nature neuroscience

Developmental regulation of human cortex transcription and its clinical relevance at single base resolution

Andrew E Jaffe[1–3], Jooheon Shin[1], Leonardo Collado-Torres[1,2], Jeffrey T Leek[2,4], Ran Tao[1], Chao Li[1], Yuan Gao[1], Yankai Jia[1], Brady J Maher[1,5,6], Thomas M Hyde[1,5–8], Joel E Kleinman[1,9] & Daniel R Weinberger[1,4–7,9]

Transcriptome analysis of human brain provides fundamental insight into development and disease, but it largely relies on existing annotation. We sequenced transcriptomes of 72 prefrontal cortex samples across six life stages and identified 50,650 differentially expression regions (DERs) associated with developmental and aging, agnostic of annotation. While many DERs annotated to non-exonic sequence (41.1%), most were similarly regulated in cytosolic mRNA extracted from independent samples. The DERs were developmentally conserved across 16 brain regions and in the developing mouse cortex, and were expressed in diverse cell and tissue types. The DERs were further enriched for active chromatin marks and clinical risk for neurodevelopmental disorders such as schizophrenia. Lastly, we demonstrate quantitatively that these DERs associate with a changing neuronal phenotype related to differentiation and maturation. These data show conserved molecular signatures of transcriptional dynamics across brain development, have potential clinical relevance and highlight the incomplete annotation of the human brain transcriptome.

The transcriptome of the human brain changes markedly across development and aging, with the largest gene expression changes occurring during fetal life, tapering into infancy[1,2]. Developmental brain disorders often involve genes that are differentially expressed in fetal as compared with postnatal life[3,4]. While exploration of the brain transcriptome has been an important approach to understanding brain development and brain disease, previous transcriptome characterizations have used primarily microarray technologies based on probe sequences that capture only a limited proportion of transcriptome diversity. Technological advances in RNA sequencing (RNA-seq) now permit a flexible and potentially unbiased characterization of the transcriptome at high resolution and coverage[5]. Yet existing published RNA-seq-based characterizations of brain development have used gene- and/or exon-level count-based summarizations[4,6,7], which require an accurate and complete gene annotation. Such feature-based read counts lack the ability to reliably identify new transcriptional activity, but they generally limit the inherent difficulty in transcript assembly and characterization based on short-read sequencing technologies[8].

We have implemented a method for RNA-seq analysis at single base resolution to more fully characterize transcription dynamics, which exploits the benefits of both count- and transcript-based methods. We describe herein the results of deep coverage sequencing of the poly(A)$^+$ transcriptomes of human dorsolateral prefrontal cortex (DLPFC) samples across six important life stages: fetal (second trimester), infant, child, teen, adult and late life. We implemented an annotation-agnostic differential expression analysis to exploit the power of RNA-seq without the difficulties of transcript assembly[9]. This method, called derfinder, identifies differential expression at base-pair resolution and forms differentially expressed regions (DERs) by joining adjacent differentially expressed bases. We tested for differences in average expression across the six age groups and used statistical permutation to calculate a measure of genome-wide significance for each DER[10]. A DER represents a differentially expressed (here, across age groups) unspliced segment of RNA that can originate from a full-length or, potentially, spliced transcript. The derfinder approach therefore interrogates transcript-level changes in gene expression via differentially expressed segments using only coverage-level RNA-seq data. This approach allows an unconstrained and unbiased search of the transcriptome to identify fragments of interest for more detailed molecular characterization of corresponding full-length transcripts.

After applying this approach to a discovery data set of 36 brain samples, we carried forward DERs that had significant differential expression in a replication data set of 36 more DLPFC samples. Significant and replicated DERs were mapped onto existing reference transcriptomes in databases such as Ensembl[11], UCSC[12] and Gencode[13] to characterize their locations in the genome. We further related the expression levels within DERs to a wide range of publicly available resources, including RNA-seq data from 16 human brain

[1]Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, Maryland, USA. [2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA. [3]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA. [4]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [5]Department of Psychiatry, Johns Hopkins School of Medicine, Baltimore, Maryland, USA. [6]Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, Maryland, USA. [7]Department of Neurology, Johns Hopkins School of Medicine, Baltimore, Maryland, USA. [8]Department of Biological Sciences, Johns Hopkins School of Medicine, Baltimore, Maryland, USA. [9]These authors contributed equally to this work. Correspondence should be addressed to A.E.J. (andrew.jaffe@libd.org) or D.R.W. (drweinberger@libd.org).

Received 27 August; accepted 14 November; published online 15 December 2014; doi:10.1038/nn.3898

regions[14], the developing mouse cortex[15], and a variety of other cell[16] and tissue[17] types to understand these patterns in a broader context (**Fig. 1**). Lastly, we identify significant enrichment for functional epigenomic marks associated with gene expression and for disease-associated genetic loci from recent GWAS. The results highlight conserved signatures of gene expression across development and aging in the human brain, including many non-exonic sequences that appear to be mature mRNAs, and identify biological fingerprints of age-associated changes in neuronal phenotypes and CNS disorder–associated genes.

RESULTS

Extensive transcriptional changes across brain development

We identified 50,650 DERs associated with development and aging that were both genome-wide significant in our discovery data set (at family-wise error rate ≤ 5%) and were also differentially expressed in a second independent sample of 36 human brains distributed across the same age ranges (at $P < 0.05$; see Online Methods and **Supplementary Table 1**). These DERs represent 8.63 megabases (Mb) of expressed sequence (**Supplementary Table 2**), annotated to 5,985 unique RefSeq genes (and 6,549 unique Ensembl) genes. There were, on average, 7.51 DERs annotated to each RefSeq gene (median = 4; interquartile range, 2–10). Only 1,454 genes contained a single DER (24.3%).

The RefSeq genes containing DERs were strongly enriched for many general developmental and metabolic processes, including organelle organization (GO:0006996; 976 of 2,368 genes, $P = 7.13 \times 10^{-29}$), regulation of gene expression (GO:0010468; 1,314 of 3,442 genes, $P = 8.62 \times 10^{-23}$) and regulation of transcription, DNA-dependent (GO:0006355; 1,127 of 2,916 genes, $P = 3.78 \times 10^{-21}$) (**Supplementary Table 3a**). A more focused gene ontology analysis using the 1,000 most significant DERs revealed more specific enrichment for neuron projection morphogenesis (GO:0048812; 49 of 575 genes, $P = 4.98 \times 10^{-11}$), neuron development (GO:0048666; 61 of 838 genes, $P = 1.29 \times 10^{-10}$), axonogenesis (GO:0007409; 43 of 509 genes, $P = 1.08 \times 10^{-9}$) and nervous system development (GO:0007399; 100 of 1,784 genes, $P = 3.84 \times 10^{-10}$) (**Supplementary Table 3b**).

Most DERs had their highest expression (adjusted for sequencing depth) in the fetal developmental period ($N = 41{,}405$; 81.7%), followed by adolescent ($N = 3{,}104$; 6.1%) and adult ($N = 2{,}621$; 5.2%). The genes containing DERs most highly expressed from infancy through adulthood were consistently enriched for synaptic transmission (GO:0007268; P value range 5.0×10^{-12}–5.5×10^{-24}), cell-cell signaling (GO:0007267; P value range 4.0×10^{-7}–1.7×10^{-17}) and other related signaling processes (**Supplementary Table 3d–g**). Notably, genes containing DERs most expressed in later life (age ≥50 years) were not enriched for these signaling processes, but instead were enriched for processes related to cellular respiration and energy-related processes (**Supplementary Table 3h**).

Principal component analysis (PCA) of the normalized coverage estimates across the 50,650 DERs revealed that the first principal component represented a linear scaling (either positive or negative) of expression across the lifespan (72% of variance explained; **Supplementary Fig. 1a**). The second and third principal components explained less variance (combined 15.1%) and represent dynamic expression from infanthood to adolescence with relatively similar levels of expression in fetal life and adulthood (**Supplementary Fig. 1b**,**c**). However, almost all DERs had much higher correlation to the first principal component (49,698; 98.1%) than the second or third principal components (605 and 346, representing 1.2% and 0.7%, respectively), suggesting that most DERs represent 'scaling' of gene expression—that is, unidirectional change—across the lifespan.

Several of the genes containing the most significant DERs showed patterns consistent with the canonical biology of brain development (**Supplementary Fig. 2**). These included the high expression of previously identified developmentally important genes during fetal life, such as *SOX11* (**Fig. 1**), which encodes a transcription factor involved in the regulation of embryonic development[18], and *DCX*, which is involved in the migration and organization of neuroblasts[19]. Expression of *SLC6A1* (*GAT1*), a sodium- and chloride-dependent GABA transporter that removes GABA from the synaptic cleft, followed the well-studied early developmental expression of the GABAergic system[20]. DERs overlapping *NRGN* and *CAMK2A*, two calcium binding proteins important for learning and memory and neuropsychiatric disorders[21,22], became most highly expressed in infant and teenage life periods, respectively. Several DERs that had their highest expression during postnatal life have been implicated in brain disorders thought to be developmental, including *RGS4*, a G protein signaling regulator associated with schizophrenia[23] that had its highest expression during adolescence, and *CNTNAP1*, a contactin-associated protein associated with autism[24] that had its highest expression during adulthood.

Many of the genes associated with DERs also showed developmental regulation across the lifespan using previously published microarray

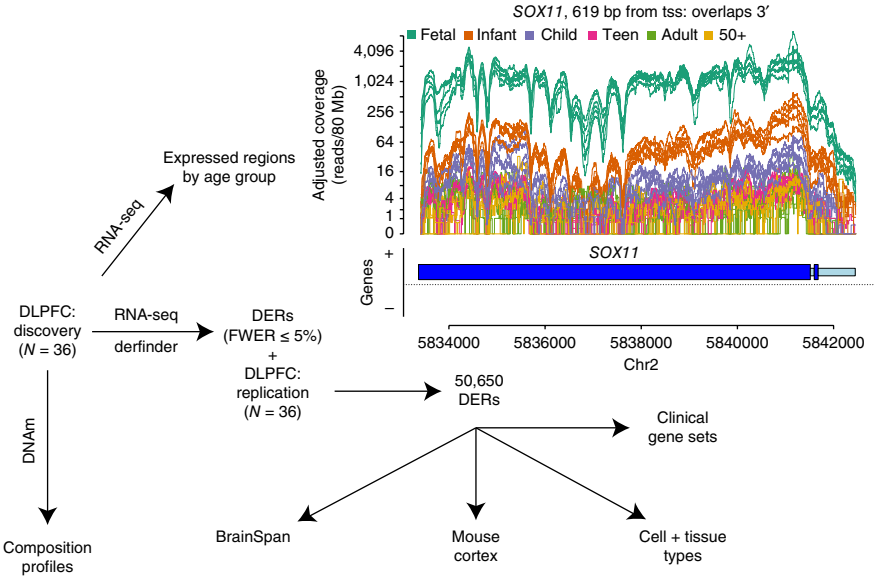

Figure 1 Design of the project. We performed RNA sequencing (RNA-seq) on 36 DLPFC samples from across the lifespan and implemented the derfinder method to identify DERs. These DERs were replicated in an independent DLPFC sample and explored across other brain regions, in the developing mouse cortex, in diverse cell and tissue types, and in the context of disease-associated gene sets. Top right, an example of a DER (see **Supplementary Table 2** for links to visualizations all DERs). We also quantified the cell composition of these DLPFC samples and defined regions of expression across the genome by age group. FWER, family-wise error rate; DNAm, DNA methylation; tss, transcriptional start site.

data on 269 individuals without psychiatric disorders[1] (obtained from GSE30272; see Online Methods), which both confirms the developmentally regulated genes identified with the DERs and highlights the gains made by using sequencing-based approaches over microarrays. Many individuals in the present RNA-seq study discovery data set (28 of 36) were interrogated in this array-based data set. Most (4,955 of 5,985, or 82.8%) of the DER-associated genes were present in the processed microarray data, and almost all of these genes were differentially expressed across the lifespan: 4,920 (99.3%), 4,684 (94.5%) and 4,304 (86.9%) were significant at $P < 0.05$, $P < 10^{-6}$ and $P < 10^{-11}$, respectively. Of the 1,030 genes showing significant differential expression only in the RNA-seq data, 432 genes were removed during quality control steps performed by Colantuoni *et al.*[1], suggesting that they may be more difficult to measure using oligonucleotide probes, and the remaining 598 were not included in the microarray design. These genes did not differ in functionality from those included on the microarray (all GO enrichment P values > 10^{-6}).

Widespread differential expression of unannotated sequence

Surprisingly, many of the age-associated DERs, while contained within genes, contained expressed sequence annotated as intronic: 21,033 significant regions (41.5%) overlapped at least one Ensembl-annotated intron (minimum overlap = 20 base pairs; see Online Methods). Furthermore, 4,214 regions (8.3%), which we term "intergenic," did not map to any Ensembl annotated genes (that is, exonic or intronic regions); 29,813 regions (58.9%) crossed at least one annotated exon (**Supplementary Fig. 3**). Not surprisingly, the exonic DERs had, on average, much higher expression across all samples than DERs annotating to non-exonic sequence (140.8 normalized reads as compared to 14.0 and 8.2 normalized reads for intergenic and intronic DERs, respectively; $P < 10^{-100}$ via linear regression) and were longer (190.3 bp versus 150.4 and 139.4 bp, respectively, $P < 10^{-20}$). Nevertheless, of the 3,056 Ensembl genes containing intron-annotated DERs, 1,765 (57.7%) genes contained both intronic and exonic DERs. These intronic changes are not likely to be due to technical artifacts, and we observed significant enrichment ($P < 10^{-100}$) of long non-coding RNAs in the intergenic DERs (Online Methods). There were similar percentages of overlapping annotated features using the UCSC hg19 knownGene (based on RefSeq) database (19,575, 6,676 and 26,886 for introns, intergenic and exons, respectively) and Gencode v19 (21,107, 3,994 and 30,016, respectively), further suggesting that the transcriptome contained in commonly accessed databases is notably incomplete, at least across human brain development.

The widespread differential expression across development and age of previously annotated intronic sequence may be due to an abundance of nuclear pre-mRNA present in the total RNA. We therefore sought to better distinguish pre-mRNA from spliced exonic mRNA by sequencing nuclear and cytosolic preparations from another six independent brain samples (three fetal and three adult; **Supplementary Table 4**). Quantifying the relative concentration of mRNA in the cytosolic and nuclear mRNA fractions provided initial evidence that our differentially expressed regions were present in the cytosol: the mean concentration ratios of cytosolic to nuclear RNA were 204.0 ng/μl:17.6 ng/μl (11.6×) in the fetal samples and 137.0 ng/μl:17.6 ng/μl (7.7×) in the adult samples, showing that most polyadenylated RNA in total polyadenylated RNA originates from the cytosol. We sequenced each mRNA fraction from each sample to characterize the widespread differential expression observed in the total RNA. The relative $\log_2$ fold changes of expression, comparing fetal to adult levels, were highly correlated across total and cytosolic poly(A)$^+$ mRNA DERs ($\rho = 0.914$), including expression of annotated intronic ($\rho = 0.664$) and intergenic ($\rho = 0.820$) regions (**Supplementary Fig. 4**). There was especially high concordance in the directionality of the non-exonic fetal versus adult fold changes: 96.4% were directionally consistent overall between cytosolic and total poly(A)$^+$ mRNA. These results implicate developmental regulation of a potentially large subset of intron-containing mRNA in the cytosolic fraction of the human frontal cortex.

Age-associated DERs lack regional specificity

We next explored the representation of our age-associated DERs in other brain regions, including other cortical and subcortical nuclei and cerebellum, using publicly available BrainSpan data[14], which included RNA-seq data across prenatal and postnatal developmental periods in 16 brain regions. Our DLPFC-identified DERs showed consistent age-related changes across each brain region with little inter-regional variability. The first principal component of only the BrainSpan normalized mean coverage data across the 50,650 DERs (explaining 59% of the variability) strongly correlated with age, particularly fetal versus postnatal, and not brain region (**Fig. 2**). The second principal component (explaining 8.7% of the variability) strongly correlated with RNA quality (**Supplementary Fig. 5**). Subsequent lesser principal components differentiated the neocortical regions from the subcortical region and cerebellum (**Supplementary Fig. 6**). Within a secondary PCA on only non-exonic DERs, the first principal component remained age (here explaining 40.6% of the variance; **Supplementary Fig. 7**). There was also significant correlation between $\log_2$ fold changes comparing fetal samples to adults in our DLPFC data set and the same fetal versus postnatal comparison within each brain region, including within previously annotated intronic and intergenic sequences (**Table 1**). The high correlations between fetal versus adult comparisons in our DLPFC samples and the BrainSpan DLPFC samples constitute an independent validation of our identified DERs, including the non-exonic sequences.

Age-associated DERs are conserved in the mouse cortex

We further examined our DERs, particularly the preponderance of non-exonic expression, by exploiting genetic synteny in mice to validate differential expression using a cross-species approach. We downloaded and renormalized publicly available data from mouse cerebral cortex, comparing embryonic day (E) 17 ($N = 4$) to adult ($N = 3$) C57BL/6 mice[15], which had previously been interrogated for differences in gene-level expression across development. We used the liftOver tool[12] to map the DERs to the mouse genome (mm10), of

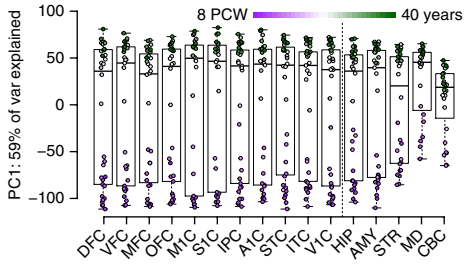

Figure 2 Age-associated differentially expressed region (DER) expression patterns across multiple brain regions. PCA was performed on normalized coverage estimates across all DERs using all BrainSpan samples. Each point is a sample colored by age: purple, prenatal; green, postnatal; white, birth. PC, principal component; var, variance; PCW, postconception weeks. The dashed vertical line separates the brain regions into neocortical and non-neocortical. Abbreviations as in **Table 1**. For each box plot, center bold line indicates median, box limits indicate the 25th and 75th percentiles, and the whiskers extend to 1.5 times the box limit percentiles.

Table 1 Correlation of fetal versus adult fold changes across brain regions within DERs

BrainSpan region	All (N = 50,560)	Intragenic (N = 4,221)	Intronic (N = 16,616)	Exonic (N = 29,813)
DFC	0.863	0.702	0.490	0.895
VFC	0.851	0.684	0.429	0.888
MFC	0.858	0.705	0.485	0.891
OFC	0.845	0.674	0.360	0.891
M1C	0.841	0.675	0.388	0.882
S1C	0.830	0.657	0.326	0.878
IPC	0.849	0.681	0.464	0.882
A1C	0.860	0.698	0.517	0.888
STC	0.871	0.720	0.576	0.894
ITC	0.852	0.694	0.473	0.881
V1C	0.867	0.701	0.534	0.894
HIP	0.828	0.660	0.397	0.862
AMY	0.845	0.677	0.444	0.872
STR	0.788	0.607	0.428	0.816
MD	0.699	0.528	0.266	0.731
CBC	0.627	0.434	0.230	0.673

Spearman correlation coefficients were calculated between $\log_2$ fold changes comparing fetal to postnatal expression in the DLPFC discovery data set and each brain region in the BrainSpan database across the DERs (All), and within the DERs annotated to specific Ensembl features. DFC, dorsolateral prefrontal cortex; VFC, ventrolateral prefrontal cortex; MFC, anterior (rostral) cingulate (medial frontal cortex); OFC, orbital frontal cortex; M1C, primary motor cortex (M1, Brodmann area 4); S1C, primary somatosensory cortex (S1, areas 3, 1 and 2); IPC, posteroinferior (ventral) parietal cortex; A1C, primary auditory cortex (core); STC, posterior (caudal) superior temporal cortex (Tac); ITC, inferolateral temporal cortex (Tev, area 20); V1C, primary visual cortex (striate cortex, V1, area 17); HIP, hippocampus (hippocampal formation); AMY, amygdaloid complex; STR, striatum; MD, mediodorsal nucleus of thalamus; CBC, cerebellar cortex.

which 37,428 mapped (73.9%, average synteny = 88.7%) and 25,372 had an average coverage >5 reads in at least one sample (22,195, 423 and 2,764 in human-annotated exonic, intergenic and intronic sequence, respectively), suggesting that a subset of these DERs are expressed in the developing mouse cortex. We identified significant correlation between the relative differences in fetal and adult human expression compared to E17 versus adult mouse expression in these syntenic regions (**Fig. 3**, $\rho = 0.771$, $P < 10^{-100}$ via Z-score). The magnitude and directionality of the expression changes in the mouse were similar to many of the human DERs (directionality concordance = 84.1% overall), including those annotated as intronic and intergenic, suggesting these age-associated DERs represent conserved expression signatures in the developing mammalian brain.

Age-associated DERs expressed in other cells and tissues

We also explored the cell type specificity of these DERs, and respective intronic and intergenic expression, using publicly available RNA-seq data from human stem cells[16] and somatic adult tissues[17]. After realigning and processing these public data sets, we observed that most DERs had on average >5 reads in at least one stem cell type (86.4%) or tissue type (84.0%), including non-exonic brain-expressed sequences (75.3% and 67.1% of non-exonic DER expression in at least one stem cell or tissue group, respectively). Furthermore, 53.3% of all DERs and 26.5% of non-exonic DERs were expressed in all five stem cell conditions in the data set (embryonic stem (ES) cells, BMP4-treated ES cells, then differentiation to mesenchymal, mesendodermal and neural progenitor cells) with coverage >5 reads, whereas only 0.4% of the DERs were expressed at this same coverage level in all 16 tissue types.

PCA identified global expression similarities of these age-associated DERs between the fetal brain sample data and the stem cell and somatic tissue data (**Fig. 4**, via principal component 1). Notably, it was the postnatal brain samples that appeared qualitatively different from the diverse cell and tissue types with respect to these DERs (**Fig. 4a**). While the DERs overlapping intronic and intergenic Ensembl-annotated sequence aligned with the stem cells in its first principal component (**Fig. 4b**), these non-exonic DERs appeared to be particular to the fetal human brain. We then contrasted these patterns to the clustering of the global transcriptome based on read counts for all Ensembl-annotated genes (**Supplementary Data 1**). Here principal component 1 distinguished the brain (fetal and postnatal) from non-brain (stem cell and somatic tissue) samples and principal component 2 distinguished developmentally active tissues (fetal brain and stem cells) from somatic postnatal tissues, including postnatal brain (**Fig. 4c**). Gene-level expression patterns across the entire transcriptome highlight tissue-specific features, whereas the DERs target more general developmental transitions. Thus, although the overall transcriptomes of cells at different stages of early differentiation are clearly distinct, the DERs reflect common features of these differentiating cells.

Age-associated DERs overlap open chromatin

We next sought to better characterize the DERs with regard to functionality, using publicly available histone data for human fetal brain[25]. We performed peak calling on chromatin immunoprecipitation and sequencing (ChIP-seq) data on six histone tail marks (H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9ac and H3K9me3) and DNase I hypersensitive site sequencing (DNase-seq) data in fetal brain[25,26] (see Online Methods) and calculated the overlap with the DERs (see Online Methods). There was highly significant overlap (at empirical $P < 10^{-100}$; see Online Methods for permutation procedure) between the DERs and histone marks associated with active chromatin, including H3K36me3 (odds ratio (OR) = 13.32), H3K4me1 (OR = 3.00), H3K4me3 (OR = 5.66), and H3K9ac (OR = 4.82). Notably, approximately half of the exonic (48.9%; 14,582 of 29,813) and intronic (49.4%; 8,204 of 16,616) DERs were within 1 kb of a significant H3K36me3 peak; a smaller proportion of the intergenic DERs were also within 1 kb (22.7%; 960 of 4,221). There was also significant overlap ($P < 10^{-100}$) between open chromatin and the DERs (OR = 3.13, using the DNase-seq data). By contrast, there was little enrichment for histone marks associated with repression, including H3K27me3 (OR = 1.04) and H3K9me3 (OR = 1.43). These effects were largely consistent between the DERs annotated to exonic and intronic sequence, but weakened for the DERs annotated

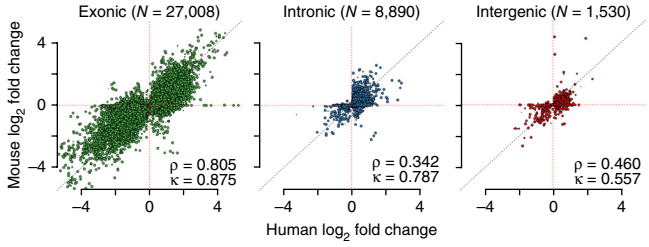

Figure 3 Cross-species comparison of differentially expressed regions (DERs). Significant DERs were ported to the mouse genome mm10 and RNA-seq coverage was extracted from the reprocessed study by Dillman *et al.*[15] comparing E17 to adult C57BL/6 mice. Log_2 fold changes comparing depth-adjusted mean differences between fetal and adult human samples were highly correlated with E17 versus adult mouse samples within each DER, stratified by human-annotated (**a**) exonic, (**b**) intronic and (**c**) intergenic sequence. Any DER with both exonic and intronic sequence was classified as exonic. Each point represents a single DER; its area indicates the proportion of the DER's width that was successfully ported over, where the largest points represent 100% mapping. ρ, Spearman correlation; κ, directionality concordance (for example, higher or lower expression in fetal relative to adult in both species).

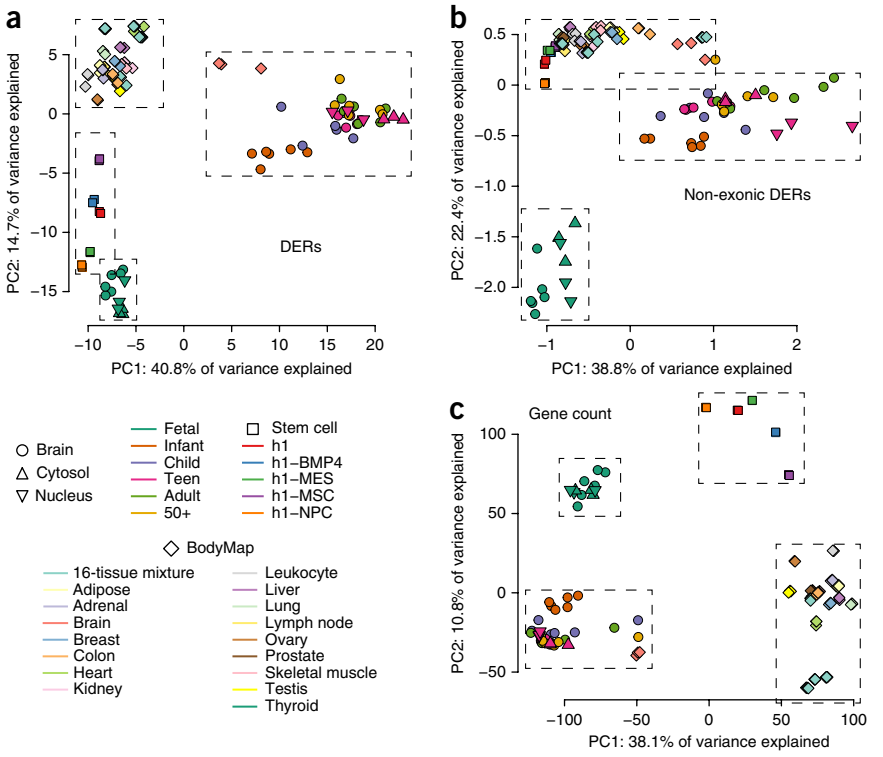

Figure 4 Clustering analysis of differentially expressed regions (DERs). PCA of (**a**) all significant DERs, (**b**) non-exonic sequence within the DERs and (**c**) gene counts from Ensembl annotation. PCA was performed on $\log_2$-adjusted coverage estimates across several data sets, including our human brain samples and publicly available differentiating stem cell and somatic tissue data. Colors and shapes for each point represent data set and condition. Dashed boxes were added by visual classification of biological groups. h1, ES condition; h1-BMP4, BMP4-treated ES; h1-NPC, neural progenitor cell; h1-MES, mesendodermal stem; h1-MSC, mesenchymal stem. BodyMap refers to data from the Illumina BodyMap project[17].

to intergenic sequence (**Supplementary Table 5**), demonstrating that the DERs largely reside in actively transcribed regions in the human fetal brain.

Age-associated DERs overlap CNS disease-associated loci

We sought to identify potential overlap between the DERs and genetic loci conferring risk for neurodevelopmental disorders, starting with schizophrenia—specifically, the 108 genome-wide significant loci from the latest Psychiatric Genomics Consortium (PGC) genome-wide association study (GWAS) of over 150,000 subjects[27]. Specifically, 42 loci (of the 108; 38.9%) overlapped at least one DER, which was statistically significant via permutation analysis ($P = 0.0013$; see Online Methods and **Table 2**). Stratifying the list of DERs by annotation class yielded more significant overlap for exonic ($P = 1.2 \times 10^{-4}$) and intronic ($P = 2.9 \times 10^{-4}$) DERs but non-significant overlap for intergenic DERs ($P = 0.053$). These effects represented odds ratios of approximately 2.0 for all, exonic and intronic DERs and 1.8 for intergenic DERs (see Online Methods).

We also assessed the overlap between the genes containing DERs and a series of pre-defined gene sets associated with other neurodevelopmental disorders, including autism, intellectual disability and syndromal neurodevelopmental disorders[3]. There was significant enrichment for genes associated with intellectual disability ($P < 10^{-4}$), and marginal association with autism ($P = 0.017$, genes in the SFARI database[28]) and genes associated with syndromal neurodevelopmental disorders ($P = 0.027$). These associations were in line with a previously published report on genes showing differential expression comparing fetal to postnatal life using microarray data[3]. Overall, these results implicate the genes containing DERs as enriched in those associated with diverse neurodevelopmental disorders.

Lastly, we conducted several analogous analyses in other disorders not typically associated with neurodevelopment, including brain-related (Alzheimer's disease and Parkinson's disease) and non-brain-related (type 2 diabetes; see Online Methods) disorders. We identified significant overlap between the age-related DERs and Parkinson's disease[29] ($P = 0.0039$) marginal overlap with Alzheimer's disease[30] ($P = 0.039$) and no overlap with type 2 diabetes[31] ($P = 0.25$). Notably, while only a small fraction of DERs were most highly expressed in adult life or later (8.4%), 4 of 7 Alzheimer's disease–associated and 5 of 11 Parkinson's disease–associated genetic loci overlapped at least one such later life DER ($P = 7.19 \times 10^{-5}$ and 1.01×10^{-4} respectively), in contrast to those associated with schizophrenia and other neurodevelopmental syndromes, for which the enrichment was primarily for DERs highly expressed in fetal life.

Fetal brain has the largest fraction of the expressed genome

We used the coverage-level RNA-seq data in our 36 discovery brain samples to barcode regions of expression within each age group (essentially a one-group generalization of the derfinder procedure) regardless of differential expression signal. After normalizing each sample to an 80-million-read library size, we identified contiguous regions where the average within-group expression levels were ≥5 normalized reads. While we identified a similar number of expressed sequences across the six age groups, the fetal samples had the largest fraction of the genome expressed across all six age groups — approximately 4% — and had the lowest proportion of expressed sequences overlapping Ensembl-annotated exons (**Table 3**). Surprisingly, each age group had a very similar proportion of all annotated Ensembl exons and introns covered (55–58%). Lastly, we observed that most PGC risk loci associated with schizophrenia[27] contained expressed sequence in the DLPFC, one of the brain regions consistently implicated in schizophrenia[32]. We observed similar metrics and inference using a threshold of ≥10 reads as a sensitivity analysis. On the basis of these results, we have created a custom UCSC Track Hub[33] called "LIBD Human DLPFC Development" that illustrates the coverage-level sequencing data within each age group (**Supplementary Fig. 8**).

Table 2 Enrichment of DERs among GWAS-positive regions

Trait	All	Exon	Intron	Intergenic
Schizophrenia	0.0013	0.0001	0.0003	0.0530
Alzheimer's disease	0.0385	0.2778	0.0117	0.6016
Parkinson's disease	0.0039	0.0100	0.0035	0.0882
Type 2 diabetes	0.2500	0.1029	0.4307	0.1200

Shown are empirical *P* values determined by permutation assessing significant overlap between DERs and locations of GWAS-positive loci for schizophrenia, Alzheimer's disease, Parkinson's disease and type 2 diabetes.

Table 3 Expressed sequences/regions by age group defined by five or more adjusted reads across consecutive bases, adjusted for library size

	Age group					
	Fetal	Infant	Child	Teen	Adult	Age ≥50
No. of regions	459,426	481,029	413,202	365,903	437,935	420,294
No. in DERs	46,813	37,618	33,958	31,818	32,849	31,563
Coverage (Mb)	121.8	107.5	97.1	90.5	92.9	91.4
Genome covered	4.1%	3.6%	3.2%	3.0%	3.1%	3.0%
Exonic	44.0%	46.8%	54.0%	58.8%	53.1%	54.1%
Intronic	77.1%	72.8%	71.1%	70.2%	69.9%	68.9%
Intergenic	11.9%	13.3%	12.9%	12.5%	12.9%	13.4%
Exons (Ensembl)	55.2%	56.8%	56.9%	55.3%	56.5%	55.8%
Introns (Ensembl)	57.6%	58.1%	57.7%	55.4%	57.2%	56.0%
108 PGC2 for SZ	83	84	83	82	83	88
Intronic, ≥10 reads	73.2%	65.6%	64.6%	64.4%	63.7%	62.4%

Exonic, intronic and intergenic rows give the percentages of the expressed regions overlapping annotated features; exons and introns rows give the converse, being the proportion of all Ensembl features (313,836 unique exons and 266,102 unique introns) covered by expressed sequences in each age group. The 108 PGC2 for SZ row gives the number of latest PGC schizophrenia-associated loci overlapping at least one expressed sequence in DLPFC. Lastly, we show, as a sensitivity analysis, the percentage of expressed regions when defined using ten or more adjusted reads.

These data can allow easy visualization of our data integrated with the diverse functionality of the UCSC Genome Browser.

Expression changes across development associate with a changing neuronal phenotype

Changes in gene expression across the lifespan may reflect a combination of changes within individual cellular populations and composition changes of varying cell types in the underlying brain tissue. In particular, a comparison of fetal frontal cortex, which contains predominantly neurons and neural progenitor cells (NPCs), and adult prefrontal cortex, which contains a mixture of neurons and glia, may reflect primarily these changing cell constituents. We therefore performed an *in silico* estimation[34] of neuronal, non-neuronal and neural progenitor cell composition using DNA methylation (DNAm) data from our brain samples projected onto publicly available DNAm data derived from cell lines (**Supplementary Table 6**), including ES cell–derived NPCs[35], and adult cortex tissue separated by fluorescence-activated cell sorting into neuronal and non-neuronal components using the NeuN antibody[34,36]. These composition estimates (the relative proportion of each cell type in each brain sample; **Supplementary Fig. 9a–c**) quantitatively confirmed the proliferation of non-neuronal cells across the lifespan ($P = 5.56 \times 10^{-5}$) and the loss of remaining NPCs at birth ($P = 6.01 \times 10^{-17}$).

We then correlated these cell type proportions with expression levels across individuals within each DER. Most DERs were significantly associated with only the NPC relative composition estimate (92.2% of DERS, Bonferroni-corrected $P_{bonf} < 0.05$, **Supplementary Fig. 9d**) and not the NeuN$^-$ estimate (1.6% of DERs, $P_{bonf} < 0.05$). Multivariate statistical modeling incorporating both NPC and NeuN$^-$ proportions (which are negatively correlated at $\rho = -0.53$) indicated that the vast majority of DERs associated only with the loss of NPCs ($N = 43{,}917$), and very few DERs associated only with NeuN$^-$ ($N = 6$). These results suggest that the widespread expression changes in human brain[1,2] at birth are more reflective of a changing neuronal phenotype—specifically, the differentiation of neural precursor and progenitor cells into mature neurons—than a rise in non-neuronal cell types.

DISCUSSION

We have identified widespread changes in the transcriptomes of the developing human prefrontal cortex, typically involving many genes previously implicated in brain development. However, unlike previous characterizations that rely on existing annotation, we observed extensive age-dependent expression of sequences previously annotated as intronic and intergenic in commonly accessed genomic databases (Ensembl, Gencode and UCSC). The majority of these DERs are most highly expressed in the fetal brain and decrease in expression across the lifespan. These developmental expression changes were largely present in cytosolic RNA from independent brain samples, were present in 15 other brain regions across development, were conserved across mouse development using synteny, and showed considerable overlap with differentiating neural progenitor cells. We further identified enrichment for active chromatin marks and for genomic regions associated with risk of schizophrenia and other neurodevelopmental disorders. Our *in silico* data suggest that most of these DERs, regardless of annotation (exonic, intronic or intergenic), reflect a changing neuronal phenotype, depicting differentiation and maturation across human brain development.

These developmental expression changes at single base resolution complement recent approaches characterizing the entire brain transcriptome within particular age groups, such as fetal[7,37] or postnatal[38]—for example, comparing expression changes across brain regions[39]. On the basis of our integration with BrainSpan data, we identified regions that do not appear to be regionally regulated, but rather appear to be generic developmental switches in the brain. This is in contrast to those genes recently reported by Pletikos *et al.*[39] as possibly related to regional parcellation. For example, while most of the genes identified as regionally associated by Pletikos *et al.*[39] were expressed in our data as based on gene level measures (reads per kilobase per million mapped > 1)—87.0% of adult, 81.3% of fetal and 88.2% of infant genes—only a smaller subset were present in the DER-overlapping 5,985 RefSeq genes: 44.4% of adult, 38.2% of fetal, and 29.4% of infant regionally associated genes. In contrast, those genes overlapped by DERs were not likely to be differentially expressed by region: of the 5,985 genes that overlapped DERs, only 5.1% were present in the adult regional association gene list, 16.3% of fetal and 0.09% of infant. We therefore hypothesize that genes associated with regional specificity are a separate subset from those associated with overall developmental processes, perhaps reflecting developmental changes arising from shifting cellular phenotypes in the latter case and regional changes representing different underlying cellular connectivities in the former.

The significant enrichment between the age-associated DERs and genetic loci associated with schizophrenia offers support for the neurodevelopmental hypothesis of the disorder[40]. The state-of-the-art GWAS study of schizophrenia, involving over 150,000 subjects, identified 108 independent loci associated with risk for illness, and these loci contain approximately 340 potential gene candidates. Because many of the candidates that map to these loci may not be participating in the population level association, a more finely grained analysis of the DERs that map to these loci may help eliminate some of the genes in these loci from the candidate list. Still, the mechanisms by which genes associated with schizophrenia lead to the emergence of the clinical syndrome in early adult life have been increasingly linked to early developmental processes involving both prenatal and postnatal factors[40]. Our evidence from the DER analysis supports this assumption. Similar enrichment of DERs was found for gene

sets associated with risk for autism, intellectual disability and various neurodevelopmental encephalopathy syndromes, all of which involve obvious early developmental clinical phenomena, thus supporting further clinical relevance of the DERs we have identified. Notably, while there was enrichment between DERs and loci implicated in neurodegenerative disorders, these genomic loci showed greater enrichment for DERs that reflect increased gene expression in adult life rather than fetal life.

While the age-associated DERs identified using a conservative statistical threshold occupied a relatively small proportion of the genome (8.63 Mb, 0.3% of the genome), we observed a much larger proportion of the genome being expressed across all age groups, particularly among fetal samples (121.8 Mb, 4.0%). As there were extensive differences among these proportions (for example, 4.0% in fetal brain versus 3.1% in adult brain), our derfinder approach depended on differential expression across six age groups, rather than focusing on fetal versus nonfetal expression differences, which are widespread[1,2]. We note that these differences in the proportion of genome expressed could result from the more diverse cellular phenotypes in the fetal brain samples, particularly the residual NPC signature. We ran derfinder with especially conservative parameters (for example, the single-base threshold), sacrificing statistical power in exchange for reducing the number of false positive DERs, an important distinction given the extent of newly identified transcriptional activity outside of previously defined exonic regions. The public availability of our data allow re-analyses with varying statistical thresholds and *post hoc* tests, particularly within individual genes of interest. We note that our DERs are, by definition, elements of transcripts and not full mRNAs. The limitations of relatively short sequence read length makes full transcript assembly challenging, but the DERs provide entry points to explore targeted transcript assembly with other methods. We also note that our RNA capture approach using poly(A) RNA pulldown has limitations, particularly with respect to uncovering noncoding RNAs, many of which are not polyadenylated, and observable 3′ biases.

Future biological experiments may better characterize the functions of these DERs, particularly the intronic and intergenic regions. Earlier RNA-seq characterization in commercially available fetal and adult brain mRNA also identified widespread intronic expression, which was hypothesized to facilitate co-transcriptional splicing[41]. The generation of more ChIP-seq-based functional histone tail marks in fetal brain can potentially generate more specific activity classes[42]. Also, translating ribosome affinity purification (TRAP)-based assays may elucidate potential translation of DERs in particular cellular systems. For example, we find preliminary evidence in the mouse genome that at least 15% of the intronic and intergenic DERs, and almost all exonic DERs, are likely incorporated into translated protein products on the basis of one small data set consisting exclusively of E14.5 mouse forebrain[43]. The translatomes from more diverse cell types in human tissue at various stages of development and cell lines may identify additional functional roles of our DLPFC-identified DERs. Similarly, we find little overlap between the DERs and reported lncRNAs from mouse neural stem cells of the subventricular zone[44] (only 2–3% of DERs, regardless of annotation), suggesting that lncRNA databases may be incomplete for human brain and that specialized subpopulations of cells may have unique transcriptomic signatures difficult to ascertain in tissue homogenates.

This study is the first, to our knowledge, to quantitatively estimate the influence of cellular composition changes on transcriptome dynamics across brain development, particularly when comparing prenatal and postnatal samples. Our results suggest that many reported differences in expression occurring across birth, and their subsequent association with or enrichment in brain disorders[4,6], may be driven principally by changing neuronal phenotypes, rather than by the commonly considered rise of non-neuronal cell types. The observation that many DERs result from a shifting cellular landscape cannot fully explain the widespread expression of non-exonic sequences, as a subset of these regions are more highly expressed in non-fetal samples. However, further research—for example, via the Epigenomics Roadmap Project[25]—will better refine the composition profiles in bulk tissue, particularly in the uniform generation of more numerous replicates (for example, NPCs) and cell types.

We anticipate that these data, both processed and raw, will be a resource for interrogating expression change across the lifespan. Our custom UCSC Track Hub can be used to visually discover transcriptional activity in candidate genes, and can be integrated with the other functional genomics tracks. The approach taken here explored one specific question within this rich data set, and our results underscore the complexity of gene expression and cellular differentiation that occurs during brain development and the incomplete nature of current transcriptome annotation.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. BioProject: PRJNA245228. The Track Hub is currently available at: http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&hubUrl=https://s3.amazonaws.com/DLPFC_n36/humanDLPFC/hub.txt. Values for gene reads per kilobase per million mapped (RPKM) are available in **Supplementary Data 1**. R code from analyses is available in **Supplementary Data 2**.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful for the vision and generosity of the Lieber and Maltz families, who made this work possible. Human brain material was acquired from the Offices of the Chief Medical Examiner of the District of Columbia and those of the Commonwealth of Virginia, Northern District, and processed and stored at the NIH Clinical Center in Bethesda, Maryland. We thank the families who donated to this research and we thank R. Straub for criticism of the data analyses. This work was supported by the Lieber Institute for Brain Development. A.E.J. was partially supported by 1R21MH102791, L.C.-T. was supported by CONACyT México (351535) and J.T.L. was supported by 1R01GM105705-01A1.

AUTHOR CONTRIBUTIONS

All authors contributed to the writing of the manuscript, plus the following individual contributions: A.E.J. designed the study, performed data analyses on summarized DERs: BrainSpan, mouse, cell and tissue types, histone tail– and disease-associated enrichments, and cell composition. J.S. performed data analysis involving processing the RNA-seq data. L.C.-T. performed data analysis involving the initial global derfinder approach. J.T.L. performed data analysis involving the initial global derfinder approach. R.T. performed RNA extractions and cytosolic separations. C.L. performed RNA extractions and cytosolic separations. Y.G. created sequencing libraries and oversaw the data generation for the discovery data. Y.J. created sequencing libraries and oversaw the data generation for the validation data. B.J.M. assisted in the biological interpretation of the computational findings. T.M.H. provided brain tissue and demographic data and assisted in biological interpretation of the computational findings. J.E.K. oversaw the project, provided brain tissue and demographic data, and assisted in biological interpretation of the computational findings. D.R.W. designed the project, oversaw the project and assisted in biological interpretation of the computational findings.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Colantuoni, C. *et al.* Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature* **478**, 519–523 (2011).
2. Kang, H.J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–489 (2011).
3. Birnbaum, R., Jaffe, A.E., Hyde, T.M., Kleinman, J.E. & Weinberger, D.R. Prenatal expression patterns of genes associated with neuropsychiatric disorders. *Am. J. Psychiatry* **171**, 758–767 (2014).
4. Gulsuner, S. *et al.* Spatial and temporal mapping of *de novo* mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518–529 (2013).
5. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
6. Parikshak, N.N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008–1021 (2013).
7. Willsey, A.J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997–1007 (2013).
8. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**, 1177–1184 (2013).
9. Frazee, A.C., Sabunciyan, S., Hansen, K.D., Irizarry, R.A. & Leek, J.T. Differential expression analysis of RNA-seq data at single-base resolution. *Biostatistics* (2014).
10. Jaffe, A.E. *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.* **41**, 200–209 (2012).
11. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42**, D749–D755 (2014).
12. Hinrichs, A.S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
13. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
14. BrainSpan. *Atlas of the Developing Human Brain.* http://www.brainspan.org/ (2011).
15. Dillman, A.A. *et al.* mRNA expression, splicing and editing in the embryonic and adult mouse cerebral cortex. *Nat. Neurosci.* **16**, 499–506 (2013).
16. Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
17. Farrell, C.M. *et al.* Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Res.* **42**, D865–D872 (2014).
18. Wang, Y., Lin, L., Lai, H., Parada, L.F. & Lei, L. Transcription factor Sox11 is essential for both embryonic and adult neurogenesis. *Dev. Dyn.* **242**, 638–653 (2013).
19. Curtis, M.A. *et al.* Human neuroblasts migrate to the olfactory bulb via a lateral ventricular extension. *Science* **315**, 1243–1249 (2007).
20. Hyde, T.M. *et al.* Expression of GABA signaling molecules KCC2, NKCC1, and GAD1 in cortical development and schizophrenia. *J. Neurosci.* **31**, 11088–11095 (2011).
21. Frankland, P.W., O'Brien, C., Ohno, M., Kirkwood, A. & Silva, A.J. Alpha-CaMKII-dependent plasticity in the cortex is required for permanent memory. *Nature* **411**, 309–313 (2001).
22. Krug, A. *et al.* The effect of neurogranin on neural correlates of episodic memory encoding and retrieval. *Schizophr. Bull.* **39**, 141–150 (2013).
23. Morris, D.W. *et al.* Confirming RGS4 as a susceptibility gene for schizophrenia. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **125B**, 50–53 (2004).
24. Wang, K. *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* **459**, 528–533 (2009).
25. Bernstein, B.E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
26. Ji, H. *et al.* An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.* **26**, 1293–1300 (2008).
27. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
28. Banerjee-Basu, S. & Packer, A. SFARI Gene: an evolving database for the autism research community. *Dis. Model. Mech.* **3**, 133–135 (2010).
29. Nalls, M.A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* **46**, 989–993 (2014).
30. Lambert, J.C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
31. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
32. Callicott, J.H. *et al.* Complexity of prefrontal cortical dysfunction in schizophrenia: more than up or down. *Am. J. Psychiatry* **160**, 2209–2215 (2003).
33. Raney, B.J. *et al.* Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* **30**, 1003–1005 (2014).
34. Houseman, E.A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
35. Kim, M. *et al.* Dynamic changes in DNA methylation and hydroxymethylation when hES cells undergo differentiation toward a neuronal lineage. *Hum. Mol. Genet.* **23**, 657–667 (2014).
36. Guintivano, J., Aryee, M.J. & Kaminsky, Z.A. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics* **8**, 290–302 (2013).
37. Miller, J.A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199–206 (2014).
38. He, Z., Bammann, H., Han, D., Xie, G. & Khaitovich, P. Conserved expression of lincRNA during human and macaque prefrontal cortex development and maturation. *RNA* **20**, 1103–1111 (2014).
39. Pletikos, M. *et al.* Temporal specification and bilaterality of human neocortical topographic gene expression. *Neuron* **81**, 321–332 (2014).
40. Kleinman, J.E. *et al.* Genetic neuropathology of schizophrenia: new approaches to an old question and new uses for postmortem human brains. *Biol. Psychiatry* **69**, 140–145 (2011).
41. Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.* **18**, 1435–1440 (2011).
42. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
43. Hupe, M., Li, M.X., Gertow Gillner, K., Adams, R.H. & Stenman, J.M. Evaluation of TRAP-sequencing technology with a versatile conditional mouse model. *Nucleic Acids Res.* **42**, e14 (2014).
44. Ramos, A.D. *et al.* Integration of genome-wide approaches identifies lncRNAs of adult neural stem cells and their progeny *in vivo*. *Cell Stem Cell* **12**, 616–628 (2013).

ONLINE METHODS

Post-mortem brain samples. Post-mortem human brain tissue was obtained by autopsy primarily from the Offices of the Chief Medical Examiner of the District of Columbia and those of the Commonwealth of Virginia, Northern District, all with informed consent from the legal next of kin (protocol 90-M-0142 approved by the NIMH/NIH Institutional Review Board). Brains tissue was stored and dissected at the Clinical Center, NIH, Bethesda, Maryland and at the Lieber Institute for Brain Development in Baltimore, Maryland. Brain material was transferred to the Lieber Institute under an approved Material Transfer Agreement where RNA was extracted and all RNA sequencing performed. Additional post-mortem fetal, infant, child and adolescent brain tissue samples were provided by the National Institute of Child Health and Human Development Brain and Tissue Bank for Developmental Disorders (http://www.BTBank.org/) under contracts NO1-HD-4-3368 and NO1-HD-4-3383. The Institutional Review Board of the University of Maryland at Baltimore and the State of Maryland approved the protocol, and the tissue was donated to the Lieber Institute for Brain Development under the terms of a material transfer agreement. Clinical characterization, diagnoses, and macro- and microscopic neuropathological examinations were performed on all samples using a standardized procedure. Details of tissue acquisition, handling, processing, dissection, clinical characterization, diagnoses, neuropathological examinations, RNA extraction and quality control measures were described previously in Lipska *et al.*[45]. The Brain and Tissue Bank cases were handled in a similar fashion (http://medschool.umaryland.edu/btbank/methods.asp). Toxicological analysis was performed on every case, and subjects with evidence of macro- or microscopic neuropathology, drug use, alcohol abuse or psychiatric illness were excluded.

We selected 6 samples per age group for our discovery data set, balancing for sex (4 male, 2 female) and RNA integrity number (RIN, mean = 8 per group), as our larger collection of fetal samples typically had higher RNA quality (for example, in Colantuoni *et al.*[1]). We then selected 36 more samples, also consisting of 6 samples across the 6 age groups as above (fetal, infant, child, teen, adult, and >50), to serve as a replication cohort. Additional demographic information for our discovery and validation data sets is available in **Supplementary Table 1**, including accession numbers in the Sequencing Read Archive (SRA) for the discovery samples.

RNA extraction and sequencing. Post-mortem tissue homogenates of dorsolateral prefrontal cortex gray matter (DLPFC) approximating BA46/9 in postnatal samples and the corresponding region of PFC in fetal samples were obtained from all brains. Total RNA was extracted from ~100 mg of tissue using the RNeasy kit (Qiagen) according to the manufacturer's protocol. The poly(A)-containing RNA molecules were purified from 1 μg DNase-treated total RNA and, following purification, fragmented into small pieces using divalent cations under elevated temperature. Reverse transcriptase and random primers were used to copy the cleaved RNA fragments into first-strand cDNA, and the second-strand cDNA was synthesized using DNA polymerase I and RNase H. We performed the sequencing library construction using the TruSeq RNA Sample Preparation v2 kit by Illumina. Briefly, cDNA fragments undergo an end repair process using T4 DNA polymerase, T4 polynucleotide kinase and Klenow DNA polymerase with the addition of a single adenosine using a Klenow polymerase lacking 3′ to 5′ exonuclease activity, and then ligated to the Illumina paired-end (PE) adapters using T4 DNA ligase. An index/barcode was inserted into Illumina adapters, allowing samples to be multiplexed in one lane of a flow cell. These products were then purified and enriched with PCR to create the final cDNA library for high throughput DNA sequencing using an Illumina HiSeq 2000.

RNA sequencing data processing. The Illumina Real Time Analysis (RTA) module performed image analysis, base calling and ran the BCL converter (CASAVA v1.8.2), generating FASTQ files containing the sequencing reads. These reads were aligned to the human genome (UCSC hg19 build) using the spliced-read mapper TopHat (v2.0.4) using the reference transcriptome to initially guide alignment, on the basis of known transcripts of Ensembl Build GRCh37.67 (the "-G" argument in the software)[46]. The total number of aligned reads across the autosomal and sex chromosomes (dropping reads mapping to the mitochondrial chromosome) per sample are provided in **Supplementary Table 1**.

Derfinder analysis. We implemented the derfinder pipeline, available from http://bioconductor.org/packages/release/bioc/html/derfinder.html, on the 36 discovery samples (**Supplementary Table 1**). Base-level coverage data (the number of reads crossing each base in the genome) was created from the aligned reads (BAM files). The statistical model was fit at every base, after performing coarse filtering to remove bases without at least 5 reads in at least 1 sample:

$$y_{ij} = \alpha_i + \beta_i \text{Group}_j + \gamma_i M_j + \varepsilon_{ij} \quad (1)$$

for coverage y_{ij} at base i for sample j, where Group_j is a categorical indicator variable for the six age groups and M_j is the scaled and log-transformed total number of mapped reads per sample, which adjusts for differences in library size between samples. This model is compared to the null model

$$y_{ij} = \alpha_i + \gamma_i M_j + \varepsilon_{ij} \quad (2)$$

by constructing an F-statistic F_i, and the vector of these F-statistics is then thresholded across the genome. Contiguous regions above the threshold form candidate differentially expressed regions (DERs), ranked by their area statistic (average F-statistic times region width), described in Jaffe *et al.*[10]. We used a per-base cutoff of $F = 20.509$, which corresponded to a per-base P value $< 10^{-8}$ for our given statistical model and sample size. Empirical P values were calculated by permuting the age group variable, keeping the coverage and library size fixed, 1,000 times and rerunning the full procedure within each permuted data set, recording the null area statistics. R code is available at https://github.com/lcolladotor/libd_n36/. The family-wise error rate (FWER) for each candidate DER was calculated on the basis of the null distribution of the maximum area statistic within each permutation[47]. Our initial F-statistic cutoff was quite conservative: 246 of 1,000 permutations did not result in a single genome-wide F-statistic greater than the threshold. We retained the 63,135 significant DERs at a FWER ≤ 5%.

We then assessed the DERs in an independent but analogous data set of 36 samples. Average coverage per DER was calculated within each of these replication samples, and then we calculated one F-statistic per DER using equations (1) and (2) above, where y_{ij} is now the sample-specific average coverage within the DER. We retained DERs that were at least marginally significant ($P < 0.05$) in this replication data set, yielding 50,560 (80.1%) genome-wide significant DERs that were also differentially expressed in this independent DLPFC data set, which were used for the analyses described below. Unreplicated DERs, as compared to replicated DERs, were narrower (83.0 bp versus 170.3 bp, $P < 10^{-100}$), had smaller areas (mean 2,633.9 versus 7,034.9, $P < 10^{-100}$) and therefore lower ranks, and lower coverage (mean 6.6 reads versus 108.7 reads, $P < 10^{-100}$), assessed via linear regression.

Gene annotations. We constructed "genomic state" objects for Ensembl version p12, UCSC build hg19 knownGene, and Gencode v19 for rapid annotation of DERs, which, briefly, assigns a single state (exonic, intronic or intergenic) to each base in the genome on the basis of the gene annotation. For a given base, we prioritize exon > intron > intergenic, such that any exonic sequence in any transcript, even if other transcripts were annotated as intronic, was assigned the exon state. Any intronic sequence not overlapping annotated exons was assigned the intron state, and the remaining genome is assigned the intergenic state. We required 20 base pairs (bp) of overlap between significant DERs and Ensembl annotation to be considered overlapping. The 100-bp mappability/alignability and Encode-excluded tracks were obtained from the UCSC Track Browser (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=141011952&g=wgEncodeMapability). LncRNA and microRNA tracks were obtained from the respective UCSC hg19 tracks as implemented in the TxDb.Hsapiens.UCSC.hg19.lincRNAsTranscripts[48] and TxDb.Hsapiens.UCSC.hg19.knownGene[49] R/Bioconductor packages. Pseudogenes were identified from the latest PseudoPipe Human Database, version 61 (ref. 50).

Technical exploration of widespread differential expression of novel transcriptional activity. RNA-seq data processing and analysis involves a number of well-documented technical biases[51–54], but we found little evidence for the significant DERs originating from technical or computational artifacts. For example, 93.7% of DERs had average alignability/mappability measurements of 100-bp reads

doi:10.1038/nn.3898

greater than 99%, only 61 and 7 regions were in tracks excluded by the Duke site and Data Analysis Center of the Encode project, consisting mainly of BSR/beta satellite repeats, respectively, and only 1.9% of regions mapped to known pseudogenes. We did observe evidence of 3′ bias in the entire set of DERs mapping within genes (the average proportion of nearest exon number to the total number of exons was 0.65, where 1 means the DER was in the last exon and 0.5 means the DER was in the middle exon), a well-described aspect of poly(A) RNA-seq[55]. However, there was substantial variability in this exonic location proportion when stratified by gene: 43.8% of genes had a DER before their middle exon (that is, the minimum exonic proportion was less than 0.5, by gene) while 52.3% of genes had a DER at the last exon (that is, the maximum exonic proportion was 1.0, by gene). Analyzing the sequence composition, the introns containing a DER had only an average 1.4-fold enrichment for poly(A) ($P = 1.58 \times 10^{-3}$) and poly(T) ($P = 8.61 \times 10^{-5}$) repeats for almost all run lengths beyond 6 bp, as compared to sequences of introns that do not contain a differentially expressed region, assessed by logistic regression, adjusting for intron length. The average GC content of the exonic DERs was significantly higher than the intronic and intergenic DERs (0.492 in exonic, as compared to 0.454 and 0.449 in intergenic and intronic, respectively; $P < 10^{-100}$) assessed via linear regression , although there was a wide range of values (interquartile range spanned ~0.15 for each annotation category) and the GC content for all three annotation class was higher than for the background genome (~0.42, as based on the hg19 build). Only 23 regions cross an annotated miRNA, but each also overlapped an annotated intron or exon, which is an important negative control given that our poly(A)$^+$ RNA library preparation should not capture these short RNAs. Lastly, of the DERs annotated as intergenic by Ensembl, 12.4% cross a known lncRNA (as determined using the TxDb.Hsapiens.UCSC.hg19.lincRNAsTranscripts database[48]), as compared to 3.7% of all DERs ($P < 10^{-100}$) assessed via a Chi-squared test.

Purification of cytosolic and nuclear RNA. We separated total RNA into nuclear and cytosolic fractions using the Cytoplasmic and Nuclear RNA Purification Kit by Norgen (cat. no. 21000, 37400) following the manufacturer's protocol with an extra step of DNase I treatment in the cytosolic fraction in three independent adult and three independent fetal samples. Sequencing libraries were constructed as above, using the poly(A) protocol. These were then sequenced on one lane of an Illumina HiSeq 2000, generating approximately 25 M reads per sample. One sample over-clustered in the sequencer, generating ~100 M reads, but its expression was highly correlated with the expression of other samples of the same type (after adjusting for library size), and it was therefore included in downstream analyses; see **Figure 4**. Further demographic material for these independent validation samples is provided in **Supplementary Table 4**.

BrainSpan RNA-seq analysis. Normalized sample-level RNA-seq coverage data were obtained in the bigwig file format (http://download.alleninstitute.org/brainspan/MRF_BigWig_Gencode_v10/) and matched to phenotype data indicating the brain region and age of each sample. Mean coverage levels for each sample within each DER were computed, and $\log_2$ fold changes comparing fetal to postnatal samples were calculated within each of the 16 brain regions that had at least 10 individuals (see **Table 1**). Principal component analysis (PCA) on the $\log_2$(normalized coverage + 1) matrix is visualized in **Figure 2** and **Supplementary Figures 5**, **6** and **7**. Spearman correlation was used to compare fetal versus adult coverage in our DLPFC samples to the fetal versus nonfetal coverage within each brain region.

Mouse RNA-seq analysis. We downloaded raw single-end 80-bp sequencing reads in the FASTQ file format from the study by Dillman *et al.*[15], available from the Sequence Read Archive (SRA)[56] at accession code SRX172890. Reads were aligned to the mouse genome (build mm10) using TopHat (version 2.0.9)[46], first aligning to the reference transcriptome ("-G" option, as described above). Significant DERs identified in the developing human brain (UCSC hg19) were mapped to the mouse genome (UCSC GRCm38/mm10) using the liftOver tool[12] implemented in the rtracklayer R/Bioconductor package[57]. Note that single human regions could result in multiple smaller subregions during the liftOver process, which were used to extract coverage-level data from the aligned mouse data, rather than the absolute range of the lifted over region. $\log_2$ fold changes were calculated as $\log_2$(mean adjusted fetal coverage + 1) − $\log_2$(mean adjusted adult coverage + 1), where each sample was normalized by the total number of mapped reads (in millions) and then averaged within each age group. Spearman correlations and directionality concordances were calculated for each human-annotated Ensembl feature, comparing the fold changes in mouse and human.

Public RNA-seq data processing. We downloaded raw sequencing reads from the Illumina BodyMap project[17] from SRA at accession code ERP000546 in the FASTQ file format. Note that each tissue sample had one replicate sequenced in a paired-end configuration (50-bp reads) and another replicate sequenced using single-end reads. Paired-end reads were therefore treated as single-end reads for alignment with TopHat (using the "-G" option, as described above) to obtain base-level coverage estimates (which does not use paired-end information), resulting in three measurements per tissue replicate. We note that single- and paired-end replicates clustered at the DER and gene count level (**Fig. 4**). Additionally, all samples labeled as "16 tissue mixture" had very low alignment rates (range: 16.4–40.6%); these were much higher in the single-tissue samples (range: 86.5–96.0%).

We also downloaded 101-bp paired-end raw sequencing reads from the UCSC Epigenome Project on differentiating stem cells[16] from SRA at accession SRP000941. These were aligned to the hg19 genome using TopHat as described above.

Cross-tissue analysis. Gene counts for the Lieber Institute post-mortem brain data and publicly available sample data were computed using the featureCounts program[58] using the Ensembl Homo_sapiens.GRCh37.73 gtf file, which were converted into the reads per kilobase per million mapped (RPKM) normalized count. Both raw and normalized coverage estimates (by total mapped reads) were extracted at the significant replicated brain DERs (N = 50,560) and the subset of DERs that did not overlap an Ensembl-annotated exon (N = 20,837). Raw coverage counts were used to confirm coverage of >5 reads across tissue and cell line group means.

Principal component analysis (PCA) was performed on the normalized coverage levels (scaled with $\log_2$ and an offset of 32) of the total set of DERs (**Fig. 4a**) and the subset of DERs that were non-exonic (**Fig. 4b**). PCA was performed on the gene RPKMs (**Supplementary Data 1**), scaled with $\log_2$ with an offset of 1 (**Fig. 4c**). $\log_2$ fold changes were calculated as above for all samples (our brain data and the publicly available data), relative to our adult (ages 20–50) adjusted coverage levels.

We further performed co-expression analyses within the three expression summarizations (individual DERs, the subset of non-exonic DERs and the overall gene counts) within the combined cell and tissue type data. To better understand the global patterns described in the main text, we computed fold changes for mean adjusted expression levels for each tissue and cell type relative to the mean of the adult (total RNA) brain samples. The pairwise Spearman correlations and concordances (both invariant to scaling) were computed for each cell and tissue type. Notably, there was high correlation ($\rho = 0.603$) and concordance ($\kappa = 0.738$) between the fetal brain sample and neural progenitor cell (NPC) fold changes in the DERs which was the only non-brain sample with concordance > 70% (other groups with high concordance were infant brain, and then the cytosolic and nuclear fractions of fetal brains). Conversely, these fetal brain samples were explicitly discordant with the other somatic non-brain tissues (all relative to adult brain expression levels). These results are consistent with a recent report[59], who found that NPCs had significantly correlated gene expression levels measured on microarrays to first trimester, but not second trimester, frontal cortex. The combination of these results suggests that cortex-derived DERs may represent a more general early developmentally conserved feature of the transcriptome.

Enrichment with chromatin marks and disease-associated loci. We downloaded the aligned reads (BED files) from the Epigenome Roadmap Project from the following GEO accession numbers: GSM621393, GSM669625, GSM806937, GSM806945, GSM916061, GSM621410, GSM806938, GSM806946, GSM706850, GSM806934, GSM806942, GSM621457, GSM669624, GSM806935, GSM806943, GSM669623, GSM621427, GSM806936, GSM806944, GSM916054, GSM1027328, GSM530651, GSM595913, GSM595920, GSM595922, GSM595923, GSM595926, GSM595928, GSM665804, GSM665819, GSM878650, GSM878651, GSM878652, GSM669944, GSM706851, GSM806948 and GSM817243. These were fetal brain epigenomic data from H3K27me3, H3K36me3, H3K4me1,

H3K4me3, H3K9ac, H3K9me3, chromatin accessibility and input. CisGenome was used to call one set of significant peaks, comparing each set of biological replicates per mark to the inputs using the default settings[26]. We tiled the hg19 genome into 1-kb bins, dropping bins in the known gaps (centromeres, telomeres, etc.), and then counted how many bins overlapped both a DER and ChIP-seq peak, only a DER, only a ChIP-seq peak, or neither. Each mark therefore generated a 2×2 table that summed to the number of genome-wide bins ($N = 2{,}861{,}069$), and we computed the odds ratio of each 2×2 table. Significance was assessed with a chi-squared test.

We performed a similar analysis for the PGC2 schizophrenia GWAS results using the chr:start-end of the 108 genomic loci from **Supplementary Table 3** of that publication[27]. First we calculated the observed proportion of 108 genomic loci that overlapped at least one DER. Then we performed permutation analysis to determine if this overlap was statistically significant: for a given permutation, we sampled 108 regions of the same widths from the genome (after removing the gaps as described above). Performing this permutation procedure 100,000 times resulted in 100,000 null overlap proportions. We then calculated an empirical *P* value, defined as the number of null proportions greater than the observed proportion. An R package for this analysis is available from GitHub[60]. The observed proportions were based on a list of (i) all DERs, (ii) exonic DERs, (iii) intronic DERs and (iv) intergenic DERs. The odds ratios for enrichment were calculated as above, using 1-kb genomic bins and counting the number of bins that overlapped PGC loci and DERs.

An analogous procedure was performed on genome-wide significant and replicated rs numbers available from main or supplementary tables for Alzheimer's disease[30], Parkinson's disease[29] and type 2 diabetes[31]. For each list of rs numbers, we used the SNAP tool[61] to find all SNPs with $R^2 > 0.6$ in Caucasian 1000 Genomes samples (mirroring the summary statistics from the schizophrenia associations) and then created a linkage disequilibrium–based locus for each index SNP. These loci were lifted over to hg19 and then used to assess the overlap with the significant DERs, both together and stratified by annotated feature.

Lastly, enrichment for disease-associated genes was calculated by first obtaining gene sets for neurodevelopmental gene sets as defined by Birnbaum *et al.*[3] directly from their Supplementary Table 1. We computed the proportion of genes in each gene set that contained at least 1 DER and assessed the significance of these observed proportions using permutation analysis. Specifically, we defined expressed genes using the featureCounts RPKM output (as described above) greater than 1.0 and resampled the same number of genes per gene set from the expressed genes (by symbol). For each permutated gene set, we calculated the proportion of null genes containing at least 1 DER and then calculated empirical *P* values based on 1,000 permutations (as above).

Expressed sequence analysis. Base-level coverage counts per sample were normalized to an 80-million-read library size (by dividing by 80 million, akin to the computation of RPKM) to identify contiguous regions above some coverage level that we defined as "expressed". Average normalized coverage levels were averaged within each age group, and these mean age group coverages were smoothed using a running mean operation with a window size of 7 bases to improve sensitivity and specificity[10,62] by reducing the number of very short 'expressed' regions (unlike in the multi-group derfinder procedure, which did not utilize smoothing). These smoothed age group means were thresholded at a coverage level of 5 reads, a threshold that we previously validated using PCR and that corresponds roughly to a one-sided *P* value <0.05 for a one-sample *t*-test with a sample size of 6, the number of samples per group here. We used a threshold of 10 reads as a sensitivity analysis that had similar results compared to using 5 reads.

Track Hub description. The track hub covers the entire genome at base-level resolution and displays the following by default: (i) the 50,560 significant DERs in dense visibility, (ii) the *F*-statistic for group differences, with the cutoff used to determine DERs, and (iii) the mean expression levels across the six samples in each of the six age groups, adjusted to an 80-million-read library size for easier interpretability and colored to match **Figure 1**. Additional tracks are available but hidden by default. These consist of the average adjusted expression within the fetal and infant nuclear and cytosolic mRNA fractions.

Composition analysis using DNA methylation (DNAm) data. We implemented *in silico* estimation of the relative proportions of three cell types (ES-derived NPCs from culture[35], and adult cortex neuronal and non-neuronal cells from tissue[36]) using epigenome-wide DNAm data using a recently published algorithm[34]. All data were obtained using the Illumina HumanMethylation450 microarray platform. After normalizing the publicly available data together using the preprocessQuantile function in the minfi Bioconductor package[63], we picked the cell type–discriminating probes as outlined by Jaffe and Irizarry[64]. This resulted in 227 unique probes that distinguished the three cell types (**Supplementary Table 6**). We then normalized the DNAm data from our 36 discovery samples and estimated the composition of our samples from the methylation profiles of the homogenate cell types at the 227 probes using nonlinear mixed modeling[34]. Composition estimates were regressed against the normalized and $\log_2$-transformed expression levels within each DER across the 36 samples, and we obtained a moderated *t*-statistic and corresponding *P* value[65] for each cell type and DER. The Bonferroni-adjusted *P* value was set at 0.05/50,560, or $P < 9.89 \times 10^{-7}$.

A **Supplementary Methods Checklist** is available.

45. Lipska, B.K. *et al.* Critical factors in gene expression in postmortem human brain: Focus on studies in schizophrenia. *Biol. Psychiatry* **60**, 650–658 (2006).
46. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
47. Churchill, G.A. & Doerge, R.W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971 (1994).
48. Carlson, M. TxDb.Hsapiens.UCSC.hg19.lincRNAsTranscripts: annotation package for TxDb object(s). (2014).
49. Carlson, M. TxDb.Hsapiens.UCSC.hg19.knownGene: annotation package for TranscriptDb object(s). (2014).
50. Zhang, Z. *et al.* PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* **22**, 1437–1439 (2006).
51. Hansen, K.D., Brenner, S.E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* **38**, e131 (2010).
52. Dohm, J.C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105 (2008).
53. Pickrell, J.K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
54. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* **12**, R22 (2011).
55. Levin, J.Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7**, 709–715 (2010).
56. Wheeler, D.L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **36**, D13–D21 (2008).
57. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).
58. Liao, Y., Smyth, G.K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
59. Brennand, K. *et al.* Phenotypic differences in hiPSC NPCs derived from patients with schizophrenia. *Mol. Psychiatry* published online, doi:10.1038/mp.2014.22 (14 April 2014).
60. Collado-Torres, L. & Jaffe, A.E. enrichedRanges: identify enrichment between two sets of genomic ranges v. 0.0.1. https://github.com/lcolladotor/enrichedRanges/ (2014).
61. Johnson, A.D. *et al.* SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938–2939 (2008).
62. Aryee, M.J. *et al.* Accurate genome-scale percentage DNA methylation estimates from microarray data. *Biostatistics* **12**, 197–210 (2011).
63. Aryee, M.J. *et al.* Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays. *Bioinformatics* **30**, 1363–1369.
64. Jaffe, A.E. & Irizarry, R.A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15**, R31 (2014).
65. Smyth, G.K. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds. Gentleman, R. *et al.*) 397–420 (Springer, New York, 2005).