RESOURCE
https://doi.org/10.1038/s41593-018-0223-0

nature neuroscience

Corrected: Publisher Correction

Enhancers active in dopamine neurons are a primary link between genetic variation and neuropsychiatric disease

Xianjun Dong[1,2], Zhixiang Liao[1,2], David Gritsch[1,2], Yavor Hadzhiev[3], Yunfei Bai[1,4], Joseph J. Locascio[1,2,5], Boris Guennewig[6,7,8], Ganqiang Liu[1,2], Cornelis Blauwendraat[9], Tao Wang[1,2], Charles H. Adler[10], John C. Hedreen[11], Richard L. M. Faull[12], Matthew P. Frosch[13], Peter T. Nelson[14], Patrizia Rizzu[9], Antony A. Cooper[7,8], Peter Heutink[9], Thomas G. Beach[15], John S. Mattick[7,8], Ferenc Müller[3] and Clemens R. Scherzer[1,2,5,16,17]*

Enhancers function as DNA logic gates and may control specialized functions of billions of neurons. Here we show a tailored program of noncoding genome elements active in situ in physiologically distinct dopamine neurons of the human brain. We found 71,022 transcribed noncoding elements, many of which were consistent with active enhancers and with regulatory mechanisms in zebrafish and mouse brains. Genetic variants associated with schizophrenia, addiction, and Parkinson's disease were enriched in these elements. Expression quantitative trait locus analysis revealed that Parkinson's disease-associated variants on chromosome 17q21 *cis*-regulate the expression of an enhancer RNA in dopamine neurons. This study shows that enhancers in dopamine neurons link genetic variation to neuropsychiatric traits.

To date the majority of disease- and trait-associated variants emerging from genome-wide association studies (GWAS) of neurologic and psychiatric diseases lie within nonprotein coding sequences. Several lines of evidence suggest that a proportion of such variants are involved in transcriptional regulatory mechanisms, including modulation of enhancer elements[1]. Many regulatory elements are cell-type-preferential[2,3], and therefore sequence variants with functional consequences are expected to manifest their effects more strongly in the cell type(s) most relevant to a specific disease phenotype.

Here we focused on systematically identifying all noncoding regulatory elements transcriptionally active in a morphologically, functionally, and biochemically distinct neuronal archetype: dopamine neurons of the substantia nigra pars compacta in human midbrain. We hypothesized that genetic variation associated with diseases involving dopaminergic neurotransmission exerts its effects through modulation of enhancers functionally active in this particular type of neurons. Perturbations of the dopaminergic system are important in the pathogenesis and treatment responses of many increasingly prevalent complex genetic diseases, including Parkinson's disease (PD), which affects 0.5 million people[4], schizophrenia, which affects 2.2 million people[5], and addiction, which affects 23.5 million people[6] (all numbers are for the United States alone). In healthy people, these dopaminergic neurons shape how we conduct our everyday lives, encoding activities related to motivation and reward. Signals from these neurons to the striatum have a profound impact on action learning and automatic movements, while projections to hippocampus and prefrontal cortex influence memories and behavior[7].

Our analysis is powered by an integrated hardware–software solution for comprehensively detecting noncoding transcription in one single and minuscule RNA sample and mapping the variation in noncoding transcription to genetic variation within dopamine neurons across multiple individuals. This method combines the base-pair resolution and a comprehensive genome-wide view afforded by ultradeep, total RNA-sequencing with the positional and cytoarchitectural precision afforded by traditional light microscopy.

Results

Identification of noncoding elements actively transcribed in dopamine neurons of human brain. To systematically identify noncoding elements actively transcribed in dopamine neurons of human brain, we used laser-capture microdissection total RNA-sequencing (lcRNAseq). Beyond traditional mRNA sequencing,

[1]Precision Neurology Program, Harvard Medical School and Brigham & Women's Hospital, Boston, MA, USA. [2]Center for Advanced Parkinson's Disease Research of Harvard Medical School and Brigham & Women's Hospital, Boston, MA, USA. [3]Institute of Cancer and Genomic Sciences, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK. [4]State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China. [5]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA. [6]Sydney Medical School, Brain and Mind Centre, The University of Sydney, Sydney, New South Wales, Australia. [7]Division of Neuroscience, Garvan Institute of Medical Research, Sydney, New South Wales, Australia. [8]St Vincent's Clinical School, UNSW Sydney, Sydney, New South Wales, Australia. [9]German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany. [10]Department of Neurology, Mayo Clinic, Scottsdale, AZ, USA. [11]Harvard Brain Tissue Resource Center, McLean Hospital, Harvard Medical School, Boston, MA, USA. [12]Centre for Brain Research, University of Auckland, Auckland, New Zealand. [13]C.S. Kubik Laboratory for Neuropathology, Massachusetts General Hospital, Boston, MA, USA. [14]Sanders-Brown Center on Aging, University of Kentucky, Lexington, KY, USA. [15]Banner Sun Health Research Institute, Sun City, AZ, USA. [16]Ann Romney Center for Neurologic Diseases, Brigham and Women's Hospital, Boston, MA, USA. [17]Program in Neuroscience, Harvard Medical School, Boston, MA, USA. *e-mail: cscherzer@rics.bwh.harvard.edu

all polyadenylated and non-polyadenylated transcripts were ultradeeply sequenced using ribodepleted RNAs from ~40,400 neurons laser-captured from 99 human postmortem brains and seven non-neuronal cell-type samples with an average of 178 million reads per sample, yielding 2.0×10^{10} pair-ended RNA-seq reads (Supplementary Table 1). Melanized neurons from the midbrain substantia nigra pars compacta of 86 high-quality human brains (dopamine neurons), pyramidal neurons from layers V/VI of the middle temporal cortex of ten brains, and pyramidal neurons from the primary motor cortex of three brains (pyramidal neurons) were laser-captured as described[8–10] (Fig. 1a). Human fibroblasts from four individuals and peripheral blood mononuclear white cells (non-neuronal cells) from three individuals were analyzed in the same pipeline (Supplementary Figs. 1a and 2 and Supplementary Table 2). Cumulatively, we found that at least 64.4% of the human genome was transcribed to produce detectable RNAs in dopamine neurons of the human brain (Fig. 1b and Methods), consistent with observations from the Encyclopedia of DNA Elements project (ENCODE) in cultured cells[11]. More than half of these reads (54.7%) mapped to intergenic or intronic regions (Fig. 1c), indicating a massive hidden layer of active noncoding transcription in human brain neurons.

Enhancer RNA (eRNA) expression is a marker for active enhancers[12–14] and can be used to estimate enhancers active in a particular cell type at given time[13]. Genetic enhancer elements control the cell-type-specific activation of gene expression. We designed a sophisticated method to systematically identify noncoding elements, including known and novel candidate enhancers that are significantly expressed in dopamine neurons, pyramidal neurons, and non-neuronal cells, using a stringent six-step filter (Fig. 1d and see Methods for details). We required aggregated reads for each cell type to achieve local peak read densities ('summits') with detection P values < 0.05 compared to randomly sampled background; without overlap with exons from annotated genes and transcription start site-proximal regions; with a minimal element length of 100 bp; and without splicing junction reads (to avoid multiexon noncoding RNAs). We then rigorously determined the statistical significance of each of these candidate transcribed noncoding elements across multiple independent samples of the same cell type (for example, across 86 independent samples for dopamine neurons) with a family-wise adjusted $P \leq 0.05$ taken as evidence of statistically significant expression.

We discovered 71,022, 37,007, and 19,690 transcribed noncoding elements (TNEs) in dopamine neurons, pyramidal neurons, and non-neuronal cells, respectively, with detection P values equal or better than the Bonferroni-corrected significance thresholds of 7.0×10^{-7}, 5.1×10^{-7}, and 6.6×10^{-7} for each of the three cell types, respectively (Supplementary Table 3). The length distribution of TNEs peaked around 400 bp (Supplementary Fig. 3a), consistent with that of the eRNAs previously reported by FANTOM5[13] and of activity-regulated eRNAs found in mouse cortical neurons[12]. Unlike promoter regions, TNEs showed a GC content distribution similar

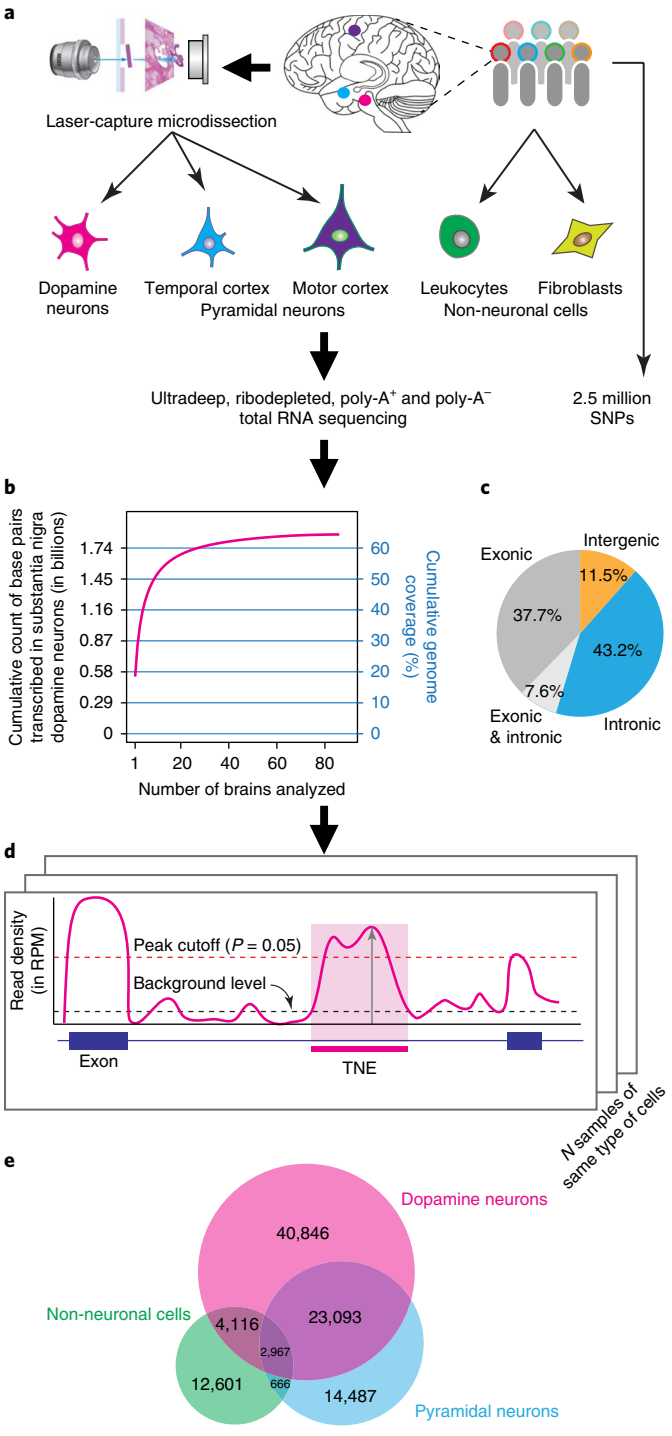


Fig. 1 | Identification of noncoding elements actively transcribed in dopamine neurons of human brain. a, lcRNAseq was used to systematically identify noncoding elements transcribed in dopamine neurons and pyramidal neurons of human brains. We analyzed dopamine neurons of the substantia nigra from 89 high-quality autopsy brains, pyramidal neurons from temporal cortex of ten brains and from motor cortex of three brains, and fibroblasts from four individuals and peripheral blood mononuclear white cells (non-neuronal cells) from three individuals. **b**, Cumulatively, 64.4% of the human genome was transcribed in dopamine neurons of human brain. **c**, More than half (54.7%) of reads mapped to intergenic or intronic genome sequences. **d**, Schematic of the method for identifying TNEs: a stringent six-step filter was applied to the RNA-seq reads aggregated from dopamine neurons, pyramidal neurons, and non-neuronal cells. Briefly, a putative TNE site was defined as a genomic region with RNA-seq reads density higher than the background transcriptional level (black dashed line) and the peak unique reads per million (RPM; vertical arrow) achieving a local detection $P \leq 0.05$. TNEs were required to exceed 100 bp; known genes and splice junctions were excluded. Finally, TNEs were required to achieve Bonferroni-corrected expression $P \leq 0.05$ across all samples of one cell type (indicated by multiple copies of the schematic) compared to length-matched, randomly selected background regions using a binomial distribution. See Methods and Supplementary Fig. 14 for details. **e**, Venn diagram with TNEs detected in dopamine neurons, pyramidal neurons, and non-neuronal cells.

to that of random genomic background regions, and this is inconsistent with PCR bias (Supplementary Fig. 3b). The vast majority of TNEs (92%) localized to intronic regions (Supplementary Fig. 3c) and tended to be positionally biased toward the 5′ end of gene body, a pattern opposite to that of partial RNA degradation, which preferentially degrades 5′ ends (Supplementary Fig. 3d). TNEs accounted for 31.42% and 32.35% of reads transcribed in dopamine and pyramidal neurons, respectively, compared to 21.08% in peripheral cells; and 26.38% of dopamine neuron TNEs were also presented in pyramidal neurons (Fig. 1e; Fisher's exact test $P < 2.2 \times 10^{-16}$, odds ratio = 3.22), but only 7.85% in peripheral cells. Subprograms of protein-coding mRNAs and noncoding RNAs (ncRNAs) expressed in dopamine neurons, pyramidal neurons, and non-neuronal cells were also characterized (Supplementary Fig. 4, Supplementary Table 4, and Methods).

TNEs identify putative enhancers active in dopamine neurons. Of the 71,022 TNEs active in dopamine neurons, 23,625 (33%) coincided with enhancers defined by one or more genomic or epigenetic features (Fig. 2 and see Methods). These features included DNase I hypersensitivity sites (DHS)[15], characteristic histone modifications (such as high H3K27ac, high H3K4me1, and low H3K4me3)[16], capped analysis of gene expression (CAGE)[13]-defined enhancers, transcriptional coactivator P300[17] binding sites, transcription factor 'hotspots'[18], and sequence conservation[19]. Of the 71,022 TNEs, 20,505 coincided with chromatin-state-defined putative active enhancers from Roadmap Epigenomics[20] and 1,212 TNEs coincided with CAGE-defined putative active enhancers[13]. The overlap was significantly higher than expected by chance alone ($P < 2.2 \times 10^{-16}$ by permutation test; Supplementary Table 5).

We performed two experiments to directly benchmark TNE to putative enhancers predicted by two other methods applied to the same source (Fig. 2b). Of the TNEs called by our pipeline in the human cortex dataset from PsychENCODE[21], 44.1% (14,904 of the 33,762 TNEs) overlapped with a strong transposase-accessible chromatin assay sequencing (ATAC-seq) peak (which maps chromatin accessibility[22]) identified in the same samples (Fig. 2b). This was a significantly higher than expected by chance ($P < 2.2 \times 10^{-16}$ by permutation test). In SK-N-SH (human neuroblastoma cell line) cells, 21.7% of called TNEs (11,465 of 52,733) overlapped with putative enhancer features (Fig. 2b; for example, H3K27ac, H3K4me3, transcriptional regulator CCCTC-binding factor (CTCF) chromatin immunoprecipitation sequencing (ChIP-seq), P300 ChIP-seq, DNase I hypersensitivity, and transcription factor hotspots) delineated by ENCODE in this cell line ($P < 2.2 \times 10^{-16}$ by permutation test), similarly to the 25% overlap previously reported between CAGE-defined and chromatin state-predicted putative enhancers[13].

We grouped 71,022 dopamine TNEs into three classes according to the presence or absence of supporting features (see Methods). Specifically, 11,835 TNEs coincided with multiple supportive features (designated class I TNEs), i.e., a known DHS site plus at least one of five additional external features (enhancer chromatin state (chromHMM), CAGE-enhancer, P300 peak, transcription factor binding sites hotspot, and highly conserved noncoding elements between human and zebrafish; Fig. 2c,d). A second set of 11,790 TNEs was supported by at least one of the five external features, but lacked additional DHS evidence (designated class II TNEs; Fig. 2c,d). A third set of 47,397 TNEs had no previously reported supporting external features (termed class III TNEs; Fig. 2c,d). Bidirectional transcription of select dopamine TNEs was seen using CAGE in substantia nigra of four of the same brains used for lcRNAseq (Supplementary Fig. 5a). Moreover, transcription factor binding sites were enriched in TNE sites, based on in silico analysis of ChIP-seq peaks and motif scanning (Supplementary Fig. 5b–d and Supplementary Note).

Replication of TNEs in independent cohorts. We replicated pyramidal neuron TNEs in three independent cohorts representing 36, 498, and 795 human brain samples, respectively (Fig. 3a), and additionally confirmed select TNEs with two secondary methods (Fig. 3b and Supplementary Fig. 5a). Of the 37,007 pyramidal neuron TNEs discovered, 34,077 (92.1%) were replicated in an independent cohort of pyramidal neurons laser-captured from layer V/VI of 36 new human autopsy brains (Fig. 3a). We identified 14,679 (39.7%) and 10,718 (29%) of 37,007 pyramidal neuron TNEs from ribodepleted total RNA-seq data of frontal cortex (PsychENCODE[21]) and four cortex areas (Accelerating Medicines Partnership–Alzheimer's Disease Consortium (AMP-AD)), respectively (Fig. 3a). Select brain cell-type-specific TNEs were confirmed with a secondary method, quantitative PCR (qPCR), in laser-captured dopamine neurons (Fig. 3b). As expected, qPCR analysis of control samples lacking template or reverse transcriptase showed no expression of TNE. Finally, we confirmed a subset of dopamine neuron TNEs by performing CAGE on four substantia nigra homogenate samples (Supplementary Fig. 5a and Methods).

TNE signatures accurately cluster dopamine and pyramidal neurons. A majority (57.5%; 40,846 of 71,022) of the detected TNEs were exclusively expressed in human dopamine neurons. They were not detected in pyramidal neurons or non-neuronal cells. Thirty-nine percent (14,487 of 37,007) of pyramidal neuron TNEs were exclusive to this cell type; 64% (12,601 of 19,690) of non-neuronal TNEs were exclusively expressed in non-neuronal cells (Fig. 1e). A signature based on cell-type-exclusive TNEs clustered 106 individual samples with 99.1% accuracy (Fig. 3c), similar to the classification accuracy afforded by mRNAs and ncRNAs (Supplementary Fig. 4). Normalized counts for the 100 most-abundant exclusive TNEs in each cell type are visualized in Fig. 3d. Cell-type-preferential expression of three dopamine neuron-exclusive TNEs, three pyramidal neuron-exclusive TNEs, and one TNE common to both dopamine and pyramidal neurons (in intron 4 of the PD gene *SNCA*[23,24]; Supplementary Fig. 6a) were confirmed by qPCR in addition to lcRNAseq (Fig. 3b and Supplementary Fig. 6b). These TNEs were in close proximity to histone marks typical of active enhancers[25] as well as multiple transcription factor occupancy sites[25] (Fig. 3b).

In vivo validation of TNE enhancer activity in zebrafish, mice, and neuronal cells (Fig. 4). To determine whether TNEs can function as enhancers, we tested 15 TNEs (Supplementary Table 6) in vitro in human SK-N-MC neuroblastoma cells and non-neuronal HeLa cells. TNE sequences were inserted into a modified pGL4.10 vector (as in ref.[13]), for example, upstream of an EF1a basal promoter separated by a synthetic poly-A signal or transcriptional pause site to avoid promoter effects. Eleven of the 15 TNEs (73%) significantly increased reporter activity in neuronal cells compared to control inserts representing random background sites (Fig. 4b). Eight TNEs induced more than a two-fold increase in reporter signal (Fig. 4b), and all but one TNE exhibited considerably higher enhancer activity in the neuronal cells compared to HeLa cells (Fig. 4b).

VMP1-TNE (chr17:57,863,430–57,864,538) is located in intron 7 of the human *VMP1* gene, a key regulator of autophagy. The *VMP1*-TNE site is evolutionary conserved among vertebrates and actively transcribed in human brain dopamine neurons, pyramidal neurons, and non-neuronal cells. *VMP1*-TNE was a class I TNE with a bimodal distribution of RNA-seq reads (centered on the DHS peak; Fig. 4a), bidirectional CAGE signal (Fig. 4a), occupancy by 90 TFs (Fig. 4a), and high levels of H3K4me1 and H3K27ac (Fig. 4a), and it was predicted as putative enhancer by ChromHMM[26] in Roadmap Epigenomics[20]. It was highly active in neuroblastoma and HeLa cells in culture (Fig. 4b). To assess the activity of *VMP1*-TNE in vivo, transient transgenic reporter assays were carried out in zebrafish embryos. The PCR-amplified sequence was cloned upstream of

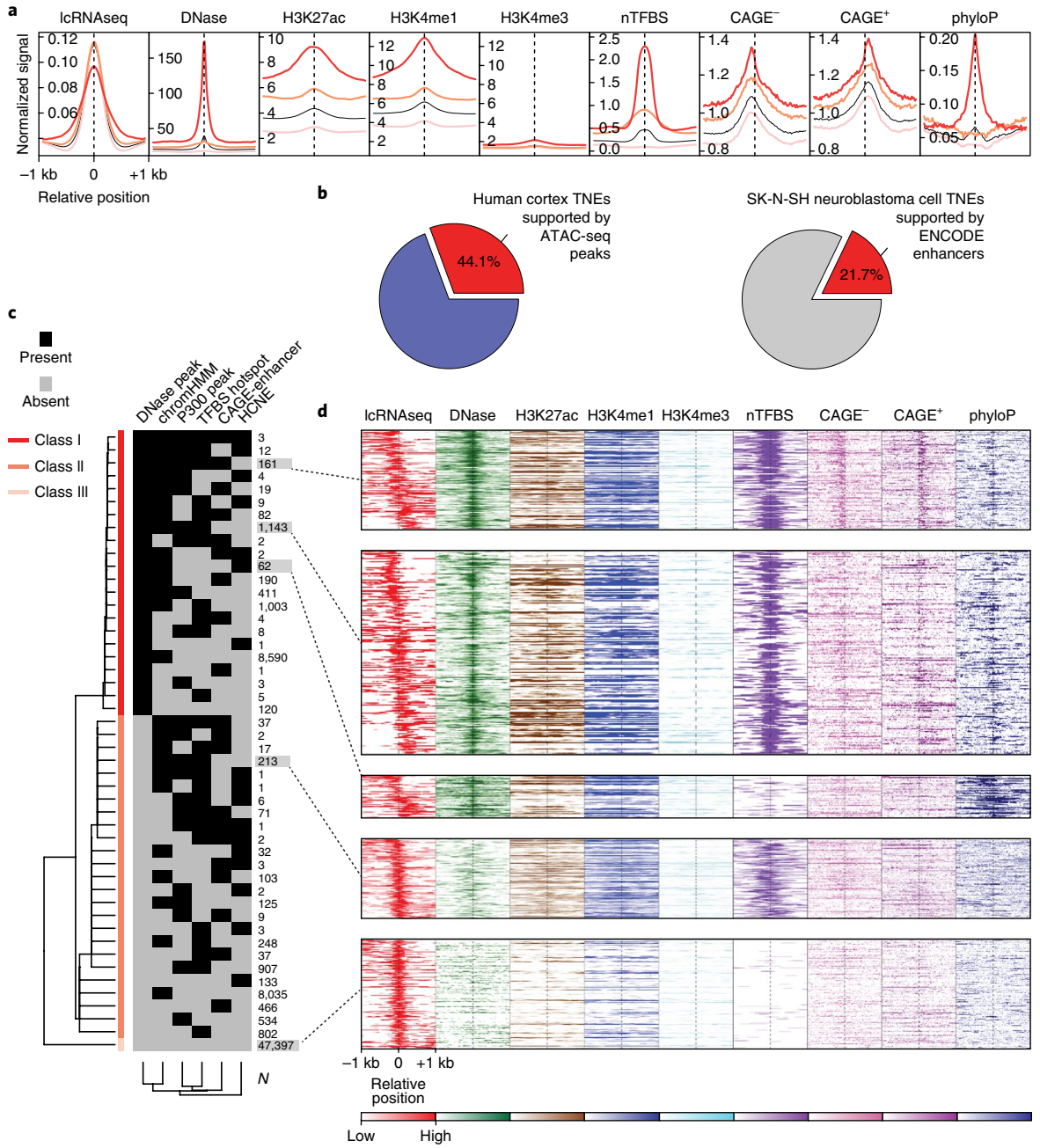


Fig. 2 | TNEs identify putative enhancers active in dopamine neurons. a, TNEs in dopamine neurons are, in part, supported by epigenetic and genetic features of active enhancers. The aggregation plot visualizes normalized signals for DHS, histone modification marks (H3K27ac, H3K4me1, and H3K4me3), transcription factor binding site hotspots (nTFBS), CAGE signals (forward and reverse strand), and sequence conservation score (phyloP) in the [−1,000, +1,000]-bp regions centered on the DHS peak (if any) or middle position of TNEs. TNEs with multiple supportive enhancer features are shown as red line (class I); TNEs with one supportive feature are shown with a pink line (class II), and TNEs previously not identified by published histone- or CAGE-enhancers are shown in light pink (class III). The black line indicates the aggregation plot for all three TNE classes combined. **b**, TNEs significantly overlapped with putative enhancers predicted by other methods applied to the same human brain samples and to the same human dopaminergic SK-N-SH neuroblastoma cell line ($P < 2.2 \times 10^{-16}$ by permutation test). **c**, We clustered 71,022 dopamine neuron TNEs into three classes (rows) according to the presence or absence of supporting features of active enhancers (columns) such as DHS, P300 peak, ChromHMM, transcription factor binding sites (TFBS) hotspot, CAGE-enhancers, and sequence conservation from external databases as described above. HCNE, highly conserved noncoding elements. **d**, Heatmaps visualize TNEs with distinct combinations of epigenetic and genetic enhancer features. The relative abundance of TNEs (rows) in dopamine neurons as measured by lcRNAseq (red; relative abundance is color-coded) and physically co-localizing signals of supportive epigenetic and genetic enhancer features (the relative abundance of each supportive enhancer features is color-coded) is shown for the same [−1,000, +1,000]-bp window as above.

zebrafish *gata2*[27] minimal promoter, linked to an *mRuby2* reporter gene in a modified pDB896 vector. A similarly sized sequence amplified from a nonconserved intergenic region with very low or no signal for enhancer marks was used to generate a control construct. Embryos injected with Has.VMP1-TNE:gata2:mRuby2 (Fig. 4c–g) reporter construct showed reproducible enrichment

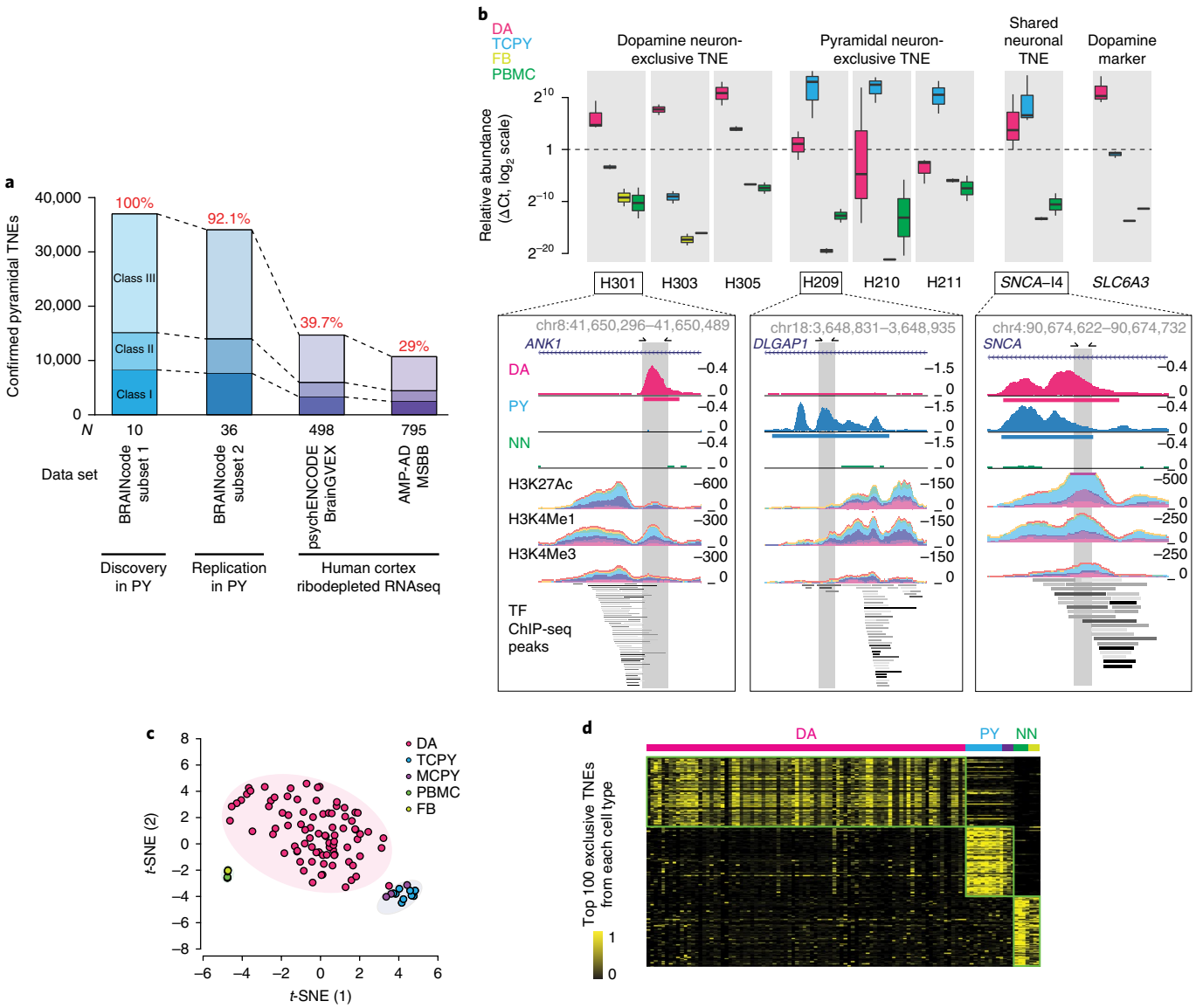


Fig. 3 | TNE signatures accurately cluster dopamine and pyramidal neurons. a, Of the 37,007 pyramidal neuron TNEs discovered, 34,077 (92.1%) were replicated in a second cohort of pyramidal neurons (PY) laser-captured from layer V/VI of 36 new human autopsy brains that had not been used for the discovery study. Moreover, using our TNE detection pipeline, we deconvoluted expression of 14,679 (39.7%) and 10,718 (29%) of pyramidal neuron TNEs in two ribodepleted total RNA-seq datasets representing cortex homogenates from 498 and 795 individuals, respectively. AMP-AD, Accelerating Medicines Partnership–Alzheimer's Disease Consortium; MSBB, Mount Sinai VA Medical Center Brain Bank; psychENCODE BrainGVEX (https://www.synapse.org/#!Synapse:syn4590909); **b**, Cell-type-preferential expression of three dopamine neuron-exclusive TNEs, three pyramidal neuron-exclusive TNEs, and one TNE common to both dopamine and pyramidal neurons (in intron 4 of the PD gene *SNCA*; see also Supplementary Fig. 6a) was confirmed additionally by qPCR (box plots, top panel) in addition to lcRNAseq (pile graphs, bottom panels; Supplementary Fig. 6b). Relative mRNA abundance of a classical dopamine-neuron marker, the dopamine transporter gene *SLC6A3* (*DAT*), was assayed as positive control (dopamine neuron samples (DA), $n=3$; temporal cortex pyramidal neurons (TCPY), $n=3$; primary human fibroblasts (FB), $n=2$; human peripheral blood mononuclear white cells (PBMC), $n=2$). Pile graph tracks show RNA-seq read densities in DA, PY, and non-neuronal cells (NN), as well as corresponding histone enhancer marks and transcription factor (TF) ChIP-seq peaks from ENCODE[25]. Box plots indicate the median (bold line), the 25th and 75th percentiles (box edges), and the most extreme data point no more than 1.5× the interquartile range from the box (whiskers). **c**, A signature based on cell-type-exclusive TNE clustered 106 individual samples with 99.1% accuracy by *t*-distributed stochastic neighbor embedding (t-SNE). **d**, The heatmap visualizes normalized counts for the 100 most-abundant TNEs exclusively expressed in DA, PY, and NN.

of enhancer activity in a specific subset of telencephalic neurons near the eyes and in cardiac cells proximal to the atrioventricular canal compared to embryos carrying control construct (Has.control:gata2:mRuby2; Fig. 4c–g and Supplementary Table 7), consistent with the expression pattern of *miR-21* (http://zfin.org/), the putative target gene in the synteny block as suggested by comparative genomics (Supplementary Fig. 7).

The VISTA consortium has established one of the largest repositories of in vivo enhancer screens during mouse development[28]. Sequences overlapping with 96 dopamine neuron TNEs were evaluated by VISTA[28], 63 (65.6%) of which were positive enhancers in vivo in mice, considerably more than expected by chance alone ($P=3.91\times10^{-3}$ by Fisher's exact test; Fig. 4h). The enrichment for VISTA-validated enhancers was similar for class I and III TNEs

(Supplementary Fig. 6c). Notably, 35 of these 63 (55.6%) VISTA-validated TNEs drove reporter gene expression in neuronal tissues, particularly midbrain, hindbrain, and the neural tube. For example, a neuron-specific TNE located in the intron of autism susceptibility candidate 2 gene (*AUTS2*) enhanced reporter activity specifically in the midbrain in 11 of 15 mouse embryos tested[28] (Fig. 4i and Supplementary Fig. 6d). In comparison, of 31 exclusively non-neuronal TNE evaluated by VISTA, 14 (45%) were positive enhancers, and only 9 (29%) were active in neuronal tissues. Collectively, these test cases show that select TNE sites enhance reporter gene expression in human neuronal cells and in neurons in the brains of zebrafish and in mice.

Variants associated with diseases of the dopamine system are over-represented in TNE actively transcribed in dopamine neurons. GWAS variants for 61 diseases and traits were significantly enriched within noncoding elements functional in dopamine neurons with *P* values below the Bonferroni-corrected significance threshold of 9.64×10^{-6} by Fisher's exact test (for example, 0.01 divided by 1,037, the total number of traits in the NHGRI GWAS catalog[29]) compared to random background (Fig. 5a and Methods). By contrast, only 43 traits were significantly enriched in promoters, 11 of them in exons (Fig. 5a). Consistent with our hypothesis, variants associated with 11 diseases and medications perturbing the dopamine system were significantly enriched in dopamine neuron TNE sites (Fig. 5a,b). These included variants associated with schizophrenia ($P=1.75\times10^{-40}$), PD ($P=5.05\times10^{-9}$), addiction ($P=1.33\times10^{-8}$), and bipolar disorder ($P=5.05\times10^{-6}$). Moreover, pharmacogenetic variants associated with response to antipsychotics were enriched in these TNE sites ($P=4.39\times10^{-14}$). Classical antipsychotics are dopamine receptor antagonists that are the standard treatment for schizophrenia. Variants associated with response to iloperidone, a specific antipsychotic medication for schizophrenia, were also enriched ($P=1.94\times10^{-6}$). Variants associated with response to the dopamine reuptake inhibitor methylphenidate (used to treat attention deficit hyperactivity disorder) were enriched in dopamine neuron TNE sites ($P=8.74\times10^{-9}$). By contrast, none of these trait variants were enriched in promoters or exons. Notably, traits relating to sleep phenotypes, which are modulated by dopamine neurons (for example, refs.[30,31]) and perturbed in PD[32], were highly enriched in these TNE sites ($P=2.6\times10^{-55}$; Fig. 5a,b). Strikingly, cardiovascular traits (Fig. 5a,b); diseases and traits clustering around obesity, weight, and diabetes (Fig. 5a,b); and brain-volume-related traits (Fig. 5a,b) were also over-represented in dopamine neuron TNEs compared to random genomic background. The enrichment density for dopamine system traits was similar for each of the three TNE classes (Supplementary Fig. 8a).

Dopamine neuron TNEs harbor a higher density of GWAS variants linked to traits of the dopamine system than enhancer predictions without cell-type-specificity. GWAS single-nucleotide polymorphism (SNP) density analyses showed a higher density of GWAS variants for dopamine system traits in TNE active in midbrain dopamine neurons compared to FANTOM5-predicted and ChromHMM-predicted putative enhancers, exons, promoters, introns, intergenic regions, and length-matched random regions (Supplementary Fig. 9).

Expression quantitative trait locus analysis reveals transcribed noncoding elements in synapse genes as main cell-autonomous effectors of *cis*-acting genetic variation. Expression quantitative trait locus (eQTL) analysis for TNEs, ncRNAs, and mRNAs was—for the first time to our knowledge—performed across cell-type-specific transcriptomes from 84 human brains (Fig. 5c). We measured or imputed 4,283,750 SNPs and associated them with normalized TNE expression using Matrix eQTL[33] (see Methods). Of these, 8,676 *cis*-acting TNE eQTLs achieved a false discovery rate of less than or equal to 0.05, comprising 3,461 unique expression-associated SNPs (eSNPs) and 151 unique TNEs (Fig. 5c). On average, 23 eSNPs were associated with expression changes in one TNE. Furthermore, 3,381 ncRNA eQTLs were significant (FDR $\leq$ 0.05), comprising combinations of 3,320 unique eSNPs and 52 unique expressed ncRNA genes (Fig. 5c and Supplementary Fig. 10). By contrast only 1,150 mRNA eQTLs reached statistical significance (FDR $\leq$ 0.05), comprising combinations of 676 unique eSNPs and 46 unique associated expressed protein-coding genes (Fig. 5c and Supplementary Fig. 10).

These 151 *cis*-regulated TNEs physically localized to introns of 102 host genes. These host genes were highly enriched in Gene Ontology (GO) terms related to synapse function ($P<4.79\times10^{-7}$ by enrichment analysis using the hypergeometric test; see Methods and see Supplementary Table 8 for full results) and in medical subject heading terms for brain disorders with $P=5.1\times10^{-10}$ (Supplementary Table 9). Mutations of several of these synapse-related host genes can cause abnormal brain development and function (Supplementary Fig. 10 and Supplementary Note). Taken together, this gene-regulatory analysis indicates that genetic variation is linked to variation in the activity of putative enhancers in synapse genes, including several loci linked to Mendelian brain disorders.

PD-associated variants *cis*-regulate a noncoding element in the *KANSL1* gene. Leveraging 495,085 SNPs associated with one or more of 1,037 human diseases or traits (19,188 disease-associated SNPs from the NHGRI-EBI GWAS catalog[29], extended via imputation of proxy SNPs with $r^2 \geq 0.8$), we identified 1,989 disease-associated SNPs that influence expression of 19 TNEs, 4 ncRNAs, and 5 mRNAs in *cis*. To distinguish coincidental co-localizations of GWAS and eQTL associations, we used regulatory trait concordance (RTC) scores[34], which assess whether a *cis*-eQTL and a trait association are tagging the same underlying functional effect. Applying a stringent RTC threshold of 0.85, we identified 23 disease-associated TNE eQTLs for which the trait and TNE expression associations may be tagging the same effect in dopamine neurons (Fig. 5c and Supplementary Table 10). Of these, 17 disease-associated eQTLs were identified for ncRNAs and 1 for mRNAs.

Eight of these 23 TNE-eQTLs linked PD-associated variants to a putative eRNA expressed from intron 2 of the *KANSL1* gene with *P* values as low as 1.57×10^{-7} (Supplementary Table 10). The corresponding RTC scores were 0.91–1.00, indicating that the GWAS-derived disease variants explain the eQTL observation. Six of the eight PD-associated eSNPs mapped to the exact same 712,000-bp-long linkage disequilibrium (LD) block on chromosome 17q21 (here termed LD2; Fig. 5d) and were significantly associated with upregulation of the same *KANSL1*-TNE1 in carriers of risk alleles (6.46×10^{-7}). Two additional eSNPs mapped to a nearby LD block (LD3; Fig. 5d). Conditional eQTL analysis adjusting for the lead GWAS variant rs17649553 suggested that some eSNPs in LD2 and one in LD3 might carry an independent signal (Methods, Supplementary Fig. 11, and Supplementary Table 11). Chromosome 17q21 is the second-most-important GWAS peak for sporadic PD (after *SNCA*) and unequivocally associated with susceptibility for PD, with *P* values as low as 2.23×10^{-48} in a meta-GWAS of more than 100,000 cases and controls[35]. There is precedent that copy-number variation in the *KANSL1* locus causally impacts brain function, as microdeletions of the locus cause Koolen de Vries syndrome, a neurological disease with severe learning disability and developmental delay. In addition to upregulating the *KANSL1*-TNE, the same PD-associated variants in LD2 (but not those localized to LD3) were associated with downregulation of an expressed pseudogene, *LRRC37A4P* ($P=2.36\times10^{-7}$; Supplementary Table 10). *LRRC37A4P* is localized near *KANSL1* under the chromosome 17q21 GWAS peak. By contrast, eQTL associations for *MAPT*

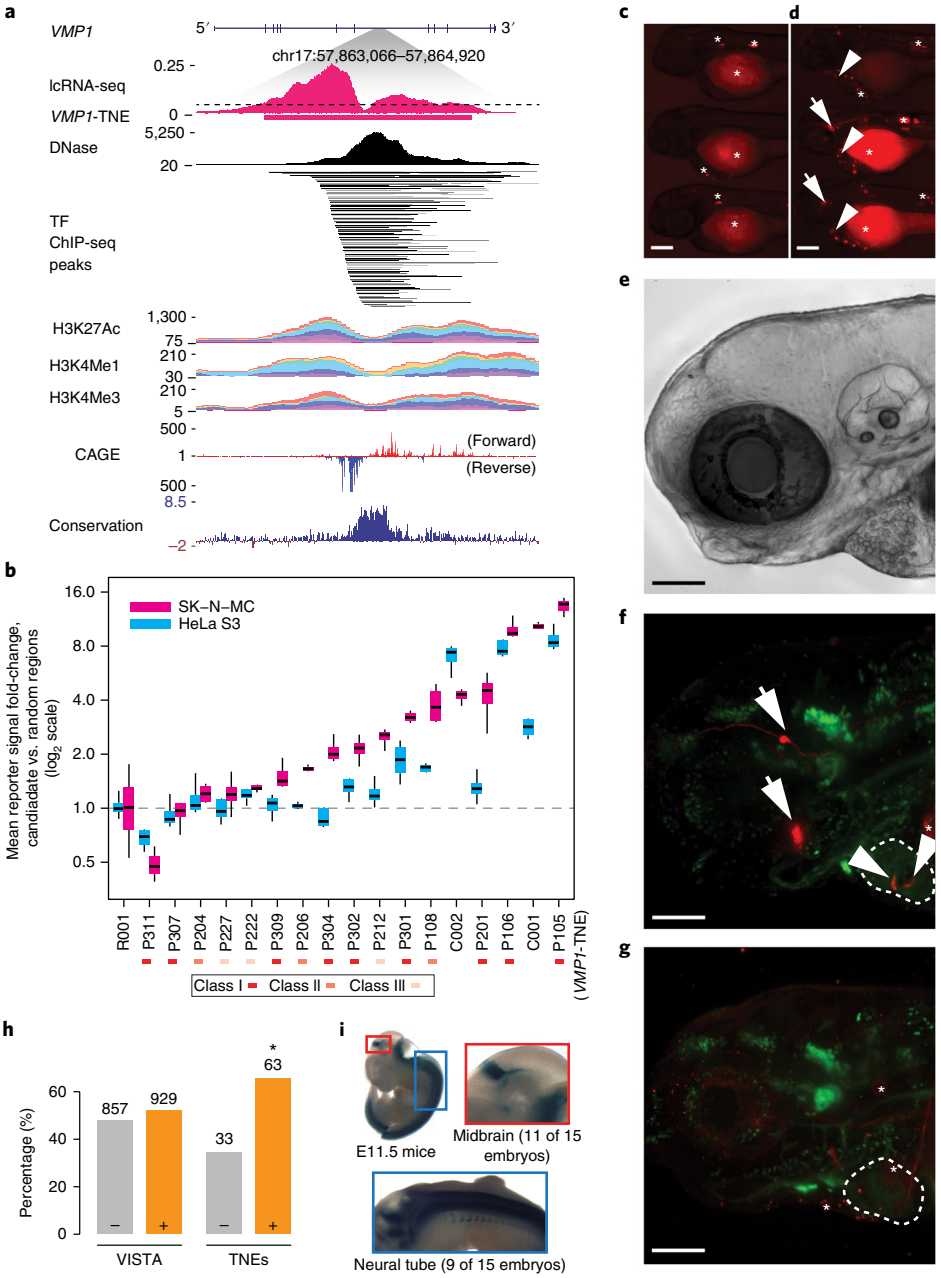


Fig. 4 | In vivo validation of TNE enhancer activity in zebrafish, mice, and neuronal cells. a, *VMP1*-TNE in intron 7 of human *VMP1*. This evolutionary conserved TNE is supported by classical epigenetic features: a putative active enhancer including open chromatin (DNase)[20], high levels of H3K4me1 and H3K27ac[25], and bidirectional CAGE signal[13]. **b**, Enhancer reporter assays for TNE-defined putative enhancers in HeLa S3 (cyan) and SK-N-MC neuroblastoma line (magenta) cells. TNE regions are labeled Pxxx (see Supplementary Note for details); C001 and C002, enhancers from FANTOM5 (positive controls)[13]; R001, random genomic background region (negative control); $n=4$ transfections were independently performed and analyzed for each TNE in HeLa S3 and another $n=4$ independent transfections in SK-N-MC cells. $^{*}P<0.05$; $^{**}P<0.01$; $^{***}P<0.001$; two-tailed Student's *t* test. Box plots as in Fig. 3. **c**–**g**, Group view of embryos injected with control:gata2:mRuby2 and *VMP1*-TNE:gata2:mRuby2. Enhanced reporter activity was observed in the embryos injected with *VMP1*-TNE-containing reporter construct (**d**) compared to the control element (**c**). In addition, embryos injected with *VMP1*-TNE:gata2:mRuby2 (**d**) show tissue-specific reporter expression in a group of telencephalic neurons in proximity to the eye (arrows) and atrioventricular canal (arrowheads). Background (ectopic) activity (stars) predominantly in skin yolk muscle and autofluorescence from blood and eye pigmentation (stars) was observed in both *VMP1*-TNE:gata2:mRuby2- and control:gata2:mRuby2-injected embryos. Scale bars, 200 µm. (**e**) A brightfield reference image of the embryo regions shown in **f**,**g**. (**f**,**g**) High-magnification view on the head and heart region of ETvmat2:gfp transgenic embryos injected with control:gata2:mRuby2 and *VMP1*-TNE:gata2:mRuby2. GFP reporter expression in these embryos was used as marker for the heart ventricle (dashed line). Expression in the telencephalic neurons (arrows) and atrioventricular canal (arrow heads) in *VMP1*-TNE:gata2:mRuby2 injected embryos can be seen. Stars, ectopic activity. The experiment was repeated independently with similar results four times, with 478 and 408 embryos screened in total for *VMP1*-TNE and control constructs, respectively (see Supplementary Table 7 for details). Scale bars, 100 µm. **h**, Putative enhancers evaluated in mice by VISTA[28]. Of 1,789 putative enhancers tested in mice by VISTA (left two bars), 929 (52%) were active in mice. By contrast, 63 (66%) of 96 TNE-defined putative enhancers were found to be active enhancers in mice ($^{*}P=3.91\times10^{-3}$, hypergeometric test). **i**, Reporter activity of a neuron-specific intronic TNE located in the *AUTS2* gene is seen in the midbrain (red insert) and neuronal tube (blue insert) of mouse embryonic day (E) 11.5 embryos by VISTA[28]. Embryos have an average crown-rump length of 6 mm.

mRNA, a biological candidate in this region, did not reach genome-wide significance (Fig. 5e–g). The inverse eQTL relations between the lead GWAS-derived SNP, rs17649553, and *KANSL1*-TNE1 and *LRRC37A4P*, respectively, were confirmed by a second method, cell-type-specific qPCR (Supplementary Fig. 12a). Moreover, this association was independently replicated in a second cohort of neurons laser-captured from 31 high-quality control brains (Supplementary Fig. 12b and Supplementary Table 12). Third, the rs17649553–*LRRC37A4P* eQTL association was further confirmed in 56 substantial nigra and 96 frontal cortex samples from the Genotype-Tissue Expression Consortium (GTEx; Supplementary Fig. 12c,d), which used a poly-A^+ selection protocol that does not allow for assaying *KANSL1*-TNE1 RNA.

Discussion

eRNA expression is a feature of active enhancers[12–14] and can be used as a marker to estimate their activity in a particular cell type[13]. Genome elements with enhancer chromatin marks that are transcribed into eRNAs have substantially higher validation rates in in vitro enhancer assays than enhancers defined exclusively by chromatin states[13]. Moreover, in transgenic mouse reporter assays, over half of putative enhancers identified on the basis of deep RNA-sequencing functioned as enhancers with reproducible activity in the predicted tissue[36]. Many chromatin-defined enhancers are not regulatorily active in a particular cellular state, but may be active in other cells[37] or are premarked for fast regulatory activity upon stimulation[38].

We showed a highly specific program of enhancer elements that are actively transcribed in physiologically and morphologically distinct, disease-relevant dopamine and pyramidal neurons, in situ, in human brains. Nearly two-thirds (64.4%) of the genome were cumulatively transcribed in dopamine neurons, including 71,022 noncoding elements, many of which were consistent with histone-state and CAGE-defined active enhancers, as well as with in vivo regulatory functions in zebrafish and mouse neurons. We provided mechanistic evidence that some of these elements function as enhancers of transcription in zebrafish brain, in the midbrain of mice, and in human cultured neuronal cells using genetics and reporter assays. Moreover, multiple independent lines of evidence—including chromatin state, CAGE expression, and transcription factor binding analyses— support the view that these transcribed noncoding elements are putative enhancers specifically active in dopamine neurons.

Variants associated with 11 diseases or medications perturbing the dopamine system were enriched in dopamine neuron-specific TNE sites, much more so than in promoters and or exons (Fig. 5a,b). Risk alleles associated with major disorders of dopaminergic neurotransmission, schizophrenia, PD, addiction, and bipolar disorder overlocalized to active TNE sites. Compellingly, even pharmacogenetic variants linked to treatment response were enriched in active enhancers. These observations suggest that GWAS variants might modulate enhancers active in dopamine neurons and thereby regulate the transcriptional program underlying susceptibility for these neuropsychiatric diseases. Finally, risk alleles associated with sleep-related phenotypes were enriched in TNE sites ($P=2.6\times10^{-55}$). Indeed, dopamine neurons have a role in sleep regulation[30,31], and REM sleep behavior disorder is an early sign of PD[32].

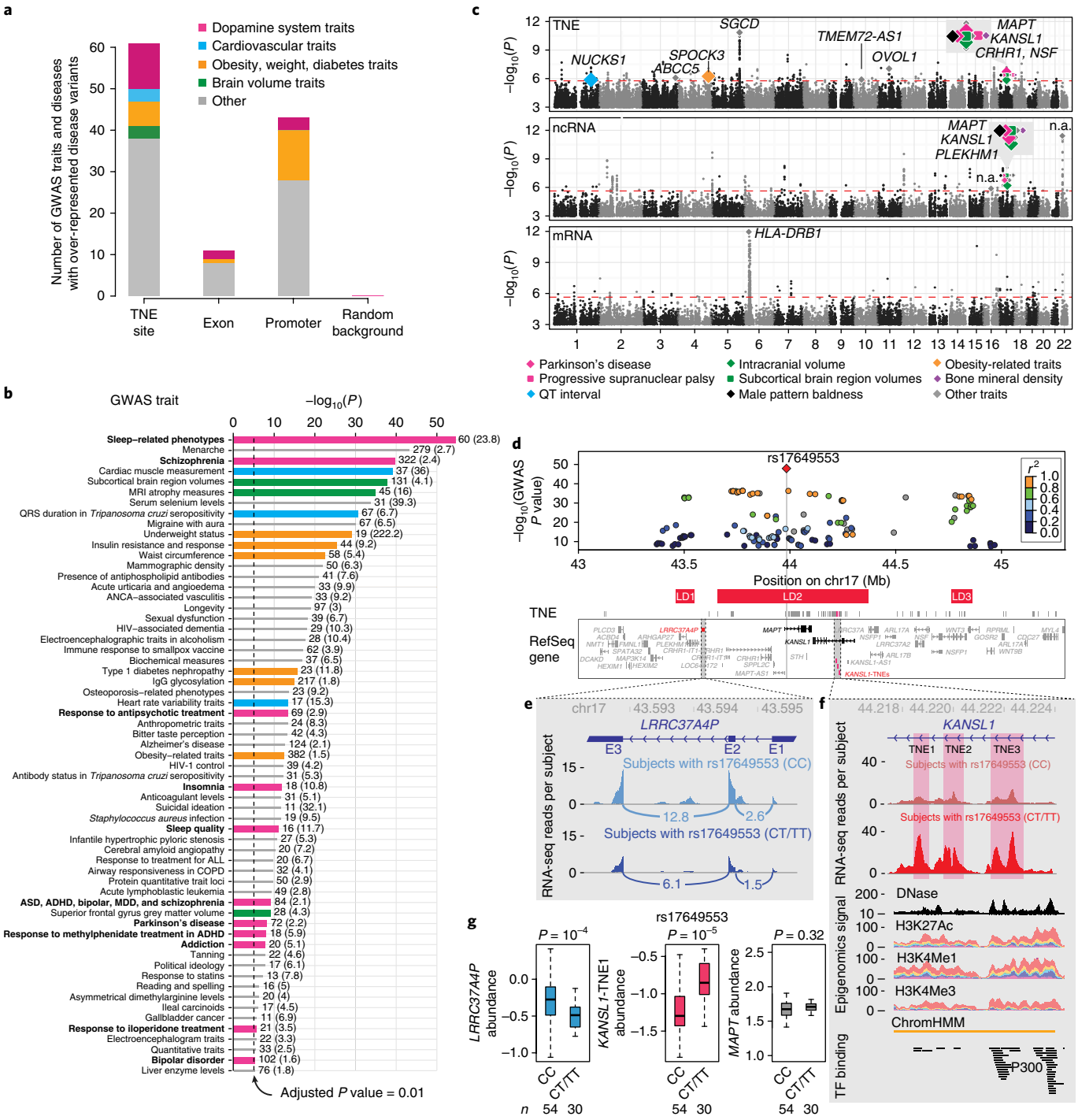
eQTL analysis for putative eRNAs was performed across cell-type-specific transcriptomes from 84 human brains (Fig. 5c). We thereby uncovered transcribed noncoding elements in synapse genes as a main cell-autonomous effector of *cis*-acting genetic variation in dopamine neurons. Notably, the number of TNE eQTLs greatly surpassed the number of mRNA eQTLs and ncRNA eQTLs identified.

The second most significant GWAS locus for sporadic PD is located on chromosome 17q21. This locus shows unequivocal evidence for association with PD. The regulated gene has not been established, but *MAPT* has been commonly assumed to be the

Fig. 5 | Putative enhancers active in dopamine neurons link genetic variation to neuropsychiatric disease. a, GWAS diseases and traits with variants significantly enriched in dopamine TNEs, exons, promoters, and random background regions. Variants for 61 diseases and traits were enriched within TNE-defined putative active enhancers in dopamine neurons with *P* values below the Bonferroni-corrected significance threshold of 9.64×10^{-6} by one-sided Fisher's exact test compared to 71,022 random genomic background regions (see Methods). The largest share of traits ($n=11$) enriched within putative active enhancers clustered around perturbations of the dopamine system (pink). By contrast, only 43 traits were enriched in promoters (including two involving the dopamine system), 11 in exons, and none in random background regions. **b**, Diseases and traits significantly enriched in TNE-defined putative enhancers in dopamine neurons. Variants associated with eleven diseases or medications perturbing the dopamine system (horizontal pink bars) were dramatically over-localized in dopamine neuron TNEs. The number of disease-variants colocalizing to dopamine TNEs for each trait as well as odds ratios (in parenthesis) are shown next to each bar. *X* axis, *P* values by one-sided Fisher's exact test ($-\log_{10}$ scale); *y* axis, diseases and traits. ADHD, attention deficit hyperactivity disorder; ASD, autism spectrum disorder; MDD, major depressive disorder; ANCA, antineutrophil cytoplasmic antibody; ALL, acute lymphoblastic leukemia; COPD, chronic obstructive pulmonary disorder; QRS duration, a feature on an electrocardiogram. **c**, eQTL analysis reveals transcribed noncoding elements in synapse genes as main cell-autonomous effectors of *cis*-acting genetic variation in human brain dopamine neurons. Manhattan plots for TNE eQTLs (top), ncRNA eQTLs (middle), and mRNA eQTLs (bottom) are shown. Diamonds, eSNPs with RTC ≥ 0.85; colors indicate different groups of diseases. Gene symbols of the host loci for these eSNPs are shown (n.a., intergenic regions). *Y* axis, *P* values of eSNP-transcript associations from Matrix-eQTL linear regression model ($n=84$). A false-discovery rate (FDR) of 0.05 was considered significant (red dashed line); QT interval, another feature on an electrocardiogram. **d**, PD-associated variants *cis*-regulate a noncoding element in the *KANSL1* gene in dopamine neurons. The locus plot visualizes *P* values (*y* axis) and chromosomal location (*x* axis) of genetic variants associated with susceptibility for PD in the chromosome 17q21 GWAS peak[35] (chr17:43,000,000–45,300,000 in hg19). Red diamond, lead susceptibility variant rs17649553; other PD-associated variants in the locus are represented by circles (r^2 with the lead SNP is color-coded). Forty-five RefSeq genes are physically localized under this GWAS peak (box). LD blocks are shown as red horizontal bars. *P* values extracted from http://www.pdgene.org for GWAS meta-analysis of 13,708 PD cases and 95,282 controls are shown. **e**,**f**, Increased expression of *KANSL1*-TNE1 (right, red) and decreased expression of *LRRC37A4P* (left, blue) in dopamine neurons of individuals carrying one (CT) or two (TT) copies of the risk allele (CT/TT) compared to individuals without the risk allele (CC). **e**, Pile graphs of average RNA-seq reads density aligning to the exon (E) 3–E2 and E2–E1 junctions of *LRRC37A4P* are visualized for individuals carrying (navy) or not carrying (cyan) the risk allele. **f**, Genomic landscape of *KANSL1*-TNE1 expression in noncarriers (red-brown) and carriers (red) of the risk allele. *KANSL1*-TNE1 is expressed from intron 2 of the *KANSL1* gene (chr17:44,218,414–44,219,042). Note nearby *KANSL1*-TNE2 and -TNE3 showing similar expression patterns. *KANSL1*-TNE1 is expressed from a chromatin-state-defined enhancer. Lower tracks display DNase uniformed signal (Roadmap[20]); stacked H3K27ac, H3K4me1, and H3K4me3 signals (ENCODE[25]); chromHMM-predicted enhancer based on histone modifications (Roadmap[20]); and transcription factor ChIP-seq peak clusters (ENCODE[25]). **g**, Boxplots representing the eQTL relation between the lead PD-associated rs17649553 variant and transcript abundance of *KANSL1*-TNE1, *LRRC37A4P*, and *MAPT* in dopamine neurons. Box plots as in Fig. 3b. *P* values from Matrix-eQTL linear regression model, $n=84$ biologically independent samples.

prime candidate. Using eQTL analysis, we provide striking evidence pointing at regulation of a putative eRNA expressed from intron 2 of the *KANSL1* gene as a gene-regulatory mechanism for this susceptibility locus. The *KANSL1* locus is important for normal brain function. Microdeletions cause Koolen de Vries syndrome, a neurological disease with severe learning disability and developmental delay[39]. The *KANSL1*-TNE1 eQTL association was confirmed by cell-type-specific qPCR and replicated in an independent cohort. By contrast, eQTL associations for *MAPT* did not reach statistical significance in dopamine neurons ($P = 0.32$). Long-read sequencing and larger datasets will be required to comprehensively illuminate the relation between structural variation and transcriptional function in this complex locus.

The *KANSL1*-TNE1 eQTL appears to be a 'super-eQTL' of variants associated with eight dopaminergic, radiographic, pulmonary, and dermatologic traits all localized to the same LD2 block on chromosome 17q21 and all associated with *KANSL1*-TNE1 upregulation (Fig. 5c). Six of these seemingly disparate traits are clinically implicated in multisystem features of PD. Progressive supranuclear palsy (trait 2) leads to neurodegeneration of dopamine neurons (Fig. 5c and Supplementary Table 10). Men with early-onset male pattern baldness (trait 3) have a 28% higher risk of developing PD[40]. Genetic variants for intracranial volume (traits 4 and 5) are related to PD[41], and PD patients are prone to reduced bone mineral density (trait 6)[42,43]. Thus, PD and seven clinically related traits with variants localizing to an LD block on chromosome 17q21



are associated with *KANSL1*-TNE1 expression through a uniform gene-regulatory mechanism.

This study was powered by innovations both in wet and dry lab methods and provides an online resource of mRNAs, ncRNAs, and TNE expression in dopamine and pyramidal neurons, as well as dopamine neuron-specific mRNA, ncRNA, and TNE eQTLs (BRAINcode, http://www.humanbraincode.org). Our method allows detection of the full complement of mRNAs, ncRNAs, and active enhancers in a single and minuscule RNA sample and combines the base-pair resolution and a comprehensive genome-wide view afforded by ultradeep total RNA-sequencing with the positional and cytoarchitectural information afforded by traditional light microscopy. It can be transferred to other morphologically or regionally defined brain and peripheral cells of critical relevance to health and disease. Moreover, the three-in-one approach (detecting three types of RNAs: TNEs, mRNAs, and ncRNAs) offers simplicity and noise reduction compared to approaches relying on separate methodologies, experiments, and source materials for assaying enhancers and mRNAs. lcRNAseq offers advantages to RNA sequencing of brain region homogenates (a suspension of all types of glial, neuronal, immune, and vascular cells resident in a tissue block) or of sorted nuclei without precise information on their three-dimensional origins in human brain and morphological features[44,45]. Conversely, fluorescent in situ sequencing (FISSEQ) and other in situ hybridization-based methods preserve valuable positional information, but the number of transcripts probed has been limited[46].

This analysis showed that putative enhancers active in dopamine neurons link genetic variation to neuropsychiatric traits. It has clear applications for the genetics of more than 20 million patients in the United States alone with perturbed dopamine systems, in narrowing the search window for functional associations and therapeutic nodes, and for defining the regulatory networks that underpin this archetype of a human brain neuron.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41593-018-0223-0.

Received: 11 September 2017; Accepted: 23 July 2018;
Published online: 17 September 2018

References

1. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* **10**, 184–194 (2009).
2. Birney, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007). ENCODE Project Consortium et al..
3. Heintzman, N. D. et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
4. Kowal, S. L., Dall, T. M., Chakrabarti, R., Storm, M. V. & Jain, A. The current and projected economic burden of Parkinson's disease in the United States. *Mov. Disord.* **28**, 311–318 (2013).
5. Cloutier, M. et al. The economic burden of schizophrenia in the United States in 2013. *J. Clin. Psychiatry* **77**, 764–771 (2016).
6. National Institute of Drug Abuse. Treatment Statistics. *DrugAbuse.gov* https://www.drugabuse.gov/sites/default/files/drugfacts_treatmentstats.pdf (2011).
7. Hassan, A. & Benarroch, E. E. Heterogeneity of the midbrain dopamine system. *Neurology* **85**, 1795–1805 (2015).
8. Zheng, B. et al. PGC-1α, a potential therapeutic target for early intervention in Parkinson's disease. *Sci. Transl. Med.* **2**, 52ra73 (2010).
9. Liang, W. S. et al. Neuronal gene expression in non-demented individuals with intermediate Alzheimer's disease neuropathology. *Neurobiol. Aging.* **31**, 549–566 (2010).
10. Elstner, M. et al. Neuromelanin, neurotransmitter status and brainstem location determine the differential vulnerability of catecholaminergic neurons to mitochondrial DNA deletions. *Mol. Brain* **4**, 43 (2011).
11. Djebali, S. et al. Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
12. Kim, T.-K. et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
13. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
14. Core, L. J. et al. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).
15. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
16. Heintzman, N. D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
17. Visel, A. et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
18. Yip, K. Y. et al. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome. Biol.* **13**, R48 (2012).
19. Engström, P. G., Fredman, D. & Lenhard, B. Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. *Genome. Biol.* **9**, R34 (2008).
20. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
21. Akbarian, S. et al. The PsychENCODE project. *Nat. Neurosci.* **18**, 1707–1712 (2015).
22. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 1–9 (2015).
23. Mittal, S. et al. β2-Adrenoreceptor is a regulator of the α-synuclein gene driving risk of Parkinson's disease. *Science* **357**, 891–898 (2017).
24. Scherzer, C. R. et al. GATA transcription factors directly regulate the Parkinson's disease-linked gene alpha-synuclein. *Proc. Natl. Acad. Sci. USA.* **105**, 10907–10912 (2008).
25. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
26. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
27. Ellingsen, S. et al. Large-scale enhancer detection in the zebrafish genome. *Development* **132**, 3799–3811 (2005).
28. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–D92 (2007).
29. Welter, D. et al. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
30. Jiang, Y. et al. A genetic screen to assess dopamine receptor (DopR1) dependent sleep regulation in *Drosophila*. *G3 (Bethesda)* **6**, 4217–4226 (2016).
31. González, S. et al. Circadian-related heteromerization of adrenergic and dopamine D_4 receptors modulates melatonin synthesis and release in the pineal gland. *PLoS Biol.* **10**, e1001347 (2012).
32. Breen, D. P. et al. Sleep and circadian rhythm regulation in early Parkinson disease. *JAMA Neurol.* **71**, 589–595 (2014).
33. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
34. Nica, A. C. et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6**, e1000895 (2010).
35. 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).
36. Wu, H. et al. Tissue-specific RNA expression marks distant-acting developmental enhancers. *PLoS Genet.* **10**, e1004610 (2014).
37. Mercer, E. M. et al. Multilineage priming of enhancer repertoires precedes commitment to the B and myeloid cell lineages in hematopoietic progenitors. *Immunity* **35**, 413–425 (2011).
38. Ostuni, R. et al. Latent enhancers activated by stimulation in differentiated cells. *Cell* **152**, 157–171 (2013).
39. Koolen, D. A. et al. Mutations in the chromatin modifier gene *KANSL1* cause the 17q21.31 microdeletion syndrome. *Nat. Genet.* **44**, 639–641 (2012).
40. Li, R. et al. Six novel susceptibility loci for early-onset androgenetic alopecia and their unexpected association with common diseases. *PLoS Genet.* **8**, e1002746 (2012).
41. Adams, H. H. H. et al. Novel genetic loci underlying human intracranial volume identified through genome-wide association. *Nat. Neurosci.* **19**, 1569–1582 (2016).
42. Torsney, K. M. et al. Bone health in Parkinson's disease: a systematic review and meta-analysis. *J. Neurol. Neurosurg. Psychiatry* **85**, 1159–1166 (2014).
43. Ding, H. et al. Unrecognized vitamin D3 deficiency is common in Parkinson disease: Harvard Biomarker Study. *Neurology* **81**, 1531–1537 (2013).

44. Saliba, A. E., Westermann, A. J., Gorski, S. A. & Vogel, J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.* **42**, 8845–8860 (2014).
45. Lake, B. B. et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586–1590 (2016).
46. Lee, J. H. et al. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.* **10**, 442–458 (2015).

Acknowledgements

We thank H. Suzuki and T. Suzuki of RIKEN for providing the modified pGL4.10_mod3_EF1α vector and consultation. We are grateful to C. Vanderburg of the Advanced Tissue Resource Center, Massachusetts General Hospital, for his expertise and support. We thank Z. Weng at the University of Massachusetts Medical School for sharing additional data from the ENCODE consortium. We thank A. Sandelin and R. Andersson, both from Copenhagen University; A. Regev, Broad Institute; and M. Feany, Brigham & Women's Hospital, for insightful comments and guidance. We thank C. Liu, A. Shieh, and T. Goodman for assisting in extracting the RNA-seq and ATAC-seq data in the BrainGVEX dataset. We gratefully acknowledge the Banner Sun Health Institute, Massachusetts Alzheimer's Disease Research Center at Massachusetts General Hospital, Harvard Brain Tissue Resource Center at McLean Hospital, University of Kentucky ADC Tissue Bank, University of Maryland Brain and Tissue Bank, Pacific Northwest Dementia and Aging Neuropathology Group at University of Washington Medicine Center, and Neurological Foundation of New Zealand for providing human brain tissue. This study was funded in part by NIH grant U01 NS082157 and the US Department of Defense (to C.R.S.); NIH R01AG057331 (to C.R.S.) funded RNA-seq of pyramidal neurons; with additional support from the Michael J. Fox Foundation (MJFF) (to C.R.S. and C.H.A., respectively); the Australia NHMRC GNT1067350 (to A.A.C. and J.S.M.); NIA P30 AG028383 (to P.T.N.); UK Wellcome Trust Investigator award (to F.M.); NINDS U24 NS072026 National Brain and Tissue Resource for Parkinson's Disease and Related Disorders (to T.G.B. and C.H.A.); NIA P50 AG005134 (to M.P.F.). The MSBB data were generated as part of the AMP-AD Consortium from postmortem brain tissue collected through the Mount Sinai VA Medical Center Brain Bank and were provided by E. Schadt from Mount Sinai School of Medicine. PsychENCODE data were generated as part of the PsychENCODE Consortium, supported by grants U01MH103339, U01MH103365, U01MH103392, U01MH103340, U01MH103346, R01MH105472, R01MH094714, R01MH105898, R21MH102791, R21MH105881, R21MH103877, and P50MH106934 awarded to S. Akbarian (Icahn School of Medicine at Mount Sinai), G. Crawford (Duke), S. Dracheva (Icahn School of Medicine at Mount Sinai), P. Farnham (USC), M. Gerstein (Yale), D. Geschwind (UCLA), T.M. Hyde (LIBD), A. Jaffe (LIBD), J.A. Knowles (USC), C. Liu (UIC), D. Pinto (Icahn School of Medicine at Mount Sinai), N. Sestan (Yale), P. Sklar (Icahn School of Medicine at Mount Sinai), M. State (UCSF), P. Sullivan (UNC), F. Vaccarino (Yale), S. Weissman (Yale), K. White (U Chicago), and P. Zandi (JHU).

Author contributions

X.D. performed data analysis with important contributions from D.G., B.G., G.L., C.B., and T.W. T.G.B., C.H.A., M.P.F., P.T.N., J.C.H., R.L.M.F., and C.R.S. obtained and clinically and neuropathologically characterized patient samples. Z.L. and D.G. were responsible for laser-capture and RNA-seq data production. Z.L. and Y.B. performed validation experiments. F.M. and Y.H. designed and performed zebrafish experiments. C.B., P.R., and P.H. performed CAGE experiments. C.R.S. and X.D. wrote the paper with input from all other authors. C.R.S., J.S.M., F.M., A.A.C., and J.J.L. oversaw data analysis and interpretation. C.R.S. conceived, designed, analyzed, and interpreted the study.

Competing interests

C.R.S. has collaborated with Pfizer and Sanofi; has consulted for Sanofi; has served as Advisor to the Michael J. Fox Foundation, NIH, and Department of Defense; is on the Scientific Advisory Board of the American Parkinson Disease Association; received funding from the NIH, the US Department of Defense, the Michael J. Fox Foundation, and the American Parkinson Disease Association; and is named as co-inventor on two US patent applications on biomarkers for PD held in part by Brigham & Women's Hospital. B.G. is the founder of Pacific Analytics PTY LTD, Australia and is a founding member of the International Cerebral Palsy Genetics Consortium, a member of the Australian Genomics Health Alliance, and is on the Scientific Advisory Board of Iggy Get Out!, Australia. T.G.B. provides consultancies to Prothena and GSK; is on the Advisory Board of Vivid Genomics; and has contracted research with Avid Radiopharmaceuticals, Navidea Biopharmaceuticals, and Aprinoia Therapeutics. The other authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41593-018-0223-0.

Correspondence and requests for materials should be addressed to C.R.S.

Methods

Sample collection and processing. We started with 107 high-quality frozen postmortem human control brain samples identified from Banner Sun Health Institute, Brain Tissue Center at Massachusetts General Hospital, Harvard Brain Tissue Resource Center at McLean Hospital, University of Kentucky ADC Tissue Bank, University of Maryland Brain and Tissue Bank, Pacific Northwest Dementia and Aging Neuropathology Group (PANDA) at University of Washington Medicine Center, and Neurological Foundation of New Zealand Human Brain Bank. Detailed quality measures and demographic characteristics of these samples are shown in Supplementary Table 1. Median RNA integrity numbers (RIN) were 7.8, 7.8, and 7.2 for substantia nigra samples (used to laser-capture dopamine neurons), temporal cortex (used to laser-capture temporal cortex pyramidal neurons), and motor cortex samples (used to laser-capture motor cortex pyramidal neurons), indicating high RNA quality. Median postmortem intervals were exceptionally short, with 3 h for substantia nigra, 3 h for temporal cortex, and 13 h for motor cortex samples, further consistent with highest sample quality (Supplementary Table 1).

The 107 brain samples represented 93 subjects without clinicopathological diagnosis of a neurodegenerative disease meeting the following stringent inclusion and exclusion criteria. Inclusion criteria: (i) absence of clinical or neuropathological diagnosis of a neurodegenerative disease, for example, PD according to the UKPDBB criteria[47], Alzheimer's disease according to NIA-Reagan criteria[48], or dementia with Lewy bodies by revised consensus criteria[49]; for the purpose of this analysis incidental Lewy body cases (not meeting clinicopathological diagnostic criteria for PD or other neurodegenerative disease) were accepted for inclusion; (ii) PMI $\leq$ 48 h; (iii) RIN[50] $\geq$ 6.0 by Agilent Bioanalyzer (good RNA integrity); and (iv) visible ribosomal peaks on the electropherogram. Exclusion criteria were: (i) a primary intracerebral event as the cause of death; (2) brain tumor (except incidental meningiomas); (3) systemic disorders likely to cause chronic brain damage. We also included eight non-brain tissue samples as controls, including five samples of peripheral blood mononuclear cell (PBMC) and three fibroblasts (FB), provided by Harvard Biomarker Study and Coriell Institute. This study was approved by the Institutional Review Board of Brigham and Women's Hospital.

We then performed laser-capture microdissection (LCM) on the brain samples to extract neurons from different brain regions. LCM was performed similarly to previously reported procedures by us and others[8,51–53]. For each substantia nigra sample, 300–800 dopamine neurons, readily visualized in HistoGene-stained frozen sections based on hallmark neuromelanin granules, were laser-captured using the Arcturus Veritas Microdissection System (Applied Biosystems). For each temporal cortex (middle gyrus) or motor cortex sample, about 300 pyramidal neurons were outlined in layers V/VI by their characteristic size, shape, and location in HistoGene-stained frozen sections and laser-captured using the Arcturus Veritas Microdissection System (Applied Biosystems). Total RNA was isolated and treated with DNase (Qiagen) using the Arcturus Picopure method (Applied Biosystems), yielding approximately 7–8 ng RNA per subject. Total RNA was linearly amplified into 5–10 µg of double-stranded cDNA using the validated, precise, isothermal RNA amplification method implemented in the Ovation RNA-seq System V2 (NuGen)[54,55]. Unlike PCR-based methods that exponentially replicate original transcript and copies, with this method only the original transcripts are linearly replicated[54,55], and amplification is initiated at the 3′ end as well as randomly, thus allowing for amplification of both mRNA and nonpolyadenylated transcripts[54,55]. Sequencing libraries were generated from 500 ng of the double-stranded (ds) cDNA using the TruSeq RNA Library Prep Kit v2 (Illumina), according to the manufacturer's protocol. The cDNA was fragmented, and end repair, A-tailing, adaptor ligation were performed for library construction. Sequencing library quality and quantity control was performed using the Agilent DNA High Sensitivity Chip and qPCR quantification, respectively. Libraries were sequenced (50 or 75 cycles, paired-end) on Illumina HiSeq 2000 and 2500 at the Harvard Partners Core.

Genotyping and imputation. Each sample was genotyped using the Infinium Omni2.5Exome-8 BeadChips (Illumina), which includes more than 2.5 million tagged SNPs from the HapMap and 1000 Genomes Project. The total 98 samples from 93 subjects were genotyped in three batches, with technical replicates for five subjects. We computed the pairwise IBD of genotypes between replicates using PLINK2, and reached an average proportion of 0.9991 IBD. Thus, we kept unique sample and replicates in batch 1 for further quality control analysis.

We applied PLINK2[56] (v1.9beta) and in-house scripts to perform rigorous subject and SNP quality control (QC; Supplementary Fig. 13a) that included (i) SNP GC score filtering, (ii) subject call rates, (iii) gender misidentification, (iv) genotype call rates, (v) Hardy–Weinberg equilibrium testing, (vi) test mishaps, (vii) heterozygosity outliers, and (viii) IBS/IBD filtering. In total, we excluded 5,249 SNPs with GC < 0.25; 1,955 SNPs not in the genome assembly we used (hg19); 20,049 SNPs with call rates < 95%; 57 SNPs with Hardy–Weinberg equilibrium $P < 10^{-6}$; 1,295,546 SNPs with MAF < 0.05; and two subjects with IBS/IBD PI_HAT > 0.9. In total, 91 subjects with 1,235,673 SNPs passed QC.

We employed SHAPEIT2[57] (v2.5) to perform pre-phasing and then IMPUTE2[57] (v2.3.1) to impute the post-QC genotyped markers in autosomal chromosomes using reference haplotype panels from the 1000 Genomes Project (Phase 3), which include a total of 77.8 million SNPs in 2,504 individual samples. For genotyped markers in chromosome X, we used the 1000 Genomes Project Phase I Integrated Release Version 3 as reference haplotype in 1,092 individuals. The genotyped calls of imputed genotypes with posterior probability < 0.9 were marked as missing, and we kept biallelic genotypes for further analysis. After genotype imputation, we filtered out imputed SNPs with MAF < 0.05 and info metric < 0.5 that had been compared in a previous review[58], which resulted in 4,889,047 imputed SNPs. In total 6,124,720 SNPs were passed to downstream eQTL analysis.

RNA sequencing data analysis pipeline. RNA-seq raw files in FASTQ format were processed in a customized pipeline. For each sample, we first filtered out reads that failed vendor check or were too short (<15 nt), after removing the low-quality ends and possible adaptor contamination using fastq-mcf with options "-t 0 –x 10 –l 15 –w 4 –q 10 –u". We then checked the quality using FastQC and generated *k*-mer profile using kpal[59] for the remaining reads. Reads were then mapped to the human genome (GRCh37/hg19) using Tophat[60] (v2.0.8) by allowing up to two mismatches and 100 multiple hits. Reads mapped to ribosomal RNAs or to the mitochondrial genome were excluded from downstream analysis. Gene expression levels were quantified using FPKM (fragments per kilobase of transcript per million mapped reads). Only uniquely mapped reads were used to estimate FPKM. To calculate normalized FPKM, we first ran Cuffquant[61] (v2.2.1) with default arguments for genes annotated in GENCODE (v19), and then ran Cuffnorm with parameters "-total-hits-norm –library-norm-method quartile" on the CBX files generated from Cuffquant.

Sample QC based on RNA-seq data. We performed sample QC similarly to 't Hoen PA et al.[62]. In brief, we ran *k*-mer profiling for filtered reads using kpal[59], and calculated the median profile distance for each sample. Samples with distances clearly different from the rest of the samples were marked as outliers (Supplementary Fig. 1c). We also calculated pair-wise Spearman correlations of gene expression quantification across samples and measured the median correlation (*D*-statistics) for each sample (Supplementary Fig. 1b,d). Samples with *D*-statistics markedly different from the rest of the samples were deemed outliers. Moreover, we tested for concordance between reported clinical sex and sex indicated by the expression of female-specific *XIST* gene and male-specific Y-chromosome gene (Supplementary Fig. 1e). Samples from the first batch with a relative low sequencing depth were also excluded. In addition to these samples used for cell-type-specific transcriptome analyses, we analyzed various additional control samples (for example, amplification controls, tissue homogenate) and technical replicates (Supplementary Fig. 1f–h). In the end, 106 of 115 samples passed QC and were used for downstream analysis (Supplementary Fig. 1a).

Defining the cumulative transcribed region by RNA-seq. Previously, ENCODE reported that, in cell lines, 62.1% (cumulatively) of the genome was transcribed with at least five mapped reads (Supplementary Table 11 of ref.[11]). In our study, we rigorously accounted for sequencing depth and thus considered a genomic sequence as transcribed only if it had a read coverage of more than 0.05 RPM (unique reads per million). This corresponds to approximately 10 mapped reads (considering that for each sample we had, on average, 178 million mapped reads). With this rigorous definition, we showed that the cumulative coverage of transcribed regions in the dopamine neuron samples is 64.4%.

Defining catalogs of expressed ncRNAs and mRNAs. Normalized expression values of the 106 samples that passed QC were used as input. We first excluded genes with FPKM of zero in all 106 samples. Next, surrogate variable analysis and batch adjustment was performed using the sva[63] and ComBat[64] packages in R. In brief, the FPKM values were $\log_{10}$-transformed after adding a pseudocount of 0.0001. FPKM values within each group were adjusted for age, sex, and RIN, as well as hidden covariates, using frozen surrogate variable analysis (sva[63]). ComBat[64] was used to adjust for batch effects. Median expression values for each gene were calculated for each cell type. To rigorously exclude low-abundance genes, genes with median adjusted FPKM < 0.01 in a cell type were not considered expressed in that cell type. GENCODE genes meeting these criteria were used to create a detailed catalog of mRNAs and ncRNAs expressed in a cell type.

Genes 'exclusive' to dopamine neurons, pyramidal neurons, or non-neuronal cells, respectively, were defined as those that achieved a median adjusted FPKM $\geq$ 0.01 in only one of these three cell types (with adjusted FPKM < 0.01 in each of the other two cell types). We used the t-SNE package in R for *t*-distributed stochastic neighbor embedding analysis and the heatmap2 package for clustering and visualization purposes of cell-type-exclusive ncRNAs and mRNAs.

Definition of TNE regions. A schematic of the TNE identification pipeline is shown in Fig. 1d and a flow chart in Supplementary Fig. 14a. TNE identification analysis was performed separately for dopamine neurons, pyramidal neurons, and non-neuronal cells. We first calculated the reads density values (in RPMs) at each genomic nucleotide position for all samples. We then calculated the aggregation signal for each cell type by computing the trimmed mean (for example, trimming the 10% highest and lowest data points) of RPMs across the total *n* samples

from the cell type of interest for each nucleotide position. We then scanned this aggregation signal in UCSC BegGraph format with a six-step filter:

(1) Scan each nucleotide position to filter for (keep for analysis) genomic regions with RPMs higher than the background level. The background level is defined as the average read density across the nuclear genome (i.e., the sum of all RPMs in a cell type divided by the total number of base pairs comprising the nuclear genome, for example, 3,095,677,412 for hg19). The borders of the selected genomic regions for each candidate TNE site were thus defined by the first and the last nucleotide for each TNE site that met the RPM cutoff;
(2) For each candidate region from step 1, require the summit RPM (i.e., maximal RPM in the region) to achieve a detection $P \leq 0.05$ compared to transcriptional background noise. The transcriptional background was defined by randomly selecting 1,000,000 single nucleotide positions outside of the EXCLUSION region (see Methods) and calculating the distribution of their RPMs. The background signal was fitted to a normal distribution using the fitdist(x,'norm') function in R. See Supplementary Fig. 14b for the distribution of background signals. Neighboring regions were merged into one region if the genomic distance between them was less than 100 bp;
(3) Exclude any regions overlapping with the EXCLUSION regions defined below (for example, known genes, CAGE-defined promoters, and genomic gap regions);
(4) Require candidate regions to be longer than 100 bp;
(5) Exclude candidate regions containing junction sites supported by more than ten spliced reads in each of at least five samples. Junction sites were combined from the junctions.bed files of Tophat output;
(6) For candidate regions meeting these criteria, we then required statistically significant expression across samples. We first computed the mean RPM values of each candidate region and then estimate the significance (P value) compared to expression noise observed in random background regions of the sample. P values were computed by comparing the expression levels to the random background distribution of each sample, for example, $P = 1 - \mathrm{Fn}(x)$, where $\mathrm{Fn}(x)$ is the empirical cumulative distribution function of expression levels of the same number of background regions with matched length randomly picked up beyond the EXCLUSION regions. Then for each candidate region, we computed the number of samples 'called' with $P \leq 0.05$ and calculated the probability of observing this number of called samples by chance alone using a binomial distribution with the population probability set at 0.05. Finally, we rigorously corrected the binomial P values for each candidate region for the total number of tests performed using Bonferroni corrections. Candidate regions with Bonferroni-corrected $P \leq 0.05$ were considered significantly expressed in the given cell type.

Regions excluded from the construction of random background regions. We defined 'EXCLUSION' as a set of regions to exclude when constructing the random background regions. The EXCLUSION regions included any known transcribed regions (i.e., [−500, +500] bp of annotated exons from GENCODE (v19)[65], UCSC known genes, lincRNA from NONCODE (v4)[66], and rRNA from repeatMasker), FANTOM5 CAGE-defined promoters (i.e., [−500, +500] bp regions flanking the CAGE-predicted TSS), and genomic gap regions in the UCSC hg19 assembly.

The n values of background regions picked for the analysis of dopamine neurons, pyramidal neurons, or non-neuronal cells equaled the n values of 71,022, 37,007, and 19,690 TNEs detected in each of the three cell types, respectively. Background regions were randomly picked from the human genome (without the EXCLUSION regions) with length distributions matched to each TNE set.

Defining exclusive vs. shared TNEs. TNEs were further annotated into 'shared' and 'exclusive' classes depending on whether they overlapped (i.e., by at least 1 nt) with TNEs detected in the other cell types. Cell-type-exclusive TNEs were exclusively detected in one cell type. They did not overlap with TNEs detected in another cell type. TNEs detected in more than one cell type were termed shared. Infrequently, a dopamine neuron TNE overlapped with more than one pyramidal neuron TNE. Thus, in Fig. 1e the intersections between dopamine neuron TNEs and pyramidal neuron TNEs (or dopamine neuron TNE and non-neuronal TNE) show the number of dopamine neuron TNEs that physically overlap with any pyramidal neuron TNE (or non-neuronal TNE). Similarly, the intersections between pyramidal neuron TNEs and non-neuronal TNEs (not shared with dopamine neurons) show the number of pyramidal neuron TNEs that physically overlap with any non-neuronal TNE. The area-proportional Venn diagram was generated using eulerAPE[67].

Characterization of TNEs using regulatory annotations. To explore the possible role of TNEs in gene regulation, we characterized TNEs with various known regulatory data in human brain (if available) or cell lines. For example, we used chromHMM 'enhancer' states in any of the ten human brain tissues in the Roadmap Epigenomics Project for histone-defined enhancers[20,26]. Enhancers are marked as the E6, E7, or E12 states from the 15-state chromHMM segmentation defined by five core marks: H3K4me3, H3K4me1, H3K36me3, H3K27me3, and H3K9me3. The ten brain tissues are hippocampus middle, substantia nigra, anterior caudate, cingulate gyrus, inferior temporal lobe, angular gyrus, dorsolateral prefrontal cortex, germinal matrix, fetal brain female, and fetal brain male. We used DNase-seq peak called in fetal brain of the Roadmap Epigenomics Project[20] for DNase hypersensitivity sites. For TF binding, we used the TF ChIP-seq peak clusters (wgEncodeRegTfbsClusteredV3 from UCSC Genome Browser) from the ENCODE project[25,68], which contains the most comprehensive TF ChIP-seq repository (to date). Other regulatory data include EP300 binding peaks from the ENCODE project[23], CAGE-defined enhancers from the FANTOM5 project[13], and sequence conservation score (phyloP) based on 100 vertebrate genomes comparison[69].

By converting these features into binary codes (1 or 0) according to their presence or absence in TNE regions, we further built a simple classifier using these binary codes. For example, we defined a TF binding hotspot as a region containing at least 5 distinct TFs ChIP-seq peaks. An epigenomic enhancer was present if any of the chromHMM enhancer states (E6, E7, E12) overlapped with the region. For conservation, we overlapped TNEs with HCNEs (highly conserved noncoding elements) as defined in Ancora[19] and defined 'being conserved' as a TNE overlapping with an HCNE between human and zebrafish with at least 70% similarity and 50 nt in length. We built the weighted classifier with relative two-fold higher weight for DNase signal and implemented it in R using the function daisy().

GWAS SNP enrichment analysis. We first downloaded the GWAS-associated SNPs from the NHGRI-EBI GWAS catalog[70] (v1.0, downloaded on 4 November, 2015), which includes 19,188 SNP–disease/trait associations after successfully porting it back to the hg19 assembly. We then extended this set to 495,085 autosomal associations by including proxy SNPs imputed from the 1000 Genomes project. Proxy SNPs were extracted using SNAP[71] from either of three populations in the 1000 Genomes Pilot 1 dataset with distance limit of 250 kb and linkage disequilibrium (LD) r^2 threshold of 0.8. Nonassociated SNPs were extracted from dbSNP (build 137). We calculated the number of trait-associated and nonassociated SNPs that physically localized (or did not localize) to TNE, promoters (unique locations of all GENCODE v19 protein-coding gene TSSs ± 200 bp), exons (unique locations of all GENCODE v19 protein-coding gene transcript inner exons), or random regions (100,000 genomic regions of 400 bp randomly selected beyond the TNEs, FANTOM5 permissive enhancers, and EXCLUSION regions defined above), respectively. Only diseases/traits with more than three associated SNPs localizing to TNEs were considered for this analysis. For each genomic feature associated with a disease/trait with an odds ratio > 1, we performed a Fisher's exact test. P values equal to or below 9.64×10^{-6} (i.e., 0.01 divided by 1,037, the total number of diseases/traits tested in NHGRI-EBI GWAS catalog as of 4 November, 2015) were considered statistically significant.

Validating enhancer activity in HeLa S3 and neuroblastoma cells. PCR primers for the amplification of TNE-defined enhancer candidates and control regions from genomic DNA were designed using the Primer3web (v4.0.0)[72], and restriction sites SalI and BamHI were separately added to 5′ end of the sense and antisense primer. Combined primer sequences were prevalidated with UCSC's In-Silico PCR web tool and synthesized by Thermo Fisher Scientific. All primers sequences are listed in Supplementary Table 6.

The modified vector pGL4.10_mod3_EF1α was kindly provided by RIKEN, and its structure is also described in Supplementary Fig. 9d in their publication[13]. In brief, an EF1a basal promoter fragment was inserted into HindIII and NheI sites of the promoterless pGL4.10 (Promega) to construct the pGL4.10EF1a vector, and then the BamHI- and SalI-containing fragment (as the enhancer insertion site) was removed and reinserted at the SpeI site located upstream of the synthetic poly(A) signal/transcriptional pause site to generate modified versions of the pGL4.10EF1a vector. The poly(A) site was inserted between the enhancers insertion site and the basal promoter is to avoid read-through from the enhancer, since we expect that many of our test elements are transcribed.

The PCR reaction was performed in 50 µL reaction buffer to amplify each sequence of interest from 100 ng of human cerebellum tissue gDNA using a One Taq DNA polymerase Kit (New England Biolabs). The PCR product was digested with BamHI and SalI (New England Biolabs), and the restriction DNA fragment (insert) was isolated using agarose gel electrophoresis and purified by the MinElute Gel Extraction Kit (Qiagen). The pGL4.10_mod3_EF1α vector was also digested with BamHI and SalI, and the double-digested DNA (vector) was isolated and purified in the same way as the insert. Using T4 DNA Ligase (New England Biolabs), 100 ng of insert and 20 ng of vector were ligated in 10 µL reaction buffer; 1 µL of ligation reaction buffer was transferred to 100 µL of DH5α-competent cells (Invitrogen). Positive colonies were selected by colony PCR and correct insertion in the plasmid was confirmed by sequencing. Cloned plasmids for transfections were purified using the QIAamp DNA Midi Kit (Qiagen).

HeLa-S3 cells were cultured in MEM (Gibco) supplemented with 10% FBS (Gibco), 100 U/mL penicillin and streptomycin (Gibco). SK-N-MC neuroblastoma cells were cultured in DMEM (Gibco) supplemented with 10% FBS (Nichirei Bioscience Inc.), MEM (WAKO) supplemented with 10% FBS (Gibco), and 100 U/mL penicillin and streptomycin (Gibco). A total of 7.5×10^3 cells per well

of HeLa-S3 cells and 4×10^4 cells per well SK-N-MC cells were seeded in 96 well plates 24 h before transfection.

W cotransfected 190 ng of plasmids inserted with the PCR products and 10 ng of pGL4.73 Renilla luciferase plasmid (Promega) into HeLa-S3 and SK-N-MC cells, respectively, using Lipofectamine 2000 (Invitrogen), according to the manufacturer's instructions. Each transfection was performed independently three times. After 24 h, the luciferase activities were measured by a Gen5 Microplate Reader (BioTek) using the Dual-glo luciferase assay system (Promega) according to the manufacturer's instructions.

Validating enhancer activity in zebrafish. Selected TNEs with potential enhancer activity and one negative control element (a nonconserved intergenic sequence region with very low or no signal for enhancer marks such as DNase I hypersensitivity, H3K4me1, or H3K27ac) were amplified from human genomic DNA using primers (Supplementary Table 6). PCR products were purified using NucleoSpin Gel and a PCR Clean-up Kit (Macherey-Nagel) and cloned upstream of the zebrafish *gata2* promoter[73] linked to an *mRuby2* reporter gene into a modified pDB896 vector (a gift from D. Balciunas, Temple University). The cloning procedures were performed using an In-Fusion HD Cloning Kit (Clontech) according to the manufacturer's instructions, into a BamHI linearized vector. Plasmid DNA was purified using a Qiagen-tip 20 miniprep kit (Qiagen) and verified by restriction digest and sequencing.

Zebrafish stocks (*Danio rerio*) were kept and used according to Home Office regulations (UK) at the University of Birmingham. For these experiments, the enhancer trap transgenic line ETvmat2:GFP[74] was used. Adults were crossed pairwise and eggs were collected and injected within 20 min after fertilization. Microinjection solutions contained 20 ng/μL of plasmid DNA and 0.1% of phenol red (Sigma). Injections were performed through the chorion and into the cytoplasm of zygotes using an analog microinjector MINJ-1 (Tritech Research). About 150–200 eggs were injected per construct, and experiments were replicated at least three times. Embryos were kept in E3 Medium containing 50 μg/mL gentamicin (Thermo Fisher Scientific) and 0.003% phenylthiourea (Sigma) at 28.5 °C.

Injected embryos were screened for expression during the first 5 d postfertilization and group images were taken on Zeiss Axio Zoom V16 stereo microscope. Selected embryos showing specific expression pattern were imaged at the relevant developmental stage on a Zeiss Lightsheet Z1 microscope with 20× objective and 0.5 optical zoom. Stacks containing 250–300 slices with 2-μm thickness were acquired, and maximum intensity projections were made using Zeiss ZEN Black Software.

eQTL analysis pipeline. The eQTL analysis was performed for both GENCODE genes and TNEs using the 84 subjects for which lcRNAseq data from dopamine neurons as well as genotyping data were available. For genes, we first filtered for genes with FPKM > 0.05 in at least ten individuals, then transformed FPKM to rank-normalized gene expression. In brief, the FPKM values were $\log_{10}$-transformed (adding a pseudocount of 0.01). The measurements for each gene were transformed into a normal distribution while preserving relative rankings (quantile normalization) and the mean and s.d. of the original measurement. For TNEs, the expression distribution was close to a normal distribution and thus quantile normalization was not indicated. Moreover, our TNE identification method already selected for TNEs pervasively expressed across multiple individuals. We then performed surrogate variable analysis (SVA) with the sva R package[63] to adjust for the effects of known covariates, including batch, age, gender, RIN, PMI, and read-length. Adjusted expression values extracted from fsva() function were used for downstream eQTL analysis. We used RLE (relative log expression) plots to visually inspect the effects of covariate adjustment. We also filtered out SNPs with missing values or with MAF ≤ 0.05 in the 84 subjects. Matrix-eQTL[33] was applied for *cis*-eQTL analysis, with the *cis* window defined as 1 megabase between the SNP and the nearest end of a gene or TNE. Nominal *P* values were generated for SNP–gene pairs in linear regression mode. See Supplementary Fig. 13b for detail.

TNE–host gene function enrichment analysis. We found that 151 *cis*-regulated TNEs physically localized to introns of 102 host genes. Gene-set enrichment analysis was performed using the C5 gene sets (GO terms) implemented in the MSigDB database using the hypergeometric test. Each gene set contained genes annotated to the same GO term. For each gene set, the hypergeometric test was performed for $k - 1$, K, $N - K$, and n, where k is the number of TNE host genes that are part of a GO term gene set, K is the total number of genes annotated to the same GO term gene set, N is the total number of all known human genes, and n is the number of genes in the query set. The top 50 GO terms enriched in these TNE host genes are shown in Supplementary Table 8 (all with FDR $q < 0.05$).

We also evaluated whether there was specific enrichment among *cis*-regulated TNEs in genes associated with brain disorders. We used diseases in MeSH C10 (nervous system diseases) or F03 (mental disorders) for brain disorders, and associated disease to genes using GenDisNet database. The disease–gene association was extracted from DisGeNet[75] (http://www.disgenet.org/) filtered with GDA > 0.1. For all annotated protein-coding genes, we performed Fisher's exact test based on whether a gene was associated with brain disorder and a gene hosted a *cis*-eQTL TNE (Supplementary Table 9).

TF binding motif enrichment analysis. For TFs with ENCODE ChIP-seq, we extracted their peak coordinates from the wgEncodeRegTfbsClusteredV3 file downloaded from UCSC Genome Browser, which contains 4,380,444 TF binding peaks from 161 TFs in total. We also downloaded 579 nonredundant TF motifs in vertebrate from JASPAR (version 2018)[76] and then scanned the whole genome with the motifs using the program FIMO[77] (default parameters with $P < 10^{-4}$) to get 418,034,884 putative binding sites. For each TF, Fisher's exact test was performed to determine whether observed occurrences of TF peaks (for ENCODE ChIP-seq) or binding sites (for JASPAR motifs) in TNEs were significantly enriched more than expected. In brief, for each TF in JASPAR, we assigned the full set of putative binding regions of JASPAR motifs to one cell of the 2×2 table according to whether a region was bound by the TF or not and whether it overlapped with TNE or not. So, each TF had a 2×2 table for Fisher's exact test. This was similarly done for ENCODE TF ChIP-seq peaks.

We also tested the TF motif enrichment against random genomic sequences that were GC- and length-matched. We first extracted the GC- and length-matched random genomic background regions using the GC_compo (http://opossum.cisreg.ca/GC_compo/), and then tested motif enrichment using the AME program in the MEME suite for all 579 nonredundant TF motifs in vertebrates from JASPAR CORE 2018.

Causality analysis for TNE, ncRNA, and mRNA eQTLs. We used the relative trait concordance (RTC) method to integrate QTL and GWAS data to detect potential disease-causing *cis*-regulatory effects according to the method described in ref.[34]. Using this method, an RTC score of 1 or near-1 indicates a potentially causal *cis*-regulatory effect.

To reduce redundancy in the output of the RTC analysis due to SNPs in strong LD, we pruned the result using the following rules. If multiple eSNPs shared the same LD block with a GWAS SNP, we only took the eSNP with best RTC score for each GWAS variant–transcript pair. If multiple eSNPs achieved the exact same top RTC score and they included the GWAS-derived variant itself, we selected the GWAS variant as the top eSNP. If multiple variants achieved the exact same top RTC score (but did not include the GWAS-derived variant itself), then we arbitrarily picked one of these top-scoring eSNPs as a representative eSNP. The pruned result is shown in Supplementary Table 10.

Three haplotype blocks were defined for the chr17q21 locus by plink2 (plink –blocks–blocks-max-kb 1000) using the CEU subpopulation ($n = 99$) in the 1000 Genome Project. Conditional eQTL analysis was performed for the chr17q21 locus by including the rs17649553 genotype as an additional covariate. All eQTL pairs for genes/TNEs and SNPs in the locus (chr17:43,000,000–45,300,000 in hg19) are displayed in Supplementary Fig. 11. The majority of significant eQTL SNPs became insignificant after conditional analysis of rs17649553, except for 31 SNPs in *KANSL1* (green dots on the top right corner) and one SNP in *NSF* (red dot on the top right corner). The 31 SNPs are in the same LD block as rs17649553.

Confirming TNE and mRNA expression by qPCR. Quantitative PCR was performed using SYBR Green Master Mix (Life Technologies) on an ABI 7900HT instrument (Applied Biosystems). Primer sequences are shown in Supplementary Table 6. To confirm the expression of lcRNAseq-derived TNE and mRNAs in dopamine neurons and pyramidal neurons, relative abundances of target TNE or mRNAs were evaluated by qPCR in linearly amplified laser-captured, microdissected samples from human substantia nigra or temporal cortex, as well as in linearly amplified human fibroblast and PBMC samples (Fig. 3b). TNE and mRNA expression was further confirmed in SK-N-MC human neuroblastoma cells and Human Universal Reference RNA (not shown). The human reference gene *GUSB* was used to normalize for RNA loading. Control samples lacking template and those lacking reverse transcriptase showed virtually no expression of these target TNEs, and mRNAs indicating that primer dimers or DNA contamination did not materially influence results. Expression values were analyzed using the comparative threshold cycle method[24]. Equal amplification efficiencies for target and reference transcripts were confirmed using melting curve analysis.

qPCR evaluation of chromosome 17q21 eQTL in a second, independent cohort of 31 individuals. Postmortem brain samples from 31 individuals were analyzed. These individuals were without a clinical or neuropathological diagnosis of neurodegenerative disease and met the inclusion and exclusion criteria described in the "Sample collection and processing" section. These new brain samples were obtained from Banner Sun Health Institute, Brain Tissue Center at Massachusetts General Hospital, and University of Kentucky ADC Tissue Bank. Pyramidal neurons were laser-captured from the middle temporal gyrus of each of the 31 individuals and linearly amplified as described in the "Sample collection and processing" section. These samples showed exceptional quality, as documented by a median RNA integrity number 7.7 and a median postmortem interval of 2.9 h (Supplementary Table 12). Relative expression abundances of the two target transcripts, *KANSL1*-TNE1 and *LRRC37A4P*, were assayed using SYBR Green qPCR (Life Technologies). The geometric mean of two reference genes, *EIF4A2*

and *RPL13*, was used to control for RNA loading. Control samples lacking template and those lacking reverse transcriptase showed virtually no detectable expression. Relative expression abundance of each of the target genes was compared in subjects carrying one or two risk alleles (CT or TT) and those without risk allele (CC) at rs17649553. A two-tailed Student's homoscedastic *t* test was used to determine statistical significance. Data are visualized in Supplementary Fig. 12.

Technical confirmation of lcRNAseq eQTL results in laser-captured dopamine neurons by qPCR. We confirmed the lcRNAseq-based dopamine neuron eQTLs for *KANSL1*-TNE1 and *LRRC37A4P*, respectively, using SYBR Green qPCR (Life Technologies). The geometric mean of two reference genes, *EIF4A2* and *RPL13*, was used to control for RNA loading. For this confirmatory experiment, laser-captured dopamine neuron samples from 35 substantia nigra samples (also used for lcRNAseq) were analyzed. Data are visualized in Supplementary Fig. 12.

Postmortem brain CAGE methods. Four human postmortem brains (healthy controls) were obtained from University of Maryland, University of Washington, and McLean Hospital, with the same inclusion/exclusion criteria as described above. Substantia nigra tissue samples were used for cap analysis gene expression (CAGE). We extracted 5 µg of total RNA from each sample using the RNeasy RNA Kit (Qiagen) with RNA integrity number (RIN) > 6. Use of postmortem samples for expression analysis was approved by the IRB of Brigham & Women's Hospital.

Libraries were constructed using a published CAGEseq protocol adapted for next-generation sequencing[78]. Briefly, cDNA was synthesized from total RNA using random primers, and this process was carried out at high temperature in the presence of trehalose and sorbitol to extend cDNA synthesis through GC-rich regions in 5′ untranslated regions. The 5′ ends of messenger RNA within RNA–DNA hybrids were selected by the cap-trapper method and ligated to a linker so that an EcoP15I recognition site was placed adjacent to the start of the cDNA, corresponding to the 5′ end of the original mRNA. This linker was used to prime second-strand cDNA synthesis. Subsequent EcoP15I digestion released the 27-base pair (bp) CAGEseq reads. After ligation of a second linker, CAGEseq tags were polymerase-chain reaction amplified, purified, and sequenced on the HiSeq 2000 (Illumina) using standard protocol for 50-bp single-end runs.

CAGEseq data were filtered for CAGEseq artifacts using TagDust[79] (version 1.12), removal of reads mapping to known ribosomal RNA genes and low-quality reads, mapping to the human genome (hg19) using Burrows–Wheeler Aligner (version 0.5.9) for short reads. Reads mapping to autosomes were used to minimize gender and normalization biases for subsequent analysis. Normalization was done based on the amount of reads per million sequence reads.

Data collection, statistical analysis, and data presentation. Sample sizes were based on the total number of available high-quality brain samples that met inclusion and exclusion criteria. No statistical methods were used to predetermine sample sizes, but our sample sizes are consistent with those recommended by the Genotype-Tissue Expression Consortium[80]. No randomization of data collection was performed in this study. Brains were selected based on predefined inclusion and exclusion criteria (see above). Sample outliers were rationally identified as described in the section "Sample QC based on RNA-seq data." TNEs were defined in a rigorous six-step process as detailed in the section "Definition of TNE regions." Data were not excluded based on arbitrary post hoc considerations. Data collection and analysis were not performed blind to the conditions of the experiments. Data distribution was assumed to be normal, but this was not formally tested, except that the normality of transcriptional background signal was checked by visual inspection.

R (The R Foundation for Statistical Computing, Vienna, Austria) was used for other statistical tests. Box plots were used to present multigroup comparisons. In all box plots, center line represents the median value; box limits, first and third quartiles; whiskers, the most extreme data point that is no more than 1.5 times the interquartile range from the box.

Statistical tests used in each figure: Fig. 4b, two-tailed Student's *t* test; Fig. 4h, hypergeometric test; Fig. 5b, one-sided Fisher's exact test; Fig. 5c, linear regression model in Matrix-eQTL; Fig. 5d, meta-GWAS from http://www.pdgene.org/; Fig. 5g, two-sided Student's *t* test; Supplementary Fig. 5b–d, one-sided Fisher's exact test; Supplementary Fig. 6c, hypergeometric test; Supplementary Fig. 10a, linear regression model in Matrix-eQTL; Supplementary Fig. 10b, linear regression model in Matrix-eQTL (for three-group comparisons) or two-sided Student's *t* test (for two-group comparisons); Supplementary Fig. 11, linear regression model in Matrix-eQTL; Supplementary Fig. 12, two-sided Student's *t* test.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability. Custom code associated with this study is available upon reasonable request.

Data availability

RNA-seq and genotyping raw data have been deposited in dbGAP under accession number phs001556.v1.p1. The supporting data and eQTL results for the BRAINcode project can be queried at http://www.humanbraincode.org through a user-friendly interface. Other data supporting the findings of this study are available upon reasonable request.

References

47. Hughes, A. J., Daniel, S. E., Kilford, L. & Lees, A. J. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. *J. Neurol. Neurosurg. Psychiatry* **55**, 181–184 (1992).
48. The National Institute on Aging and Reagan Institute Working Group on Diagnostic Criteria for the Neuropathological Assessment of Alzheimer's Disease. Consensus recommendations for the postmortem diagnosis of Alzheimer's disease. *Neurobiol. Aging.* **18 Suppl**, S1–S2 (1997).
49. Bonanni, L., Thomas, A., Onofrj, M. & McKeith, I. G. Diagnosis and management of dementia with Lewy bodies: third report of the DLB Consortium. *Neurology* **66**, 1455 (2006). author reply 1455.
50. Schroeder, A. et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* **7**, 3 (2006).
51. Unni, V. K., Ebrahimi-Fakhari, D., Vanderburg, C. R., McLean, P. J. & Hyman, B. T. Studying protein degradation pathways in vivo using a cranial window-based approach. *Methods* **53**, 194–200 (2011).
52. Ingelsson, M. et al. No alteration in tau exon 10 alternative splicing in tangle-bearing neurons of the Alzheimer's disease brain. *Acta Neuropathol.* **112**, 439–449 (2006).
53. Liu, G. et al. Metal exposure and Alzheimer's pathogenesis. *J. Struct. Biol.* **155**, 45–51 (2006).
54. Kurn, N. et al. Novel isothermal, linear nucleic acid amplification systems for highly multiplexed applications. *Clin. Chem.* **51**, 1973–1981 (2005).
55. Faherty, S. L., Campbell, C. R., Larsen, P. A. & Yoder, A. D. Evaluating whole transcriptome amplification for gene profiling experiments using RNA-seq. *BMC Biotechnol.* **15**, 65 (2015).
56. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
57. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
58. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
59. Anvar, S. Y. et al. Determining the quality and complexity of next-generation sequencing data without a reference genome. *Genome. Biol.* **15**, 555 (2014).
60. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
61. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
62. 't Hoen, P. A. C. et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.* **31**, 1015–1022 (2013).
63. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, 1724–1735 (2007).
64. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
65. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
66. Zhao, Y. et al. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.* **44 D1**, D203–D208 (2016).
67. Micallef, L. & Rodgers, P. eulerAPE: drawing area-proportional 3-Venn diagrams using ellipses. *PLoS One* **9**, e101717 (2014).
68. Wang, J. et al. Factorbook. org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res.* 41, (D171–D176 (2013).
69. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
70. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45 D1**, D896–D901 (2017).
71. Johnson, A. D. et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938–2939 (2008).
72. Untergasser, A. et al. Primer3–new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).
73. Meng, A., Tang, H., Ong, B. A., Farrell, M. J. & Lin, S. Promoter analysis in living zebrafish embryos identifies a cis-acting motif required for neuronal expression of GATA-2. *Proc. Natl. Acad. Sci. USA.* **94**, 6267–6272 (1997).
74. Wen, L. et al. Visualization of monoaminergic neurons and neurotoxicity of MPTP in live transgenic zebrafish. *Dev. Biol.* **314**, 84–92 (2008).

75. Piñero, J. et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45 D1**, D833–D839 (2017).
76. Mathelier, A. et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **44 D1**, D110–D115 (2016).
77. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
78. Takahashi, H., Lassmann, T., Murata, M. & Carninci, P. 5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protoc.* **7**, 542–561 (2012).
79. Lassmann, T., Hayashizaki, Y. & Daub, C. O. TagDust–a program to eliminate artifacts from next generation sequencing data. *Bioinformatics* **25**, 2839–2840 (2009).
80. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

nature research

Corresponding author(s): Clemens Scherzer

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	All code for the data collection is available at https://github.com/sterding/BRAINcode.
Data analysis	PLINK2 (v1.9beta), SHAPEIT2 (v2.5), IMPUTE2 (v2.3.1), fastq-mcf, FastQC, kpal, Tophat(v2.0.8), Cuffquant (v2.2.1), sva, ComBat, UCSC Kent Utilities, eulerAPE, Primer3web (v4.0.0), Marix-eQTL, TagDust (v1.12), BWA (v0.5.9), R (v3.4.4)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

BRAINcode RNA-seq and genotyping raw data have been deposited in dbGAP under accession number phs001556.v1.p1. The processed data and eQTL results for

BRAINcode project can be queried at http://www.humanbraincode.org through a user-friendly interface. The BRAINcode data sets submitted to dbGAP include the raw data used for figures: Fig. 1b-e, Fig. 2a,c,d; Fig. 3a,3c-d; Fig. 4a; and Fig. 5a-c,5e-g.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were based on the total number of available high-quality brain samples that met inclusion and exclusion criteria. No statistical methods were used to pre-determine sample sizes but our sample sizes are consistent with those recommended by the Genotype-Tissue Expression Consortium (GTEX consortium, Nature, 2017).
Data exclusions	Inclusion criteria: (1) absence of clinical or neuropathological diagnosis of a neurodegenerative disease e.g. Parkinson's disease according to the UKPDBB criteria45, Alzheimer's disease according to NIA-Reagan criteria, dementia with Lewy bodies by revised consensus criteria. For the purpose of this analysis incidental Lewy body cases (not meeting clinico-pathological diagnostic criteria for PD or other neurodegenerative disease) were accepted for inclusion. (2) PMI ≤ 48 hours; (3) RIN48 ≥ 6.0 by Agilent Bioanalyzer (good RNA integrity); (4) visible ribosomal peaks on the electropherogram. Exclusion criteria were: (1) a primary intracerebral event as the cause of death; (2) brain tumor (except incidental meningiomas); (3) systemic disorders likely to cause chronic brain damage. We also included eight non-brain tissue samples as controls, including five samples of peripheral blood mononuclear cell (PBMC) and three fibroblasts (FB), provided by Harvard Biomarker Study and Coriell Institute. This study was approved by the Institutional Review Board of Brigham and Women's Hospital.
Replication	Attempts at replication were successful. Replication of TNE was performed in four independent cohorts as delineated in Fig. 3. Moreover, select TNE were confirmed by a second method, qPCR, as shown in Fig. 3. The inverse eQTL relation between the lead GWAS-derived SNP rs17649553 and KANSL1-TNE1 and LRRC37A4P, respectively, was confirmed by a second method, cell type-specific qPCR (Supplementary Fig. 12a). Moreover, this association was independently replicated in a second cohort of neurons laser-captured from 31 high-quality control brains (Supplementary Fig. 12b, Supplementary Table 12). Furthermore, the rs17649553-LRRC37A4P eQTL association was further confirmed in 56 substantial nigra and 96 frontal cortex samples from GTEx (Supplementary Fig. 12c,d), which used a polyA+ selecting protocol that does not allow for assaying KANSL1-TNE1 RNA.
Randomization	Allocation was not random and covariates (such as age, sex, PMI) were adjusted in the analysis.
Blinding	All samples were from controls (see eligibility criteria above). Blinding to case/control status is not applicable.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HeLa and SK-N-MC cell lines were obtained from ATCC.
Authentication	HeLa and SK-N-MC cells were used from ATCC and their identity was confirmed by microsatellite testing.
Mycoplasma contamination	All cell lines tested are negative for mycoplasma contamination.

Commonly misidentified lines (See ICLAC register)	SK-N-MC cells were used from ATCC and their identity was confirmed by microsatellite testing.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Zebrafish (Danio rerio) were used. Both males and females, adults, and embryos were used.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve samples collected from the field.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Characteristics are shown in Supplemental Table 1. Briefly, the mean age at death (standard deviation) was 81 (10.2) for autopsy brains used for lcRNAseq of nigral dopamine neurons. The male:female ratio was 2:1. The median post-mortem interval (stdev) was 3 hours (6.6 hours). The median (stdev) RIN number was 7.8 (0.8).
Recruitment	We started with 107 high-quality, frozen postmortem human control brain samples identified from Banner Sun Health Institute, Brain Tissue Center at Massachusetts General Hospital, Harvard Brain Tissue Resource Center at McLean Hospital, University of Kentucky ADC Tissue Bank, University of Maryland Brain and Tissue Bank, Pacific Northwest Dementia and Aging Neuropathology Group (PANDA) at University of Washington Medicine Center, and Neurological Foundation of New Zealand Human Brain Bank.