



HAL
open science

Long non-coding RNA repertoire and open chromatin regions constitute midbrain dopaminergic neuron - specific molecular signatures

J. Gendron, C. Colace-Sauty, N. Beaume, H. Cartonnet, J. Guegan, D. Ulveling, C. Pardanaud-Glavieux, I. Moszer, H. Cheval, P. Ravassard

► To cite this version:

J. Gendron, C. Colace-Sauty, N. Beaume, H. Cartonnet, J. Guegan, et al.. Long non-coding RNA repertoire and open chromatin regions constitute midbrain dopaminergic neuron - specific molecular signatures. *Scientific Reports*, 2019, 9, pp.1409. 10.1038/s41598-018-37872-1 . hal-02017603

HAL Id: hal-02017603

<https://hal.sorbonne-universite.fr/hal-02017603>

Submitted on 13 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.



L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

SCIENTIFIC REPORTS



OPEN

Long non-coding RNA repertoire and open chromatin regions constitute midbrain dopaminergic neuron - specific molecular signatures

J. Gendron, C. Colace-Sauty, N. Beume, H. Cartonnet, J. Guegan, D. Ulveling, C. Pardanaud-Glavieux, I. Moszer , H. Cheval & P. Ravassard 

Midbrain dopaminergic (DA) neurons are involved in diverse neurological functions, including control of movements, emotions or reward. In turn, their dysfunctions cause severe clinical manifestations in humans, such as the appearance of motor and cognitive symptoms in Parkinson's Disease. The physiology and pathophysiology of these neurons are widely studied, mostly with respect to molecular mechanisms implicating protein-coding genes. In contrast, the contribution of non-coding elements of the genome to DA neuron function is poorly investigated. In this study, we isolated DA neurons from E14.5 ventral mesencephalons in mice, and used RNA-seq and ATAC-seq to establish and describe repertoires of long non-coding RNAs (lncRNAs) and putative DNA regulatory regions specific to this neuronal population. We identified 1,294 lncRNAs constituting the repertoire of DA neurons, among which 939 were novel. Most of them were not found in hindbrain serotonergic (5-HT) neurons, indicating a high degree of cell-specificity. This feature was also observed regarding open chromatin regions, as 39% of the ATAC-seq peaks from the DA repertoire were not detected in the 5-HT neurons. Our work provides for the first time DA-specific catalogues of non-coding elements of the genome that will undoubtedly participate in deepening our knowledge regarding DA neuronal development and dysfunctions.

Midbrain dopaminergic (DA) neurons account for the majority of DA neurons in the adult brain^{1,2}. They are mostly located within two structures, the *substantia nigra pars compacta* (SNpc) and the ventral tegmental area (VTA), both originating from the ventral mesencephalon during development. DA neurons from the SNpc project to the dorsolateral striatum and caudate putamen, thereby forming the nigrostriatal pathway involved in the control of voluntary movements. Their progressive but massive neurodegeneration in Parkinson's Disease is responsible for the appearance of the motor symptoms that principally include rigidity, bradykinesia and tremor^{3,4}. DA neurons from the VTA participate to the mesocorticolimbic pathway, associated with emotion and reward, as they innervate the ventral striatum and prefrontal cortex. Dysfunctions of these neurons have been linked to several human pathologies, in particular schizophrenia, depression and drug addiction⁵. Moreover, DA neurons from the VTA also degenerate in Parkinson's Disease but to a lesser extent than the SNpc neurons⁶⁻⁸.

An increasing number of studies highlight the diversity of the midbrain DA neuronal subtypes from the molecular to the electrophysiological levels, not only between SNpc and VTA, but also within these two structures⁹⁻¹². Thus, using single cell RNA-seq to identify cell-specific molecular signatures, it has recently been shown that adult midbrain DA neurons are subdivided into five subtypes that arise from only two populations of embryonic DA neurons⁹. Importantly, these two fetal subtypes do not each give rise to the SNpc or the VTA, but both participate to the emergence of these adult midbrain regions. In terms of spatial organization however, embryonic

InsERM U 1127, CNRS UMR 7225, Sorbonne Universités, UPMC Univ Paris 06 UMR S 1127, Institut du Cerveau et de la Moelle épinière, ICM, F-75013, Paris, France. J. Gendron and C. Colace-Sauty contributed equally. H. Cheval and P. Ravassard jointly supervised this work. Correspondence and requests for materials should be addressed to H.C. (email: helene.cheval@upmc.fr) or P.R. (email: philippe.ravassard@upmc.fr)

DA neurons segregate into two populations that will each define the future SNpc and VTA¹³. This spatial distribution occurs after the radial migration of differentiating DA neurons from the ventricular zone to the mantle layer of the ventral mesencephalon, creating an intermingled pool of DA neurons that will later constitute the SNpc and the VTA. Then, from gestation day 14.5 to 15.5 (E14.5–E15.5) SNpc neurons migrate tangentially, creating a spatial subdivision between the midbrain DA neurons in mouse embryos¹⁴. Therefore at stage E14.5, mesencephalic DA neurons constitute a roughly spatially and molecularly homogeneous population, suggesting that this embryonic stage constitutes a developmental crossroad before the important diversification of DA neuronal subsets.

So far, molecular signatures defining DA neuronal subtypes have been obtained using transcriptomic data only focused on protein-coding genes^{9,10}. However, recent developments suggest that non-coding elements of the genome such as long non-coding RNAs (lncRNAs) or active regulatory sequences, including promoters or enhancers, constitute repertoires displaying a greater cell specificity than protein-coding genes^{15–20}. lncRNAs are increasingly scrutinized for their multiple regulatory functions from the epigenetic to the post-translational levels^{21–24}, and for their involvement in crucial developmental and cellular processes, such as neuronal differentiation^{17,25–28}. Importantly, genetic mapping of single nucleotide polymorphisms (SNPs) in human pathologies demonstrated that the majority of the SNPs fall into non-coding regions^{29,30}. Consistent with this observation, literature linking lncRNAs as well as active regulatory sequences to human diseases, including Alzheimer disease, Parkinson's Disease, Schizophrenia, drug addiction, cancer, or Diabetes, is growing^{17,31–40}.

In this study, we seek to expand our knowledge on the molecular signatures displayed by mesencephalic DA neurons at E14.5 before their divergence into specific cellular subtypes involved in many physiological and pathological mechanisms. We used high throughput RNA-seq and ATAC-seq and identified novel lncRNAs and active regulatory sequences specific from this population.

Results

Efficient enrichment in DA neurons from mouse E14.5 ventral mesencephalon. To ensure the cell-specificity of our DA neuronal population, we FACS-purified cells originating from E14.5 ventral mesencephalon of transgenic mice expressing GFP under the control of the rat Tyrosine Hydroxylase (*Th*) promoter⁴¹ (Fig. 1a). Cells from the sorted populations were either cultured for 90 minutes and assessed for Th expression, or used to carry out deep RNA-seq and ATAC-seq. Immunofluorescence experiments revealed 81% of Th⁺ cells in the GFP⁺ population, and no Th⁺ cells in the GFP[−] population (Fig. 1b), consistent with previous data from the literature⁴¹. In parallel, *Th* mRNA expression was analysed by RT-qPCR prior to RNA-seq, showing a 90 fold enrichment in GFP⁺ cells compared to the GFP[−] control population (Fig. 1c). We generated cDNA libraries from polyadenylated RNA and mapped ~800 million paired-end sequence reads from a total of 3 independent RNA-seq datasets originating from the GFP⁺ cells (288,046,125 reads for the first dataset; 233,351,531 for the second dataset and 394,620,452 for the third). We performed a pathway analysis on the 1500 most expressed protein-coding transcripts obtained (Fig. 1d), excluding mitochondrial genes, and observed a strong enrichment in genes associated with the terms “Parkinson's Disease” (p-value = 2.874×10^{-13}) and “Dopamine receptor-mediated signalling pathway” (p-value = 2.862×10^{-6}), that appeared within the first 3 occurrences. Terms associated with other neurotransmitter systems, such as serotonergic (5-HT), GABAergic and glutamatergic receptors signalling pathways, also emerged from this analysis pathway, yet with a much less significant p-value. Accordingly, based on fragments per kilobase per million of reads (FPKM), DA lineage marker genes, from progenitors to differentiated cells, were strongly expressed (Fig. 1e), in contrast with marker genes from glutamatergic, noradrenergic, 5-HT neurons and from pericytes, radial glial like cells and endothelial cells that constitute non-neuronal cell types present in the E14.5 ventral mesencephalon⁹. GABAergic markers were however highly expressed, suggesting a slight contamination of our DA population by GABAergic neurons. This GABAergic population potentially covers the 20% Th[−] cells observed in the FACS-sorted GFP⁺ population (Fig. 1b). Overall, the transcriptomic data, along with Th expression analyses at the protein and mRNA levels, indicate that we predominantly isolated DA neurons from E14.5 ventral mesencephalon. RNA-seq and ATAC-seq performed using this approach therefore constitute relevant tools to identify the DA repertoires of lncRNAs and active regulatory sequences.

The lncRNA repertoire of mesencephalic DA neurons. Identification of lncRNAs relied on the following criteria: i) length ≥ 200 pb, ii) expression ≥ 1 FPKM for at least one out of the 3 RNA-seq datasets and at least 2 non-null FPKM values out of the 3 RNA-seq replicates, as well as iii) low protein-coding potential as assessed by CPAT⁴². From the list of lncRNAs obtained using these parameters, we identified two categories of transcripts depending on the presence or the absence of an ATAC-seq peak at the transcription start site (TSS). Indeed, we observed that the presence of ATAC-seq peaks at TSS correlates with genes actively transcribed, independent of the level of expression (Fig. 2a). Nevertheless, sequencing polyadenylated RNA often induces a bias towards less mapped reads in the first exons, especially for long transcripts as clearly illustrated in Supplementary Fig. S1. Therefore we also took in account transcripts with multiple exons that fulfilled the first 3 criteria, but that were not associated with open chromatin at their putative TSS, in order to keep lncRNAs whose first exon(s) had potentially not been correctly sequenced. In contrast with multiexonic transcripts with identified junctions between exons, the likelihood to confuse unannotated monoexonic transcripts with transcriptional background, or even sequencing artefacts led us to discard monoexonic transcripts not associated with an ATAC-seq peak at their TSS. Finally, in situations where several isoforms were identified, we only took into account the most expressed isoform. Using the combination of all of the above criteria, the list of selected transcripts was defined as repertoire of lncRNAs. This way we identified 1,294 lncRNAs, of which 939 had not yet been annotated (Fig. 2b). We used a previously described classification to categorize lncRNAs¹⁷, and found 660 “intergenic”, 401 “divergent”, 217 “overlapping antisense” and 16 “convergent” lncRNAs as described in the scheme Fig. 2b. Amongst the unannotated lncRNAs, 73.1% carried a single exon, whereas monoexonic lncRNAs constituted only 16.9% of the annotated lncRNAs pool (data not shown). Interestingly, in the first 100 most expressed lncRNAs from

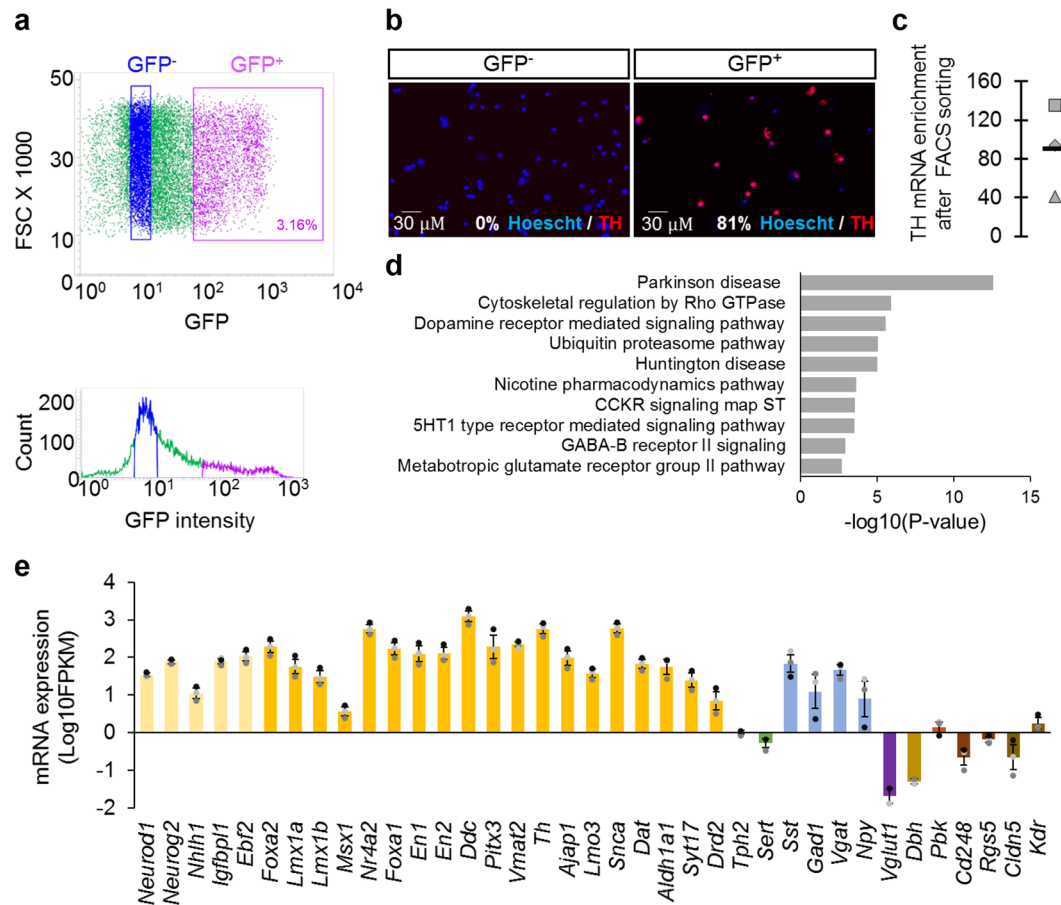


Figure 1. Protein-coding transcriptome of DA neurons isolated from ventral mesencephalons of E14.5 embryos expressing GFP in *Th*⁺ cells. **(a)** FACS sorting of GFP⁺ and GFP⁻ cell populations for subsequent RNA-seq and ATAC-seq analyses. **(b)** *Th* expression (red) assessed by immunofluorescence on GFP⁺ and GFP⁻ cells cultured for 90 minutes after FACS. Nuclei were stained using Hoechst (blue). **(c)** *Th* mRNA relative expression of 3 independent GFP⁺ cell populations used for RNA-seq (triangle, diamond and square), compared to their matching control GFP⁻ cell populations. The bar represents the mean of the 3 enrichment values. mRNA expression was normalized relatively to *Tbp* mRNA expression. **(d)** Pathway analysis (Panther 2016) performed on the 1500 most expressed protein-coding genes obtained from 3 independent RNA-seq datasets, excluding mitochondrial genes. **(e)** mRNA expression in Log₁₀(FPKM) of numerous cellular subtype marker genes. Each circle represents mRNA expression of a marker gene from 1 RNA-seq, and the bar represents the mean of the 3 values. Error bars show standard error of the mean. The color code is as follow: pale yellow, dopaminergic progenitors; yellow, differentiating and differentiated dopaminergic neurons; green, serotonergic neurons; blue, GABAergic neurons; purple, glutamatergic neurons; mustard, noradrenergic neurons; brown, radial glial like cells; dark brown, pericytes; brown-orange, endothelial cells.

the mesencephalic DA repertoire, 37 were novel transcripts and 16 of them had only one exon (Supplementary Table S1). Consistent with the literature^{16,17,22,24,43}, we found that lncRNAs were weakly expressed compared to protein-coding genes, with no overt difference between intergenic lncRNAs and the other lncRNAs categories (grouped as “Antisense” in the Fig. 2c). Using lncRNAs’ closest upstream and downstream genes, or in the case of the overlapping antisense lncRNAs, their overlapped genes, we performed a pathway analysis on the top 20% most expressed lncRNAs from this and found “Dopaminergic Neurogenesis” (adjusted p-value = 0.02151) amongst the only two significantly enriched terms (Fig. 2d). In addition to the fact that 72.6% of the transcripts identified were not previously annotated, this strongly suggests that this DA lncRNA repertoire reflects a high degree of specificity associated with DA neurons.

In order to examine this remarkable tissue specificity further, we compared the DA repertoire of lncRNAs to a repertoire generated at the same stage from hindbrain serotonergic (5-HT) neurons, a monoaminergic neuronal subtype close to DA neurons⁴⁴. Mesencephalic DA neurons and hindbrain 5-HT neurons originate from each side of the mid-hindbrain organizer, and are distributed within anatomically very close nuclei. Both these monoaminergic neurons display a similar developmental pattern regarding kinetics of progenitor specification, migration or differentiation, and project in numerous common brain areas. Importantly, 5-HT neurons also degenerate in Parkinson’s Disease and have been associated with some motor symptoms such as resting tremors, but also non motor symptoms, including anxiety and depression⁴⁵. Thus, using the same approach, we took

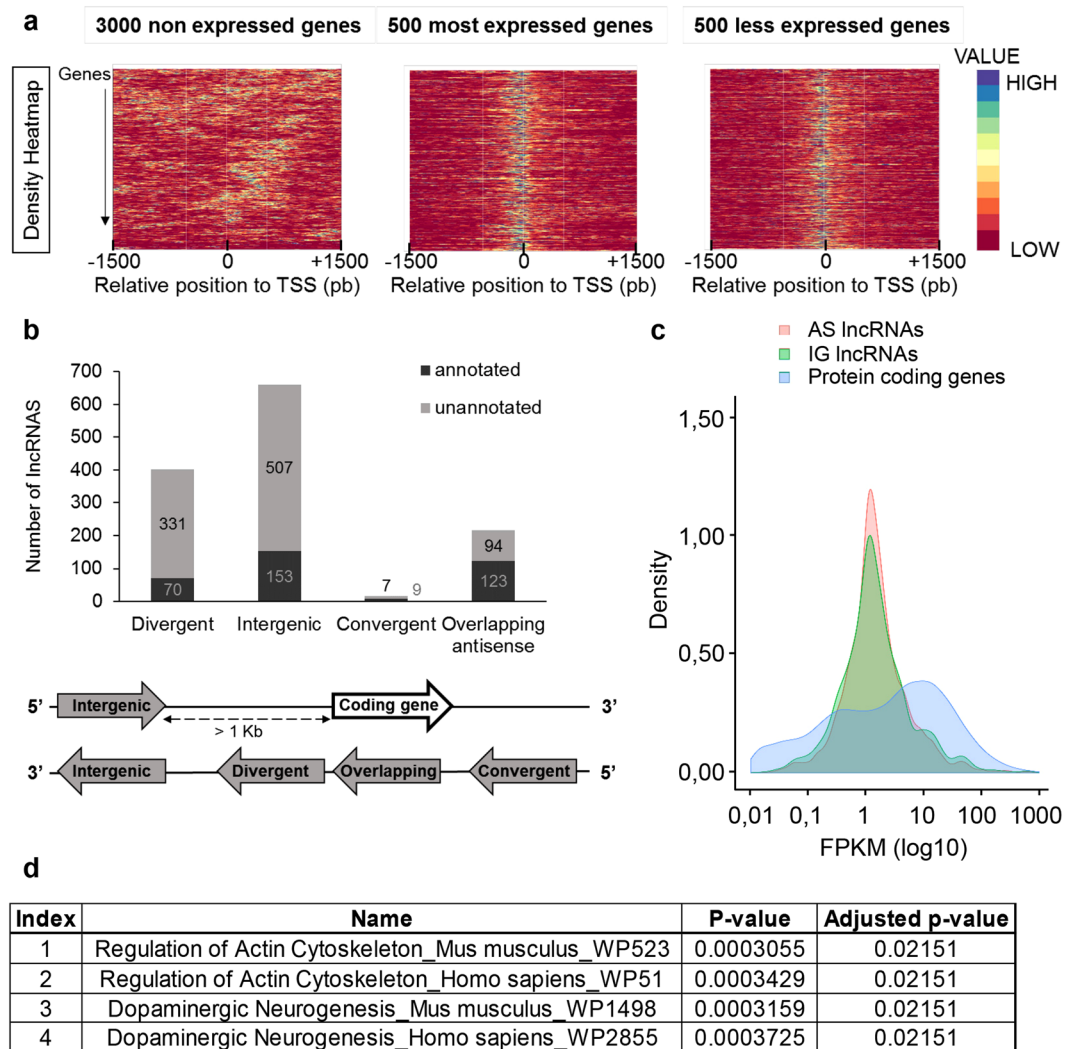


Figure 2. Identification of the lncRNAs expressed in mesencephalic DA neurons. **(a)** Density heatmaps representing the height of ATAC-seq peaks relative to the TSS position of non-expressed (left), highly (middle) or weakly (right) expressed genes in FACS-purified DA neurons. **(b)** Number of lncRNAs depending on their categories as represented below the graph. Dark grey, lncRNAs annotated in Ensembl; light grey, lncRNAs unannotated. **(c)** Density plot of the coding genes described in Ensembl and the lncRNAs (AS: antisense, IG: intergenic). **(d)** Pathway analysis (Wikipathway 2016, Enrichr) performed on the neighbouring genes to the top 20% most expressed lncRNAs using 3 independent RNA-seq datasets.

advantage of Masch1 CRE X Rosa YFP mice that express YFP in 5-HT neurons⁴⁶, to FACS-purify YFP cells from the rhombomeres 1 to 3 of E14.5 embryos (Supplementary Fig. S2a). Following FACS-sorting, tryptophane hydroxylase 2 (Tph2), the neuronal rate limiting enzyme of serotonin biosynthesis, was used as marker. Tph2 immunofluorescence demonstrated an enrichment of 5-HT neurons in the YFP⁺ population, containing 98% of Tph2⁺ cells, compared to the YFP⁻ population which exhibited 24% of Tph2⁺ cells (Supplementary Fig. S2b). *Tph2* mRNA expression, analysed by RT-qPCR prior to RNA-seq, indicated a 16.25 fold enrichment in YFP⁺ cells compared to the YFP⁻ control population (Supplementary Fig. S2c). Regarding the RNA-seq, cDNA libraries from polyadenylated RNA were produced and about 500 million paired-end sequence reads were mapped in the totality of the 3 independent datasets generated (239,852,552 reads for the first dataset; 199,165,019 for the second dataset and 155,506,770 for the third). Sequencing revealed a high expression of 5-HT marker genes in FPKM, but also of genes expressed in DA and GABAergic neurons (Supplementary Fig. S2d). However, principal component analysis comparing the RNA-seq datasets obtained with the FACS-sorted cells from E14.5 ventral mesencephalons and from rhombomeres r1-3, confirmed that they constitute 2 distinct cell populations, with the replicates from each cell type forming 2 separate clusters (Supplementary Fig. S3). Using all criteria described previously, we identified in the 5-HT cell population a repertoire of 1,293 lncRNAs, among which 806 had not yet been annotated (Supplementary Fig. S4). We specifically found 594 “intergenic”, 14 “divergent”, 551 “overlapping antisense” and 134 “convergent” lncRNAs. Since the libraries of 5-HT neuron RNA-seq experiments were not performed in a stranded-specific manner, we could not infer the strand for unannotated monoexonic transcripts.

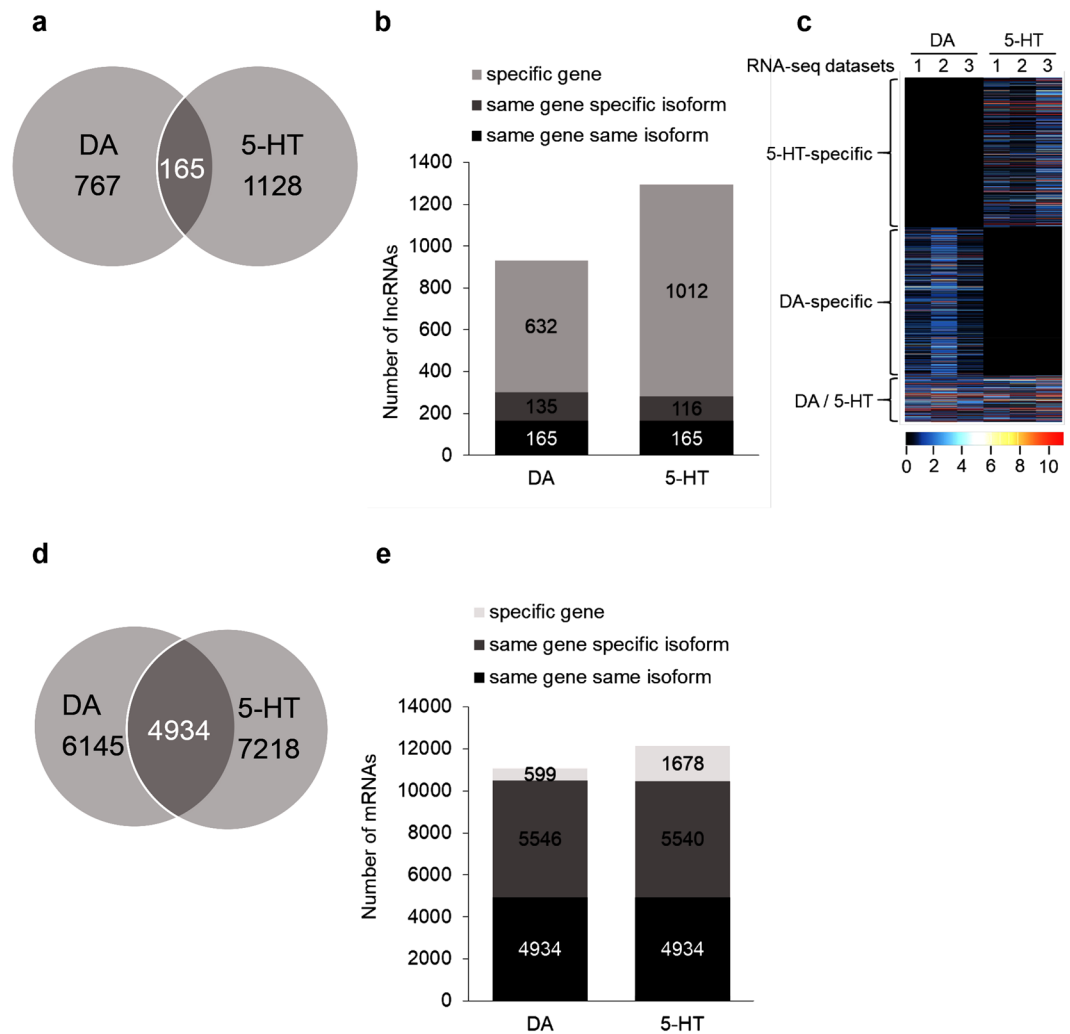


Figure 3. Comparison of the DA and 5-HT repertoires of lncRNAs and protein-coding mRNAs. **(a)** Venn diagram of overlap of both lncRNA repertoires. **(b)** Distribution of lncRNAs transcripts in the DA and 5-HT repertoires depending on their cell specificity. **(c)** Heatmap representing expression in FPKM of each transcripts from both repertoires in the DA and 5-HT datasets (3 samples each). **(d)** Venn diagram of overlap of both protein-coding mRNAs repertoire. **(e)** Distribution of protein-coding transcripts in the DA and 5-HT repertoires depending on their cell specificity.

As our strategy was to discard transcripts lying within 1 kb from a protein-coding gene on the same strand (see material and methods), this led us to discard all unannotated monoexonic transcripts located at less than 1 kb from a protein-coding gene for the 5-HT repertoire. This probably explains the relative low number of divergent lncRNAs in 5-HT neurons compared to the DA repertoire (Fig. 2b).

The 168 remaining monexonic transcripts were mostly intergenic (located at a distance above 1 kb from a protein coding gene, $n = 139$) and unannotated (only 4 were already annotated). Moreover, we observed an elevated number of antisense transcripts in the 5-HT repertoire ($n = 551$) compared to the DA repertoire ($n = 217$), with the number of annotated transcripts in the 5-HT repertoire ($n = 265$) even higher than the total number of DA antisense transcripts (see Figs 2 and S4). Since the identification of such annotated lncRNAs is not impacted by the difference of library preparation (stranded *versus* non-stranded), this indicates that the elevated number of antisense transcripts in the 5-HT repertoire reflects a distinctive cellular feature rather than a technological bias.

To assess the specificity of DA and 5-HT lncRNA repertoires, we focused on categories of lncRNAs that were represented in both datasets. Thus, for unannotated monoexonic lncRNAs only intergenic transcripts were considered. We extracted 767 lncRNAs specific to DA neurons, 1128 specific to 5-HT neurons and 165 expressed in both cell types (Fig. 3a). Interestingly, common lncRNAs displayed higher expression level than cell-specific transcripts (Fig. 3c). In our methodology to establish the lncRNA repertoire of a given cell type, we only took in account the most expressed transcript when several isoforms of a same lncRNA were present. Therefore, we then evaluated the possibility that some specific lncRNAs may in fact have an equivalent, which would be a different isoform, in the other cell type. Thus, to compare isoform usage between DA and 5-HT samples, we looked at correspondence between lncRNAs expressed in the repertoires of both cell types. We thus defined three possible

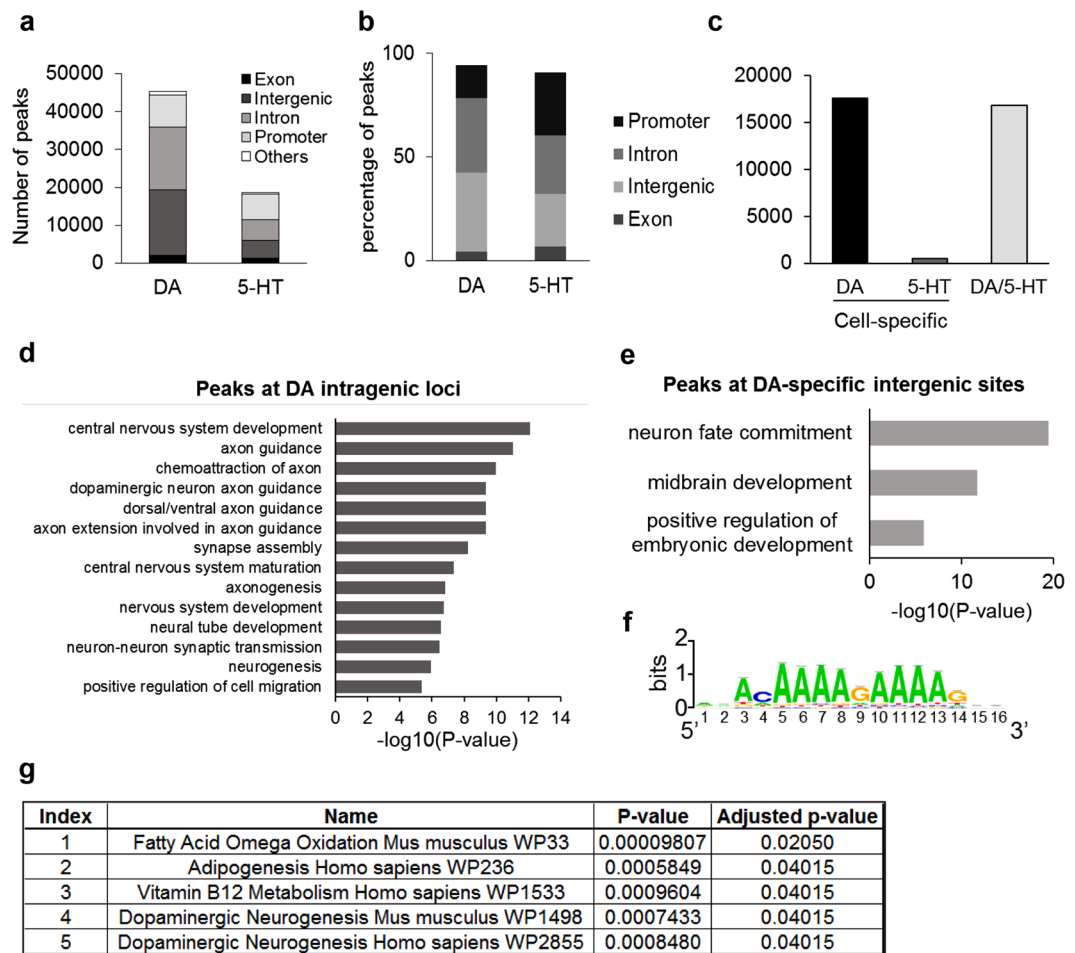


Figure 4. Analysis of DNA regulatory regions using ATAC-seq. **(a)** Number of ATAC-seq peaks depending on their genomic loci. $n = 3$ ATAC-seq datasets for the DA repertoire; $n = 2$ ATAC-seq datasets for the 5-HT repertoire. **(b)** Percentage of ATAC-seq peaks depending on their genomic loci. The category named ‘Others’ regroups ATAC-seq peaks found at 3’UTR regions and TTS. **(c)** Number of ATAC-seq peaks specific to the 3 DA datasets (DA), specific to the 2 5-HT datasets (5-HT), and present in the 5 datasets (DA/5-HT). **(d)** GO Biological process on ATAC-seq peaks linked to exons and introns from the DA-specific ATAC-seq repertoire (Enrichr). **(e)** GO Biological process on intergenic ATAC-seq peaks from the DA-specific ATAC-seq repertoire (GREAT). **(f)** Sox3 binding motif frequently found at loci associated with DA-specific intergenic ATAC-seq peaks (RSAT). **(g)** Pathway analysis (Wikipathway 2016, Enrichr) performed on the DA-specific ATAC-seq peaks associated with promoters.

levels of specificity: “specific gene” when the lncRNA is present in one cell type only, “same gene specific isoform” in such case this isoform is seen only in one cell type but the other cell type expresses another isoform for the same gene, and “same gene same isoform” when the same isoform of the lncRNA is used in both cell types (namely, the 165 lncRNAs). Figure 3b shows the repartition of all three categories in DA and 5-HT neurons. In both cases, “specific gene” is the most important category by far. The same analysis on protein-coding transcripts demonstrated that the proportion of transcripts specifically expressed in one or another cell type is much lower than for the lncRNAs: 55.5% of specific mRNAs in DA neurons and 59.4% in 5-HT neurons *versus* 82.3 and 87% of specific lncRNAs in DA and 5-HT neurons respectively (Fig. 3a,d). Importantly, focusing on the gene level rather than the transcript level, only 5.4% protein-coding transcripts expressed in DA neurons and 13.8% in 5-HT neurons are cell type specific genes, whereas this proportion is much higher for lncRNAs (68 and 78%; Fig. 3b,e). Moreover, we noticed once again the predominance of novel lncRNAs in the DA- and 5-HT-specific repertoires (78% and 73% respectively), whereas the transcripts expressed in both neuronal subtypes were mostly previously annotated (90%). Altogether, these data highlight the high cell-specificity of the lncRNAs.

Mapping open chromatin regions in mesencephalic DA neurons. We used ATAC-seq technology to identify potential regulatory regions of the chromatin active in DA neurons. We distinguished 45,402 ATAC-seq peaks present simultaneously in the 3 ATAC-seq datasets obtained from mesencephalic DA neurons (Fig. 4a), distributed within promoters (19%), intragenic (40%, comprising exons and introns) and intergenic (38%) loci (Fig. 4b). In comparison, we found 18,658 ATAC-seq peaks representing open chromatin regions in both datasets

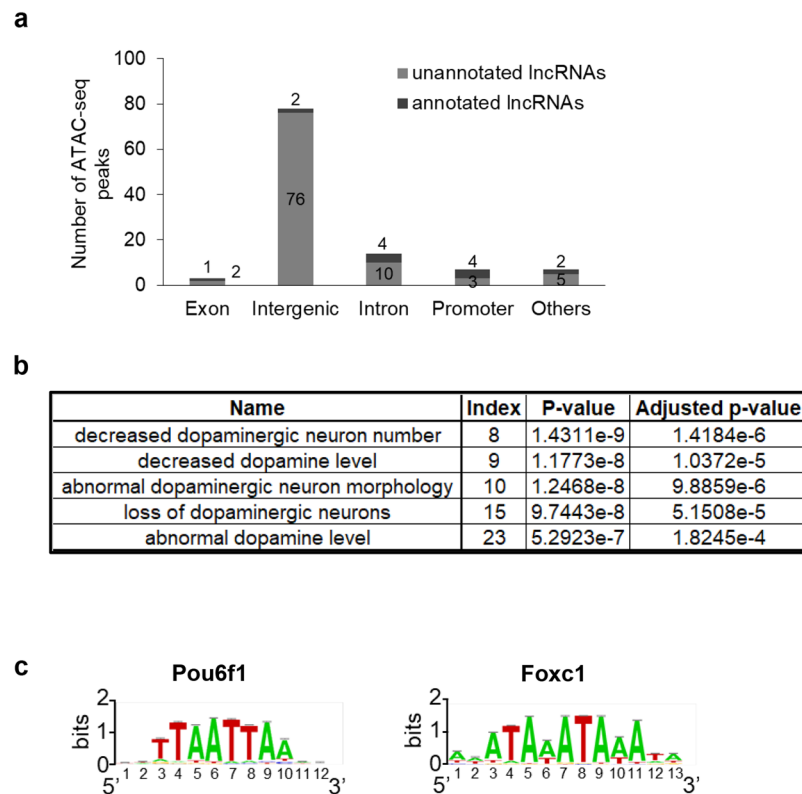


Figure 5. Analysis of DA-specific ATAC-seq peaks coinciding with promoters of DA lncRNAs. **(a)** Number of DA-specific ATAC-seq peaks associated with DA lncRNAs, depending on their genomic loci. Dark grey, lncRNAs annotated in Ensembl; light grey, lncRNAs unannotated. **(b)** MGI Phenotype ontology enrichment analysis performed on these DA-specific ATAC-seq peaks associated with DA lncRNAs (GREAT). **(c)** Binding motifs found at loci associated with DA-specific ATAC-seq peaks associated with DA lncRNAs.

originating from hindbrain 5-HT neurons. In contrast to the DA repertoire of regulatory regions, the majority of 5-HT ATAC-seq peaks were first associated with promoters (37%), then with intragenic (35%) and intergenic (25%) loci (Fig. 4a,b).

Among both these lists of potentially active regulatory regions, 16,856 of them were found in all of the DA and 5-HT datasets, and were considered as common (Fig. 4c). Interestingly, we identified 17,616 ATAC-seq peaks present in the 3 DA datasets but absent in the 5-HT datasets, indicating that 39% of the DA open chromatin regions were specific to this neuronal subtype (Fig. 4c). Conversely, only 513 ATAC-seq peaks were detected in all the 5-HT data sets and not within the DA peaks, constituting a small fraction of 3% of the 5-HT repertoire that was cell-specific. This correlated with the transcriptomic data (Supplementary Fig. S2d) that suggested that the 5-HT cell populations expressed common genes with DA neurons.

Gene Ontology (GO) enrichment analysis performed on genes associated with DA-specific intragenic ATAC-seq peaks, i.e. only detected in the 3 DA datasets, revealed numerous terms linked to biological processes involved in central nervous system development and maturation (Fig. 4d), including axon guidance and notably “dopaminergic neuron axon guidance” (p -value = 4.363×10^{-10} ; Fig. 4d). We then focused on intergenic ATAC-seq peaks that include regulatory regions such as distal enhancers and completed a similar analysis on adjacent genes relative to DA-specific intergenic ATAC-seq peaks. We found a significant gene enrichment in biological processes involving neuronal development, including “midbrain development” (p -value = 1.95×10^{-12} ; Fig. 4e), confirming the cell specificity of these ATAC-seq peaks. In addition, we discovered that these intergenic open chromatin regions were significantly enriched with a DNA-binding motif associated with the transcription factor Sox3 (e -value = 1.2×10^{-30}), which has notably been associated with neurogenesis⁴⁷ (Fig. 4f). Finally, a pathway analysis demonstrated that the term “Dopaminergic neurogenesis” was significantly enriched (p -value = 0.0007433) in genes whose promoters were associated with DA-specific ATAC-seq peaks (Fig. 4g). Altogether, these data not only indicate that we identified the mesencephalic DA map of open chromatin regions, but also substantiate the cell-specificity of this repertoire obtained by ATAC-seq.

By cross-analysing both DA repertoires of lncRNAs and open chromatin regions, we identified 109 DA-specific ATAC-seq peaks overlapping the TSS of lncRNAs (Fig. 5a). Interestingly, 96 out of these 109 lncRNAs were identified for the first time (unannotated) and the majority of them were intergenic. Using Mouse Genome Informatics (MGI) Phenotype ontology, we found that numerous terms describing phenotypes linked to DA neurons were enriched with genes neighbouring these specific lncRNAs (Fig. 5b). *De novo* motif discovery with DA-specific promoters (Fig. 5c) suggested that they are enriched for the motifs bound by Pou6f1 (e -value = 9.5×10^{-3}), a

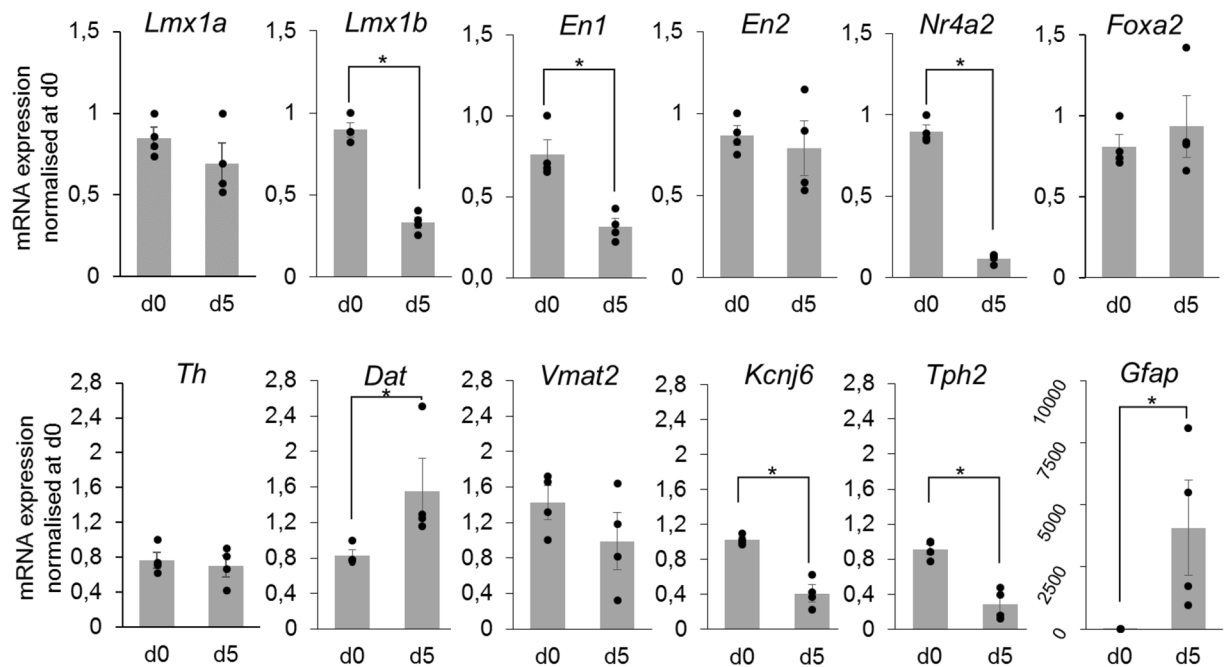


Figure 6. mRNA expression of DA marker genes of DA progenitors (*Lmx1a*, *Lmx1b*, *En1*, *En2*, *Nr4a2*, *Foxa2*), differentiated DA neurons (*Th*, *Dat*, *Vmat2*, *Kcnj6*), 5-HT neurons (*Tph2*) and astrocytes (*Gfap*) in primary cultures from E14.5 ventral mesencephalons at day 0 and after 5 days. Grey columns represent the mean value of 4 independent culture experiments depicted by black circles. mRNA expression was normalized relatively to *Tbp* mRNA expression. Values represented correspond to the mean value of the 3 replicates from each experiment, normalised to the value of 1 reference experiment at day 0. Error bars show standard error of the mean. *p-value $\leq 0,05$.

transcription factor expressed in post-mitotic neurons⁴⁸, and *Foxc1* (e-value = 1.5×10^{-4}) whose expression has been shown to be downregulated in the midbrain DA neurons from patients with Parkinson's Disease⁴⁹.

Expression analysis of lncRNAs in a primary culture of E14.5 ventral mesencephalons. After dissection and dissociation of E14.5 ventral mesencephalons, we cultured cells for 5 days (d5) in order to study selected lncRNAs from the DA repertoire. First, using RT-qPCR to assess marker genes of the DA lineage (Fig. 6), we observed that the cell population obtained presented DA progenitors at d0 and d5, with some markers such as *Lmx1b*, *En1* and *Nr4a2* showing a decreased expression at d5. Markers of differentiated DA neurons were also expressed at both time points, such as *Th*, *Dat*, *Vmat2* and *Kcnj6*. Interestingly, the increase in *Dat* expression suggested some degree of DA neurons maturation in culture. Moreover, *Kcnj6* has been shown to be expressed more abundantly in DA neurons from the SNpc than the VTA⁵⁰, and thus its decreased expression from d0 to d5 implied that at least a fraction of DA neurons present in the culture potentially displayed a VTA identity. Expression of the 5-HT marker gene *Tph2* decreased from d0 to d5, however as expected we noticed a massive increase in *Gfap* expression, reflecting proliferation of astrocytes. We then evaluated whether we could detect lncRNAs from the DA repertoire in this system. Selection of lncRNAs was based on literature curation searching for: (i) implication of their adjacent coding-genes, or themselves, in neuronal development and differentiation, ideally in the DA lineage; (ii) potential involvement in brain pathology. Examples of genomic organization of selected lncRNAs is presented in Supplementary Fig. S5. We analysed expression of 28 lncRNAs (Table 1 and Fig. 7), most of which were stably expressed from d0 to d5. Some displayed an increase in expression, such as *Shhg1*, which has been shown to play a role in cell proliferation in cancer^{51–53}, and has been associated with patients suffering from Parkinson's Disease³⁵. Others, including the novel lncRNA *lnc-En1-1_3* whose closest gene is *En1*, decreased from d0 to d5.

Discussion

In this study, we identified and characterized the DA repertoires of lncRNA loci and open chromatin regions from ventral mesencephalons at E14.5. We found 1,294 lncRNAs expressed in DA neurons, among which 939 had not been previously described. Most of these transcripts were intergenic or divergent. As a comparison, we also identified 1,293 lncRNAs expressed in hindbrain 5-HT neurons, comprising 806 novel transcripts. Regarding their position relative to the closest gene, 5-HT lncRNAs were predominantly intergenic and overlapping anti-sense. Moreover, both repertoires reflected the two distinct populations since only 165 lncRNAs were found in common. In parallel, ATAC-seq analysis allowed for the identification of 45,402 open chromatin regions in E14.5 DA neurons, more than twice the number of ATAC-seq peaks observed in 5-HT neurons (18,658). These putative active regulatory regions were distributed within intergenic, exonic, and intronic loci, as well as within promoter

	exons	locus	category	closest coding genes		overlapped gene
				downstream	upstream	
1810026B05Rik-1_1	5	chr7:80688957-80703006	intergenic	Chd2	A830073O21Rik	—
lnc-Enah-1_2	2	chr1:183950276-183959960	divergent	Srp9	Enah	—
Rmst	4	chr10:91618180-91628236	intergenic	Gm16484	Nedd1	—
2010320M18Rik	1	chr8:73300708-73301478	divergent	Mast3	Pik3r2	—
lnc-Plxna2-1	1	chr1:196444409-196445898	divergent	Camk1g	Plxna2	—
Gm2694	4	chr8:89996573-90049469	divergent	4933402J07Rik	Cbln1	—
Kantr	3	chrX:148729370-148762038	intergenic	Kdm5c	Tspyl2	—
lnc-5930403N24Rik-1	1	chr10:36859252-36861073	overlapping antisense	Marcks	Lama4	5930403N24Rik
2900009J06Rik	2	chr1:129650242-129670663	overlapping antisense	Tmem163	Ccnt2	Acmsd
lnc-Nkain2-1	1	chr10:32609904-32611996	overlapping antisense	Rnf217	Trdn	Nkain2
lnc-Hpgds-1	1	chr6:65100926-65101303	intergenic	C130060K24Rik	Hpgds	—
lnc-Pik3c2a-1	1	chr7:123587063-123587841	divergent	Nucb2	Pik3c2a	—
1700045I19Rik	1	chrX:160198072-160199261	intergenic	Ap1s2	Grpr	—
C130071C03Rik-1	1	chr13:83867069-83868718	intergenic	Tmen161b	Mef2c	—
lnc-En1-1_3	8	chr1:122518213-122586980	intergenic	Insig2	En1	—
1810044D09Rik	2	chr6:91390997-91391746	intergenic	Chchd4	Wnt7a	—
lnc-U6-1_11	6	chr15:89875438-89905660	intergenic	Alg10b	Syt10	—
lnc-Gm6768-1	1	chr12:120591340-120591768	intergenic	Macc1	Itgb8	—
Shhg4	4	chr18:35713064-35717970	intergenic	Matr3	Gm5239	—
lnc-Slc25a24-1	2	chr3:108923456-108925959	divergent	4930443G12Rik	Slc25a24	—
lnc-U6-1_4	1	chr1:4677948-4679320	intergenic	Sox17	Mrpl15	—
A930011O12Rik	2	Chr14:65208662-65212786	intergenic	Kif13b	MsrA	—
lnc-Zfp541-1_1	1	chr7:16633660-16634413	intergenic	Zfp541-1	Gltscr1	—
lnc-BC032203-1	1	chr17:46993166-46994043	intergenic	BC032203	A330017A19Rik	—
Shhg1	10	chr19:8797802-8800934	divergent	Wdr74	Slc3a2	—
lnc-Cpsf2-1	3	chr12:103206308-103213872	divergent	Atxn3	Cpsf2	—
Shhg5-1	5	chr9:88415894-88417721	intergenic	Syncrip	Zfp949	—
2700069I18Rik-1_2	1	chr3:5219240-5219623	overlapping antisense	Gm10748	Pxmp3	Zfhx4

Table 1. Brief description of the selected lncRNAs analysed in Fig. 7. For each lncRNA, the number of exons, locus, category, closest coding-genes and overlapped genes are provided. In bold are represented the DA-specific lncRNAs relatively to the 5-HT repertoire.

regions. Comparing regions of open chromatin from DA and 5-HT neurons, we observed that more than a third of the DA repertoire were not found in the 5-HT neurons, whereas most of the 5-HT ATAC-seq peaks were also mapped in DA neurons. We overlapped the DA repertoires of lncRNAs and open chromatin regions and identified specific regions mostly associated with novel DA lncRNAs. Finally we selected lncRNAs expressed in the DA datasets and analysed their expression in a primary culture of ventral mesencephalons.

Consistent with data from the literature, our study highlights the high degree of cell-specificity of both lncRNAs repertoire and map of open chromatin, that actually represent more accurate molecular signatures associated with cellular subtypes than protein-coding genes (Fig. 3)^{15–19}. Regarding lncRNAs, we indeed observed that the majority of the lncRNAs expressed in DA neurons were described for the first time, and the comparison with lncRNAs identified from 5-HT neurons highlighted even more this cell-specificity as only a small fraction of transcripts were expressed in both neuronal subtypes. However, in contrast with many studies^{19,54,55}, we did not eliminate transcripts carrying a single exon in our identification criteria but retained monoexonic transcripts with a clear ATAC-seq signal overlapping their TSS. Example of a novel monoexonic transcript is illustrated in Supplementary Fig. S5d. Using such stringent criteria, we cannot rule out that some monoexonic lncRNAs are missing in the repertoires due to the sequencing under-representation of transcript 5' ends (Supplementary Fig. S1). This technology bias is more important for long transcripts. The average length of monoexonic transcripts is of 918 nt and 745 nt in the DA and 5-HT repertoires respectively, consistent with the size range where the sequencing bias is minimal (Supplementary Fig. S1). The resulting repertoire contained 73.1% of monoexonic transcripts among the novel lncRNAs identified (and only 16.9% among the annotated lncRNAs). Therefore we cannot exclude that the cell-specificity that we observed was in part due to a bias associated with the lack of annotation of single exon lncRNAs. Nevertheless, this bias does not imply that these monoexonic lncRNAs were not specifically expressed in DA neurons, and the fact that we found a majority of cell-specific transcripts comparing DA and 5-HT lncRNAs using the same criteria to generate both repertoires, indicated that both constitute molecular signatures of the neuronal subtypes they have been generated from. Moreover, for the lncRNAs identification process and further cell-specificity analyses, we focused on transcripts expression, rather than gene expression. Thus, we noticed that several transcripts could be detected at the same locus, representing isoforms of the same gene, and decided to only consider the most expressed transcript. Using this strategy, we took into account not

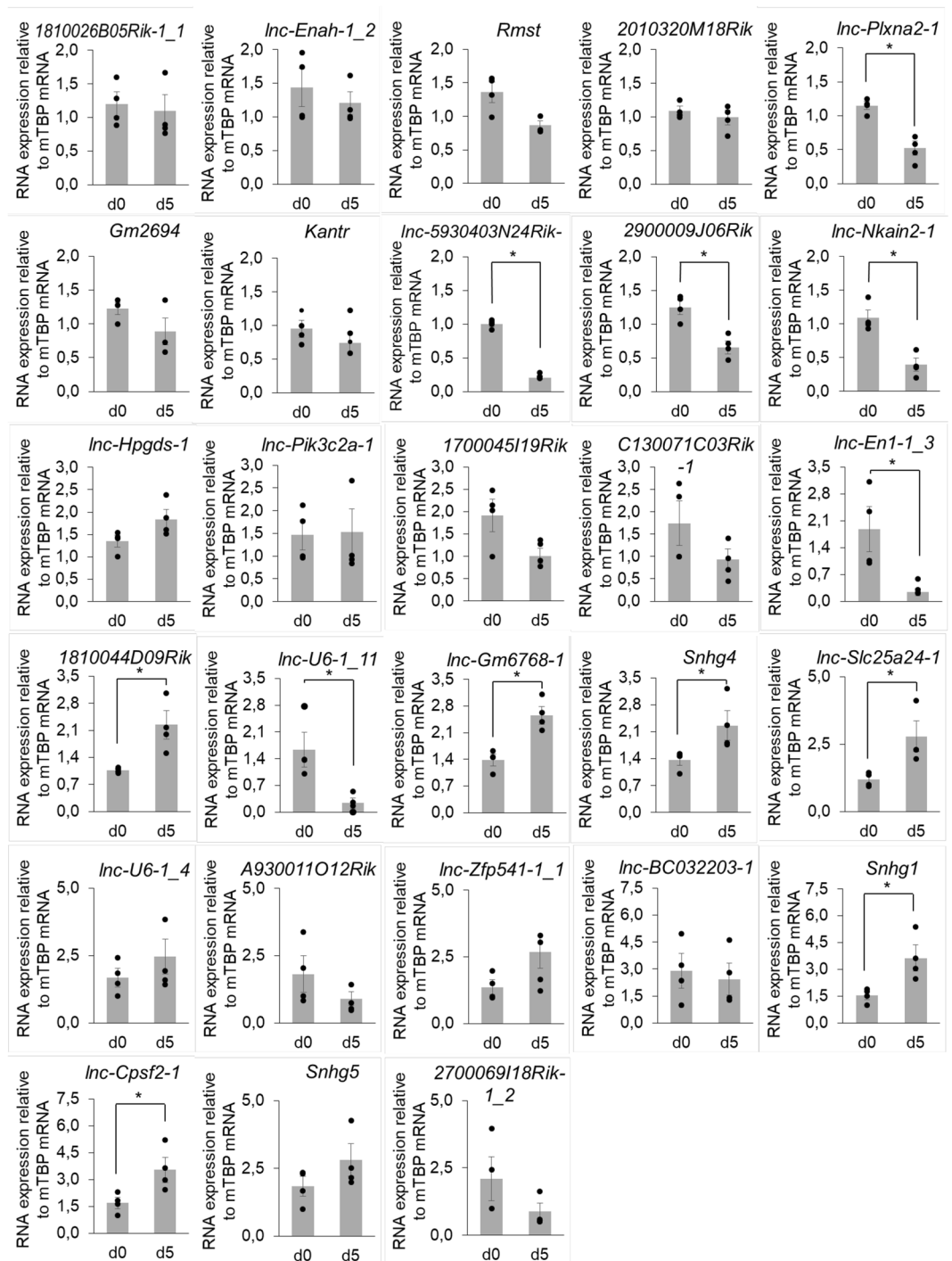


Figure 7. RNA expression of selected lncRNAs in primary cultures from E14,5 ventral mesencephalons at day 0 and after 5 days. Grey columns represent the mean value of 4 independent culture experiments depicted by black circles. mRNA expression was normalized relatively to *Tbp* mRNA expression. Values represented corresponds to the mean value of the 3 replicates from each experiment, normalised to the value of 1 reference experiment at day 0. Error bars show standard error of the mean. *p-value $\leq 0,05$.

only the transcription process, but also the splicing process that is still poorly studied regarding lncRNAs. This way, we have been able to identify cell-specific isoforms of lncRNAs, resulting in a number of cell-specific lncRNAs more precise and higher than if we had chosen an identification process based on genes.

In order to select lncRNAs actively transcribed, we used ATAC-seq data to identify TSS of our candidate transcripts. However, we observed numerous multiexonic transcripts that were expressed within our criteria, but still did not harbour an ATAC-seq peak at their TSS. Again, a possible explanation lies in the sequencing of polyadenylated RNA that often results in a decrement in mapped reads towards the first exons. It is therefore

possible that some of the expressed lncRNAs that were not displaying an ATAC-seq peak at their putative TSS were in fact not integrally sequenced. It is also possible that these lncRNAs could be detected with RNA-seq, but that there was no permissive region detectable at their promoter by ATAC-seq. This could be explained by very weak transcriptional activity producing stable transcripts and/or active transcription in a small subset of cells. To circumvent this issue, single cell deep RNA-seq and ATAC-seq for the study of lncRNAs is essential, even though the development of these techniques at this scale is still at its beginning.

We generated maps of open chromatin from DA and 5-HT neurons and observed that the DA datasets displayed far more putative regulatory regions than the 5-HT datasets. Such difference between cell types has already been described⁵⁶ and is potentially intrinsically associated with the nature of the cells studied. However, the percentage of ATAC-seq peaks specific to DA neurons was strikingly higher than the peaks only present in the 5-HT neurons. While this observation demonstrates consistency between DA repertoires of lncRNAs and open chromatin regions that both display an important cell-specificity, it shows quite a difference within the 5-HT repertoires. Indeed, many lncRNAs but only 3% of the open chromatin regions were specific to 5-HT neurons. The weak specificity level of ATAC-seq peaks compared to the DA datasets could however reflect the fact that the cell population extracted from E14.5 r1-3 rhombomeres also expressed DA marker genes. Also, because of the reduced numbers of ATAC-seq peaks in 5-HT neurons compared to DA neurons, we cannot rule out that the number of monoexonic transcripts that are intergenic was underestimated in 5-HT neurons (139 *versus* 384 in the DA repertoire, see material and methods). Indeed, we chose to discard from our analysis monoexonic transcripts that do not harbour an ATAC-seq peak at their putative TSS in our analysis. However, the difference between the DA and 5-HT ATAC-seq datasets is principally due to a preferential loss of intergenic and intronic peaks (respectively 3.6 and 3.1 times less in 5-HT neurons, Fig. 4a). Therefore, it is possible that some intergenic lncRNAs displaying a single exon in the 5-HT repertoire have been eliminated from our analysis.

Importantly, Genome Wide Association studies have allowed identification of many SNPs in human pathologies that are associated with non-coding regions of the genome^{29,30,57}, suggesting that numerous risk factors linked to diseases could alter the function of lncRNAs or enhancer regions. Our work substantiates the increasing literature showing that lncRNAs and open chromatin regions constitute very specific molecular signatures, and strengthens the need to study these elements in distinct cellular subtypes, especially in the context of human pathologies that are associated with dysfunction of specific cells, such as cancers, Diabetes or Parkinson's Disease. Interestingly, 44 Parkinson's disease risk loci have been identified from meta-analysis of genome wide association studies^{57,58}. Using synteny analysis, we found 8 lncRNAs of the DA repertoire located in the mouse syntenic regions corresponding to genomic areas of the Parkinson's disease human risk loci, (Supplementary Table S2). Among these lncRNAs, 6 were unannotated and 5 of them were specifically found in DA neurons. Although found in both DA and 5-HT neurons, 2900009J06Rik-1, one of the candidate lncRNAs we studied, is significantly more expressed in ventral mesencephalons than in rhombencephalons of E14.5 embryos (Supplementary Fig. S5). Overall, this highlights the cell-specificity of the lncRNAs potentially linked to Parkinson's Disease. Discovery of cell-specific regulatory lncRNAs or regulatory DNA sequences might therefore provide new clues towards a better comprehension of human diseases but also advancements in the search of therapeutic targets.

Material and Methods

Animals. All procedures were conducted in compliance with the European and French legislations (EU directive 2010/63/UE), and were approved by the "Direction Départementale de la Protection des Populations" under accreditation number A75-13-19.

To purify dopaminergic (DA) neurons by fluorescence-activated cell sorting (FACS) prior to RNA-seq and ATAC-seq, we used TH-GFP mice, in which GFP is expressed under the control of the *Th* promoter⁴¹. This transgenic line, maintained on C57BL/6J background, was generously given by H. Okano. To isolate serotonergic (5-HT) neurons by FACS, we were generously given by C. Parras Mash1-CRE × ROSA YFP mice, in which YFP is expressed in 5-HT neurons⁴⁶. Mice had *ad libitum* access to food and water, and were housed in cages containing up to 5 animals under temperature-controlled conditions and maintained on a 12/12 hours light/dark cycle.

To obtain E14.5 embryos, males from these transgenic lines were mated with Swiss wild-type females overnight, and pregnancies confirmed the next morning by inspection of the vaginal plug, defining embryonic day 0.5.

For primary cell culture experiments, we used E14.5 embryos of Swiss wild-type mice purchased from Charles River, France.

Tissue collection. To obtain DA neurons, 5–15 ventral mesencephalons from E14.5 embryos were dissected for each experiment. Regarding 5-HT neurons, 4–8 regions containing rhombomeres r1, r2 and r3 of the hind-brain were carefully removed from the embryonic brains for each experiment. After removing the meninges, tissue was collected in ice-cold HBSS 1X until dissociation.

Fluorescence-activated cell sorting (FACS). Tissue was mechanically dissociated into a single cell suspension in Neurobasal medium with B27 (Life Technologies) and kept at 4 °C until FACS purification. Cell suspensions were filtered with a 50 µm filter and then processed by FACS for selection of DA GFP⁺ or 5-HT YFP⁺ cells. FACS was performed on an INFLUX 500 cell sorter. Dead cells were excluded by addition of propidium iodide. Cell suspensions from ventral mesencephalons or hindbrain r1-3 of wild-type mice were used to adjust background fluorescence. For RNA-seq, single cells were collected in Lysis Buffer (RNeasy Micro Kit Qiagen) with 1% of β-mercapto-ethanol and immediately kept in dry ice at –80 °C until RNA extraction. For ATAC-seq or isolated cells culture on GFP⁺ or YFP⁺ cells, single cells were collected in Neurobasal medium with B27 supplement (Life Technologies), 2% FBS and kept at 4 °C.

Immunofluorescence. An average of 8000 cells sorted by FACS (GFP⁺/GFP⁻ and YFP⁺/YFP⁻) were independently plated on Poly-D-lysine hydrobromide (Sigma) coated Labteck, cultured for 90 min in Neurobasal medium with B27 supplement and 20% FBS. Cells that attached were fixed with 4% paraformaldehyde in PBS for 10 minutes, followed by successive washes with PBS and PBS-Tween (1X, 0.1% Tween). After incubation in blocking solution for 1 h (PBS-Tween 10% goat serum), cells were immunolabelled overnight at 4 °C using the following antibodies: mouse anti-TH (1:400; Millipore MAB318), rabbit anti-GFP-YFP (1:750; Megaprob A11122), rabbit anti-TPH2 (1:500; Novus Biologicals). The next day, samples were washed, then incubated during 1 h at room temperature with the following secondary antibodies: goat anti-rabbit Alexa 555 (1:1000; Invitrogen A-21428), goat anti-rabbit Alexa 488 (1:1000; Invitrogen A-11008), goat anti-mouse Alexa 555 (1:1000; Life Technologies A21425). Nuclei were labelled with Hoechst. All images were collected on a Leica microscope.

Primary cell culture from E14.5 ventral mesencephalons. Dissociation of E14.5 ventral mesencephalons was performed by alternating mechanical dissociation and 10 min decantation steps 4 times. Dissociated cells from E14.5 ventral mesencephalons were cultured in 12 well-plates previously coated with Neurobasal medium complemented with 1% ECM (Sigma-Aldrich), Fibronectin from bovine plasma (1:250, Sigma-Aldrich) and 1% Penicilin/streptomycin (Life Technologies). Around 1 ventral mesencephalon was used for each well. Cells were incubated for 5 days at 37 °C, 5% CO₂ with Neurobasal medium complemented with B27 + vitamin A supplement (1:50; Life Technologies), 1% L-glutamine (Life Technologies), 1% Penicilin/streptomycin, 20 mg/mL GDNF (Peprotech), 1 mM AMPc (Sigma-Aldrich) and 20% FBS (Helvetica Health Care). Twenty mg/mL of freshly made GDNF was added at days 2 and 4. Four independent experiments were conducted. For each experiment, dissociated cells were distributed within 3 wells at day 0 and cultured independently until collection at day 5 for RT-qPCR.

RNA extraction. Total RNA was extracted from DA or 5-HT neurons using an RNeasy Microkit (Qiagen) following manufacturer's instructions. RNA was treated with DNase I (Qiagen) for 20 minutes at room temperature to prevent genomic DNA contamination. For RT-qPCR, RNA concentrations were determined by spectrophotometry (Nanodrop 2000c, THERMO Scientific). For RNA-seq, a High Sensitivity RNA ScreenTape analyzer (Agilent Technologies) was used to assess RNA concentrations as well as the RNA integrity number (RIN) to verify RNA quality for all tested samples. RNA was stored at -80 °C until reverse transcription or RNA-seq.

Real time quantitative RT-PCR (RT-qPCR). Up to 500 ng RNA was used to generate a first cDNA strand (Superscript II reverse transcriptase, THERMO Fisher Scientific) with random hexamers as indicated by the manufacturer. qPCR experiments were realized on the Light Cycler 96 or 1536 real-time PCR system (Roche); with SYBER green detection. The comparative method of relative quantification ($2^{-\Delta\Delta CT}$) was used to calculate the expression levels of each target gene and mouse *Tbp* mRNA was used to normalize the expression of all samples. The list of primers used is provided in the Supplementary table S3.

RNA-sequencing (RNA-seq). Three independent FACS experiments were achieved for each neuronal subtype, providing 34,110 to 82,678 GFP⁺ cells for DA neurons and 9,207 to 11,107 YFP⁺ cells for 5-HT neurons. One ng of total RNA was used for RNA-seq. For FACS-purified samples of DA neurons, stranded library was prepared using TotalScript RNA sequencing kit (Epicentre) following manufacturer's recommendation. For purified samples of 5-HT neurons, non-stranded library were prepared using SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing (Clontech) following manufacturer's recommendation. 3 DA and 3 5-HT libraries were sequenced using NextSeq500 HighOutputKit v2 (300cycles) cartridge (FC-404-2004 Illumina).

ATAC-sequencing (ATAC-seq). Three independent FACS experiments were achieved for DA neurons analysis, providing 50,000 to 87,500 GFP⁺ cells, and 2 independent FACS experiments were performed for 5-HT neurons analysis, providing respectively 11,448 and 19,407 YFP⁺ cells. Sorted cells by FACS were collected in Neurobasal medium with B27 supplement (Life Technologies), 2% FBS and kept at 4 °C until ATAC-Seq. Cells were centrifuged at 500 g, at 4 °C during 20 min. Cells were resuspended in 25 µl of lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl₂, 0.1% IGEPAL CA-630) during 10 min at 4 °C. Then supernatant was taken out after a centrifugation at 500 g, at 4 °C during 30 min. For transposase reaction, the pellet was resuspended in 25 µl of 12.5 µl 2x TN buffer; 2 µl of Tn5; 10.5 µl d'H₂O and incubated at 37 °C for 1 h. Then 5 µl of clean-up buffer (900 mM NaCl, 300 mM EDTA, 5% SDS) were added with 2 µl of 5% SDS and 2 µl of Proteinase K, and cells were incubated for 30 min at 40 °C. Samples were then cleaned with two SPRI clean up (Agencourt © AMPure ©XP), with 68 µl of SPRI beads, eluted in 13 µl of buffer EB (Qiagen Cat No./ID: 19086). Extracted DNA concentration was measured by ScreenTape analyzer (Agilent Technologies). To generate libraries, PCR reactions were performed using the kapa PCR mix (Kapa biosystem) with 12.5 µl Kapa, 1 µl primers and 11.5 µl of sample, with the NextEra primers (1 µl /primer). PCR conditions were performed as described: 98 °C during 2 min and then 9 cycles of 98 °C during 20 s, 63 °C during 30 s, 72 °C during 1 min. Then, a new SPRI clean-up was made to do a size cut off of amplified PCR products and after that, DNA concentration was measured by ScreenTape analyzer (Agilent Technologies). Finally, a second PCR was performed with the same conditions as the first one and a last SPRI clean-up was made and library of tagged open chromatin was ready to be sequenced. Libraries were sequenced, with 75 bp paired-end reads, on an Illumina NextSeq500 platform,

Bioinformatics. RNA-seq data processing. Raw sequencing data was quality-controlled with the FastQC program. Adapter sequences were removed by Cutadapt. Low quality reads were trimmed or removed using Trimmomatic (minimum length: 40 bp). Reads were aligned to the mouse reference genome (build mm9) with the TopHat2 tool⁵⁹ (option for no multihits) and mapping results were quality-checked using RNA-SeQC. Normalization and differential analysis were performed with the DESeq2 package.

ATAC-seq data processing. Steps for quality control were identical to those used for RNA-seq data treatment (Trimmomatic, FastQC). Reads with a length below 100 bp have been removed in further analysis. Paired-end reads were mapped to the mouse genome (build mm9) with Bowtie2. Duplicate reads were discarded with the Picard tools. Peaks were called using the MACS2 program with the option callpeak. Individual peaks separated by less than 100 bp were merged with BEDOPS and features annotations were obtained from the HOMER mm9 database. We mapped 52,862; 93,056 and 94,352 peaks from the 3 DA datasets, and 38,772 and 22,880 peaks from the 2 5-HT datasets. We extracted the ATAC-seq peaks present simultaneously in the 3 ATAC-seq datasets obtained from DA neurons ($n = 45,402$) and the ATAC-seq peaks in both datasets originating from 5-HT neurons ($= 18,658$), as shown in Fig. 4a. To analyse the number of specific or common ATAC-seq peaks between the DA and 5-HT datasets, we selected the DA-specific peaks as peaks present in the 3 DA datasets but absent in both 5-HT datasets; the 5-HT-specific peaks as peaks present in both 5-HT datasets but absent in the 3 DA datasets; and the common ATAC-seq peaks between DA and 5-HT neurons as peaks present in the 5 datasets. We therefore excluded from this stringent analysis the peaks found in the 3 DA datasets but only in one 5-HT dataset ($n = 10930$), as well as the peaks found in both 5-HT datasets but only in one or two DA datasets ($n = 1289$).

Data from ATAC-seq and RNA-seq results were intersected based on overlaps between a given ATAC peak and the first/last nucleotide of a TSS/TTS, respectively.

Construction of lncRNA catalogues. Transcriptomes were assembled using the Cufflinks/Cuffmerge suite, guided by the GENCODE GTF mm9 annotation file. Quantification and normalization at the gene (XLOC) and transcript (TCONS) levels were performed with Cuffquant/Cuffnorm. A consolidated result file containing transcript attributes (e.g., exon number, Cufflinks class-code), FPKM values, intersection with ATAC peaks as defined above, and GENCODE annotations, was then produced. Annotated lncRNAs were selected from TCONS entries whose Cufflinks class-codes were different from ‘-’ (Unknown, intergenic transcript) and ‘x’ (Exonic overlap with reference on the opposite strand) and whose annotation contained one of the following biotype: “lncRNA”, “antisense”, “non_coding”. Potential novel lncRNAs were identified from TCONS entries characterized by Cufflinks class-codes ‘-’ or ‘x’. For both classes, the following criteria were used: length > 200 bp and FPKM ≥ 1 in at least 1 sample out of 3 replicates.

The closest protein-coding gene for each lncRNA was identified using the tool ‘bedtools closest’. Coding potential of transcripts was assessed using CPAT⁴² (cut-off: 0.44). Those information were included in the repertoire consolidated file. Moreover, lncRNAs located at a distance lower than 1 kb from a known gene on the same strand (or whose strand was undetermined), were eliminated.

Due to technical reasons, the libraries of 5-HT neuron RNA-seq experiments were not performed in a stranded-specific manner, resulting in the inability to infer the strand for unannotated monoexonic transcripts. Regarding transcripts displaying multiple exons, assignment of the strand was performed by identification of consensus splice sites by the TopHat2 tool. Since our strategy was to discard transcripts lying within 1 kb from a protein-coding gene on the same strand, this led us to discard all unannotated monoexonic transcripts located at less than 1 kb from a protein-coding gene for the 5-HT repertoire, leaving mostly intergenic lncRNAs (i.e. located at a distance superior to 1 kb from a protein-coding gene). Thus, for comparative analyses of specificity between DA and 5-HT repertoires, we have compared lncRNAs presenting the same characteristics in both repertoires, i.e. we excluded the monoexonic transcripts located at less than 1 kb of a protein-coding gene from the DA repertoire. This way we avoided the introduction of a biased estimation of cell specificity. Correspondences between lncRNAs of the two cell type repertoires were defined with the following strategy. A reciprocal intersection between transcript coordinates of each repertoire was computed with a threshold equal to 90% of sequence length in common; the number of exons of the pairs thus defined had to be identical; coordinates for each pair of exons had to differ from no more than 50 bp for internal exons and 500 bp for outermost exons. Then manual curation has been performed to ensure the strength of the correspondences. lncRNA labelled as “Specific gene” do not have counterpart in the repertoire of the other cell type. lncRNAs expressed as distinct isoforms in each cell type belong to the category called “same gene specific isoform”. lncRNAs expressed as the same isoform in both repertoires are labelled as “same gene same isoform”, and are non-specific.

Gene Ontology (GO), Pathway enrichment analysis and search for transcription factors motifs.

To perform GO and Pathway enrichment analysis on a list of genes, Enrichr was exploited, using GO Biological process, Panther 2016, and Wikipathway 2016 databases^{60,61}. To perform these analyses on non-coding regions, we used GREAT that analyses the annotations of the nearby genes, using GO biological processes and MGI phenotype ontology databases⁶². To search for transcription factors-associated motifs, RSAT was used⁶³.

Statistics. Statistical analyses to assess differences in gene expression from day 0 to day 5 of cell cultures were conducted using two-tailed Mann Whitney *U*-tests (GraphPad Prism 6). For each gene, values were normalised to the mean of the 3 replicates of 1 experiment at day 0. Values represented therefore correspond to the mean value of the 3 replicates from each experiment, normalised at day 0.

One-tailed Mann Whitney *U*-tests were used to compare lncRNA expression between ventral mesencephalons and r1-3 rhombomeres. Mann Whitney *U*-tests and p-values are gathered in Supplementary Tables S4.

For GO and Pathway enrichment analysis on a list of genes with Enrichr, a Fisher exact test was used. A Binomial test over genomic regions was performed for GO biological processes and MGI phenotype ontology analyses using GREAT.

Data Availability

The GEO accession number for RNA-Seq and ATAC-Seq reported in this paper is: GSE108917.

References

- German, D. C., Schlusberg, D. S. & Woodward, D. J. Three-dimensional computer reconstruction of midbrain dopaminergic neuronal populations: from mouse to man. *J Neural Transm* **57**, 243–254 (1983).
- Pakkenberg, B., Moller, A., Gundersen, H. J., Mouritzen Dam, A. & Pakkenberg, H. The absolute number of nerve cells in substantia nigra in normal subjects and in patients with Parkinson's disease estimated with an unbiased stereological method. *J Neurol Neurosurg Psychiatry* **54**, 30–33 (1991).
- Dickson, D. W. *et al.* Neuropathological assessment of Parkinson's disease: refining the diagnostic criteria. *Lancet Neurol* **8**, 1150–1157, [https://doi.org/10.1016/S1474-4422\(09\)70238-8](https://doi.org/10.1016/S1474-4422(09)70238-8) (2009).
- Kalia, L. V. & Lang, A. E. Parkinson's disease. *Lancet* **386**, 896–912, [https://doi.org/10.1016/S0140-6736\(14\)61393-3](https://doi.org/10.1016/S0140-6736(14)61393-3) (2015).
- Morales, M. & Margolis, E. B. Ventral tegmental area: cellular heterogeneity, connectivity and behaviour. *Nat Rev Neurosci* **18**, 73–85, <https://doi.org/10.1038/nrn.2016.165> (2017).
- Brichta, L. & Greengard, P. Molecular determinants of selective dopaminergic vulnerability in Parkinson's disease: an update. *Front Neuroanat* **8**, 152, <https://doi.org/10.3389/fnana.2014.00152> (2014).
- Damier, P., Hirsch, E. C., Agid, Y. & Graybiel, A. M. The substantia nigra of the human brain. II. Patterns of loss of dopamine-containing neurons in Parkinson's disease. *Brain* **122**(Pt 8), 1437–1448 (1999).
- Hirsch, E., Graybiel, A. M. & Agid, Y. A. Melanized dopaminergic neurons are differentially susceptible to degeneration in Parkinson's disease. *Nature* **334**, 345–348, <https://doi.org/10.1038/334345a0> (1988).
- La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* **167**, 566–580 e519, <https://doi.org/10.1016/j.cell.2016.09.027> (2016).
- Poulin, J. F. *et al.* Defining midbrain dopaminergic neuron diversity by single-cell gene expression profiling. *Cell Rep* **9**, 930–943, <https://doi.org/10.1016/j.celrep.2014.10.008> (2014).
- Roeper, J. Dissecting the diversity of midbrain dopamine neurons. *Trends Neurosci* **36**, 336–342, <https://doi.org/10.1016/j.tins.2013.03.003> (2013).
- Smidt, M. P. Molecular Programming of Mesodiencephalic Dopaminergic Neuronal Subsets. *Front Neuroanat* **11**, 59, <https://doi.org/10.3389/fnana.2017.00059> (2017).
- Arenas, E., Denham, M. & Villaescusa, J. C. How to make a midbrain dopaminergic neuron. *Development* **142**, 1918–1936, <https://doi.org/10.1242/dev.097394> (2015).
- Bodea, G. O. *et al.* Reelin and CXCL12 regulate distinct migratory behaviors during the development of the dopaminergic system. *Development* **141**, 661–673, <https://doi.org/10.1242/dev.099937> (2014).
- Mercer, T. R., Dinger, M. E., Sunken, S. M., Mehler, M. F. & Mattick, J. S. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci USA* **105**, 716–721, <https://doi.org/10.1073/pnas.0706729105> (2008).
- Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**, 1915–1927, <https://doi.org/10.1101/gad.17446611> (2011).
- Moran, I. *et al.* Human beta cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell Metab* **16**, 435–448, <https://doi.org/10.1016/j.cmet.2012.08.010> (2012).
- Yao, P. *et al.* Coexpression networks identify brain region-specific enhancer RNAs in the human brain. *Nat Neurosci* **18**, 1168–1174, <https://doi.org/10.1038/nn.4063> (2015).
- Akerman, I. *et al.* Human Pancreatic beta Cell lncRNAs Control Cell-Specific Regulatory Networks. *Cell Metab* **25**, 400–411, <https://doi.org/10.1016/j.cmet.2016.11.016> (2017).
- Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol* **16**, 144–154, <https://doi.org/10.1038/nrm3949> (2015).
- Orom, U. A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46–58, <https://doi.org/10.1016/j.cell.2010.09.001> (2010).
- Ulitsky, I. & Bartel, D. P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154**, 26–46, <https://doi.org/10.1016/j.cell.2013.06.020> (2013).
- Yang, G., Lu, X. & Yuan, L. lncRNA: a link between RNA and cancer. *Biochim Biophys Acta* **1839**, 1097–1109, <https://doi.org/10.1016/j.bbagr.2014.08.012> (2014).
- Jarroux, J., Morillon, A. & Pinskaya, M. History, Discovery, and Classification of lncRNAs. *Adv Exp Med Biol* **1008**, 1–46, https://doi.org/10.1007/978-981-10-5203-3_1 (2017).
- Young, T. L., Matsuda, T. & Cepko, C. L. The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina. *Curr Biol* **15**, 501–512, <https://doi.org/10.1016/j.cub.2005.02.027> (2005).
- Bond, A. M. *et al.* Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry. *Nat Neurosci* **12**, 1020–1027, <https://doi.org/10.1038/nn.2371> (2009).
- Ramos, A. D. *et al.* Integration of genome-wide approaches identifies lncRNAs of adult neural stem cells and their progeny *in vivo*. *Cell Stem Cell* **12**, 616–628, <https://doi.org/10.1016/j.stem.2013.03.003> (2013).
- Clark, B. S. & Blackshaw, S. Understanding the Role of lncRNAs in Nervous System Development. *Adv Exp Med Biol* **1008**, 253–282, https://doi.org/10.1007/978-981-10-5203-3_9 (2017).
- Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888, <https://doi.org/10.1126/science.1156409> (2008).
- Brodie, A., Azaria, J. R. & Ofran, Y. How far from the SNP may the causative genes be? *Nucleic Acids Res* **44**, 6046–6054, <https://doi.org/10.1093/nar/gkw500> (2016).
- Millan, M. J. Linking deregulation of non-coding RNA to the core pathophysiology of Alzheimer's disease: An integrative review. *Prog Neurobiol* **156**, 1–68, <https://doi.org/10.1016/j.pneurobio.2017.03.004> (2017).
- Wang, L. K., Chen, X. F., He, D. D., Li, Y. & Fu, J. Dissection of functional lncRNAs in Alzheimer's disease by construction and analysis of lncRNA-mRNA networks based on competitive endogenous RNAs. *Biochem Biophys Res Commun* **485**, 569–576, <https://doi.org/10.1016/j.bbrc.2016.11.143> (2017).
- Carrieri, C. *et al.* Expression analysis of the long non-coding RNA antisense to Uchl1 (AS Uchl1) during dopaminergic cells' differentiation *in vitro* and in neurochemical models of Parkinson's disease. *Front Cell Neurosci* **9**, 114, <https://doi.org/10.3389/fncel.2015.00114> (2015).
- Soldner, F. *et al.* Parkinson-associated risk variant in distal enhancer of alpha-synuclein modulates target gene expression. *Nature* **533**, 95–99, <https://doi.org/10.1038/nature17939> (2016).
- Kraus, T. F. J. *et al.* Altered Long Noncoding RNA Expression Precedes the Course of Parkinson's Disease—a Preliminary Report. *Mol Neurobiol* **54**, 2869–2877, <https://doi.org/10.1007/s12035-016-9854-x> (2017).
- Vucicevic, D., Schrewe, H. & Orom, U. A. Molecular mechanisms of long ncRNAs in neurological disorders. *Front Genet* **5**, 48, <https://doi.org/10.3389/fgene.2014.00048> (2014).
- Hu, J. *et al.* Systematically characterizing dysfunctional long intergenic non-coding RNAs in multiple brain regions of major psychosis. *Oncotarget* **7**, 71087–71098, <https://doi.org/10.18632/oncotarget.12122> (2016).
- Huang, X., Luo, Y. L., Mao, Y. S. & Ji, J. L. The link between long noncoding RNAs and depression. *Prog Neuropsychopharmacol Biol Psychiatry* **73**, 73–78, <https://doi.org/10.1016/j.pnpbp.2016.06.004> (2017).
- Bannon, M. J. *et al.* Identification of long noncoding RNAs dysregulated in the midbrain of human cocaine abusers. *J Neurochem* **135**, 50–59, <https://doi.org/10.1111/jnc.13255> (2015).

40. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet* **46**, 136–143, <https://doi.org/10.1038/ng.2870> (2014).
41. Sawamoto, K. *et al.* Visualization, direct isolation, and transplantation of midbrain dopaminergic neurons. *Proc Natl Acad Sci USA* **98**, 6423–6428, <https://doi.org/10.1073/pnas.111152398> (2001).
42. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* **41**, e74, <https://doi.org/10.1093/nar/gkt006> (2013).
43. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22**, 1775–1789, <https://doi.org/10.1101/gr.132159.111> (2012).
44. Niederkofler, V., Asher, T. E. & Dymecki, S. M. Functional Interplay between Dopaminergic and Serotonergic Neuronal Systems during Development and Adulthood. *ACS Chem Neurosci* **6**, 1055–1070, <https://doi.org/10.1021/acschemneuro.5b00021> (2015).
45. Grosch, J., Winkler, J. & Kohl, Z. Early Degeneration of Both Dopaminergic and Serotonergic Axons - A Common Mechanism in Parkinson's Disease. *Front Cell Neurosci* **10**, 293, <https://doi.org/10.3389/fncel.2016.00293> (2016).
46. Parras, C. M. *et al.* The proneural gene Mash1 specifies an early population of telencephalic oligodendrocytes. *J Neurosci* **27**, 4233–4242, <https://doi.org/10.1523/JNEUROSCI.0126-07.2007> (2007).
47. Bergsland, M. *et al.* Sequentially acting Sox transcription factors in neural lineage development. *Genes Dev* **25**, 2453–2464, <https://doi.org/10.1101/gad.176008.111> (2011).
48. Cui, H. & Bulleit, R. F. Expression of the POU transcription factor Brn-5 is an early event in the terminal differentiation of CNS neurons. *J Neurosci Res* **52**, 625–632, 10.1002/(SICI)1097-4547(19980615)52:6<625::AID-JNR1>3.0.CO;2-A (1998).
49. Zhang, L. *et al.* Targeted methylation sequencing reveals dysregulated Wnt signaling in Parkinson disease. *J Genet Genomics* **43**, 587–592, <https://doi.org/10.1016/j.jgg.2016.05.002> (2016).
50. Reyes, S. *et al.* GIRK2 expression in dopamine neurons of the substantia nigra and ventral tegmental area. *J Comp Neurol* **520**, 2591–2607, <https://doi.org/10.1002/cne.23051> (2012).
51. Sun, Y. *et al.* The long noncoding RNA SNHG1 promotes tumor growth through regulating transcription of both local and distal genes. *Oncogene*, <https://doi.org/10.1038/nc.2017.286> (2017).
52. Xu, C. *et al.* Systemically identifying and prioritizing risk lncRNAs through integration of pan-cancer phenotype associations. *Oncotarget* **8**, 12041–12051, <https://doi.org/10.18632/oncotarget.14510> (2017).
53. Zhang, M. *et al.* Long noncoding RNA SNHG1 predicts a poor prognosis and promotes hepatocellular carcinoma tumorigenesis. *Biomed Pharmacother* **80**, 73–79, <https://doi.org/10.1016/j.biopha.2016.02.036> (2016).
54. Ulitsky, I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat Rev Genet* **17**, 601–614, <https://doi.org/10.1038/nrg.2016.85> (2016).
55. Zhang, K., Huang, K., Luo, Y. & Li, S. Identification and functional analysis of long non-coding RNAs in mouse cleavage stage embryonic development based on single cell transcriptome data. *BMC Genomics* **15**, 845, <https://doi.org/10.1186/1471-2164-15-845> (2014).
56. Ackermann, A. M., Wang, Z., Schug, J., Naji, A. & Kaestner, K. H. Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Mol Metab* **5**, 233–244, <https://doi.org/10.1016/j.molmet.2016.01.002> (2016).
57. 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, <https://doi.org/10.1038/ng.3043> (2014).
58. Chang, D. *et al.* A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat Genet* **49**, 1511–1516, <https://doi.org/10.1038/ng.3955> (2017).
59. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36, <https://doi.org/10.1186/gb-2013-14-4-r36> (2013).
60. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128, <https://doi.org/10.1186/1471-2105-14-128> (2013).
61. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90–97, <https://doi.org/10.1093/nar/gkw377> (2016).
62. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495–501, <https://doi.org/10.1038/nbt.1630> (2010).
63. Medina-Rivera, A. *et al.* RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Res* **43**, W50–56, <https://doi.org/10.1093/nar/gkv362> (2015).

Acknowledgements

We thank Ildem Akermann for critical reading of the manuscript and Marion Benazza for technical support. We thank the iGenSeq and iCONICS platforms (Institut du Cerveau et de la Moelle Epinière) for performing high throughput sequencing and bioinformatics technical assistance respectively. We are grateful to the platform ImagoSeine-cytométrie en flux (Institut Jacques Monod) for technical support in FACS-sorting. All animal work was conducted at the ICM PHENOPARC Core Facility. The Core is supported by 2 “Investissements d’avenir” (ANR-10-IAIHU-06 and ANR-11-INBS-0011-NeurATRIS) and the “Fondation pour la Recherche Médicale”. The research leading to these results has received funding from the program “Investissements d’avenir” ANR-10-IAIHU-06. J.G. was supported by a fellowship from the French Ministry of research and by the “Fondation pour la Recherche Médicale”. NB received financial support from Edmond J. Safra foundation.

Author Contributions

J. Ge., C.S.-C., H. Ca. and C.P.-G. performed the experiments and analysed data. N.B., J. Gu., D.U. and I.M. performed the bioinformatics analysis. P.R. and H. Ch. designed the experiments, supervised the data analysis and wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-018-37872-1>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019