suitable reference transcriptome for isoform-level quantification and differential expression testing with Cuffdiff2 (Trapnell et al., 2012a).

We first analyzed the sorted markers in our whole transcriptome analyses and find they are in the expected populations. As anticipated, the expression profiles of Bcl11b, Satb2, and Tle4 were consistent with known patterns of expression *in vivo* (Figure 1G). This specificity extended more broadly to a cohort of 150 genes known to have varying degrees of subtype specific expression during development (Figure 1H and Table S2) (Arlotta et al., 2005; Hoerder-Suabedissen and Molnar, 2013; Lein et al., 2007; Lodato et al., 2014; Molyneaux et al., 2007; 2009) for which we now provide detailed profiles of temporal changes in expression. Notably, most genes that distinguish between ScPN and CThPN are not expressed until E16.5.

To evaluate the purity of the sorted populations, we investigated genes with known expression in interneurons, oligodendryoctes, astrocytes, and endothelial cells (e.g. Dlx1, Gad1, Gad2, Sox10, Gfap, and Vwf; Figure S1D-H) that normally would not be expressed in these populations. We observed they were not expressed in most samples, indicating that these populations have minimal contamination with other cell types. The singular exception was the CPN samples from E15.5 where the FACS plots demonstrated that the Satb2-high population was not a clearly distinct population for gating (Figure 1B, CPN gate) and had low yet detectable levels of expression of Dlx1, Gad1, Gad2, Sst, and Lhx6, indicating possible contamination from migrating interneurons (Figure S1D-E). Intriguingly, we found that the P1 CThPN population also expressed somatostatin (Sst), a gene whose expression is normally associated with subpopulations of interneurons (Kawaguchi and Kubota, 1997).

However, in contrast to the E15.5 CPN samples, other known markers of interneurons were not expressed in P1 CThPN at levels strikingly different from other conditions (e.g. Dlx1, Gad1, Gad2, Slc32a1; Figure S1D-E). This observation indicates that Sst expression might not be the result of contamination by interneurons but that this gene is differentially expressed in CThPN or a subpopulation of CThPN. Together, these data demonstrate that FACS purification with antibodies to cell-type-specific transcription factors can be utilized to obtain high-quality RNA to resolve transcriptional dynamics.

Identification of genes contributing to class-specific neuronal identity

We first sought to identify genes with significant differences in gene-level expression over time or between subtypes, and used these expression estimates to identify clusters of meaningful expression patterns. Cuffdiff2 was used to identify 8864 genes (8058 protein-coding and 806 lncRNAs) with significant ($q<0.0004$; Cuffdiff2 test) differential expression between cell types and developmental stages. Examining all pair-wise differences between conditions, we can describe the set of all genes that change during both neuronal differentiation over time and subtype specification during corticogenesis. These 8864 genes represent a comprehensive list of PolII transcribed elements that significantly distinguish these neuronal populations during cortical development and were utilized in select downstream analyses to investigate the dynamics of gene expression.

To disentangle gene expression changes that correlate with general neuronal maturation from those that correlate with cell type specification, we performed principal component analysis (PCA) independently on the significant protein-coding genes and lncRNAs. PCA on the protein-coding genes revealed that neuronal maturation over embryonic time was a greater source of variability than neuronal subtype differentiation (Figure S2A). In contrast, PCA on significant lncRNAs identified principal components that were mixed for both temporal and neuronal cell-type-specific contexts (Figure S2B).

To identify novel marker genes that appropriately distinguish between these three neuronal subtypes and might play functional roles in their development and function, we sought to identify those genes that exhibit the highest degree of cell type specificity. To this end, we ranked all significantly differentially expressed genes using a custom similarity score based on the Jensen-Shannon distance between the gene's normalized expression and an ideal gene with uniform expression for a given cell type across time (Methods). For each cell type, the top 25 most specific genes that meet this criteria are presented in Figure 2A. As anticipated, several genes with previously described expression in each cell type were identified, including Lhx2, Cux1, and Cux2 for CPN (Bulchand et al., 2003; Nieto et al., 2004), Oma1 and 6430573F11Rik for ScPN (Arlotta et al., 2005), and Tle4 and Ngfr in the corticothalamic/subplate population (Allen and Lobe, 1999; Allendoerfer et al., 1990). In addition to previously known markers, we identified many new genes, including lncRNAs that have not previously been described as being specifically expressed within these three neuronal populations. *In situ* hybridization for twelve such genes confirmed highly restricted patterns of expression in the developing cortex consistent with their RNA-seq expression profiles (Figure S3A-L). For example, RNA-seq data reveals that linc-Cyp7b1-3 is expressed in ScPN, and *in situ* hybridization demonstrates expression restricted to a subset

of layer V cells, consistent with restriction to ScPN (Figure S3J). Together these results suggest that many novel lncRNAs identified have consistent expression patterns in vivo.

We sought to obtain an unbiased view of the various expression profiles used during specification and maturation of these cell types to begin to discern patterns of co-regulation. We distilled the significant gene expression profiles into 20 distinct patterns of gene expression using nearest-neighbor agglomerative graph clustering on cosine similarities (Table S3; see Experimental Procedures). The resulting clusters were manually partitioned into five distinct groups: three groups of “class specific signature clusters” (CPN, ScPN, and CThPN); one group of “mixed-cell type” clusters whose genes demonstrate expression in only two cell types (i.e. corticofugal, which includes ScPN and CThPN); and one group of clusters whose genes demonstrate consistent expression between all cell types, either increasing or decreasing over time (Figure 2B). Genes within the subclass-signature clusters represent likely candidates that may contribute to neuronal cell type specification or subtype specific activities, while those genes in the latter clusters are most likely associated with broader neuronal differentiation processes shared among distinct neuronal cell types.

The organization of these data into clusters of similarly regulated genes vastly expands upon the list of known cortical neuron subtype specific genes. For example, cluster 11 contains Bcl11b (Ctip2), which plays a key role in mediating ScPN axonal projections to the spinal cord (Arlotta et al., 2005). Bcl11b shares a similar expression profile with 218 other genes within cluster 11 demonstrating elevated expression in ScPN from the earliest stages of development. Cluster 16 contains the known ScPN genes Crim1, Crym, and Diap3 (Arlotta et al., 2005) as well as 743 other genes with subcerebral specific profiles that increase in expression during development. Each of these clusters contains both lncRNAs and protein-coding genes (Figure 2B; pie chart inlays).

In order to understand the relative contribution of lncRNAs and protein-coding transcripts to cell type specificity, we compared a calculated maximum specificity score (Cabili et al., 2011) for protein-coding genes and lncRNAs across all three cell types at each time point. Given the complexity of comparing specificities when lncRNAs as a population are expressed at significantly lower levels than protein-coding genes [Figure 2C; (Cabili et al., 2011)], we examined specificity scores using two different methods to correct for the confounding influence of expression level.

First, we resampled protein-coding genes drawn from the learned empirical distribution of lncRNA maximum FPKM values for each time point. For each of the 1000 samples (Figure 2D, black lines) we performed a Kolmogorov-Smirnov (K-S) test between the cumulative densities of the lncRNA maximum specificity scores (Figure 2D, red lines) and the scores for the resampled protein-coding genes. Next, to more appropriately model cell type specificity as a continuous function of expression levels, a generalized additive model (GAM) was fit across all genes with maximum specificity as the response variable and maximum FPKM expression, cell type, and developmental time as explanatory variables (Figure 2F; Experimental Procedures).

We observed that expression level accounts for the majority of variation in cell type specificity in our data (Figure 2F). While lncRNAs were slightly more cell-type-specific than protein-coding genes (Figure 2D & 2E, by K-S test >97% of the resampled specificity distributions are different for all time points, $p<0.01$; Figure 2F & 2G, by GAM $p<3.2e^{-06}$, Pr(<|t|)), the magnitude of the difference is small. Interestingly, we also observed a significant increase ($p<0.00015$; Pr(<|t|)) in specificity for both lncRNAs and protein-coding genes over the course of cortical neuron development (Figure 2G; second panel), consistent with increased transcriptional divergence between neuronal cell types over developmental time. Together, these data indicate that lncRNAs have a small, statistically significant increase in specificity after correcting for expression; this difference is small and its biological value hard to define.

Transcriptional dynamics of lncRNAs

Our data provide a context in which to examine the relative contributions of specific lncRNAs to neuronal cell-type identity, as opposed to previously recognized tissue-level differences. We decided to examine in more detail the lncRNAs that were identified in these neurons. We assembled a total of 5195 lncRNA genes from our RNA-seq data, 1136 of which represent novel gene loci compared to the UCSC mm9 reference transcriptome (Figure 3A, detailed in Experimental Procedures). In addition, we identified 2978 known lncRNA gene loci, and 512 novel isoforms of known lncRNA genes from our data (Figure 3A). We assembled 500 lncRNAs characterized as antisense to a known gene. However, we chose to remove these antisense lncRNAs from our quantification and differential expression assays since overlapping protein-coding gene expression would confound accurate expression estimates from un-stranded libraries.

We observed 806 lncRNAs that exhibit significant ($q<0.0004$; Cuffdiff2) changes in expression over time or between cell types (Figure 3B). Of the 806 significant lncRNAs, 449 lncRNAs (55.7%; 135 CPN, 180 ScPN, 134 CThPN) are assigned to cell-type signature clusters, while 259 lncRNAs (32.1%) are associated with cell-type independent clusters. The remaining 87 lncRNAs (10.8%) can be found in the mixed cell-type clusters. The bulk of the significant lncRNAs (688; 85.4%) are intergenic to known genes, and the remaining 118 (14.6%) share a bidirectional promoter with a known protein-coding gene.

We next assessed the specificity of our discovered lncRNAs for expression within the brain relative to other tissues. We quantified the expression of the 806 significant lncRNAs across a panel of 29 publicly available RNA-seq datasets (detailed in Experimental Procedures). 49.4% of significant lncRNAs (398/806) were detected at FPKM > 2 in at least one of the tissues sampled. As expected, the number of tissues in which a given significant lncRNA was expressed with an FPKM > 2 (Figure 3C) was dramatically lower (mean 2.7; median 0), compared to the counts for significant protein-coding genes (mean 16.3; median 19).

We next asked whether we could identify any significantly regulated lncRNAs with potential human syntenic equivalents. 175 (21.7%) of the 806 significant lncRNAs have an identifiable syntenic human equivalent (Table S4), as defined by the presence of a transcribed element within the same syntenic region in the human genome (hg19; TransMap; detailed in Experimental Procedures). This fraction is consistent with previous

observations of the lincRNAs with putative human orthology (Cabili et al., 2011). Interestingly, despite syntenic transcription, these particular lncRNAs do not demonstrate any significant difference in cell type specificity within our cortical differentiation dataset relative to lncRNAs with no discernable human ortholog (Figure 3D).

Many of the identified lncRNAs could be used as neuron type specific markers for each population. For each neuronal cell type, we observed several lncRNAs with dynamic and specific expression. For example linc-Cyp7b1-3 is a multi-exonic intergenic lncRNA expressed from chromosome 3 that displays a high degree of cell type specificity for ScPN (Figures 3E and S3J). This lncRNA has a PhyloCSF score of −30372, indicating a very low probability of being a protein-coding gene. Other lncRNAs, including linc-Phf17-2 (PhyloCSF Score −15093) and linc-1700066M21Rik-1 (PhyloCSF Score −30981), demonstrate remarkable cell type specificity for either CThPN or CPN, respectively (Figures 3F-G). These few examples highlight the diversity of a large set of cell-type-specific lncRNAs and suggest that some lncRNAs could be used to classify neuronal subtypes during cortical development.

Neuronal subtype specific use of gene pathway components

We next analyzed our resource for gene sets or pathways that are differentially regulated in each cell type. We first conducted a Gene Ontology enrichment analysis using the lists of genes differentially expressed between any two neuron types at P1 (5% FDR) and additionally filtered by a maximum specificity score. Reactome gene sets [C2 reactome; MSigDB; (Subramanian et al., 2005)] were tested for enrichment using a hypergeometric test. The list of significant gene sets was fairly consistent between the three cell types (not shown), which is expected given that current annotations only cover broad processes of cellular development and function. We noticed however, that for many of the common gene sets, the individual genes driving these signatures were differentially expressed between class and time.

We therefore devised a different approach to identify genes that share similar curated annotations but are uniquely expressed in a given cell class. To achieve this, all genes were rank-ordered by their maximum specificity for any of the three neuronal populations at P1 and filtered for a minimum FPKM expression level of 2. The ranked list was used as input for a pre-ranked Gene Set Enrichment Analysis [GSEA; (Subramanian et al., 2005)] against the collection of Reactome gene sets (Figure 4A; C2 Reactome v4.0; MSigDB).

We observed several pathways that exhibited significant differential usage of genes across individual neuronal classes including gene sets containing cell surface receptors, potassium channels, and various components of the extracellular matrix. Specifically, the top four most significant Reactome gene sets involve G-protein coupled receptor (GPCR) signaling molecules including both receptors and ligands, as well as genes involved in specific downstream signaling cascades (Figure 4B).

To further explore the specific use of GPCR receptor classes across our three cell types, we generated gene sets from the curated lists of seven transmembrane (7TM) receptors from the IUPHAR database of receptors and ion channels (Sharman et al., 2012). We identified a

cohort of metabotropic glutamate receptors that are differentially used by cortical neuron subtypes including Grm1 and Grm4 (specific to CThPN), Grm2 (specific to CPN), Grm3 (specific to corticofugal), and Grm5, Grm7, and Grm8 with shared expression in both CPN and ScPN. In contrast, expression of the two $GABA_B$ receptors Gabbr1 and Gabbr2 remained consistent across all three cell types. Interestingly, we also observed a significant differential usage of several other 7TM receptor classes including specific expression of several 5-HT (serotonin) receptors amongst the corticofugal cell types, and highly specific expression of the adrenoreceptors including Adra2c in the CPN, Adrb1 in ScPN, and Adra1b and Adra2a in CThPN. In addition, we also identified a number of orphan GPCRs with specific subtype expression including Gpr158 (CPN), Opn3 and Gpr176 (ScPN), as well as Gpr3, Gpr22, and Gpr39 (CThPN).

We identified several other gene sets with a significant enrichment for cell-type-specific genes including a set containing axon guidance molecules. Analysis of a manually curated set of known cell surface ligands and receptors with demonstrated roles in axonal guidance revealed the different codes of related molecules expressed by each neuronal subtype during circuit formation (Figure 4C). Importantly, this degree of specificity is only observed for a select few gene sets while the majority, including many genes involved in basal metabolic processes, are expressed at common levels (Figure 4D). These results are consistent with a previous study that broadly suggested that cell surface proteins greatly contribute to the diversity of CNS cell types (Doyle et al., 2008). However, here we provide a high-resolution characterization of the expression of distinct subsets of genes within broader gene sets otherwise shared by closely related cortical projection neuron subtypes that begins to describe subtle differences between these classes and their dynamic changes during development.

Isoform-level resolution of transcriptional dynamics during corticogenesis

The high resolution and depth of sequencing within our resource allows us to investigate additional aspects of transcriptional regulation such as alternative splicing and alternative promoter usage during corticogenesis (Figure 5A) (Trapnell et al., 2012a). We identified 812 genes that undergo promoter switching, 1068 genes that significantly alter their protein-coding sequence, and 1181 genes that demonstrate significant shifts in their relative isoform abundance. Interestingly, 597 of these genes with an alternative regulatory event demonstrate transcript-level regulation without significant change in overall gene expression (Figure 5B). Of these genes, 371 (31.4% of isoform switching genes) represented significant alternative isoform usage, 296 (27.7% of CDS switching genes) demonstrated a significant change in CDS, and 194 (23.9% of promoter switching genes) demonstrated alternative promoter usage during corticogenesis (Table S5).

Several genes provide compelling examples of how interpretations of gene-level significant differences would often exclude genes that use RNA processing as an alternative form of regulation during cortical development. One example is the anterograde motor protein Kif1a, a causal gene of hereditary spastic paraplegia (Klebe et al., 2012), a degenerative disease that affects corticospinal motor neurons as well as other ScPN. We observed comparable levels of increasing Kif1a expression in each cell type (Figure 5C). However,

we found that while two of the alternative isoforms for Kif1a are strongly upregulated over time, the expression of one isoform that is missing a cassette exon (uc007cdg.2) drops significantly during development. Another example is Lrrtm4, an important regulator of synaptic development with highly selective expression in the brain (Siddiqui et al., 2013), which exhibits significant antithetical isoform regulation in ScPN and CThPN with one isoform increasing in ScPN while another is decreasing in CThPN (Figure 5D). Lastly, Rbm7, a putative RNA binding protein (Lubas et al., 2011), appears at the gene level to maintain stable expression in all three cell types over time (Figure 5E). However, at the isoform level we observe a complete conversion during cortical development from the expression of isoform uc009pie.2, which codes for a 265 amino acid protein, to isoform uc009pif.2, which includes an alternatively spliced exon with three in frame stop codons that results in a truncated peptide (Figure 5F). Rbm7 isoform level differences identified by RNA-seq were confirmed by quantitative RT-PCR (Figure 5G). These examples of isoform level changes in expression demonstrate the importance of examining the dynamics of the transcriptome at the sub-gene level to identify significant differences between the use of individual isoform, promoter, and transcription start sites by different neurons.

DeCoN: The Developing Cortical Neuron Subtype Transcriptome Resource

As part of this study, we have created an interactive companion website to facilitate future analysis and exploration of these data (http://rinnlab.rc.fas.harvard.edu/pyramidal/). This website will enable investigators to visualize, query, and manipulate the extensive transcriptome data presented here and serve as a repository for future transcriptomic and epigenetic data on cell-type-specific development in the neocortex. At both the gene and geneset level, we have provided several web based visualization tools to investigate this dataset using all of the analyses described here as well as additional exploratory tools including utilities for gene discovery.

When possible, these data are integrated with existing data from various external sources including Allen Brain Atlas *in situ* hybridization data (Lein et al., 2007) and processed RNA-seq reads via tracks in the UCSC Genome Browser to facilitate a thorough exploration of the data (Raney et al., 2013). The experimental methodology, rich data resource, and user-friendly interface for analysis at the gene and system levels provide a critical platform for understanding the genome-wide transcriptional dynamics of neuronal differentiation.

Discussion

To begin to characterize the transcriptional events responsible for the establishment of the neuronal diversity of the neocortex in greater detail, we developed a high throughput experimental method, combined with massively parallel RNA sequencing, and robust systems-level analyses to characterize the transcriptional dynamics of three neuronal populations during development. We built an interactive platform (DeCoN: The Developing Cortical Neuron Transcriptome Resource) to enable multidimensional data mining, exploration, and dissemination that is scalable and designed to integrate future data on additional neuronal classes and different species.

High throughput isolation of neuronal subtypes

A key feature of the CNS is the incredible diversity of cell types. Antibody based discrimination of distinct cell types has enabled high throughput study in the immune and hematopoietic systems. Similar methods are lacking for parsing the heterogeneity of the CNS and the field has lagged considerably behind efforts in other tissues due to the fragility of neurons and the fact that many cell-type-specific neuronal cell surface markers are localized to the axon or dendrites rather than on the soma and thus are lost during tissue dissociation. Here, using fixed neurons, we demonstrate that FACS purification with antibodies to cell-type-specific transcription factors can be utilized to finely discriminate between different populations of neuronal subtypes, despite dynamic profiles of transcription factor expression during development.

It is increasingly evident that cortical projection neuron subtypes are themselves heterogeneous (Hoerder-Suabedissen and Molnar, 2013; Sorensen et al., 2013). The true diversity of neocortical neurons is yet to be determined but can be more readily explored with antibodies to additional transcription factors or intracellular epitopes to further subdivide and refine the molecular taxonomy of the populations described here.

Because these methods do not require genetic labeling, this approach opens the door for investigations of specific neuronal populations in the human and non-human primate brain, which has thus far been limited to laser capture microdissection of individual cells or broad regions (Hawrylycz et al., 2013; Johnson et al., 2009; Kang et al., 2011; Oldham et al., 2008). Adopting this approach will greatly facilitate the study of disease susceptibility of selected classes of neurons in humans (Saxena and Caroni, 2011).

Notably, since neurons are fixed prior to FACS purification, they readily withstand higher-pressure FACS conditions, which have traditionally limited the ability to purify large numbers of neurons. This advantage makes possible the collection of millions of neurons of a particular subtype, enabling broad, multidimensional profiling of the transcriptome and epigenome in mouse and human models of development and disease. Additionally, with new markers identified here, it will be possible to further subdivide individual classes to explore pyramidal neuron heterogeneity.

Insights into the transcriptional complexity of closely related neuronal subtypes

The process of neuronal specification, migration, and circuit formation is complex with multiple levels of regulation. Through deep sequencing of developing pyramidal neuron transcriptomes, we can describe the extent of transcriptome changes including isoform-level transcriptional regulation as well as the differential use of multiple promoters and transcription start sites (TSS) as means to add variety and diversity to the transcriptional output.

The accurate and complete assembly of lncRNA transcripts remains a difficult problem generally complicated by the lower relative abundance, and lagging curation and annotation of full-length lncRNA transcripts. The combination of cell-type purity and deep sequencing in our data enhances the likelihood that our assembled lncRNAs represent full-length transcript reconstructions, since read coverage is the strongest predictor of assembly quality

(Steijger et al., 2013). We have attempted to minimize the impact of low read coverage on the quality of our assembled lncRNAs by requiring that any lncRNA transcript have a minimum of 3.0× coverage; a threshold previously determined to provide an estimated >80% recovery rate of well-annotated protein-coding exons using Cufflinks (Steijger et al., 2013). It must be noted however, that experimental validations of individual lncRNA transcripts currently remains the 'gold-standard' for assessing assembly quality. Additional attempts to minimize the error rates for our assembly involved the selection of a PhyloCSF score threshold of 100, which was previously demonstrated to correspond to a false negative error rate of 6% for protein-coding genes and a false positive error rate of 9.5% (Cabili et al., 2011). Most stringently, we also exclude transcripts with any significant Pfam hit in any of the three possible reading frames. The result is a catalog of assembled lncRNA transcripts from discrete neuronal populations that benefits from the increased depth of sequencing and the relative cellular homogeneity of the input materials.

Using this lncRNA catalog, we find that these genes are at least as specific as protein-coding genes of comparable expression levels in distinguishing individual pyramidal neuron subtypes. The data also highlight the importance of accounting for differences in level of expression when comparing the specificity of lncRNAs and protein-coding genes.

Though only a handful of lncRNAs have been functionally examined, increasing evidence suggests that lncRNAs have diverse biological functions, including chromatin modification, transcriptional regulation, and post-transcriptional processing (Mercer et al., 2009). LncRNAs also appear to play roles in the specification and maintenance of cell identity (Guttman et al., 2011; Sauvageau et al., 2013; Sun et al., 2013). Global transcriptome studies of whole brain or microdissected layers have identified numerous lncRNAs expressed in the cortex (Ayoub et al., 2011; Belgard et al., 2011; Mercer et al., 2008; Ponjavic et al., 2009; Zhang et al., 2014). The current data suggest that lncRNAs might contribute to subtype-specific neuronal properties, and provide a more comprehensive list of marker genes and potential therapeutic targets. Although functional analysis of lncRNAs is in its infancy and inherently challenging (Bassett et al., 2014), here we provide hundreds of new subtype specific lncRNAs as candidates to investigate the contributions of lncRNAs to the development and function of distinct neuronal populations.

Scalable platforms for data analysis

A comprehensive description of the transcriptional programs controlling neuronal specification represents a large dataset that requires new interactive tools to permit full exploration of the data by investigators. Here, we have established an interactive web platform that allows users to easily examine gene and isoform level expression data, explore clusters of related genes, discover new genes with user defined expression profiles, and links to displays of processed RNA-seq reads on the UCSC Genome Browser. Continually expanding databases of *in situ* hybridization and tissue level transcriptome data such as the Allen Brain Atlas (Lein et al., 2007) and the UCSC Genome Browser (Raney et al., 2013) are integrated into this dataset to facilitate more integrative analyses of these data.

The resolution of RNA-seq highlights the extensive regulation at the isoform level for individual neuronal subtypes, and these data are now available for investigators to examine

for individual genes or gene sets of interest. Beyond the expression of isoforms, our data provide the means to identify novel transcript variants not described here that can be assembled from this resource. Future expansion of this resource with additional transcriptomic and epigenetic data from these and other populations of cortical neurons will enable a more comprehensive understanding of the molecular programs driving corticogenesis.

Experimental Procedures

FACS-purification

All animals were handled according to protocols approved by the Institutional Animal Care and Use Committee (IACUC) of Harvard University. For each biological replicate, the somatosensory cortex of one litter of CD1 embryos or pups (six to ten per litter) was dissociated into single cell suspension as previously described (Arlotta et al., 2005), and cells were fixed immediately with 4% paraformaldehyde. The protocol for intracellular staining and RNA isolation was modified from (Hrvatin et al., 2014). Single cell suspension of fixed cells was immunostained under RNase free conditions with anti-SATB2, anti-CTIP2, and anti-TLE4. Appropriate gates for FACS were set based on relative levels of SATB2, CTIP2, and TLE4 expression to isolate CPN, ScPN, and CThPN as described in Figure 1. Additional information is in Supplemental Experimental Procedures.

Immunohistochemistry and in situ hybridization

Immunohistochemistry and in situ hybridization were preformed as previously described (Lodato et al., 2014; Molyneaux et al., 2005). In situ probes are detailed in Table S6. Additional information is in Supplemental Experimental Procedures.

RNA isolation

RNA was recovered from FACS-purified cells using the RecoverAll Total Nucleic Acid Isolation Kit (Ambion) according to manufacturer's instructions except proteinase K digestion was performed at 50°C for 3 hours. RNA concentration was quantified with Nanodrop 1000 and quality was determined with an Agilent 2100 Bioanalyzer. RNA Integrity Numbers (RIN) for all samples were between between 6.3 and 9. Typical yield was between 1 and 2 picograms RNA per sorted event, requiring approximately 100,000 cells to yield 200 ng RNA for library input.

RNA-seq, transcriptome assembly, and differential expression

Purified RNA served as input for the standard Illumina RNA-seq library preparation with poly(A) selection. 100 base pair-paired end reads were mapped to the mouse genome (mm9) using Tophat2 (Kim et al., 2013) with an average of 1.09×10^8 mapped reads (range 7.62×10^7-1.33×10^8; s.d. 1.30×10^7) and assembled into transcripts using Cufflinks (Trapnell et al., 2010). Individual assemblies were merged using Cuffmerge and UCSC coding genes as a reference. The merged assembly was used as input for our previously-described long non-coding RNA (lncRNA) identification pipeline (Cabili et al., 2011). Assembled lncRNAs were appended to the list of known UCSC protein-coding genes to establish a suitable reference transcriptome. Differential expression testing was performed between all pairs of

conditions using Cuffdiff2. Data was visualized using the cummeRbund package from Bioconductor (Gentleman et al., 2004; Trapnell et al., 2012b) and additionally integrated into the DeCoN interactive web resource. RNA-seq data is available in the NCBI Gene Expression Omnibus repository (Accession GSE63482). Additional information is in Supplemental Experimental Procedures.

PCA and Cluster Analysis

Principal component analysis was performed on individual gene-level fragments per kilobase of RNA per million reads mapped (FPKM) using the cummeRbund package in R on the lists of significant protein-coding and lncRNA genes. A 20-way clustering solution of differential gene expression profiles was obtained using nearest-neighbor agglomerative graph clustering on cosine similarities using the CLUTO utility [described in (Zhao and Karypis, 2005)]. Additional information is in Supplemental Experimental Procedures.

lncRNA and protein-coding gene specificity analysis

Maximum specificity scores were calculated for protein-coding genes and lncRNAs across all three cell types at each time point as described in (Cabili et al., 2011). To correct for the confounding influence of expression level, we 1) compared lncRNA specificity scores to resampled protein-coding genes drawn from the learned empirical distribution of lncRNA maximum FPKM values for each time point, and 2) employed a generalized additive model (GAM) using the ‘mgcv’ R-package (Wood, 2011) to appropriately model the observed non-linear relationship between expression level (FPKM) and specificity (S), and to identify any significant effects of gene biotype on specificity, and the observed increase in specificity at each time point. Additional information is in Supplemental Experimental Procedures.

Identification of syntenic positional equivalents

lncRNA human syntenic positional equivalents were defined by the presence of a transcribed element, within the same syntenic region in the human genome (hg19; TransMap). Additional information is in Supplemental Experimental Procedures.

Quantitative RT-PCR

Real-time PCR was performed using TaqMan Gene Expression Assays (Applied Biosystems) according to manufacturer’s protocol for standard cycling conditions. Additional information is in Supplemental Experimental Procedures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We are grateful to Barbara Tazon-Vega for assistance with RNA-seq library generation, Emanuela Zucaro for assistance with in situ hybridizations, David Wilkinson for assistance with website development, Stefano Stifani for sharing the TLE4 antibody, and Simona Lodato, Cole Trapnell, David Hendrickson, Douglas Melton, and members of the Rinn and Arlotta Lab for experimental support and discussions on the manuscript. LAG is a recipient of a NSF Postdoctoral Fellowship in Biology. SH is supported by the Sternlicht Director’s Fund Fellowship. JLR is supported by the NIH Directors New Innovator (DP2OD006670), P01 GM099117 and Center for Cell Circuits (P50

HG006193-01). PA is supported by NIH (NS062489, NS073124, NS078164), the New York Stem Cell Foundation, and the Harvard Stem Cell Institute and is a New York Stem Cell Foundation-Robertson Investigator.

References

Alcamo EA, Chirivella L, Dautzenberg M, Dobreva G, Fariñas I, Grosschedl R, McConnell SK. Satb2 regulates callosal projection neuron identity in the developing cerebral cortex. Neuron. 2008; 57:364–377. [PubMed: 18255030]

Allen TT, Lobe CGC. A comparison of Notch, Hes and Grg expression during murine embryonic and post-natal development. Cell Mol Biol Incl Cyto Enzymol. 1999; 45:687–708.

Allendoerfer KL, Shelton DL, Shooter EM, Shatz CJ. Nerve growth factor receptor immunoreactivity is transiently associated with the subplate neurons of the mammalian cerebral cortex. Proc Natl Acad Sci USA. 1990; 87:187–190. [PubMed: 2153287]

Arlotta P, Molyneaux BJ, Chen J, Inoue J, Kominami R, Macklis JD. Neuronal subtype-specific genes that control corticospinal motor neuron development in vivo. Neuron. 2005; 45:207–221. [PubMed: 15664173]

Ayoub AE, Oh S, Xie Y, Leng J, Cotney J, Dominguez MH, Noonan JP, Rakic P. Transcriptional programs in transient embryonic zones of the cerebral cortex defined by high-resolution mRNA sequencing. Proceedings of the National Academy of Sciences. 2011; 108:14950–14955.

Bassett AR, Akhtar A, Barlow DP, Bird AP, Brockdorff N, Duboule D, Ephrussi A, Ferguson-Smith AC, Gingeras TR, Haerty W, et al. Author response. eLife. 2014; 3

Belgard TG, Marques AC, Oliver PL, Abaan HO, Sirey TM, Hoerder-Suabedissen A, García-Moreno F, Molnár Z, Margulies EH, Ponting CP. A transcriptomic atlas of mouse neocortical layers. Neuron. 2011; 71:605–616. [PubMed: 21867878]

Bulchand S, Subramanian L, Tole S. Dynamic spatiotemporal expression of LIM genes and cofactors in the embryonic and postnatal cerebral cortex. Dev Dyn. 2003; 226:460–469. [PubMed: 12619132]

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011; 25:1915–1927. [PubMed: 21890647]

Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, Gardiner BB, Askarian-Amiri ME, Ru K, Solda G, Simons C, et al. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. Genome Research. 2008; 18:1433–1445. [PubMed: 18562676]

Doyle JP, Dougherty JD, Heiman M, Schmidt EF, Stevens TR, Ma G, Bupp S, Shrestha P, Shah RD, Doughty ML, et al. Application of a Translational Profiling Approach for the Comparative Analysis of CNS Cell Types. Cell. 2008; 135:749–762. [PubMed: 19013282]

Fertuzinhos S, Li M, Kawasawa YI, Ivic V, Franjic D, Singh D, Crair M, Sestan N. Laminar and Temporal Expression Dynamics of Coding and Noncoding RNAs in the Mouse Neocortex. CellReports. 2014:1–13.

Fishell G, Heintz N. Perspective. Neuron. 2013; 80:602–612. [PubMed: 24183013]

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biology. 2004; 5:R80. [PubMed: 15461798]

Grote P, Wittler L, Hendrix D, Koch F, Währisch S, Beisaw A, Macura K, Bläss G, Kellis M, Werber M, et al. Short Article. Dev Cell. 2013; 24:206–214. [PubMed: 23369715]

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L, et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature. 2011; 477:295–300. [PubMed: 21874018]

Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, van de Lagemaat LN, Smith KA, Ebbert A, Riley ZL, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature. 2013; 489:391–399. [PubMed: 22996553]

Heiman M, Schaefer A, Gong S, Peterson JD, Day M, Ramsey KE, Suárez-Fariñas M, Schwarz C, Stephan DA, Surmeier DJ, et al. A translational profiling approach for the molecular characterization of CNS cell types. Cell. 2008; 135:738–748. [PubMed: 19013281]

Hoerder-Suabedissen A, Molnar Z. Molecular Diversity of Early-Born Subplate Neurons. Cereb Cortex. 2013; 23:1473–1483. [PubMed: 22628460]

Hrvatin S, Deng F, O'Donnell CW, Gifford DK, Melton DA. MARIS: Method for Analyzing RNA following Intracellular Sorting. PLoS ONE. 2014; 9:e89459. [PubMed: 24594682]

Johnson MB, Kawasawa YI, Mason CE, Krsnik Z, Coppola G, Bogdanović D, Geschwind DH, Mane SM, State MW, Sestan N. Functional and evolutionary insights into human brain development through global transcriptome analysis. Neuron. 2009; 62:494–509. [PubMed: 19477152]

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AMM, Pletikos M, Meyer KA, Sedmak G, et al. Spatio-temporal transcriptome of the human brain. Nature. 2011; 478:483–489. [PubMed: 22031440]

Kawaguchi Y, Kubota Y. GABAergic cell subtypes and their synaptic connections in rat frontal cortex. Cereb Cortex. 1997; 7:476–486. [PubMed: 9276173]

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes inthe presence of insertions, deletions and genefusions. Genome Biology. 2013; 14:R36. [PubMed: 23618408]

Klebe S, Lossos A, Azzedine H, Mundwiller E, Sheffer R, Gaussen M, Marelli C, Nawara M, Carpentier W, Meyer V, et al. KIF1A missense mutations in SPG30, an autosomal recessive spastic paraplegia: distinct phenotypes according to the nature of the mutations. European Journal of Human Genetics. 2012; 20:645–649. [PubMed: 22258533]

Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature. 2007; 445:168–176. [PubMed: 17151600]

Lobo MK, Karsten SL, Gray M, Geschwind DH, Yang XW. FACS-array profiling of striatal projection neuron subtypes in juvenile and adult mouse brains. Nat Neurosci. 2006; 9:443–452. [PubMed: 16491081]

Lodato S, Molyneaux BJ, Zuccaro E, Goff LA, Chen H-H, Yuan W, Meleski A, Takahashi E, Mahony S, Rinn JL, et al. Gene co-regulation by Fezf2 selects neurotransmitter identity and connectivity of corticospinal neurons. Nat Neurosci. 2014:1–35. [PubMed: 24369367]

Lubas M, Christensen MS, Kristiansen MS, Domanski M, Falkenby LG, Lykke-Andersen S, Andersen JS, Dziembowski A, Jensen TH. Interaction Profiling Identifiesthe Human Nuclear Exosome Targeting Complex. Molecular Cell. 2011; 43:624–637. [PubMed: 21855801]

Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic regulation. Nature Publishing Group. 2013; 20:300–307.

Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev Genet. 2009; 10:155–159. [PubMed: 19188922]

Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS. Specific expression of long noncoding RNAs in the mouse brain. Proc Natl Acad Sci USA. 2008; 105:716–721. [PubMed: 18184812]

Molyneaux BJ, Arlotta P, Fame RM, MacDonald JL, MacQuarrie KL, Macklis JD. Novel subtype-specific genes identify distinct subpopulations of callosal projection neurons. J Neurosci. 2009; 29:12343–12354. [PubMed: 19793993]

Molyneaux BJ, Arlotta P, Hirata T, Hibi M, Macklis JD. Fezl is required for the birth and specification of corticospinal motor neurons. Neuron. 2005; 47:817–831. [PubMed: 16157277]

Molyneaux BJ, Arlotta P, Menezes JRL, Macklis JD. Neuronal subtype specification in the cerebral cortex. Nat Rev Neurosci. 2007; 8:427–437. [PubMed: 17514196]

Nieto M, Monuki ES, Tang H, Imitola J, Haubst N, Khoury SJ, Cunningham J, Gotz M, Walsh CA. Expression of Cux-1 and Cux-2 in the subventricular zone and upper layers II-IV of the cerebral cortex. J Comp Neurol. 2004; 479:168–180. [PubMed: 15452856]

Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Geschwind DH. Functional organization of the transcriptome in human brain. Nat Neurosci. 2008; 11:1271–1282. [PubMed: 18849986]

Ponjavic J, Oliver PL, Lunter G, Ponting CP. Genomic and Transcriptional Co-Localization of Protein-Coding and Long Non-Coding RNA Pairs in the Developing Brain. PLoS Genet. 2009; 5:e1000617. [PubMed: 19696892]

Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. Cell. 2009; 136:629–641. [PubMed: 19239885]

Ramos AD, Diaz A, Nellore A, Delgado RN, Park K-Y, Gonzales-Roybal G, Oldham MC, Song JS, Lim DA. Integration of Genome-wide Approaches Identifies lncRNAs of Adult Neural Stem Cells and Their Progeny In Vivo. Stem Cell. 2013; 12:616–628.

Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D, et al. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics. 2013

Sauvageau M, Goff LA, Lodato S, Bonev B, Groff AF, Gerhardinger C, Sanchez-Gomez DB, Hacisuleyman E, LI E, Spence M, et al. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. eLife. 2013; 2:e01749–e01749. [PubMed: 24381249]

Saxena S, Caroni P. Selective Neuronal Vulnerabilityin Neurodegenerative Diseases:from Stressor Thresholds to Degeneration. Neuron. 2011; 71:35–48. [PubMed: 21745636]

Sharman JL, Benson HE, Pawson AJ, Lukito V, Mpamhanga CP, Bombail V, Davenport AP, Peters JA, Spedding M, Harmar AJ, et al. IUPHAR-DB: updated database content and new features. Nucleic Acids Research. 2012; 41:D1083–D1088. [PubMed: 23087376]

Siddiqui TJ, Tari PK, Connor SA, Zhang P, Dobie FA, She K, Kawabe H, Wang YT, Brose N, Craig AM. An LRRTM4-HSPG Complex Mediates Excitatory Synapse Developmenton Dentate Gyrus Granule Cells. Neuron. 2013; 79:680–695. [PubMed: 23911104]

Sorensen SA, Bernard A, Menon V, Royall JJ, Glattfelder KJ, Desta T, Hirokawa K, Mortrud M, Miller JA, Zeng H, et al. Correlated Gene Expression and Target Specificity Demonstrate Excitatory Projection Neuron Diversity. Cereb Cortex. 2013

Steijger T, Abril JF, Engström PG, Kokocinski F, Abril JF, Akerman M, Alioto T, Ambrosini G, Antonarakis SE, Behr J, et al. Assessment of transcript reconstruction methods for RNA-seq. Nat Meth. 2013; 10:1177–1184.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005; 102:15545–15550. [PubMed: 16199517]

Sugino K, Hempel CM, Miller MN, Hattox AM, Shapiro P, Wu C, Huang ZJ, Nelson SB. Molecular taxonomy of major neuronal classes in the adult mouse forebrain. Nat Neurosci. 2005; 9:99–107. [PubMed: 16369481]

Sun L, Goff LA, Trapnell C, Alexander R, Lo KA, Hacisuleyman E, Sauvageau M, Tazon-Vega B, Kelley DR, Hendrickson DG, et al. Long noncoding RNAs regulate adipogenesis. Proc Natl Acad Sci USA. 2013; 110:3387–3392. [PubMed: 23401553]

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with rNA-seq. Nat Biotechnol. 2012a; 31:46–53. [PubMed: 23222703]

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols. 2012b; 7:562–578.

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. [PubMed: 20436464]

Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2011; 73:3–36.

Zhang Y, Chen K, Sloan SA, Bennett ML, Scholze AR, O'Keeffe S, Phatnani HP, Guarnieri P, Caneda C, Ruderisch N, et al. An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. J Neurosci. 2014; 34:11929–11947. [PubMed: 25186741]

Zhao Y, Karypis G. Hierarchical Clustering Algorithms for Document Datasets. Data Mining and Knowledge Discovery. 2005; 10:141–168.

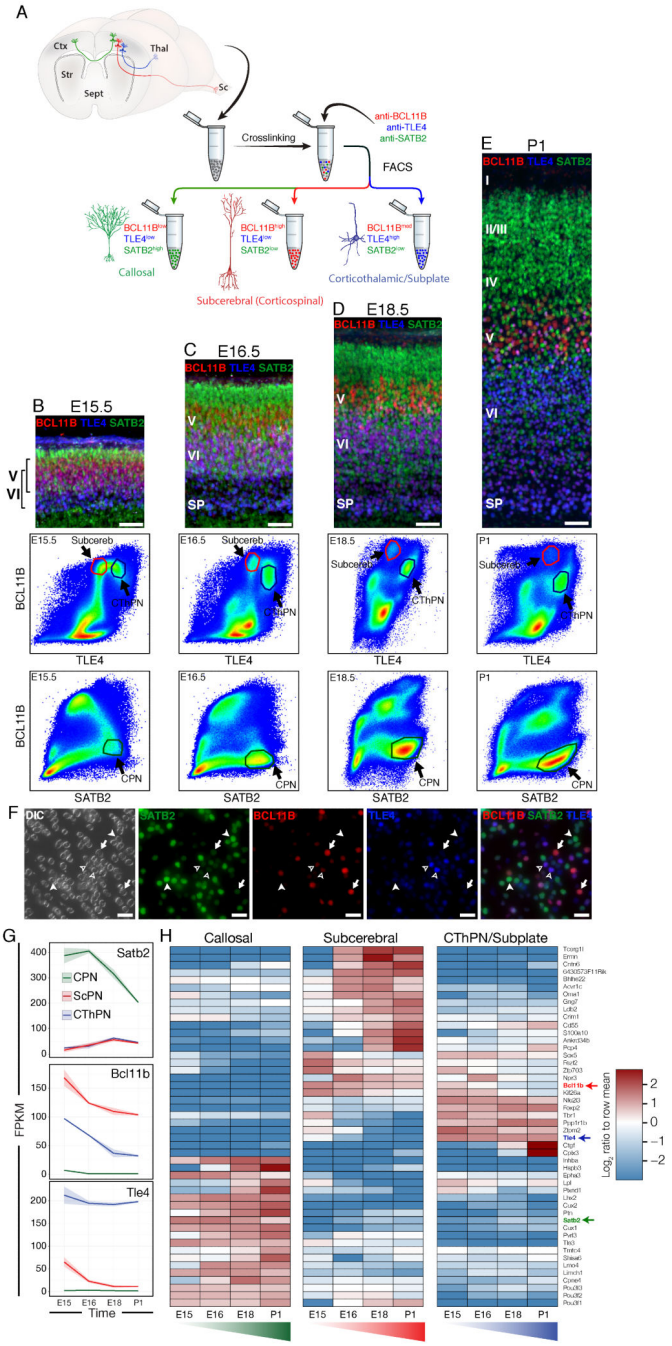

Figure 1.

(A) Schematic overview of the purification of three distinct neuronal subtypes for RNA-seq: callosal projection neurons (CPN; green), subcerebral projection neurons (ScPN; red), and corticothalamic/subplate neurons (CThPN; blue) (B-E) Immunofluorescence labeling of coronal sections of E15.5, E16.5, E18.5 and P1 mouse neocortex with antibodies to BCL11B, TLE4, and SATB2 and corresponding FACS plots from dissociated cortex highlighting the selection process to identify each cell type of interest. (F) Dissociated E18.5 cells prior to FACS with labeled CPN (arrowheads), ScPN (arrows), CThPN (open

arrowheads). (G) Gene-level RNA-seq expression profiles for Bcl11b, Satb2, and Tle4 confirm expression in specific cellular populations. Lines represent Cuffdiff2 expression estimates; shaded areas represent 95% confidence intervals. (H) A heat map of row-mean-centered gene-level expression patterns for known subtype specific genes confirms the specific identities of the three isolated neuronal populations. Scale bars: (B-E) 50 μm, (F) 20 μm. See also Figure S1.

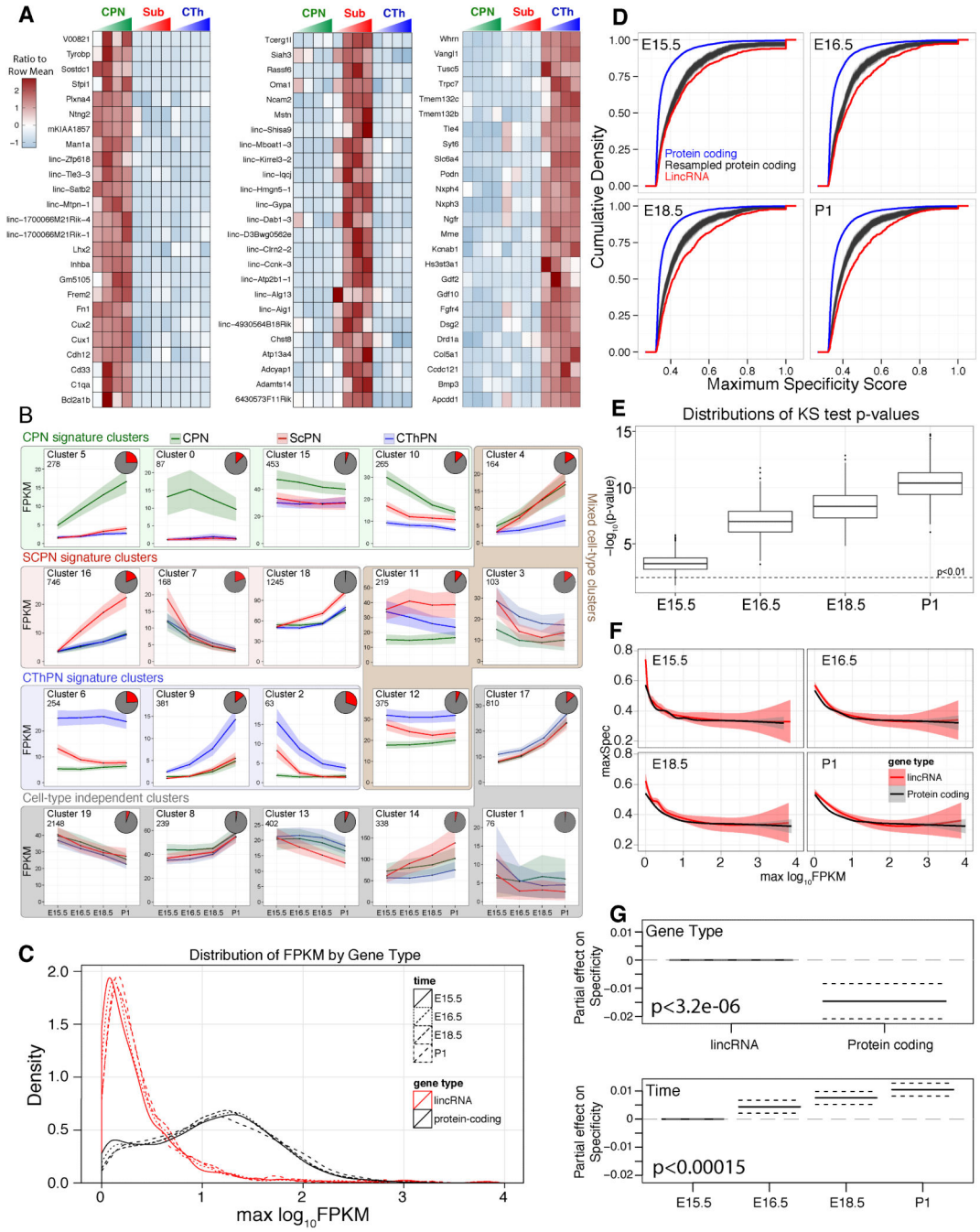

Figure 2.

Comprehensive transcriptional analysis of neuronal cell type specificity.

(A) Heat maps of the row-mean-centered expression profiles for the 25 most specific genes for each neuronal subtype. (B) A 20-way clustering solution of differential gene expression profiles. Clusters are manually grouped by cell type specificity. Inset pie charts indicate proportion of genes in each cluster that are lncRNAs (red) or protein-coding genes (gray). Shaded areas represent 95% confidence intervals. (C) Density plots of gene expression estimates (FPKM) for lncRNAs (red) and protein-coding genes (black). (D) Cumulative

density of maximum specificity scores across each condition for protein-coding genes (blue), resampled protein-coding genes drawn from a learned empirical distribution of lncRNA maximum FPKM values (black; 1000 samples individually plotted at each time point), and lncRNAs (red). (E) Distributions of KS-test p-values for resampled protein-coding gene max specificity scores versus lncRNA max specificity scores at each time point. (F) Smoothed spline illustrating the fitted inverse relationship between expression level (FPKM) and max cell-type specificity score at each time point for each gene type. (G) Significant effects on gene specificity scores attributed to the explanatory variables gene type or time after fitting a generalized additive model between specificity and expression. See also Figures S2-S3.

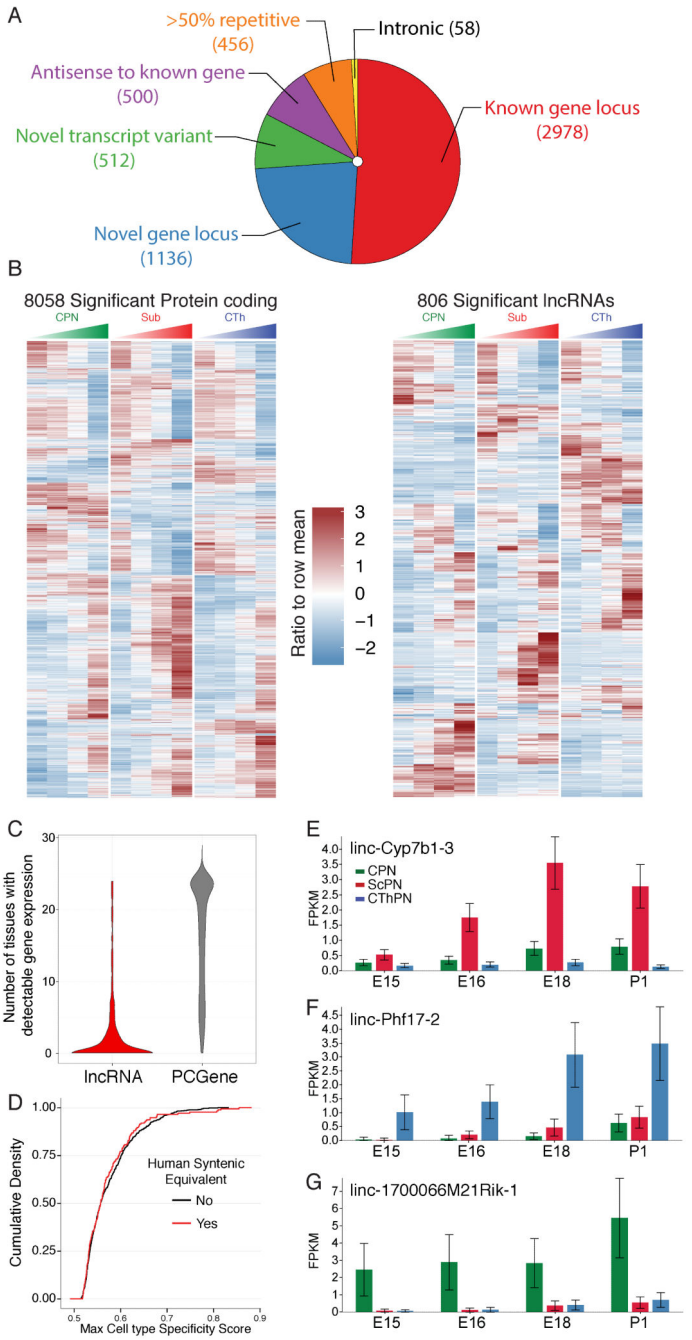

Figure 3.

Cell type specific lncRNAs in the developing neocortex

(A) Pie chart detailing the distribution of 5195 lncRNAs identified in the pyramidal neuron transcriptome. (B) Heat map of differentially expressed protein-coding genes and lncRNAs. (C) The number of tissues with detectable expression for a given significant lncRNA (FPKM >= 2) is significantly lower (mean 2.7; median 0) than the number for significant protein-coding genes (mean 16.28; median 19). (D) Cumulative densities of maximum cell-type specificity scores for lncRNAs with a human syntenic equivalent (red) and those

without (black) suggests that the presence of a lncRNA across species does not correlate with a change in cell type specificity. (E-G) Barplots of estimated expression levels for linc-Cyp7b1-3, linc-Phf17-2, and linc-1700066M21Rik-1; lncRNAs that exhibit high degree of cell type specificity for ScPN, CThPN, and CPN, respectively. Error bars represent 95% confidence intervals in expression estimates.

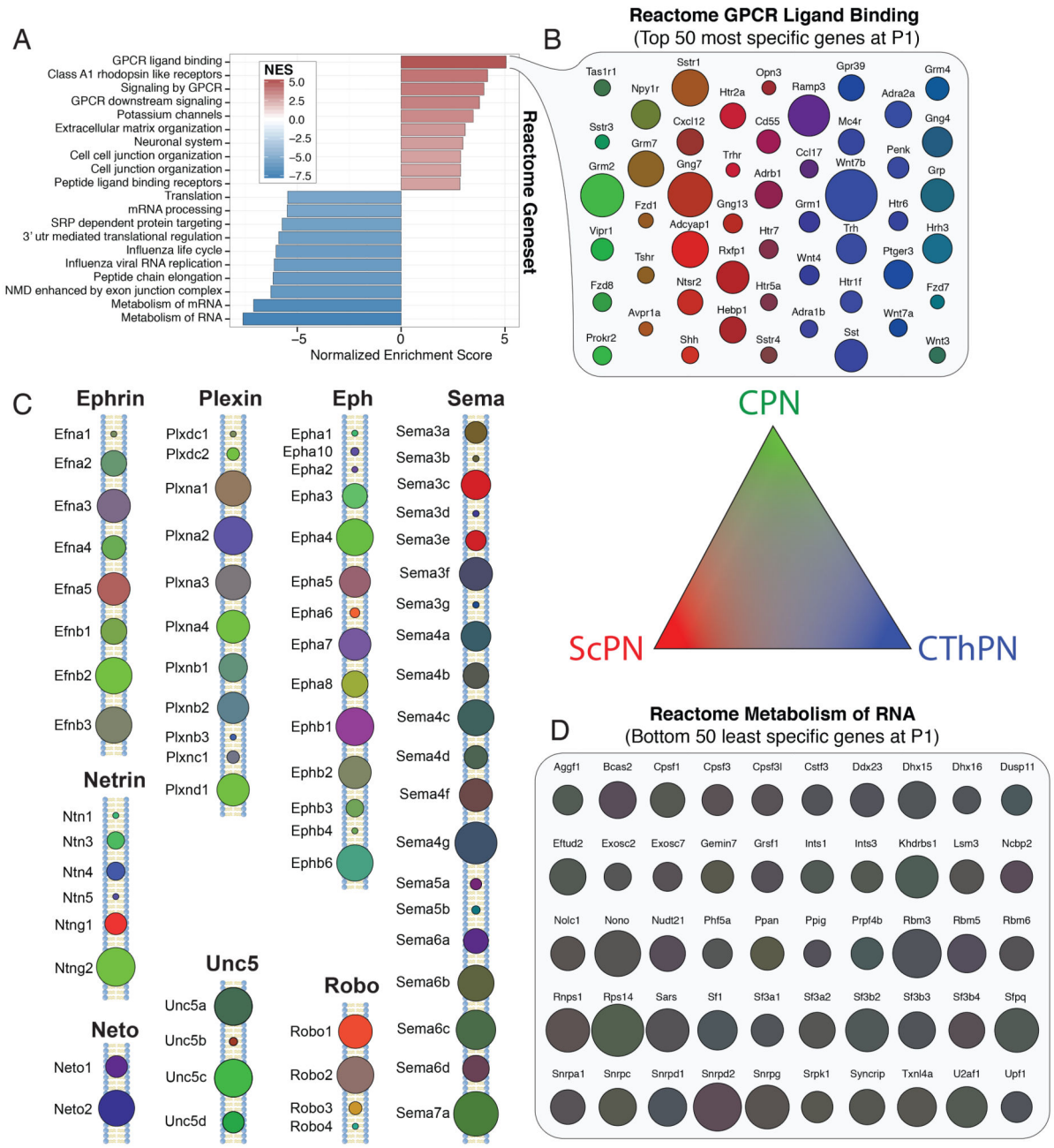

Figure 4.

GPCR and cell surface molecules are the leading indicators of neuronal diversity.

(A) Significantly ($p<0.001$) enriched and depleted gene sets from a pre-ranked Gene Set Enrichment Analysis against the Reactome collection of gene sets. (B) Dot plots of the 50 most specific genes at P1 within the GPCR ligand binding Reactome gene set. Diameter of the dots are mapped to expression estimates at P1 and color mapped to relative cell type specificity. (C) Dot plots for specific classes of axon guidance molecules reveals that individual neuronal subclasses use different codes of related molecules to inform axonal targeting decisions. (D) Dot plot of genes within the lowest-ranked Reactome gene set for

contrast. Genes within this set and other basal metabolic processes show little variation between cell types.

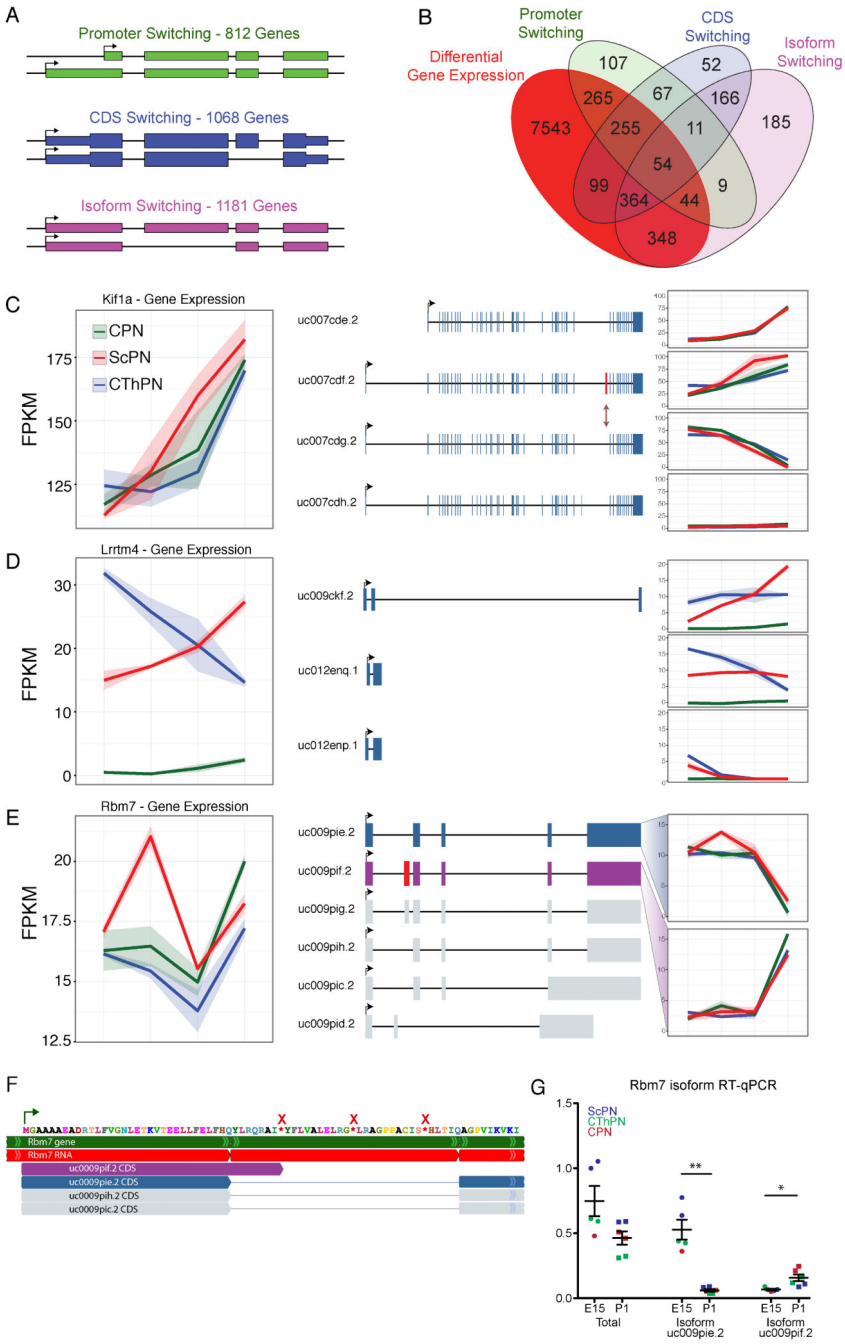

Figure 5.

Regulated programs of gene expression at the isoform level.

(A) Schematic depicting the number of genes that undergo promoter switching, alter their protein-coding sequence, or demonstrate shifts in isoform abundance. (B) Euler diagram describing the total number of significant regulatory events identified including those that take place without detectable changes in overall gene expression (red oval). (C) Kif1a gene level expression increases over time while individual isoforms undergo a dramatic shift in expression from isoform uc007cdg.2 to isoform uc007cdf.2. (D) Diametrically opposing

changes in expression are observed for two Lrrtm4 transcript variants in ScPN and CThPN. (E) Insignificant gene level expression estimates for Rbm7 belie a significant shift in expression at the isoform level. (F) This results in a switch to a significantly truncated peptide as a result of the inclusion of three in-frame stop codons for the uc009pif.2 isoform. (G) RT-qPCR confirms RNA-seq detected expression dynamics from E15 to P1 between all isoforms, uc009pie.2 ($p = 0.0035$), and uc009pif.2 ($p = 0.016$). * $p < 0.05$, ** $p < 0.01$. Shaded areas in C-E represent 95% confidence intervals.