The coding and long non-coding single-cell atlas of the developing human fetal striatum

Vittoria Dickinson Bocchi[1,2], Paola Conforti[1,2], Elena Vezzoli[1,2,†], Dario Besusso[1,2], Claudio Cappadona[1,2,¥], Tiziana Lischetti[1,2], Maura Galimberti[1,2], Valeria Ranzani[2], Raoul J.P. Bonnal[2, ±], Marco De Simone[2⁺], Grazisa Rossetti[2, ±], Xiaoling He[4], Kenji Kamimoto[5,6,7], Ira Espuny-Camacho[1,2,§], Andrea Faedo[1,2,⋔], Federica Gervasoni[2,3, ±], Romina Vuono[4,₸], Samantha A. Morris[5,6,7], Jian Chen[8], Dan Felsenfeld[8], Giulio Pavesi[1], Roger A. Barker[4], Massimiliano Pagani[2,3, ±*] and Elena Cattaneo[1,2*]

[1] Dipartimento di Bioscienze, Università degli Studi di Milano, Milan, Italy;

[2] INGM, Istituto Nazionale Genetica Molecolare, Milan, Italy;

[3] Dipartimento di Biotecnologie Mediche e Medicina Traslazionale, Università degli Studi di Milano, Milan, Italy.

[4] WT-MRC Cambridge Stem Cell Institute and Department of Clinical Neuroscience, University of Cambridge, Cambridge, UK;

[5]Department of Developmental Biology, Washington University School of Medicine in St. Louis. 660 S. Euclid Avenue, Campus Box 8103, St. Louis, MO 63110, USA

[6]Department of Genetics, Washington University School of Medicine in St. Louis. 660 S. Euclid Avenue, Campus Box 8103, St. Louis, MO 63110, USA

[7]Center of Regenerative Medicine. Washington University School of Medicine in St. Louis. 660 S. Euclid Avenue, Campus Box 8103, St. Louis, MO 63110, USA

[8] CHDI Management/CHDI Foundation, New York, U.S.A;

*Corresponding authors: elena.cattaneo@unimi.it; massimiliano.pagani@unimi.it

[†]Current affiliation: Department of Biomedical Sciences for Health, Università degli Studi di Milano, Via G. Colombo 71, 20133 Milan, Italy
[¥]Current affiliation: Department of Biomedical Sciences, Humanitas University, Via Rita Levi Montalcini 4, Pieve Emanuele 20090 Milan, Italy.
[±]Current affiliation: IFOM- FIRC Institute of Molecular Oncology, Milan, Italy.
[⁺]Current affiliation: Department of Radiation Oncology, Cedars-Sinai Medical Center, Los Angeles, CA, USA
[§]Current affiliation: Université de Liège, GIGA Stem Cells, Quartier hôpital 15, Avenue Hippocrate, B-4000 Liège, Belgium
[⋔]Current affiliation: Axxam, OpenZone, Via Meucci 3 20091 Bresso, Milan, Italy
[₸] Current affiliation: Medway School of Pharmacy, University of Kent, Chatham, Kent, United Kingdom.

Abstract

Deciphering how the human striatum develops is paramount to understand diseases affecting this region. To decode the transcriptional modules that regulate this structure during development we first catalogued 1116, de novo identified, lincRNAs and then profiled 96,789 single-cells from the early human fetal striatum. We found that D1 and D2 medium spiny neurons (MSNs) arise from a common progenitor and that lineage commitment is established during the post-mitotic transition, across a pre-MSN phase that exhibits a continuous spectrum of fate determinants. We then uncovered cell type-specific gene regulatory networks that we validated through *in silico* perturbation. Finally, we identified human-specific lincRNAs that contribute to the phylogenetic divergence of this structure in humans. In conclusion, our study has delineated the cellular hierarchies governing MSN lineage commitment.

One Sentence Summary

Bulk and single-cell RNA-seq reveal the human-specific modules governing human striatal development with the identification of previously unannotated cellular states and gene regulatory networks that define the ontogeny of medium spiny neurons.

Main Text

The striatum is a subcortical structure made up of the caudate nucleus and putamen and is important in motor control and learning, procedural behaviour as well as cognition, emotional and motivational responses. Many of these functions are associated with uniquely human abilities including the development of speech and language (*1*).

In adults, the striatum is primarily composed of medium spiny neurons (MSNs) carrying either D1-type or D2-type dopamine receptors interspersed with aspiny interneurons (*2*). This organization does not appear to vary among species (*3*). However, this apparent homogeneity in cellular composition mask functional complexity of human-specific striatal circuits.

During development, MSNs are derived from progenitors in the lateral ganglionic eminence (LGE) of the telencephalon (*4*). Most studies of this area have concentrated on rodent models and the coding signature of limited cell types. Long non-coding RNAs (lncRNA) have a higher tissue specificity than mRNAs (*5*) and a diversity that correlates more closely with brain complexity than protein-coding (PC) genes (*6*). LncRNAs may therefore, better discriminate cell states during striatal development in human (*3*).

In this study, we first conducted bulk RNA-seq to identify novel intergenic long non-coding RNAs (lincRNA) expressed in the developing human LGE that we then leveraged to define the spatial coding and non-coding map for this area that discriminates it from the surrounding neocortex (CX) and medial ganglionic eminence (MGE). We then performed single-cell RNA-seq (scRNA-seq) to characterize cell diversity in the LGE, identify LGE specific coding and non-coding cell states, define cell-type specific gene regulatory networks and decipher pivotal bifurcation points in the formation of human MSNs. Finally, we performed *in silico* perturbations by knocking out and overexpressing key transcription factors (TFs) of the MSN-lineage, together with immunohistochemistry (IHC) and fluorescence in situ hybridization (FISH) analysis to validate our findings.

lincRNA discovery and scRNA-seq profiling of the human LGE

In order to identify unannotated putative lincRNAs expressed during human striatal development, we dissected the CX, LGE and MGE from human embryos between 7 and 20 postconceptional weeks (pcw), with at least 2/3 biological replicates for each developmental stage (Table S1). We then profiled by RNA-seq the transcriptome of each area (Fig. 1A) and set up a computational pipeline (Fig. S1A) to create a corresponding catalogue of lincRNAs (Fig. S1B-E). Since our bulk RNA-seq protocol was unstranded we decided to include only a reliable set of intergenic lncRNAs (lincRNAs), as genic lncRNAs are difficult to correctly predict without strand information discriminating them from overlapping genes. To better understand the relationship between lincRNAs identified in other tissues and our newly identified lincRNAs, two catalogues derived from different human tissues and cell types (*7*, *8*), plus a lincRNA catalogue of the developing human CX (*9*) were integrated into this study.

We detected 1116 novel lincRNA loci (Fig. 1B, Table S2), among which the highest number was found in the developing LGE and the lowest in the CX (Fig. S1F), probably due to the greater extent of data available from the CX. We found that lincRNAs had an average of 2.27 exons as for all lincRNA catalogues integrated in this study, except the FANTOM catalogue which used CAGE-seq instead of classic RNA-seq and showed fewer exons (Fig. S1G). For each sampled pcw, we then defined a signature of both PC genes and lincRNAs that were uniquely expressed in the LGE, CX, MGE (Fig. 1C, Fig. S2A, Table S3). Gene ontology (GO) analysis in the LGE between 7 and 11pcw revealed an enrichment for terms related to neural differentiation, forebrain and subpallium development (Fig. S2B, Table S4), while at 20pcw, this changed to regulation of synaptic plasticity and neurotransmitter secretion, reflecting the more mature state of the 20pcw striatum (Fig. S2C, Table S4).

Pathway enrichment analysis (IPA, Ingenuity Systems) revealed huntingtin (HTT) as the most significant upstream regulator (p-value < 0.001) of the LGE at all-time points considered (Fig. 1D, Fig. S2D, Table S5). This suggests that HTT may have an important role in human striatal development.

To explore how these PC genes and lincRNAs define specific cell states during development of the human striatum, we surveyed 96,789 high quality single cells from the LGE between 7 and 11pcw (Fig. 1A, Table S1). We were able to discriminate 15 clusters and their transcriptional signatures (Fig. 1E, Table S6) that were then classified according to canonical markers (Fig. 1F). Biological replicates were well distributed in each detected cluster (Fig. S3A) and all time-points considered between 7 and 11pcw covered the identified clusters (Fig. S3B, C), indicating that during this stage of early development the same cell types are present. We report that both D1- and D2-MSNs subtypes are present in all time-points considered (Fig. S3C) and they are characterized by known markers such as *ISL1*, *EBF1* in D1- and *SIX3* and *SP9* in D2-MSNs (Fig. 1F, Fig. S4A-C).

GO analysis revealed that apical progenitors (APs) are associated with terms like cell adhesion and glial cell differentiation, mirroring the proliferating nature of this cell type. Basal progenitors (BPs) express genes related to mRNA splicing, RNA binding, and forebrain development suggesting that a network of posttranscriptional control coordinates and primes these cells for their final cell division and fate acquisition. Genes identified in pre-MSN are enriched in terms linked to nervous system development and axon guidance, while D1- and D2-MSNs show a more mature gene signature linked to synapse organization and chemical synaptic transmission (Fig. S4D, Table S7).

Marginally contaminating populations of cells included *NKX2.1*-, *LHX6*- and *LHX8*-expressing MGE interneurons, migrating interneurons, ventral neocortical cells, ventral CGE cells and endothelial cells (Fig. S5A-C). By sub-clustering the migrating interneurons we were able to observe dorsal LGE (dLGE) interneurons expressing *SP8*, *COUP-TFII* (*NR2F2*) together with *PAX6* and MGE migrating interneurons expressing *NXK2.1* and *LHX6* that were presumably passing through the sub-ventricular zone (SVZ) of the LGE at this developmental stage since the same population did not express *LHX8* (Fig. S5C, D). As expected, *ERBB4* was found in both types of migrating interneurons (Fig. S5D). The finding that these two populations cluster together despite their different sites of origin and that sub-clustering was required to reveal their diversity, suggests that their interneuron migratory class signature is stronger than their spatial/lineage of origin signature. We did not observe any markers of oligodendrocytes, astrocytes or microglia, indicating that these cells appear later in normal human striatal development and found no evidence of corridor cells that have probably migrated to the globus pallidus by 7pcw in humans.

As previously reported (*7*), in our study bulk expression levels of lincRNAs are lower than those of PC genes, and our single-cell data also show differential expression for these two biotypes (Fig. S5E). However, when comparing the different catalogues we observed that lincRNAs identified in the developing human CX (*9*) and the ones identified in this study (HFB) displayed higher transcript levels, both in bulk and single-cell data, compared to lincRNAs identified in different adult tissues (Fig. 1G). This suggests that these lincRNAs are associated with the development of the human telencephalon. Overall lincRNAs are more specific than PC genes both in bulk and single-cell measurements (Fig. S5F). We show that these lincRNAs characterize different cell states of the LGE lineage as, for example, the lincRNA *hfb_G_000907* is associated with the APs whereas *hfb_G_000258* characterizes pre-MSNs and D2-MSNs (Fig. 1H).

MSN differentiation passes through a pre-MSN cell state

To infer how fate decisions occur, we defined the cell trajectories that give rise to MSNs using velocyto (*10*, *11*). The velocity field map (Fig 2A) reflects the dynamics of MSN differentiation as these cells transition from APs to D1- and D2-MSNs, passing through a pre-MSN phase. This model shows that, in the pre-MSN phase, there is already a separation between the D1 and D2 state (Fig. 2A). We show that lincRNAs display reduced splicing kinetic compared to PC genes (Fig. 2B) reflecting their inefficient splicing (*12*). However, within the putative driver genes (Table S8), we identify a number of lincRNAs that may guide MSN differentiation. These include *hfb_G_000259* that shows pronounced velocities in BPs, pre-MSNs and D2-MSNs and *hfb_G_000707* that contributes to the D1-MSNs transition (Fig. 2C).

Overall, our data suggest that the same progenitors give rise to both D1- and D2-MSNs. Sub-clustering the AP and BP groups confirm the absence of lineage commitment at these stages (Fig. S6A, B). This is further proved by calculating the connectivity between each cluster of the LGE lineage using partition-based graph abstraction (PAGA) which shows that progenitors are highly interconnected and converge to give rise to D1- and D2-MSNs through the pre-MSN phase (Fig. S6C). To explore the potential temporal relationships between cell states in the LGE lineage we used a measure of graph distance (diffusion pseudotime, DPT) (Fig. S6D, E). We observe that cell states and DPT recapitulate the known temporal dynamics of neural maturation, with mitotic APs of the ventricular zone (VZ) having the lowest DPT score, while mitotic BPs of the SVZ have an intermediate DPT score (Fig. S6E). These findings suggest that the latter cells represent a more mature progenitor state, as reflected by the expression of *HES6* which is found in cells committing to neural differentiation. Pre-MSNs together with D1- and D2-MSNs were classified as post-mitotic (Fig. S6F) and showed the highest DPT score, which peaked in cells classified as mature D1-MSNs (Fig. S6E). This cell population was the only one to express the more mature neuronal cytoskeleton marker neurofilament-M (*NEFM)* (Fig. S6G), and most cells of this cluster are found at the most advanced pcw analysed (Fig. S3C). This temporal signature was confirmed using FISH (Fig. S6H). Finally, at the cluster level, we found that splicing kinetics and differentiation accelerates substantially after cell cycle exit (especially in pre-MSNs and D2 neurons), maintaining pace during D2 production while slowing down during D1 production (Fig. S6I). This finding is consistent with a model in which immature differentiating cells (D2-MSNs) are less transcriptionally stable and reveal greater splicing dynamics than terminally differentiated cells (D1-MSNs) (*13*).

Overall, these findings suggest that both MSN subtypes are present at this early stage of development but D1-MSNs mature and reach equilibrium at a faster rate compared to D2-MSNs.

We then focused on pre-MSNs as this cell state escaped previous identification. Sub-clustering pre-MSNs revealed that these cells exhibit either a D1- or D2-MSN blueprint with enrichment of cell-specific markers of each subtype (Fig. 2D). Although sub-clustering revealed two discrete cell-states, markers like *SIX3, SP9* and *ISL1* show a gradient of opposing expression, with both markers found in both sub-clusters at different expression levels. We therefore tested different thresholds of *SIX3* and *ISL1* co-expression in the entire LGE lineage to test how the transcripts behaved throughout differentiation. We found that high levels of *SIX3* and low levels of *ISL1* are present in the D2-MSN lineage at the pre-MSN stage, and that co-expression is lost at terminal

D2-MSN fates (Fig. S7A). The opposite trend is seen for high *ISL1* and low *SIX3* in the D1 MSN-lineage (Fig. S7A). High levels of both transcripts are instead found in a very small percentage of cells (approximately 7%) and this is also reflected at the protein level (Fig. S4B, C). This suggests that the pre-MSN phase spans along a transcriptional continuum rather than being a discrete cell state.

We then identified *OCT6* (*POU3F1*), a member of the POU-III subfamily, as a specific marker of this pre-MSN cell state (Fig. 2E), as well as being LGE specific (Fig. 2F) and exhibiting transient expression (Fig. 2G). When looking at its expression at the pre-MSN stage we also show that this gene is more enriched in pre-D1 MSNs and appears to follow an expression gradient (Fig. S7B). Examining co-expression levels of this transcript with candidate D1- and D2-MSN markers we observe that low levels of *OCT6-ISL1* characterize the D1 lineage, while high levels define the pre-D1 state (Fig. 2H). Low levels of *OCT6-SIX3* are also found in D2-MSNs, however, very few positive cells are found with high expression of both markers (Fig. 2H). To confirm these findings, we performed systematic IHC analysis of these markers at 9pcw. We show that there is specific OCT6 staining in the SVZ of the LGE (Fig. 2I, J) with none in the CX, MGE and CGE (Fig. S7C). We demonstrate that OCT6 positive cells are post-mitotic neurons as only 2% are double positive for the BP marker ASCL1 (Fig. S7D-G) and more than 80% of OCT6 cells do not express the proliferative marker Ki67 in the SVZ (Fig. S7H, I). We show that 40% of OCT6 cells are positive for the D1-MSN marker ISL1, and they are all negative for CTIP2, a marker of mature MSNs (Fig. 2J-L), confirming this marker as a key gene of pre-D1 MSNs in the SVZ.

For pre-D2 MSNs, we found approximately 10% of OCT6 and SIX3 co-expressing cells in the SVZ (Fig. S8A-D). This suggests that the low OCT6 expression level characterizing this cell state translates in only few cells detectable by IHC. This low to high gradient is also seen by IHC where

SIX3 cells are enriched dorsally and fade moving ventrally, while OCT6 shows a reverse ventral to dorsal enrichment (Fig. S8A). These patterns of OCT6-ISL1 and OCT6-SIX3 co-expression in the SVZ are also confirmed at 11pcw (Fig. S8E-H). To confirm the pre-D2 state we looked at how ASCL1, SIX3 and CTIP2 behave in the SVZ (Fig. S8I). We show that most of the SIX3 expressing cells in the SVZ lack expression of both ASCL1 and CTIP2 (Fig. S8J-L) confirming the presence of the pre-D2 state in the SVZ.

Our findings support a model where D1- and D2-MSNs derive from a common progenitor and that specific subtypes of MSNs are established once cells become post-mitotic as they pass through a primed precursor phase that exhibits heterogeneity in the expression of key MSN regulators. This suggests that fate determination in MSNs does not occur as an instantaneous shift in cell fate, but is a smooth and continuous process that falls within two gradients of expression.

Gene regulatory networks of the MSN-lineage

The human LGE domain has been the most elusive area of the basal ganglia to characterize. To bridge this gap we applied SCENIC (*14*) to our scRNA-seq data and reconstructed the gene regulatory networks (GRNs) and combinatorial codes of TFs that define the different cell states in this domain (Fig. 3A, Table S9).

Overall, GRN analysis revealed that distinct regulons classify different axes of MSN development, with D1- and D2-MSNs sharing most of the active TFs (Fig. 3B). We found a number of TFs that were not previously associated to LGE development. These include: *OTX1* for APs, *VAX1* for BPs, *POU2F2* for both types of MSNs, *NANOG* in D2-MSNs and *FOXO1* for D1-MSNs.

We then used SCENIC co-expression networks to infer potential relationships between our *de novo* identified lincRNAs and our cell-type specific TFs (Fig. S9A, Table S10), as it has been shown that lincRNAs loci contain many conserved TF binding sites (*12*), suggesting that their relationship may be functionally relevant. We found that *hfb_G_000907*, a specific lincRNA of APs, is strongly linked to *SOX* genes while, *hfg_G_000296*, another lincRNA specific for APs, shows a high connection with a PDZ-LIM domain family protein called *PDLIM5*. The BP lincRNA *hfb_G_00882* shows a correlation with *RARA*, an important mediator of retinoid signalling. For MSNs, we find common lincRNAs between D1- and D2-MSN subtypes that are in the same network with universal MSN signatures like *ZNF467* and *POU2F2*. *hfb_G_000494* a D1-MSN specific lincRNA shows a connection with *SOX8* and *IKZF1*, two players of the D1-MSN class, while many of the D2-specific lincRNAs are linked to *SP9* and *MAFB*. Overall, these observations reveal insights into how this panel of lincRNAs can be regulated.

LGE and MGE progenitors are transcriptionally distinct

Given the limited information available on APs of the LGE we then focused our attention on this particular progenitor domain. We find that *SOX3* and *TCF7L1* (both inhibitors of posteriorizing WNT signals) together with *SOX9* (which is induced by SHH signalling) are active TFs that may ventralize and maintain this early progenitor phase (Fig. 4A). Other significant TFs of this progenitor domain include *OTX1*, which has been predominantly associated with the CX, and *CREB5* which has a role in neural progenitor differentiation. These regulators are interlinked by shared target genes and are likely important to control cell fate decisions within the AP domain. The *TCF7L1* network does not share target genes with the other TFs, however it does regulate genes like *NKX2.1* and *OTX2* that play a role in maintaining a regional ventral identity (Fig. 4A). We then investigated whether transcriptional diversity exists at early developmental stages between APs of the LGE and MGE. We combined our set of LGE marker genes (compared to the

CX and MGE) identified in the bulk dataset with the signatures of the different cell states identified in the scRNA-seq dataset (Fig. 4B). We identified 199 genes that are LGE and cell state-specific (Fig. 4B, Table S11), within these, 31 were specific for APs of the LGE and for two of these candidates, the TF *SALL3* and its topologically adjacent *LINC01896* (Fig. 4C, D), we confirmed by FISH that they are expressed in APs of the LGE (and CGE) and not in the MGE (Fig. 4E, Fig. S10A-D). Our results propose a model where the neuronal fate of APs in the VZ of the developing striatum is already restricted by a specific topographical and transcriptional program that subsequently drives LGE versus MGE specific cell fates.

Gene network interference reveals key MSN fate determinants

Within the striatum, D1- and D2-MSNs contribute to the direct and indirect pathways, respectively (*15*). Here, we show that both types of MSNs are shaped by a shared set of GRNs governed by specific MSN 'master' TFs (Fig. 5A, Fig. S11A, B). To validate these observations and predict their role in cell lineage determination, we used CellOracle (*16*) to perturb, *in silico*, these key TFs and their GRNs. We tested CellOracle on *SP9*, a functionally validated gene in MSN fate determination (*17*). We show that *SP9* knockout (KO) causes a block in BP differentiation, together with a specific arrest in D2-MSN generation, that shift towards D1-MSN fates (Fig. S11C). When we tested *SP9* overexpression, the opposite trend is observed, with a D1- to D2-MSN shift (Fig. S11C). This recapitulates what is known on SP9 during striatal development (*17*) and confirms the ability of CellOracle in identifying functional TFs. We then tested *ZNF467*, a zinc finger protein whose function has not been yet characterized in any tissue. We show that its KO halts MSN differentiation at the progenitor stage, while it promotes MSN specification when overexpressed (Fig. 5B). This suggests that ZNF467 may be an important TF for the entire MSN-lineage. We then tested *OCT6* and we show that KO of this gene causes an inhibition of the pre-MSN phase, with cells reverting back to progenitor domains and also a stimulation of the transition from pre-MSN to mature MSN states (Fig. 5C). This is in line with the transient nature of *OCT6* expression and may reflect the result of removing this gene at different stages of differentiation. Overexpressing *OCT6* has the opposite effect as it stalls differentiation in the pre-MSN phase (Fig. 5C). This data confirms the probable function of this gene in the pre-MSN phase, especially in pre-D1 MSNs since we detect OCT6 protein translation mainly in the D1 lineage (Fig. 2I-L).

We then tested key TFs that may drive the D1-D2 MSN bifurcation point. We show that *IKZF1*, an essential gene in the generation of D2-MSNs in mice (*18*), may have a completely opposing role during human development as we find it associated to D1-MSNs. Perturbation of this gene causes a specific obstruction in D1-MSN maturation, while its overexpression leads to increased D1 production (Fig. 5D). Finally, we tested *MAFB*, a specific TF of the D2-lineage, that has been previously associated with the survival of MGE-derived cortical interneurons (*19*). KO of *MAFB* causes an arrest in BP differentiation and an interruption in D2-production that shifts to the D1-lineage (Fig. 5E), mimicking what is seen with *SP9* KO. Instead, overexpression of *MAFB* triggers BP maturation, a conversion from D1- to D2-MSNs and promotes a conversion from pre-MSN to interneurons (Fig. 5E), supporting the role of this gene in promoting this cell class.

Since most studies in humans have generated a general MSN or a D1-MSN-specific signature (*20*)., we next decided to characterize a highly specific D2-MSN signature. From our combined bulk and scRNAseq data (Fig. 4B) we defined 13 specific LGE and D2-MSN genes (Table S11). Among them, we validated *LINC01305* and *KCNA5* (Fig. S12A) that are specific for the mantle zone (MZ) of the LGE and are not found in the MGE or CGE (Fig. S12B-E).

Finally, to discover different populations within each major class of MSNs, we performed sub-clustering of the two classes of MSNs. We reveal two sub-clusters in both D1- and D2-MSNs (Fig. 5F), and we tested whether these sub-clusters were already specified in the patch compartment, one of the two major areas (the other being the matrix) that characterize the organization of the striatum and that is formed during early development in mouse models (*21*). A previous single-cell study on post-natal mouse striatal cells (*22*) showed that *Pdyn* and *Tshz1* are specific patch markers. Here, we find that one of the clusters in the D1 population expresses *PDYN* and low levels of *TSHZ1*, while in D2-MSNs the opposite trend is observed (Fig. 5F). This suggests that in humans these two markers define MSN lineage-specified patch cells. *PDYN* expression in the MZ confirms that this organization is present at 9pcw (Fig.5G). Overall, our data suggest that human MSNs are already compartmentalized in sub-type specific patch regions at this stage of early development and we define how specific markers are distributed in these MSN subtypes.

Human specific lincRNAs show striatal specificity

We then decided to investigate how lincRNAs define striatal evolution and development. Using liftOver (*23*) and TransMap (*24*) we identified lincRNAs that map across different species, have conserved sequence identity and are also expressed in the other species considered (Fig. 6A). Conservation scores showed that lincRNAs identified in the brain, exhibit a higher score in primates compared to the mouse, suggesting that they are less conserved in lower mammals (Fig. 6B). However, conservation scores in all the species considered increased for lincRNA expressed in the adult brain compared to the developing brain (Fig. 6B). This suggests that the transcriptional complexity of the primate brain, in terms of lincRNAs, is mainly established during fetal development and that in the adult brain this biotype probably possesses more conserved functions among different species.

We further explored two human-specific lincRNAs that were not conserved in the other species considered, and that were explicitly expressed in MSNs of the developing human LGE (Fig. 6C, D). We initially used a guilt-by-association approach (*25*), to investigate the potential function of PC genes that highly correlate with this set of lincRNAs, in order to predict the latter's potential role in MSN specification. We inferred that the MSN-specific lincRNA *hfb_G_000053* could be involved in brain development and neuronal differentiation (Fig. 6E, Table S12) while the mature D1-MSN-specific lincRNA *hfb_G_000494*, could function in synaptic organization (Fig. 6F, Table S12). We validated these unannotated lincRNAs by RT-PCR and FISH and found a specific enrichment of these lincRNAs in the MZ of the developing human striatum compared to the MGE and CGE (Fig. 6G, H, Fig. S13A-E).

These findings suggest that these two lincRNAs have a role in the development of human MSNs and may provide insight into how the striatal architecture has evolved to accommodate the expanded human behavioral repertoire.

Discussion

Recent studies of the developing human fetal brain have been crucial to decipher the basis of human brain formation, function and evolution (*26–29*).

In this study, we created the first high resolution single-cell map of the early development of the human striatum from a coding and long non-coding perspective. This atlas will pave the way for future integrations with chromatin accessibility data and single-cell proteomics that will help resolve the relationship between gene regulation, transcription, and protein production during lineage determination in MSNs, especially at the pre-MSN phase identified in this study.

The main limitation of this work is the high degree of sparsity of the single-cell data. Future protocols with increased sensitivity of RNA detection may reveal new and rare cell types and will probably enable a clearer separation of the pre-MSN phase. This study was also limited by the restricted time window (early fetal development) and brain region (LGE) that was examined. Studies of both earlier and later developmental time-points combined with single-cell measurements of the CX, MGE and CGE will enhance our understanding of lineage establishment and diversification.

Nonetheless, this study has set the foundation for understanding human striatal development at unprecedented granularity. We foresee that key TFs and lincRNAs defined in this study will be leveraged *in vitro* to generate authentic MSNs that can serve as suitable donor preparations for future clinical trials in cell replacement therapies. This dataset will also be critical to understand the neurodevelopmental component of Huntington's disease (HD) during striatal development given the recent evidence of alterations in normal cortical development in human HD fetal samples (*30*). Finally, we also expect that the newly identified human-specific lincRNAs will enable us to understand the underpinnings that characterize human striatum function.

References and Notes

1. M. A. Raghanti, M. K. Edler, A. R. Stephenson, L. J. Wilson, W. D. Hopkins, J. J. Ely, J. M. Erwin, B. Jacobs, P. R. Hof, C. C. Sherwood, Human-specific increase of dopaminergic innervation in a striatal region associated with speech and language: A comparative analysis of the primate basal ganglia. *J. Comp. Neurol.* (2016), doi:10.1002/cne.23937.

2. C. R. Gerfen, The neostriatal mosaic: multiple levels of compartmental organization. *Trends Neurosci.* **15**

(1992), pp. 133–139.

3. S. Grillner, B. Robertson, M. Stephenson-Jones, The evolutionary origin of the vertebrate basal ganglia and its role in action selection. *J. Physiol.* (2013), , doi:10.1113/jphysiol.2012.246660.

4. J. Stiles, T. L. Jernigan, The basics of brain development. *Neuropsychol. Rev.* **20** (2010), pp. 327–348.

5. M. E. Dinger, K. C. Pang, T. R. Mercer, J. S. Mattick, Differentiating protein-coding and noncoding RNA: Challenges and ambiguities. *PLoS Comput. Biol.* **4** (2008), , doi:10.1371/journal.pcbi.1000176.

6. S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, T. R. Gingeras, Landscape of transcription in human cells. *Nature*. **489**, 101–108 (2012).

7. M. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, J. L. Rinn, Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).

8. C. C. Hon, J. A. Ramilowski, J. Harshbarger, N. Bertin, O. J. L. Rackham, J. Gough, E. Denisenko, S. Schmeier, T. M. Poulsen, J. Severin, M. Lizio, H. Kawaji, T. Kasukawa, M. Itoh, A. M. Burroughs, S. Noma, S. Djebali, T. Alam, Y. A. Medvedeva, A. C. Testa, L. Lipovich, C. W. Yip, I. Abugessaisa, M. Mendez, A. Hasegawa, D. Tang, T. Lassmann, P. Heutink, M. Babina, C. A. Wells, S. Kojima, Y. Nakamura, H. Suzuki, C. O. Daub, M. J. L. De Hoon, E. Arner, Y. Hayashizaki, P. Carninci, A. R. R. Forrest, An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature*. **543**, 199–204 (2017).

9. S. J. Liu, T. J. Nowakowski, A. A. Pollen, J. H. Lui, M. A. Horlbeck, F. J. Attenello, D. He, J. S. Weissman, A. R. Kriegstein, A. A. Diaz, D. A. Lim, Single-cell analysis of long non-coding RNAs in the developing

human neocortex. *Genome Biol.* **17** (2016), doi:10.1186/s13059-016-0932-1.

10. G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, P. V. Kharchenko, RNA velocity of single cells. *Nature* (2018), doi:10.1038/s41586-018-0414-6.

11. V. Bergen, M. Lange, S. Peidli, F. A. Wolf, F. J. Theis, Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* (2020), doi:10.1038/s41587-020-0591-3.

12. M. Melé, K. Mattioli, W. Mallard, D. M. Shechner, C. Gerhardinger, J. L. Rinn, Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res.* **27**, 27–37 (2017).

13. J. Q. Wu, L. Habegger, P. Noisa, A. Szekely, C. Qiu, S. Hutchison, D. Raha, M. Egholm, H. Lin, S. Weissman, W. Cui, M. Gerstein, M. Snyder, Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 5254–9 (2010).

14. S. Aibar, C. B. González-Blas, T. Moerman, V. A. Huynh-Thu, H. Imrichova, G. Hulselmans, F. Rambow, J. C. Marine, P. Geurts, J. Aerts, J. Van Den Oord, Z. K. Atak, J. Wouters, S. Aerts, SCENIC: Single-cell regulatory network inference and clustering. *Nat. Methods* (2017), doi:10.1038/nmeth.4463.

15. D. J. Surmeier, W. J. Song, Z. Yan, Coordinated expression of dopamine receptors in neostriatal medium spiny neurons. *J. Neurosci.* (1996), doi:10.1016/S1054-3589(08)60921-7.

16. K. Kamimoto, C.M. Hoffmann, S.A. Morris, CellOracle: Dissecting cell identity via network inference and in silico gene perturbation. *bioRxiv* (2020).

17. Q. Zhang, Y. Zhang, C. Wang, Z. Xu, Q. Liang, L. An, J. Li, Z. Liu, Y. You, M. He, Y. Mao, B. Chen, Z. Q. Xiong, J. L. Rubenstein, Z. Yang, The Zinc Finger Transcription Factor Sp9 Is Required for the Development of Striatopallidal Projection Neurons. *Cell Rep.* **16**, 1431–1444 (2016).

18. R. Martín-Ibáñez, E. Crespo, N. Urbán, S. Sergent-Tanguy, C. Herranz, M. Jaumot, M. Valiente, J. E. Long, J. R. Pineda, C. Andreu, J. L. R. Rubenstein, Ó. Marín, K. Georgopoulos, G. Mengod, I. Fariñas, O. Bachs, J. Alberch, J. M. Canals, Ikaros-1 couples cell cycle arrest of late striatal precursors with neurogenesis of enkephalinergic neurons. *J. Comp. Neurol.* (2010), doi:10.1002/cne.22215.

19. E. L. L. Pai, J. Chen, S. F. Darbandi, F. S. Cho, J. Chen, S. Lindtner, J. S. Chu, J. T. Paz, D. Vogt, M. F.

Paredes, J. L. R. Rubenstein, Maf and mafb control mouse pallial interneuron fate and maturation through neuropsychiatric disease gene regulation. *Elife* (2020), doi:10.7554/eLife.54903.

20. M. Onorati, V. Castiglioni, D. Biasci, E. Cesana, R. Menon, R. Vuono, F. Talpo, R. Laguna Goya, P. A. Lyons, G. P. Bulfamante, L. Muzio, G. Martino, M. Toselli, C. Farina, R. A Barker, G. Biella, E. Cattaneo, Molecular and functional definition of the developing human striatum. *Nat. Neurosci.* **17**, 1804–1815 (2014).

21. S. M. Kelly, R. Raudales, M. He, J. H. Lee, Y. Kim, L. G. Gibb, P. Wu, K. Matho, P. Osten, A. M. Graybiel, Z. J. Huang, Radial Glial Lineage Progression and Differential Intermediate Progenitor Amplification Underlie Striatal Compartments and Circuit Organization. *Neuron* (2018), doi:10.1016/j.neuron.2018.06.021.

22. A. Saunders, E. Z. Macosko, A. Wysoker, M. Goldman, F. M. Krienen, H. de Rivera, E. Bien, M. Baum, L. Bortolin, S. Wang, A. Goeva, J. Nemesh, N. Kamitaki, S. Brumbaugh, D. Kulp, S. A. McCarroll, Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* (2018), doi:10.1016/j.cell.2018.07.028.

23. R. M. Kuhn, D. Haussler, W. James Kent, The UCSC genome browser and associated tools. *Brief. Bioinform.* (2013), doi:10.1093/bib/bbs038.

24. J. Zhu, J. Z. Sanborn, M. Diekhans, C. B. Lowe, T. H. Pringle, D. Haussler, Comparative genomics search for losses of long-established genes on the human lineage. *PLoS Comput. Biol.* (2007), doi:10.1371/journal.pcbi.0030247.

25. M. Guttman, I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn, E. S. Lander, Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*. **458**, 223–227 (2009).

26. T. J. Nowakowski, A. Bhaduri, A. A. Pollen, B. Alvarado, M. A. Mostajo-Radji, E. Di Lullo, M. Haeussler, C. Sandoval-Espinosa, S. J. Liu, D. Velmeshev, J. R. Ounadjela, J. Shuga, X. Wang, D. A. Lim, J. A. West, A. A. Leyrat, W. J. Kent, A. R. Kriegstein, Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science (80-.).* (2017), doi:10.1126/science.aap8809.

27. G. La Manno, D. Gyllborg, S. Codeluppi, K. Nishimura, C. Salto, A. Zeisel, L. E. Borm, S. R. W. Stott, E.

M. Toledo, J. C. Villaescusa, P. Lönnerberg, J. Ryge, R. A. Barker, E. Arenas, S. Linnarsson, Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell*. **167**, 566-580.e19 (2016).

28. X. Fan, Y. Fu, X. Zhou, L. Sun, M. Yang, M. Wang, R. Chen, Q. Wu, J. Yong, J. Dong, L. Wen, J. Qiao, X. Wang, F. Tang, Single-cell transcriptome analysis reveals cell lineage specification in temporal-spatial patterns in human cortical development. *Sci. Adv.* **6**, 1–16 (2020).

29. S. Zhong, S. Zhang, X. Fan, Q. Wu, L. Yan, J. Dong, H. Zhang, L. Li, L. Sun, N. Pan, X. Xu, F. Tang, J. Zhang, J. Qiao, X. Wang, A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* (2018), doi:10.1038/nature25980.

30. M. Barnat, M. Capizzi, E. Aparicio, S. Boluda, D. Wennagel, R. Kacher, R. Kassem, S. Lenoir, F. Agasse, B. Braz, J.-P. Liu, J. Ighil, A. Tessier, S. Zeitlin, C. Duyckaerts, M. Dommergues, A. Durr, S. Humbert, Huntington's disease alters human neurodevelopment. *Science (80-.).* (2020), doi:10.1126/science.aax3338.

31. C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. Van Baren, S. L. Salzberg, B. J. Wold, L. Pachter, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* (2010), doi:10.1038/nbt.1621.

32. P. Ewels, M. Magnusson, S. Lundin, M. Käller, MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. **32**, 3047–3048 (2016).

33. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**, 2114–2120 (2014).

34. J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. Van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, T. J. Hubbard, GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* **22**, 1760–1774 (2012).

35. M. Burset, I. A. Seledtsov, V. V Solovyev, Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* **28**, 4364–4375 (2000).

36. C. Trapnell, B. a Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, L. Pachter, Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new

transcripts and switching among isoforms. *Nat. Biotechnol.* **28**, 511–515 (2011).

37. M. Pertea, G. M. Pertea, C. M. Antonescu, T. C. Chang, J. T. Mendell, S. L. Salzberg, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

38. A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, L. Pachter, Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* **12** (2011), doi:10.1186/gb-2011-12-3-r22.

39. V. Wucher, F. Legeai, B. Hédan, G. Rizk, L. Lagoutte, T. Leeb, V. Jagannathan, E. Cadieu, A. David, H. Lohi, S. Cirera, M. Fredholm, N. Botherel, P. A. J. Leegwater, C. Le Béguec, H. Fieten, J. Johnson, J. Alföldi, C. André, K. Lindblad-Toh, C. Hitte, T. Derrien, FEELnc: A tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res.* **45**, 1–12 (2017).

40. A. R. Quinlan, BEDTools: The Swiss-Army tool for genome feature analysis. *Curr. Protoc. Bioinforma.* (2014), doi:10.1002/0471250953.bi1112s47.

41. W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. **8**, 118–127 (2007).

42. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).

43. G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W. H. Fridman, F. Pagès, Z. Trajanoski, J. Galon, ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*. **25**, 1091–1093 (2009).

44. G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, S. Wang, GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*. **26**, 976–978 (2010).

45. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* (2018), doi:10.1186/s13059-017-1382-0.

46. F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, F. J. Theis, PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* (2019), doi:10.1186/s13059-019-1663-x.

47. Y. Zhou, B. Zhou, L. Pache, M. Chang, A. H. Khodabakhshi, O. Tanaseichuk, C. Benner, S. K. Chanda, Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* (2019), doi:10.1038/s41467-019-09234-6.

48. M. Bastian, S. Heymann, M. Jacomy, Gephi: An Open Source Software for Exploring and Manipulating Networks. *Third Int. AAAI Conf. Weblogs Soc. Media* (2009), doi:10.1136/qshc.2004.010033.

Acknowledgements

The authors acknowledge the scientific and technical assistance of Chiara Cordiglieri and Alessandra Fasciani of the INGM Imaging Facility (Istituto Nazionale Genetica Molecolare – INGM, Milan, Italy). We also acknowledge Simon Noble from the CHDI Foundation for helping in manuscript editing. **Funding:** This work was conducted with collaboration and funding from CHDI Foundation (JSC A11103), a non- profit biomedical research organization exclusively dedicated to developing therapeutics that will substantially improve the life of individuals affected by HD, to EC. This study was also supported by the European Union funded projects NeurostemcellRepair (FP7, GA no. 602278, 2013-17) and NSC-Reconstruct (H2020, GA no 874758, 2020-23) to EC and RB. The fetal tissue collection in Cambridge is supported by NIHR funding of the Biomedical Research Centre and by the Wellcome Trust 203151/Z/16/Z. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. **Author contributions:** V.D.B. and E.C. conceived the study; V.D.B., P.C., E.V., D.B., E.C. designed the experiments; V.D.B. performed computational analysis with contributions from C.C., T.L., V.R., R.J.P.B., F.G and M.P.; C.C., V.D.B and G.P. performed the conservation analysis study; P.C. performed the bulk and single-cell experiments with contributions from D.B., I.E.C., A.F., M.D.S., G.R.; E.V. performed FISH analysis; M.G. performed immunohistochemistry analysis; V.R., R.J.P.B., F.G., M.D.S., G.R., M.P. contributed the INGM single cell and bioinformatics platforms; X.H., R.V., R.B. collected the fetal samples and performed the dissection; K.K. and S.A.M. performed *in silico* perturbation analysis with V.D.B. and R.J.P.B.; J.C., D.F., M.P., E.C. contributed to data

elaboration and presentation. All authors contributed to interpretation of the results. Figures were assembled by V.D.B. together with T.L. The manuscript was written by V.D.B., revised by D.B. and E.C. and edited and proofread by all authors. E.C. proposed the research program, secured the funding, established the collaborations and coordinated the study. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All bulk and scRNA-seq data have been deposited in the ArrayExpress database at EMBL-EBI (https://www.ebi.ac.uk/arrayexpress/) under accession no. E-MTAB-8893 (bulk RNA-seq) and E-MTAB-8894 (scRNA-seq). All other data are present in the main paper or the supplement.

Supplementary Materials

Materials and Methods

Figs. S1 to S13

Tables S1 to S12

References (31-48)

Main Figures

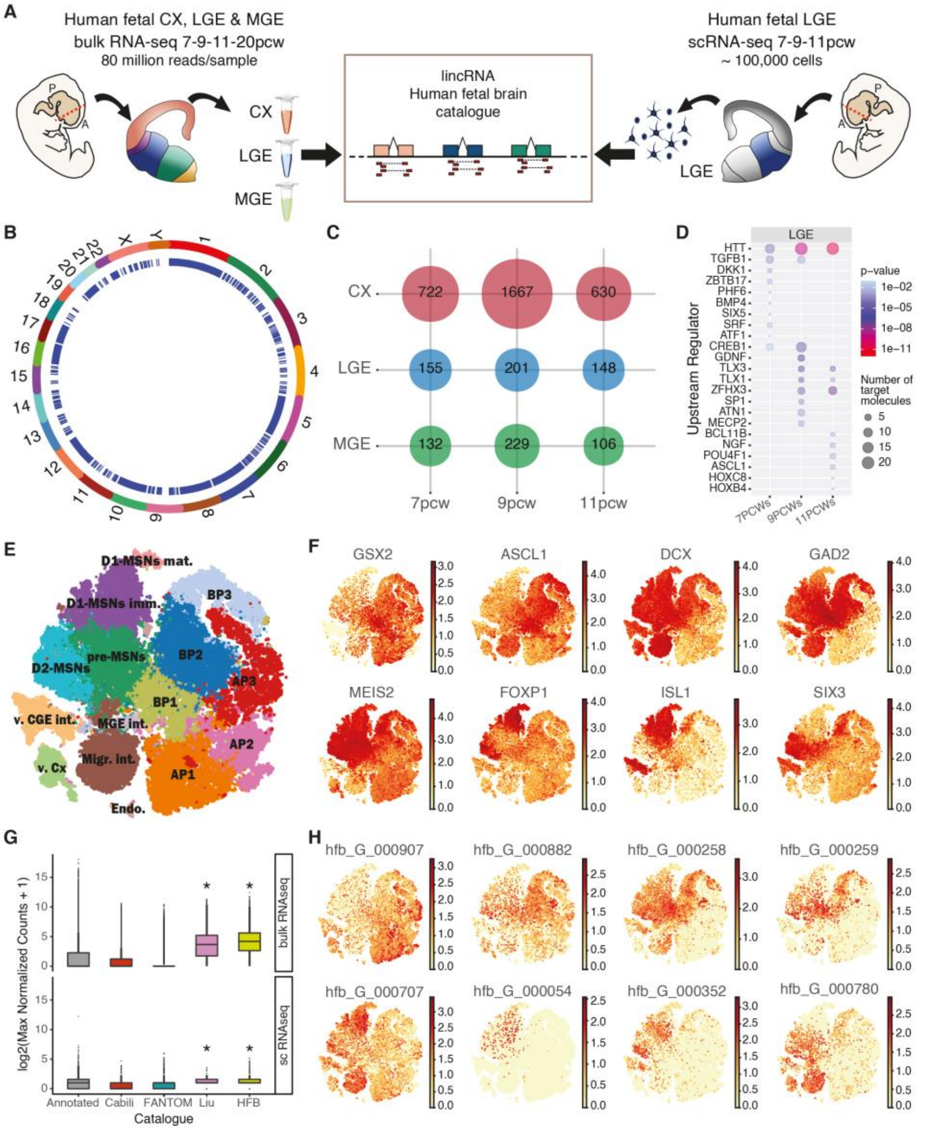

Fig. 1 The coding and non-coding transcriptional landscape of the developing human striatum. (**A**) Schematic representation of the experimental design. CX, neocortex; LGE, lateral ganglionic eminence; MGE, medial ganglionic eminence; pcw, post conceptional weeks. (**B**) Circular plot showing the genomic location of 1116 novel lincRNA. Outer circle represents chromosomes and inner circle the loci of newly identified lincRNAs. (**C**) Bubble matrix showing the number of uniquely expressed protein-coding (PC) genes and lincRNAs per area and per pcw from bulk RNA-seq data. (**D**) Upstream regulators of the bulk LGE specific signature between 7 and 11pcw. (**E**) *t*-SNE plot of 96,789 single cells from the LGE between 7 and 11pcw, colour-coded by cell type. AP, apical progenitors; BP, basal progenitors; pre-MSNs, precursor medium spiny neurons; D1-MSNs imm., immature D1 medium spiny neurons; D1-MSNs mat., mature D1 medium spiny neurons; D2-MSNs, D2 medium spiny neurons; MGE int., MGE interneurons; Migr. int., migrating interneurons; v.CGE int., ventral caudal ganglionic eminence interneurons; v.Cx, ventral neocortical neurons; Endo., endothelial cells. (**F**) Gene expression levels of early progenitors (*GSX2*), intermediate progenitors (*ASCL1*), neurons (*DCX*), GABAergic neurons (*GAD2*), LGE lineage cells (*MEIS2*), general MSNs (*FOXP1*), D1-MSNs (*ISL1*) and D2-MSNs (*SIX3*). (**G**) Distribution of maximal normalized expression of lincRNAs from different catalogues in bulk and single-cell data. Wilcoxon test with Bonferroni correction,*p<2e-16 (pairwise comparisons: Liu *vs* Fantom, Cabili, annotated lincRNAs and HFB *vs* FANTOM, Cabili, annotated lincRNAs). (**H**) Gene expression levels of highly specific lincRNAs identified *de novo* in this study.

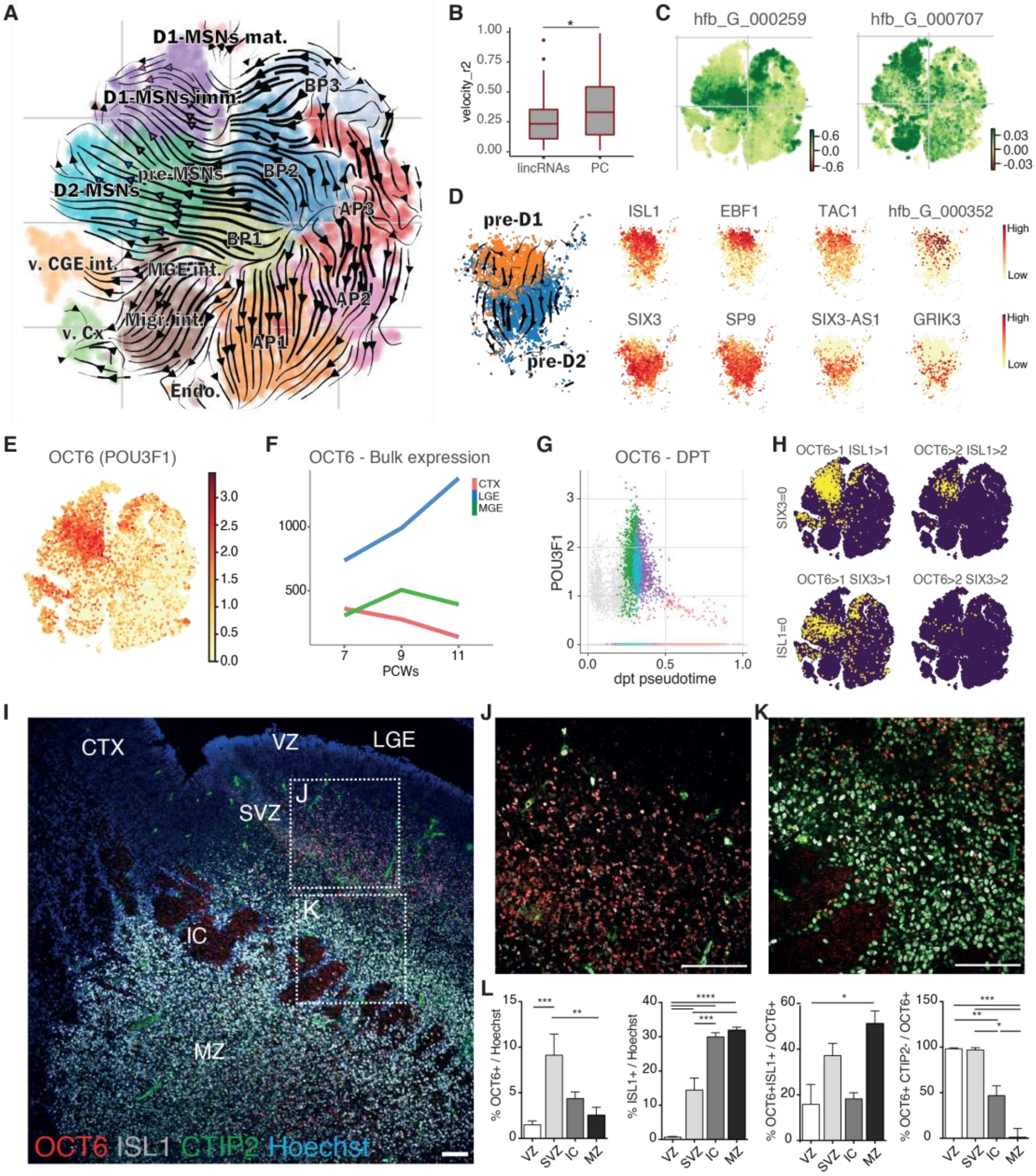

Fig. 2 Maturation and differentiation trajectory of MSNs. (**A**) Velocity estimates projected onto a two-dimensional tSNE plot of the LGE dataset. Arrow heads in the pre-MSN cluster are coloured according to commitment to either the D1 (purple) or D2 (cyan) MSN fate. (**B**) Boxplot showing splicing kinetics of lincRNAs and PC genes. Wilcoxon test with Bonferroni correction,*p<0.05. (**C**) t-SNE plot showing velocities of each gene. Positive velocities (shown in dark green) indicate that a gene is up-regulated, which occurs for cells that show higher abundance of unspliced mRNA for that gene than expected. (**D)** Sub-clusters within the pre-MSN cluster and the velocity vector fields, together with expression levels of canonical D1-MSNs (top row) and D2-MSNs markers (bottom row). (**E-G**) Single-cell (E) bulk (F) expression levels, together with single-cell expression levels (cells are coloured according to louvain cluster) plotted against pseudotime (G), of the pre-MSN specific marker *OCT6* (*POU3F1*). (**H**) *t*-SNE plot showing cells co-expressing *OCT6*/*ISL1* and *OCT6*/*SIX3* with different thresholds of gene expression (co-expressing cells are shown in yellow). (**I**) OCT6, ISL1 and CTIP2 staining of a telencephalic coronal hemisection at 9pcw (scale bar 100µm at bottom right) with (**J-K**) 40x magnification of the SVZ (I) and IC (J). (**L**) Automatic quantification of the percentage of cells positive for OCT6, ISL1 and CTIP2 with the NIS software on confocal images at 40X magnification. N = 3-6 fields for each zone: VZ, SVZ, internal capsule (IC) and MZ from 2-3 coronal slices of 1 fetus, 2-3 z stacks of each field are pre-mediated. Statistics were performed with Prism: Anova One Way, Bonferroni post test, * p< 0,05; ** p< 0,01; ***p<0,001;**** p<0,0001.

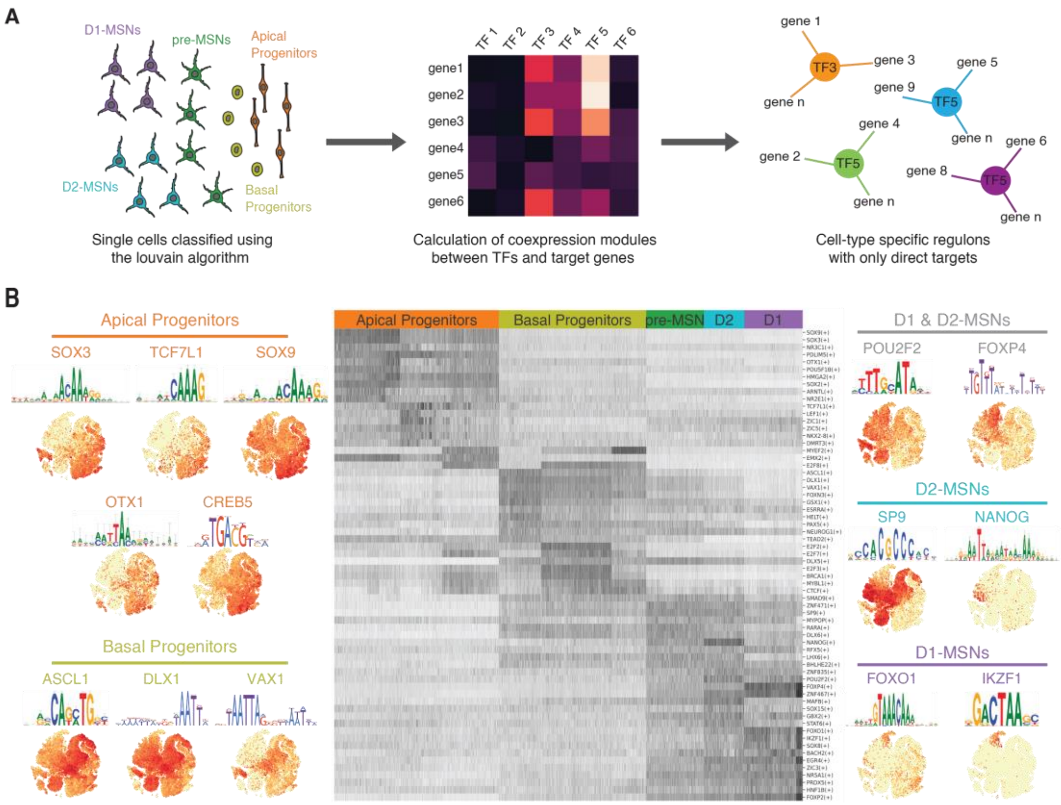

Fig. 3 The cell-specific gene regulatory networks of the LGE lineage (A) Schematic overview of the computational approach used to infer cell-type-specific gene regulatory networks using SCENIC**.** (**B**) SCENIC regulon activity matrix showing the top active transcription factors in each cell class. For each cell type, the figure depicts specific transcription factors and their associated motif and expression patterns in a *t*-SNE plot. 5

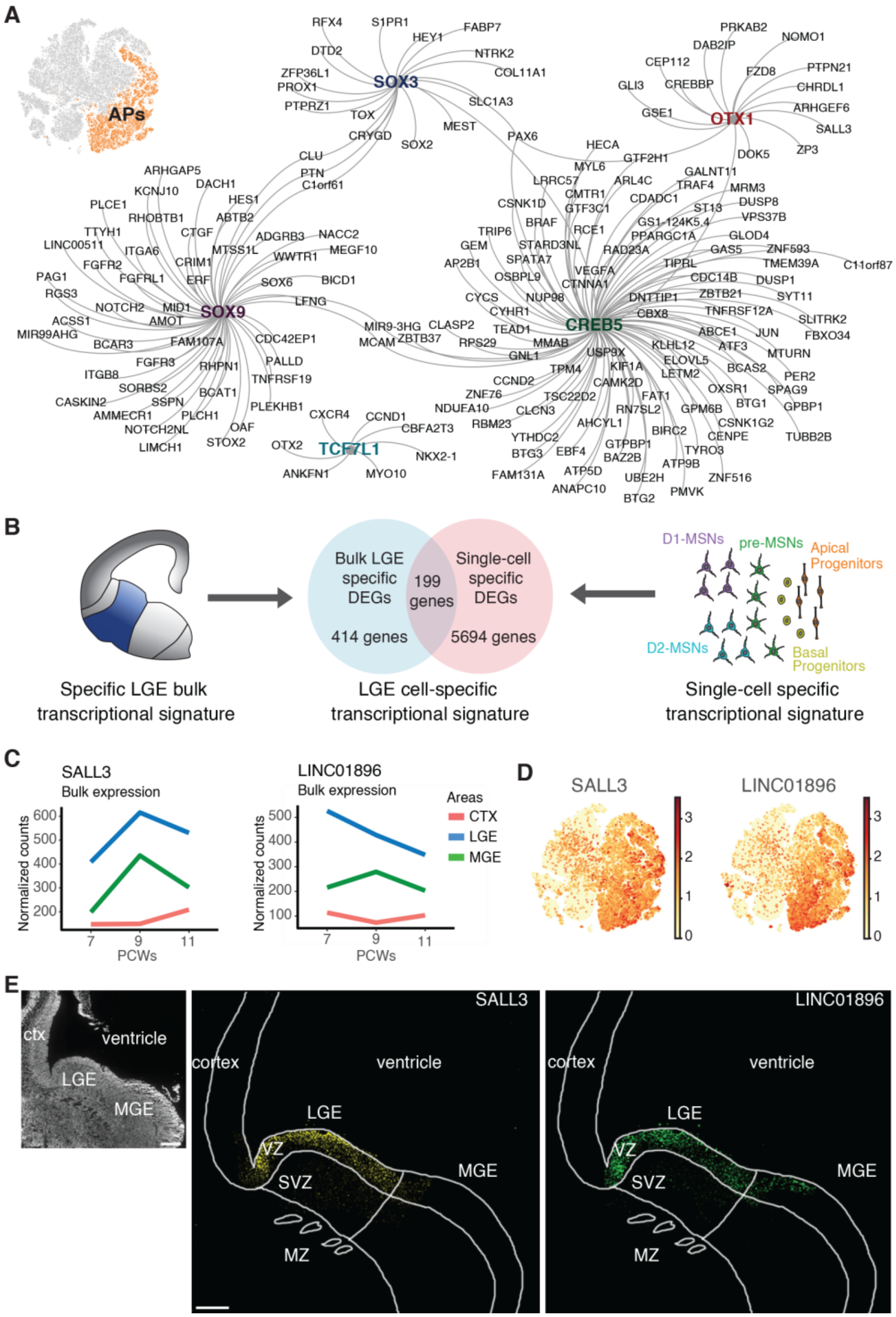

Fig. 4 The emergence of specific apical progenitors within the LGE. (**A**) Specific active regulons in APs. **(B)** Schematic overview of the computational approach used to infer LGE and cell-type-specific transcriptional signatures from bulk and scRNA-seq data. (**C**) Bulk expression levels of the LGE specific genes *SALL3* and *LINC01896*. (**D**) Single-cell expression levels of APs specific genes *SALL3* and *LINC01896* (*RP11-849I19.1*) in the LGE lineage. (**E**) FISH validation of *SALL3* and *LINC01896* on a telencephalic coronal hemisection at 9pcw (Scale bars: 500 µm, 200 µm). 5

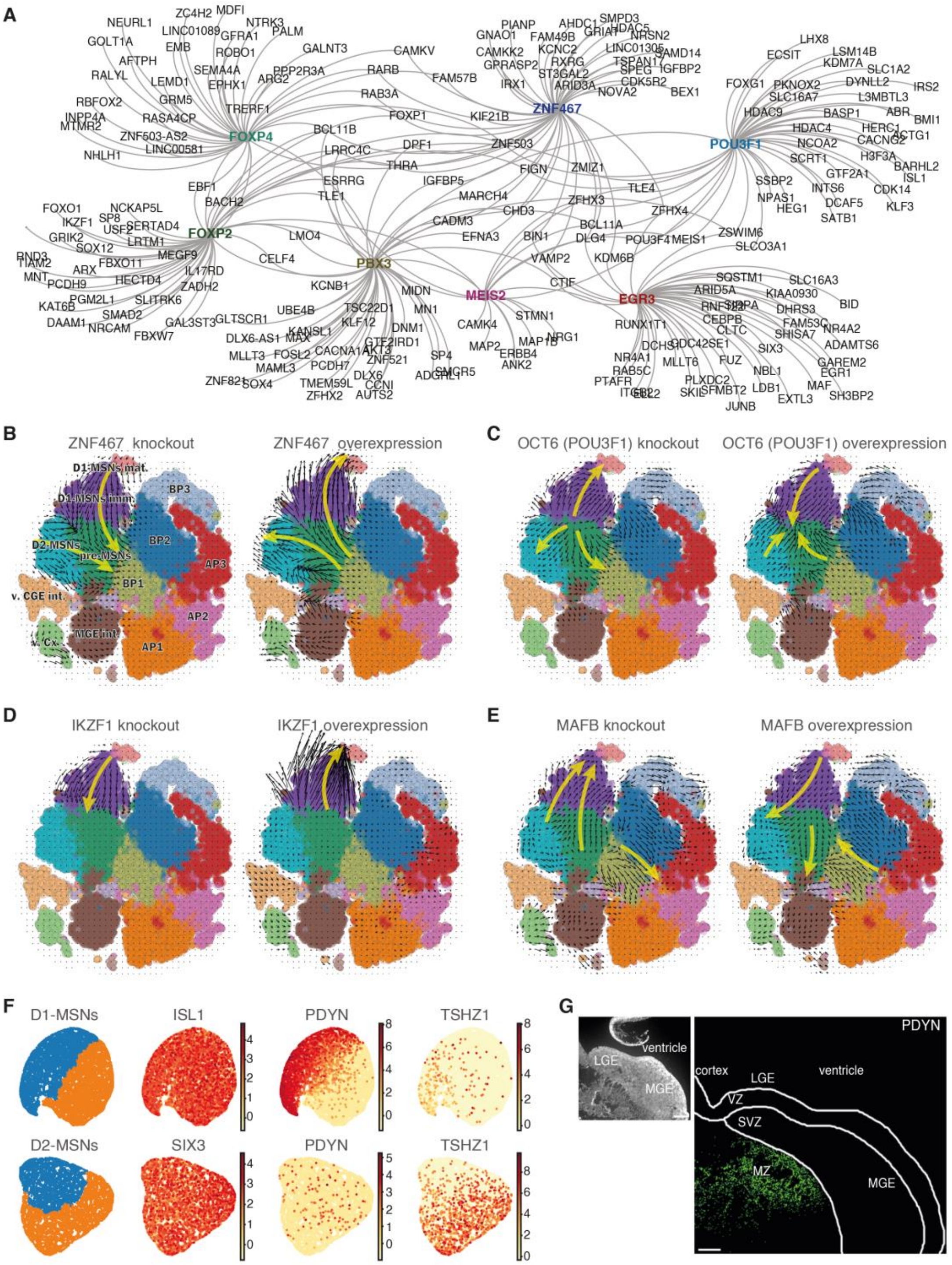

Fig. 5 *In silico* perturbation of MSN-specific gene regulatory networks. (**A**) Shared active regulons that define D1- and D2-MSNs. (**B-E**) CellOracle simulation of knockout and overexpression of the two key MSN markers *ZNF467* (B) and *OCT6* (C), together with *IKZF1* (D) a D1-MSN specific TF and *MAFB* (E) a D2-MSN TF. The effect of the perturbation in shown on the t-SNE plot with projections of cell state transition vectors for each cell. Yellow arrows were manually added to represent overall directionality. (**F**) Sub-clusters of D1- and D2-MSNs, together with expression levels of two specific patch markers (*PDYN* and *TSHZ1*) and canonical D1 (*ISL1*) and D2 (*SIX3*) MSN markers. (**G**) FISH validation of *PDYN* on a telencephalic coronal hemisection shows exclusive expression in the MZ at 9pcw (Scale bars: 500 µm, 200 µm).

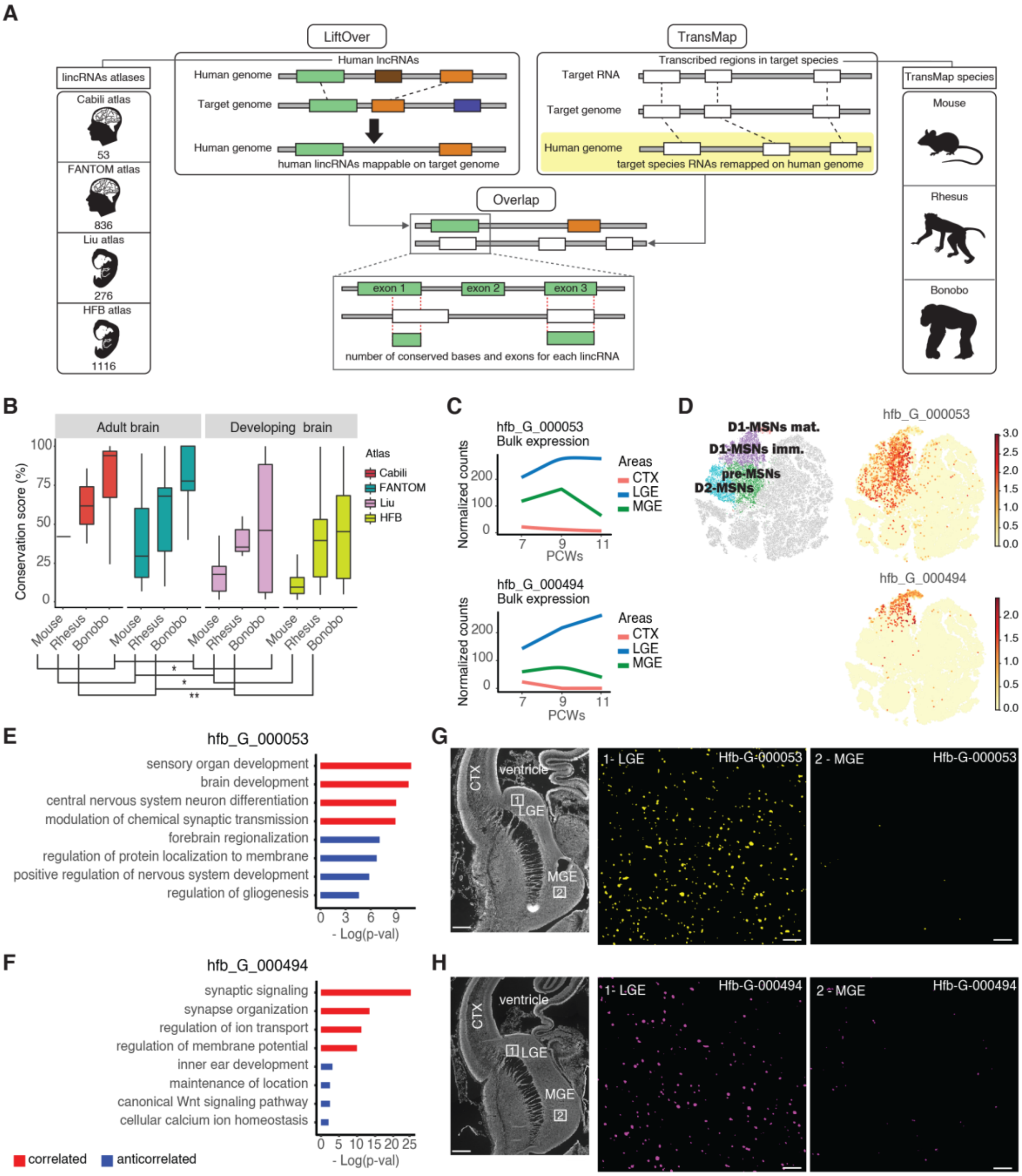

Fig. 6 The human-specific long non-coding signature of MSNs. (**A**) Computational pipeline developed to establish the conservation score of each lincRNA. (**B**) Distribution of conservation scores for different lincRNA classes during development and in adult tissues show significant differences between lincRNAs identified in the developing brain (Liu – HFB) compared to those identified in the adult brain (Cabili - FANTOM), Wilcoxon rank-sum test,*p<0.05, **p<0.002. No differences were observed between the two adult catalogues (Cabili *vs* FANTOM) or between the two catalogues of the developing brain (Liu *vs* HFB). **C.** Bulk expression levels of the human specific lincRNAs hfb_G_000053 and hfb_G_000494. (**D**) Single cell expression levels of MSNs specific lincRNAs *hfb_G_000053* and *hfb_G_000494*. (**E-F**) Enriched gene ontology terms related to genes that have a high correlation or anti-correlation pattern (p-value < 0.05 and r > 0.6 or r < -0.6) with *hfb_G_000053* (E) and *hfb_G_000494* (F). (**G-H**) FISH validation of *hfb_G_000053* (G) and *hfb_G_000494* (H) expression on a telencephalic coronal hemisection at 11pcw (Scale bars: 500 µm, 100 µm).

Science
AAAS

Supplementary Materials for

The coding and long non-coding single-cell atlas of the developing human fetal striatum

Vittoria Dickinson Bocchi, Paola Conforti, Elena Vezzoli, Dario Besusso, Claudio Cappadona, Tiziana Lischetti, Maura Galimberti, Valeria Ranzani, Raoul J.P. Bonnal, Marco De Simone, Grazisa Rossetti, Xiaoling He, Kenji Kamimoto, Ira Espuny-Camacho, Andrea Faedo, Federica Gervasoni, Romina Vuono, Samantha A. Morris, Jian Chen, Dan Felsenfeld, Giulio Pavesi, Roger A. Barker, Massimiliano Pagani* and Elena Cattaneo*

*Correspondence to: elena.cattaneo@unimi.it; massimiliano.pagani@unimi.it

This PDF file includes:

Materials and Methods
Figs. S1 to S13
Tables S1 to S12

Materials and Methods

Ethics statement

Post-mortem human fetal brain specimens under 12pcw were obtained from University of Cambridge (UK) after informed consent was obtained from all donors. All procedures were approved by the research ethical committees and research services division of the University of Cambridge and Addenbrooke's Hospital in Cambridge (protocol 96/85, approved by Health Research Authority, Committee East of England—Cambridge Central in 1996 and with subsequent amendments, the last being in November 2017) in accordance with the Human Tissue Act 2006. Post-mortem human brain specimens at 20pcw were obtained from the San Paolo Hospital, Milano after autopsy diagnostic procedures. Both documents were submitted to the Ethics Committee of the University of Milano, and ethics approval was obtained on 27 March 2013. Tissue was handled in accordance with ethical guidelines and regulations for the research use of human brain tissue set forth by the National Institute of Health (NIH) (http://bioethics.od.nih.gov/humantissue.html) and the World Medical Association Declaration of Helsinki: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/.

Human Tissue Collection

Embryonic and fetal age was extrapolated based on the date of the mother's last menstruation, ultrasound scans of the fetus in utero, crown-rump length (CRL) and visual inspection. Depending on the condition and period of the procured specimens, the neocortex, LGE and MGE were dissected. Both left and right brain regions of interest were collected from fresh tissues. Specimens were chilled on ice during dissection. The dissected samples were immediately frozen in dry ice and stored at −80 °C for later RNA extraction. For scRNA-seq, dissected LGEs were maintained in Hibernate-E media (Thermo Fisher, A1247601). Table S1 provides a complete list of all tissue samples included in this study.

RNA sample preparation and sequencing for bulk RNA-seq

Total RNA was extracted from LGE, MGE and neocortical human fetal tissues with the TRIZOL reagent (ThermoFisher Scientific, 15596018) according to the manufacturer's instructions. After DNase treatment (ThermoFisher Scientific, AM1906), optical density values of extracted RNA were measured with NanoDrop (Thermo Scientific) to confirm an A260/A280 ratio ≥ 1.9. 500 ng of total RNA from each sample was used for the preparation of an unstranded polyA library. Sequencing was performed with the Illumina HiSeq 4000 platform, with an average depth of 80 million reads (125 or 150 bp paired-end) per sample.

Cell suspension preparation for single-cell RNA-seq

Tissues were processed by dissociation using Papain Dissociation System (Worthington, LK003150) following the manufacturer’s recommendations, adjusting the incubation time based on tissue piece size. Briefly, after papain incubation, cells were resuspended in N2 media with DNase I provided in Papain Dissociation System (Worthington, LK003170) and Rock Inhibitor Y-27632 (provided by CHDI Foundation, CHDI-00197406-0001-007). Next, glass pipettes of increasingly smaller tip diameter (fire-polished) were used to dissociate the tissue to a single-cell suspension followed by a filtration with CellTrics 20um (Sysmex, 04-004-2325). Cells were pelleted and resuspended in Ultrapure BSA 0.04% (Ambion, AM2616). Cell viability (>90%) was assessed by trypan blue staining.

Library preparation and sequencing for scRNA-seq samples

Single cell sequencing libraries were prepared using the Chromium Single Cell 3′ Library & Gel Bead Kit v2 (PN- 120237), Chromium Single Cell 3′ Chip kit v2 (PN-120236) and Chromium i7 Multiplex Kit (PN-120262) according to the manufacturer's instructions. Libraries were generated starting from 8233 to 28577 cells for each sample (see Table S1) and were sequenced with the Illumina HiSeq 4000 platform, with an average depth of 50,000 reads per cell using paired end sequencing with single indexing.

TaqMan gene expression assays

For assessment of gene expression levels, TaqMan gene expression assays (Thermo Fisher Scientific) were used. 500 ng of total RNA was used for reverse transcription with a iScript cDNA Synthesis Kit (Biorad, 1708891). 1.25 ng of diluted cDNA was used for RT-qPCR to assess the expression of Hfb-G-000053 (ThermoFisher Scientific, Hs_G_000053) and Hfb-G-000494 (ThermoFisher Scientific Hs_G_000494). Gene expression levels were normalized to 18s rRNA (Hs 99999901 s1). qPCR was carried out in Bio-Rad CFX C1000 System (Bio-Rad, Foster City, CA, USA) using the TaqMan Fast Advanced Master Mix (Applied Biosystems, 4444557) according to the manufacturer's instruction. A one tailed Student t-test was used to compare groups.

FISH

Brains were dissected out, fixed (24 h, 4°C) in 4% paraformaldehyde (PFA) and incubated in sucrose 30% for 24 h at 4°C. Brains were then frozen in optimal cutting temperature (OCT) medium on dry ice. Frozen sections (15 μm) were obtained using a cryostat (Leica, Germany) and stored at − 80 °C. Sections were thawed just prior to staining and fixed with 4% PFA for 10 min followed by rinsing in PBS. Sections were then dehydrated in 5 min dehydration steps in 70% and 100% ethanol, respectively. A hydrophobic barrier was created around the sections with a hydrophobic barrier pen, and air dried for 10 min. RNAscope® Hydrogen Peroxide solution (ACD, 322335) was added to cover the sample and incubated for 10 minutes at RT. After incubation, hydrogen peroxide solution was removed and slides were washed twice with ddH_2O. Protease plus (ACD, 322331) was added to the samples and incubated in a HybEZ oven for 30 minutes at 40°C. After incubation, protease was removed by washing the slides in ddH_2O. RNAscope *in situ* hybridizations were performed according to the manufacturer's instructions, using the RNAscope® Multiplex Fluorescent Assay v2 (Advanced Cell Diagnostics) for fixed frozen tissue. The following probes with suitable combinations were used (indicated with gene target name for human, respective channel, and catalogue number all Advanced Cell Diagnostics): Dlx5 (Ch1, 569471), Dlx6 (Ch2, 569481), Dlx2 (Ch4, 483311), Isl1 (Ch3, 478591), Nkx2.1 (Ch2, 530701), Lhx6 (Ch1, 460051), Lhx8 (Ch4, 483321), Kcna5 (Ch1, 590001), LINC01305 (Ch2, 569501), Sall3 (Ch2, 590011), LINC01896 (Ch1, 589981), PDYN (Ch1, 507161). BaseScope® detection reagents v2 - RED (Advanced Cell Diagnostics) for fixed frozen tissue was used with the following probes (indicated with gene target name for human and catalogue number all Advanced Cell Diagnostics): Hfb-G-00053-Junc (817261) and Hfb-G-000494-Junc (817271). Probe hybridization took place for 2 h at 40 °C and slides were then rinsed in 1 × wash buffer, followed by four amplification steps (according to the standard protocol). Brain sections were then labelled with DAPI, and mounted with Prolong Gold mounting medium (P36930, Thermo Fisher Scientific). Slides were stored at 4 ºC before image acquisition using a Nikon Eclipse Ti-E microscope (Nikon Instruments) with a LED illumination system equipped with Andor Zyla camera (Andor Technology, Oxford Instruments, Oxford, UK).

Immunohistochemistry

Human fetal brain sections were treated as above and after dehydration in ethanol they were washed with PBS, permeabilized with 0.5% Triton X-100 (Euroclone) in PBS for 10 min, washed in PBS and retrieved with Sodium Citrate 10mM at 90°C for 30min. After antigen retrieval the sections were washed with PBS and blocked with 10% NGS (Vector) and 0.2% Triton X-100 in PBS at room temperature for 1h. Primary antibodies were diluted in solution containing 3% NGS and 0.1% Triton X-100 in PBS at 4 °C overnight. The following day sections were washed three times in PBS at room temperature. Secondary antibodies conjugated to Alexa fluorophores 488, 568 or 647 (Molecular Probe, Life Technologies) were used 1:500 in solution containing 3% NGS and 0.1% Triton X-100 in PBS at room temperature for 1 h mixed with Hoechst 33258 (5 µg/ml; Thermo Fisher Scientific) to visualize nuclei. Then the sections were washed once in PBS with 0,1% Triton X-100 and twice in PBS and finally mounted with Dako Glycergel (Aqueous Mounting Medium, Agilent) at room temperature overnight. The following day the sections were dry enough to be visualize at the microscope and then stored at 4°C.

Images were acquired with a confocal microscope (Leica SP5) and analyzed with NIS software for imaging (NIS-Elements AR v5.11) for the quantification of single or double positive cells over Hoechst positive nuclei. The confocal images taken at 40x were deconvolved with NIS to increase the resolution and then quantified by the general analyses tool following a pipeline set on parameters such as background, dimensions and circularity.

Primary antibodies and concentrations were as follows:

ASCL1 (mouse, 1:500; Becton Dickinson, cat. n. 556604), CTIP2 (rat, 1:500; Abcam, cat. n. ab18465), EBF1 (mouse, 1:1000; Santa Cruz, cat. n. sc-137065), ISLT1/2 (mouse, 1:1000; Hybridoma Bank, cat. n. 39.405), Ki67 (mouse, 1:400; Cell Signalling, cat. n. 9449), OCT6 (rabbit, 1:100; Abcam, cat. n. ab272925), OCT6 (mouse, 1:100; Merck, cat. n. MABN738), SIX3 (rabbit, 1:300; Abcam, cat. n. ab221750).

Custom reference annotation

Intergenic lncRNAs (lincRNAs) identified in the neocortex (GSE71315) together with lincRNAs identified in different tissue (Cabili:GSE30554; FANTOM: https://fantom.gsc.riken.jp/cat/) were integrated in the reference genome annotation (GENCODE release 25; version GRCh38/hg38) using Cuffcompare v.2.1.1 (*31*). Transcripts were considered as matching allowing for differences on the length of the first and last exons.

Quality control and mapping of bulk RNA-seq data

All reads were tested for QC using MultiQC (*32*). To remove any low quality reads Trimmomatic(*33*) was used. All adapters were removed together with low quality bases (below quality 3) in the leading and trailing end of the reads. Any 4 consecutive bases that had an average quality per base of 15 were removed together with reads shorter than 50 bases.

All reads were aligned to the human genome using STAR v2.5.2b (*34*), using our custom reference gene annotation (GENCODE release 25, version GRCh38/hg38). To preserve +/- strand information in unstranded RNA-seq data the outSAMstrandField option with intronMotif specified was used to retain this attribute for all alignments that contained splice junctions. Only highest confidence alignments were then input to assembly tools, with non-canonical junctions (*35*) filtered out using the option outFilterIntronMotifs with RemoveNoncanonical specified. The number of reads per gene was counted using the quantMode GeneCounts option.

De novo genome-based transcripts reconstruction

The transcriptome of each sample was assembled from the mapped reads separately by both Cufflinks (*36*) and StringTie (*37*). In particular, Cufflinks v2.2.1 was run using a likelihood based approach for fragment bias correction (*38*). In addition, to more accurately weight reads mapping to multiple locations in the genome, multi-mapped read correction (*36*) was applied. A ROC analysis was performed to determine the optimal read coverage thresholds to limit technical noise and leaky expression based on whether Cufflinks classified previously known protein coding and lincRNAs as having full read support. An average coverage threshold of 0.6 (FDR = 0.05) was used, and all assembled transcripts below this threshold were filtered out. StringTie v1.2.3 was run with default parameters, with a minimum coverage per bp of 0.7. This threshold was based on the median bp coverage for lincRNAs identified in the human developing cortex (*9*) as these lincRNAs should have an abundance similar to the ones identified in this study.

De-novo lncRNAs prediction

To classify novel transcripts into either protein-coding or lincRNAs, the alignment-free tool called FEElnc (*39*) v0.1.1 was used. The first module (FEELnc_filter) was used to remove transcripts that were shorter than 200nt, and then we filtered out unwanted/spurious transcripts that were either mono-exonic, or bi-exonic with one exon shorter than 25nt. All transcripts that overlapped exons of the known genes in the reference annotation were subsequently removed. The FEELnc_codpot module was then used to compute the coding potential score (CPS) for each of the remaining candidate transcripts. A Random Forest model was then trained to calculate an optimal CPS cut-off, with a 10 fold cross-validation on an input training set of known mRNAs and lncRNAs. The resulting CPS cut-off, maximized both sensitivity and specificity and was employed for the classification of the newly assembled transcripts.

lncRNA classification

lncRNAs were classified based on lncRNAs that either overlapped with a gene from the reference annotation (genic lncRNAs) or not (intergenic lncRNAs - lincRNAs). For this study only lincRNAs were selected and integrated in the reference annotation as the bulk RNA-seq data is unstranded and does not allow to reliably detect genic lncRNAs.

Integrating newly identified lincRNAs in the modified reference annotation

Only lincRNAs that were identified by both assemblies (Cufflinks and StringTie) and that passed the FEELnc filters were integrated in the reference annotation (GENCODE release 25; version GRCh38/hg38) using Cuffcompare v.2.1.1 (*31*). This new augmented reference annotation was then used to remap reads of bulk and scRNA-seq data.

lincRNA conservation analysis

For lincRNAs derived from the developing brain the Liu study (*9*) was combined with our dataset. For the adult brain, lincRNAs identified in the adult brain were filtered from the Cabili (*7*) and FANTOM (*8*) datasets. Only lincRNAs were considered from all atlases to avoid biases between genic vs intergenic lncRNAs in the conservation score. lincRNA conservation was evaluated using three steps. First, TransMap (*24*) alignments of UCSC, RefSeq, mRNA, and EST transcripts to the human genome (GRCh38/hg38) were downloaded from the UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgTables). Mapping coordinates on the human genome for mouse, rhesus and bonobo RNAs were then extracted to evaluate whether they overlapped with the mapping coordinates of the different human lincRNAs. Second, liftOver (*23*) was used to re-map lincRNAs to another genome with at least 60% of matching bases. Finally, human lincRNAs that had overlapping coordinates between TransMap and liftOver were identified using bedtools (*40*). For each lincRNA that showed an overlap on at least two exons, the number of matching bases were calculated and used to develop a conservation score scheme. The maximum score (100%) was assigned to a lincRNA if it was mappable on to the other genomes according to liftOver and all bases of its exons overlapped with the TransMap transcript coordinates. lincRNAs identified during development were compared to the ones in the adult brain using a Wilcoxon rank-sum test for each species.

Bulk RNA-seq differential expression analysis

To perform batch effect correction for 7pcw samples the ComBat (*41*) function was employed by using a parametric empirical Bayesian adjustment (par.prior = true) based on the 2 covariates that formed the batch. To determine a specific lincRNA/protein-coding signature for each area and pcw, DESeq2(*42*) was employed. In particular, for each pcw every combination of comparisons (CX vs LGE, CX vs MGE, LGE vs MGE, LGE vs CX, MGE vs CX and MGE vs LGE) was tested. For each of these comparisons only significantly overexpressed genes with an adjusted p-value < 0.05 were kept. To pinpoint a set of genes that were significant only for a certain region the lists of overexpressed genes from each set of comparisons (e.g. LGE vs MGE and LGE vs CX) were intersected, and only genes that were significant in both comparisons for the LGE were considered "LGE specific".

Pathway Analysis for bulk RNA-seq LGE signature

To identify Gene Ontology (GO) terms related to LGE specific genes the ClueGo plugin of Cytoscape (*43*) was used. The GO "Biological component" was employed to query these genes against the background of all genes expressed in the fetal brain tissue. Only GO with an adjusted p-value lower than 0.05 (Bonferroni step down correction) were considered (κ-score threshold = 0.3). The semantic similarity score between GO terms was calculated using GOSemSim (*44*). The resulting matrix was then subjected to hierarchical clustering to find the most represented GO terms. Upstream regulators were identified by Ingenuity Pathway Analysis (IPA; Ingenuity Systems,

https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/).

Tissue-specificity analysis

To quantify tissue specificity of lincRNAs and protein-coding genes, Jensen-Shannon divergence (JSD) was employed (*7*). For bulk data the value for each gene was calculated as its mean expression on replicates per area and pcw, while for single cell data the value for each gene was calculated as its mean expression in each subtype of cell.

Sample demultiplexing, barcode processing, gene counting and quality control of scRNA-seq data.

Droplet-based (10x) sequencing data was demultiplexed, processed, aligned and quantified using the Cell Ranger Single-Cell Software Suite (v.2.1.1, 10x Genomics) using our custom reference annotation.

All downstream analysis were done using Scanpy v1.4 (*45*), using the standard Seurat-inspired processing pipeline. All genes expressed in less than 3 cells and all cells with less than 200 detected genes and with unique gene counts under 500 or over 6000 were removed. All cells for which the total mitochondrial gene expression exceeded 5% were also removed. Data processing included dividing gene read counts by the total of each cell, multiplying by a scale factor of 10,000 and log-transforming the results. Highly variable genes were defined as having an average normalized expression between 0.0125 and 3 and a dispersion greater than 0.5. Variation in gene expression driven by the number of detected molecules and mitochondrial gene expression was regressed out to avoid these variations dominating the downstream analysis and bias the final interpretation of the data.

Clustering and annotation of scRNA-seq data

To reduce data dimensionality in cells that passed QC, the resulting 1797 highly variable genes were subjected to principal component analysis (PCA). The 10 highest significant PCs were used to construct a K-nearest neighbors (KNN) graph, on which the Louvain algorithm with a resolution=0.5 was used to define highly interconnected cell communities. We then performed t-Distributed Stochastic Neighbor Embedding (t-SNE) on the 10 PCs to obtain a two-dimensional embedding of single cells. Finally, we annotated cell communities to known cellular sub-groups based on canonical marker genes. Specific genes for each cluster were found by ranking genes with a Wilcoxon rank sum test, p-values were then adjusted for multiple testing using the Benjamini-Hochberg method. Only genes with an adjusted p-value < 0.05 were considered.

Trajectory analysis

RNA velocity analysis (*10*) was performed using the scvelo (*11*) python package (v.0.2.1) using a generalized stochastic model. The number of spliced and unspliced reads were counted directly on the cellranger output. The calculated RNA velocities were then embedded in tSNE space. All steps were performed with default parameters of the built-in functions. Potential driver genes were ranked by cell type and only genes with a spearman score greater than 0.1 were selected. PAGA (*46*) in the Scanpy Python package v.1.4 with a resolution of 0.14 was used to infer development trajectories and measure the connectivity between the clusters that were predefined using the Louvain algorithm. The diffusion pseudotime (DPT) was calculated using the tl.dpt function of the Scanpy library, and setting the earliest known cell type (AP.1) as root on the LGE lineage.

Subclustering

To produce the subclustering of the different populations, cells belonging to the clusters of interest were isolated and the Seurat analysis was repeated with clustering resolution set to 0.1 and size of the local neighborhood set to 200 for APs, BPs and migrating interneurons, 180 for pre-MSNs and to 30 for D1- and D2-MSNs.

Pathway Analysis for scRNA-seq signatures

Gene Ontology (GO) terms related to the specific signature of each population of the LGE lineage was defined using Metascape (*47*) (http://metascape.org). GO Biological Processes, Canonical Pathway (MSigDB), KEGG Pathway, BioCarta gene sets (MsigDB) and Reactome gene sets were interrogated to perform the enrichment analysis. Only GO terms with an adjusted p-value ≤ 0.01 (standard accumulative hypergeometric test) were considered and clustered. The most enriched terms are chosen as the representative term of each group. For APs and BPs only shared genes between AP1, AP2 and AP3 and then BP1, BP2 and BP3 respectively were tested for GO analysis. Upstream regulators of the first 100 ranked genes of the D1- and D2-MSNs were identified by Ingenuity Pathway Analysis (IPA; Ingenuity Systems) https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/).

Predict putative functions of lincRNAs by a guilt-by-association approach

To infer relevant biological pathways for the newly identified lincRNAs, bulk RNA-seq data (for the neocortex, LGE and MGE between 7 and 11pcw) was used to create a correlation matrix by calculating a Spearman's Rank correlation coefficient for each lincRNA:mRNA pair. Each pair showing both a p-value < 0.05 and either a correlation value r > 0.6 or r < -0.6 were considered as correlated or inversely correlated, respectively. For the two lincRNAs of interest (hfb_G_000053 and hfb_G_000494) each list of correlated/ inversely correlated genes were subsequently analysed using Metascape (*47*) (http://metascape.org).

Calculation of LGE developmental regulons

The pySCENIC pipeline (*14*) with default settings was employed to infer active TFs and their target genes. In brief, the pipeline consists of three steps. First, gene co-expression modules of TFs were calculated. Secondly, each module was selected based on the presence of a regulatory motif near a transcription start site (TSS). Modules were retained if the TF-binding motif was enriched among its targets, while target genes without direct TF-binding motifs were removed. Third, the impact of each regulon for each single-cell transcriptome was scored using the AUC score as a metric. Finally, the scores were used to calculate a Regulon Specificity Score (RSS) for each cellular population of interest. GRN plots of the different regulons were done using the Gephi software package (*48*).

Calculation of connection between regulons and lincRNAs

Each lincRNA-TFs regulatory link was calculated using GENIE3, a method used to infer gene regulatory networks within the SCENIC package. The output of GENIE3 is a table with lincRNAs and their potential regulators, and a 'importance measure' (IM), which represents the weight that the transcription factor has in the prediction of the target lincRNA. Only the links with IM > 0.001 were taken into account. To avoid spurious links we then filtered and kept links that also resulted as specifically differentially expressed in each cell population (p-value < 0.01; Wilcoxon rank sum test) that was specific for each lincRNA.

In silico perturbation of gene regulatory networks

We used CellOracle (*16*) to perturb candidate transcription factors and their gene regulatory networks. To perform all calculations, we randomly subsampled our dataset to 20k cells to limit memory use. We performed KNN imputation with 500 neighbors and 31 principal components and then followed default parameters. GRNs calculated with SCENIC were integrated and used for perturbations. To knockout a gene, expression values of candidate genes were set to 0. To overexpress a gene, we set the value of the candidate genes to twice their maximal imputed gene expression value. The probability of a cell state transition, that is based on iterative calculations of signal propagations within the GRN, was computed using default parameters.

Figures S1-S13

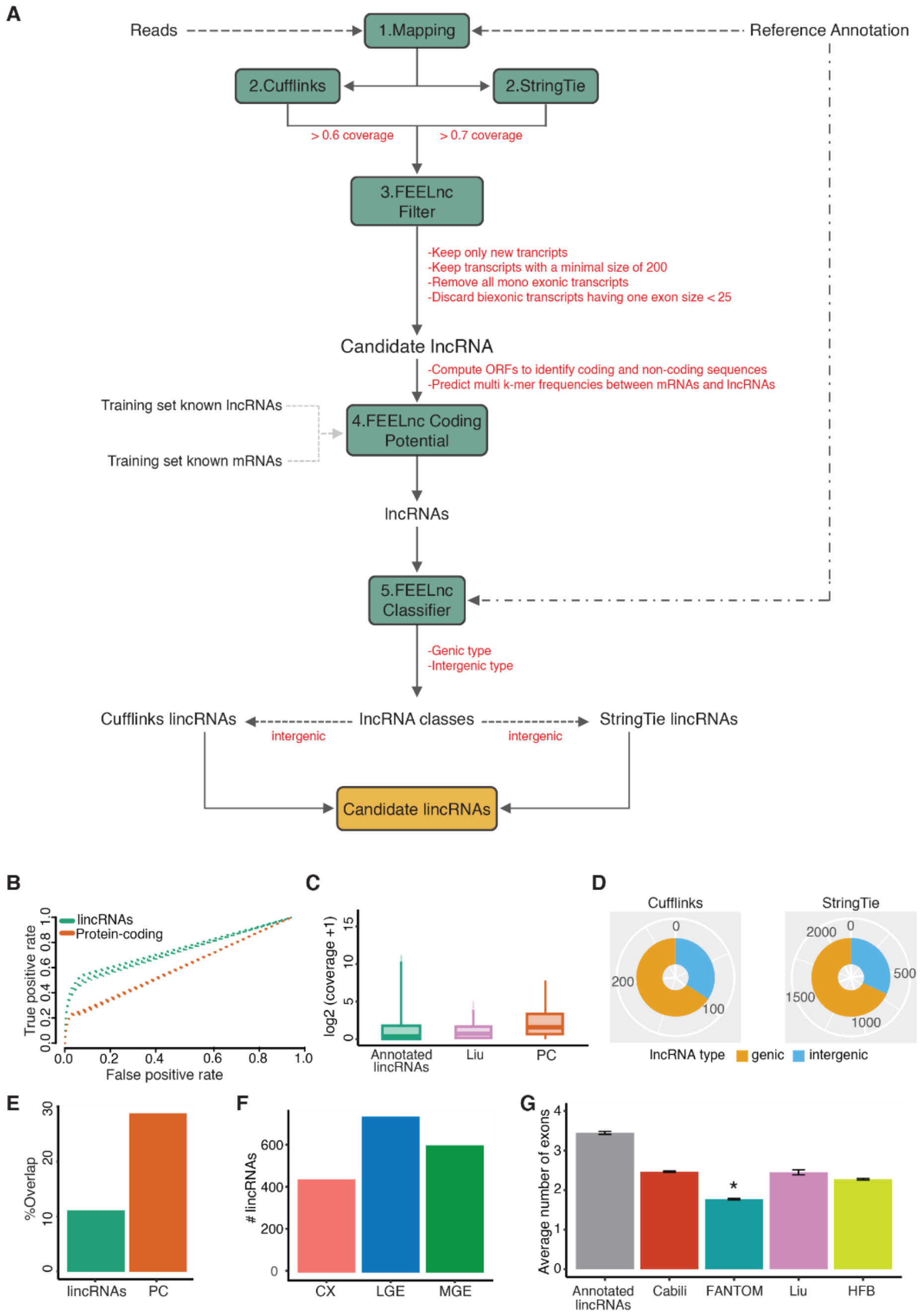

Fig. S1 ***De novo*** **analysis of lincRNAs.** (**A**) Computational pipeline for lincRNAs identification**.** All filtering steps are indicated in red. (**B**) ROC analysis of coverage thresholds required to annotate RNAs as protein-coding or lincRNAs. Coverage threshold for novel transcripts assembled with Cufflinks was selected based on an average false positive rate of 0.05 on known lincRNAs. (**C**) Distribution of RNA-Seq read counts for known protein-coding genes, general lincRNAs and lincRNAs identified in the human developing neocortex (Liu atlas) (*9*). Newly identified transcripts with StringTie were filtered to have a coverage of at least 0.7 that was the median bp coverage for lincRNAs identified in the human developing cortex. (**D**) Average number of intergenic and genic lncRNAs classified by FEELnc per sample. (**E**) Percentage of overlap between lincRNAs and protein-coding genes identified by both Cufflinks and StringTie. (**F**) Final number of lincRNAs identified in each area (**G**) Average number of exons for different lincRNAs catalogues. Wilcoxon test with Bonferroni correction,*p<2e-16.

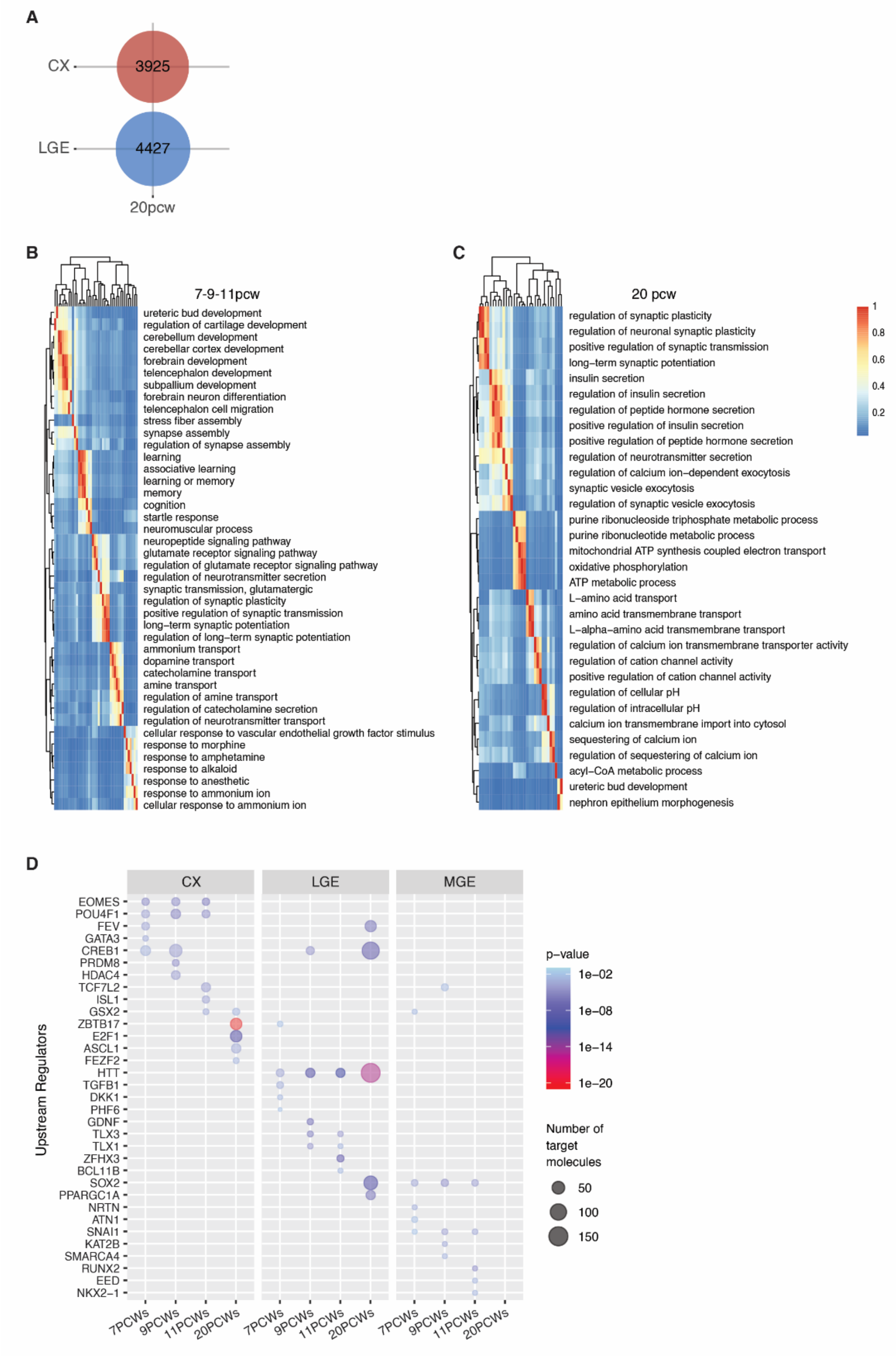

Fig. S2 The bulk RNA signature of the LGE. (**A**) Bubble matrix showing the number of uniquely expressed genes per area at 20pcw from bulk RNA-seq data. (**B-C**) Semantic similarity matrix of the LGE specific signature between 7-11pcw (D) and 20pcw (E). The semantic similarity scores of all GO-term pairs were grouped by hierarchical clustering. (GO terms with adjusted p-value < 0.05 are shown). Rows and columns show the list of enriched GO terms. The colors represent the semantic distances between GO terms. Yellow-red clusters identify groups of terms sharing semantic similarity about biological processes. (**D**) Upstream regulators of the CX, LGE and MGE bulk specific signature at 7, 9, 11 and 20pcw.

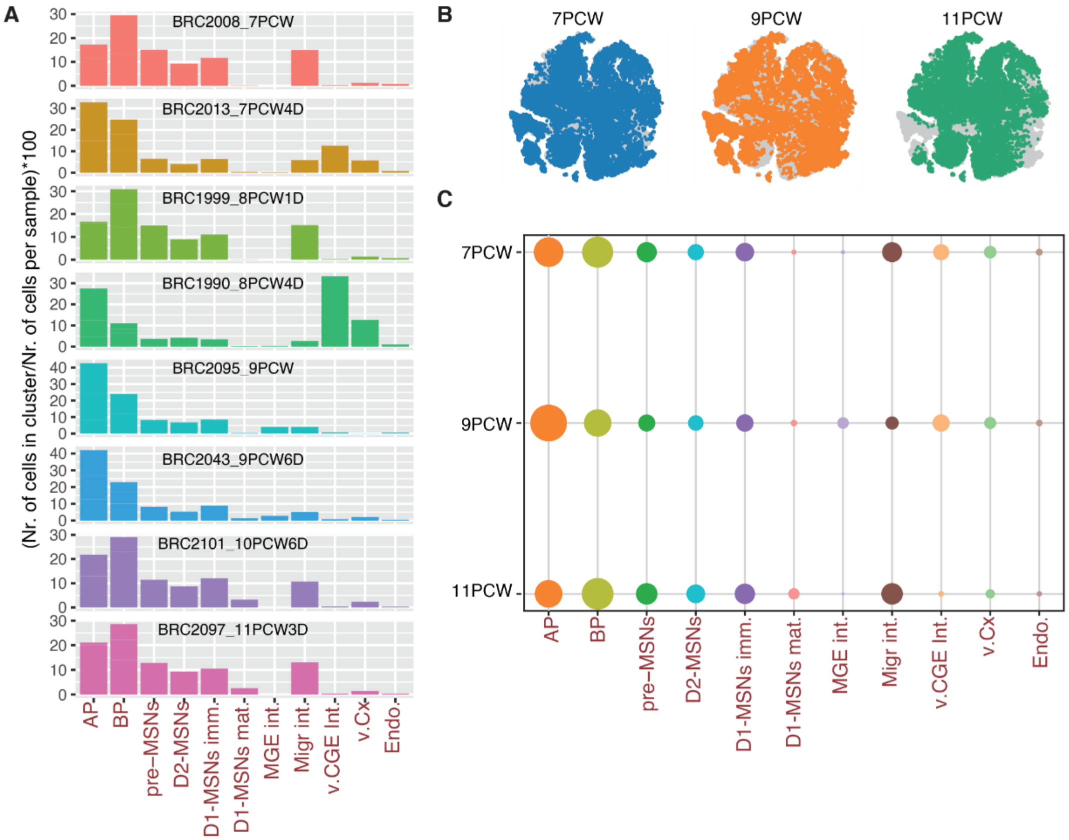

Fig. S3 Distribution of single-cells within each cluster. (**A**) Bar plot showing the number of cells that contribute to each cluster per sample (normalized by total number of cells derived from each sample). (**B**) *t*-SNE plot of all cells derived from 10x Genomics sequencing using 3′ chemistry between 7 and 11pcw colour-coded by post-conceptional week (pcw). (**C**) Bubble plot showing contribution of each pcw to each detected cluster (normalized by total cells per pcw).

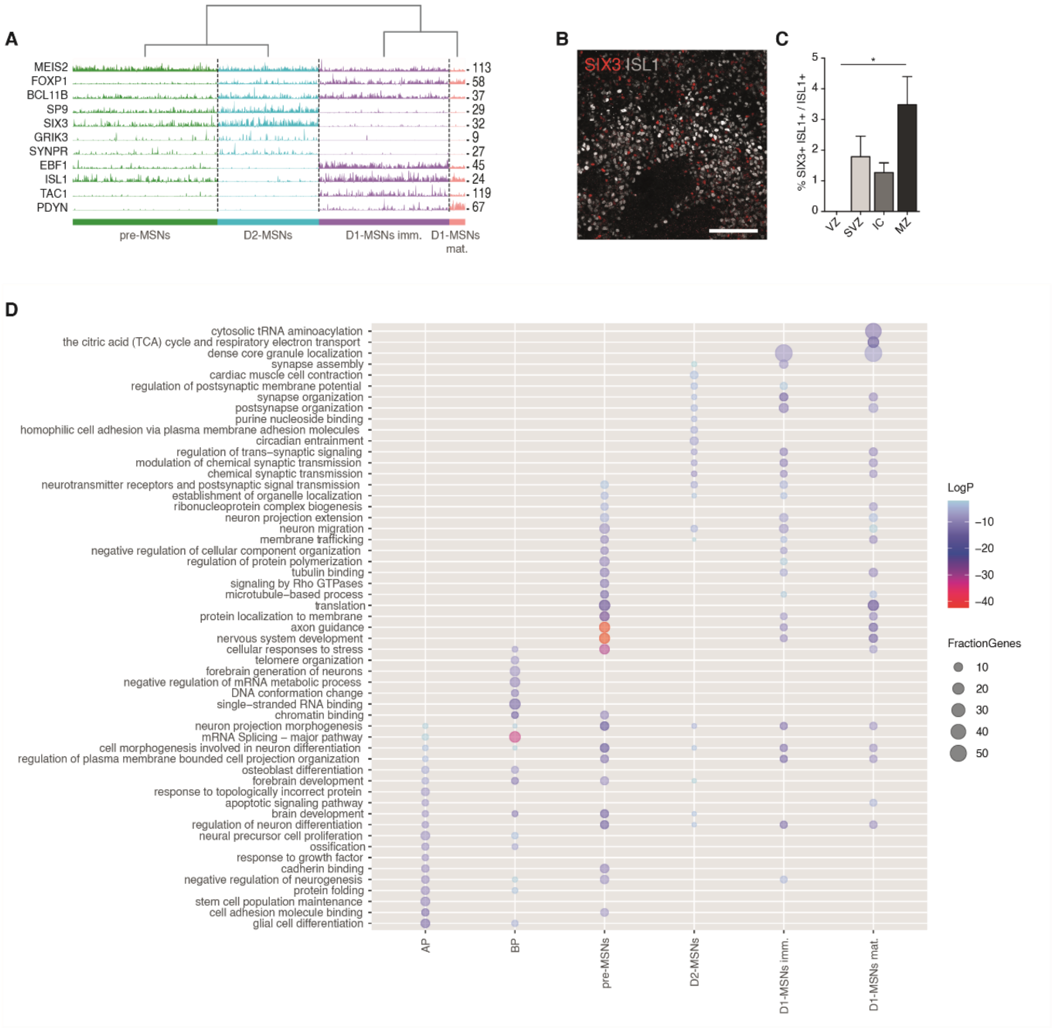

Fig. S4 MSNs signatures and active pathways in each cell state of the LGE. (**A**) Track plot showing gene expression levels of canonical MSNs marker genes (*MEIS2*, *FOXP1*, *BCL11B* and *PPP1R1B*), D2-MSNs (*SP9*, *SIX3*, *GRIK3*, *SYNPR*) and D1-MSNs (*EBF1*, *ISL1*, *TAC1*, *PDYN*). Gene expression is represented by height. (**B**) SIX3 and ISL1 staining of telencephalic coronal hemisection on the MZ of the LGE at 9pcw. Scale bar 100µm at bottom right. Confocal acquisition at 40x magnification of one fetal sample. (**C**) Automatic quantification of percentage of cells positive for SIX3 and ISL1 with NIS software on confocal images at 40X magnification. N = 3-6 fields for each zone (VZ, SVZ, IC and MZ) from 2-3 coronal slices of 1 fetus, 2-3 z stacks of each field are pre-mediated. Statistics with Prism, Anova One Way, Bonferroni post test, * p< 0,05. (**D**) Dot plot showing enrichment of Gene Ontology terms for each cell community of the LGE lineage.

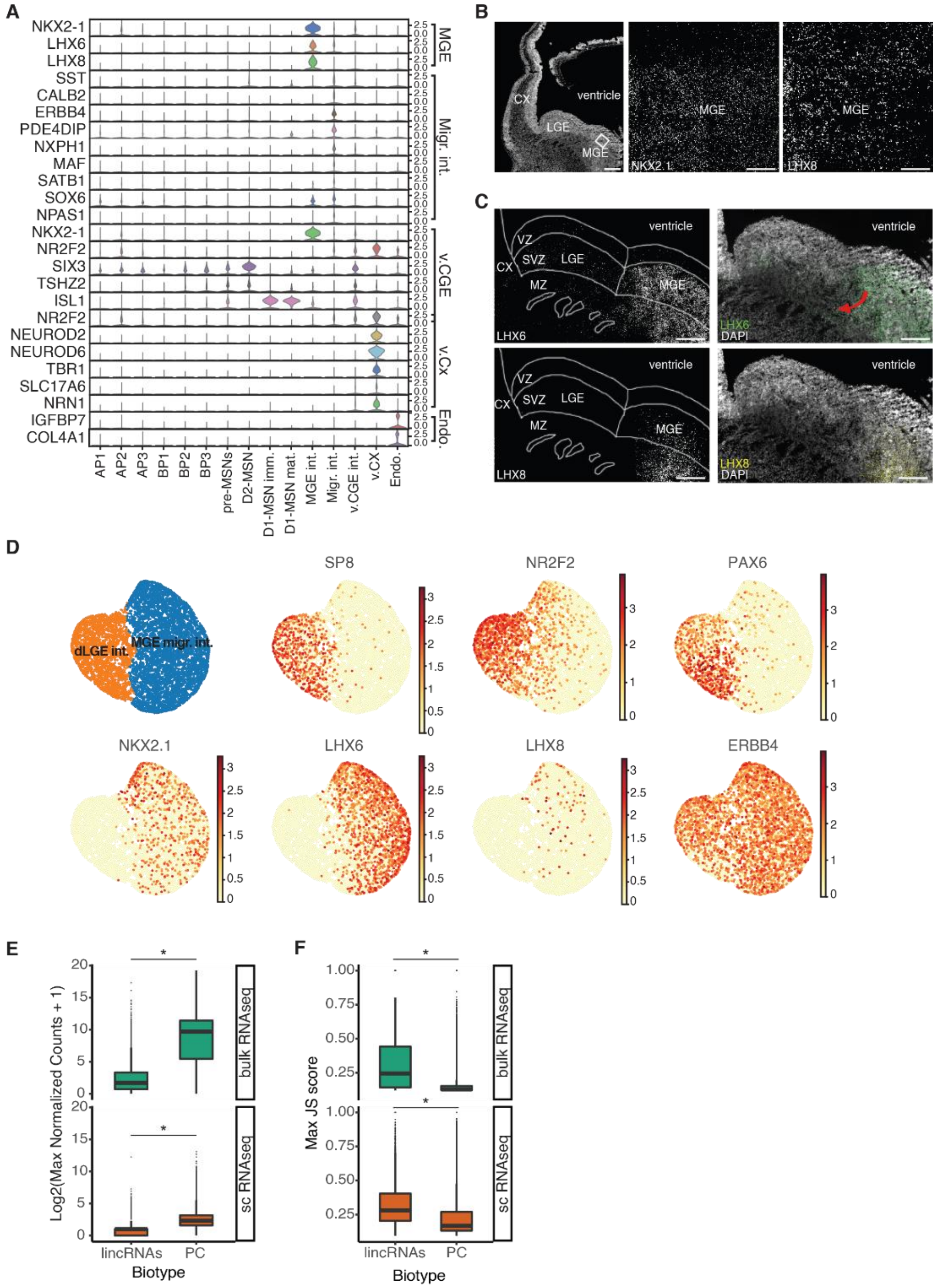

Fig. S5 Contaminating populations and lincRNA expression levels in single cells. (**A**) Violin plot showing expression levels of canonical marker genes of contaminating cells within the dissected LGE. (**B**) FISH validation at 9pcw of the MGE specific markers *NKX2.1* and *LHX8* on a telencephalic coronal hemisection (500 µm, 200 µm). (**C**) FISH validation on a telencephalic coronal hemisection at 9pcw confirming *LHX6* and *LHX8* expression in resident interneurons and LHX6^{+}/LHX8^{-} signal in migrating interneurons shown by the red arrow (300 µm). (**D**) Sub-clusters of migrating interneurons (Migr int.) shows two sub-clusters of interneurons one deriving from the dorsal (dLGE int.) and expressing *SP8*, *NR2F2* (*COUPTF2*), *PAX6* and the other from the MGE (MGE migr int.) identified by *NKX2.1* and *LHX6* expression. *ERBB4* is shared by both classes of interneurons. (**E**) Distribution of maximal normalized expression of lincRNAs and protein-coding-genes in bulk and single-cell data. (**F**) Distribution of maximal tissue specificity scores calculated for each transcript across the different areas and pcw in bulk data and across different cell states in single-cell data. Wilcoxon test with Bonferroni correction,*p<2e-16.

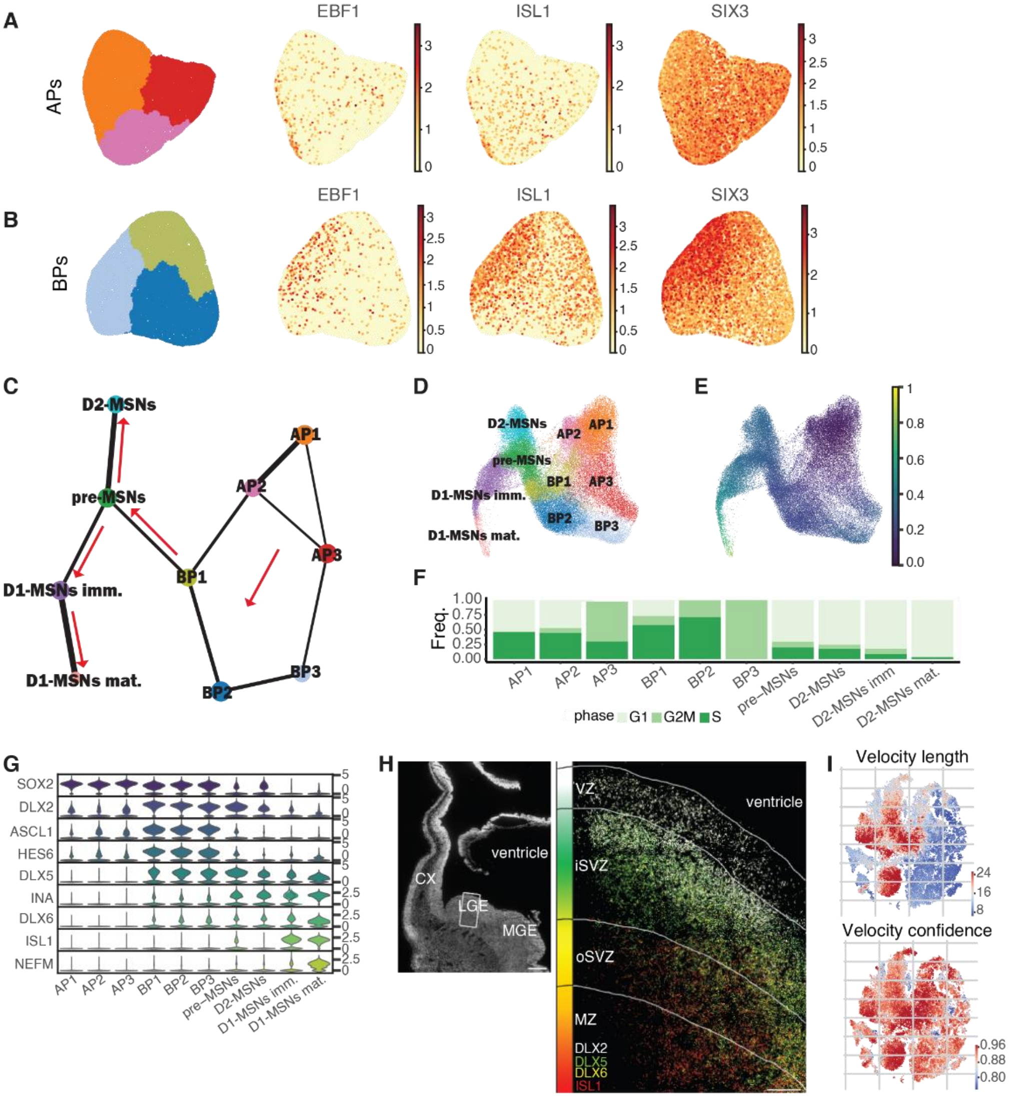

Fig. S6 Trajectory inference of the MSNs lineage. (**A, B**) Sub-clusters of APs (A) and BPs (B) show no commitment of certain cells to the D1 or D2 fate. (**C**) Abstracted graph of the LGE lineage showing edges connecting clusters with a connectivity threshold higher than 0.14. Nodes correspond to the clusters identified by the Louvain algorithm. The size of nodes is proportional to the number of cells in each cluster. Thickness of edges represents the strength of the connectivity between partitions. Arrows shows direction of differentiation according to canonical markers. (**D, E**) Single-cell embedding of the abstracted graph of the LGE lineage (D) and LGE lineage coloured by diffusion pseudotime score (E). (**F**) Frequency of each cell cycle phase in the different cell states of the LGE lineage. (**G**) Expression of canonical neuronal maturation genes along the differentiation trajectory of MSNs coloured by diffusion pseudotime score. (**H**) Multi-FISH analysis on a telencephalic coronal hemisection at 9pcw confirming the sequential activation of a set of genes within the ventricular zone (VZ), inner and outer subventricular zone (i/oSVZ) and mantle zone (MZ) in agreement with the computationally predicted differentiation trajectories (Scale bars: 500 µm, 150 µm). (**I**) The speed differentiation and the coherence of the vector field of the LGE single-cell dataset.

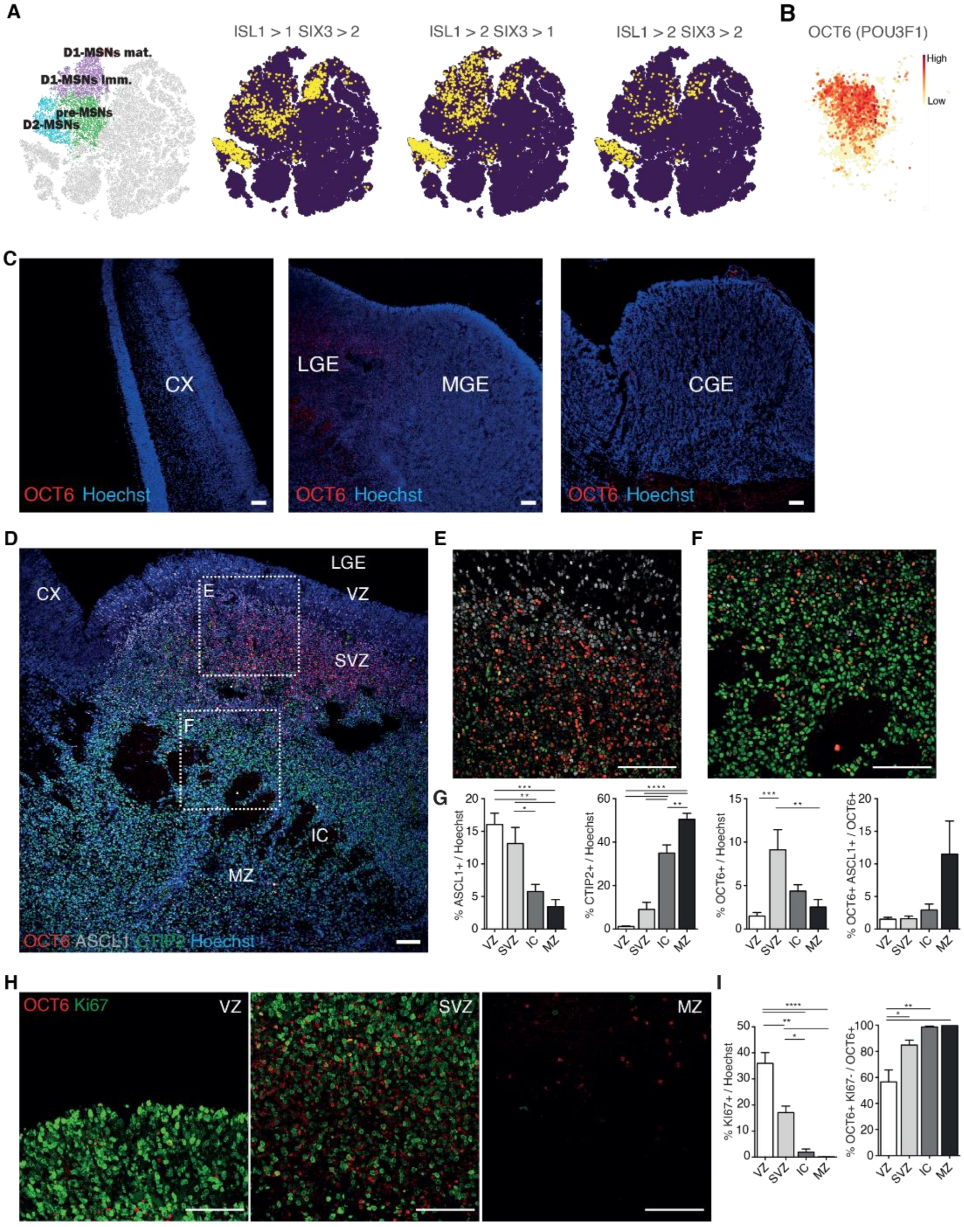

Fig. S7 IHC validation of the pre-D1 MSN state. (**A**) *t*-SNE plot showing cells co-expressing *ISL1* and *SIX3* with different thresholds of gene expression (co-expressing cells are shown in yellow). (**B**) OCT6 expression in pre-MSN sub-clusters. (**C**) OCT6 staining of telencephalic coronal hemisection on the CX, MGE and CGE at 9pcw. Scale bar 100µm at bottom right. Confocal acquisition at 10x magnification of one fetal sample. (**D**) OCT6, ASCL1 and CTIP2 staining of a telencephalic coronal hemisection at 9pcw (scale bar 100µm at bottom right) with (**E, F**) 40x magnification of the SVZ (E) and the internal capsule (IC) (F). (**G**) Automatic quantification of the percentage of cells positive for OCT6, ASCL1 and CTIP2 with the NIS software on confocal images at 40X magnification. (**H**) 40x magnification of OCT6 and KI67 staining in the VZ, SVZ and MZ of the LGE. (**I**) Automatic quantification of the percentage of cells positive for OCT6 and Ki67 with the NIS software on confocal images at 40X magnification. N = 3-6 fields for each zone (VZ, SVZ, IC and MZ) from 2-3 coronal slices of 1 fetus, 2-3 z stacks of each field are pre-mediated. Statistics were performed using Prism: Anova One Way, Bonferroni post test, * p< 0,05; ** p< 0,01; ***p<0,001;**** p<0,0001.

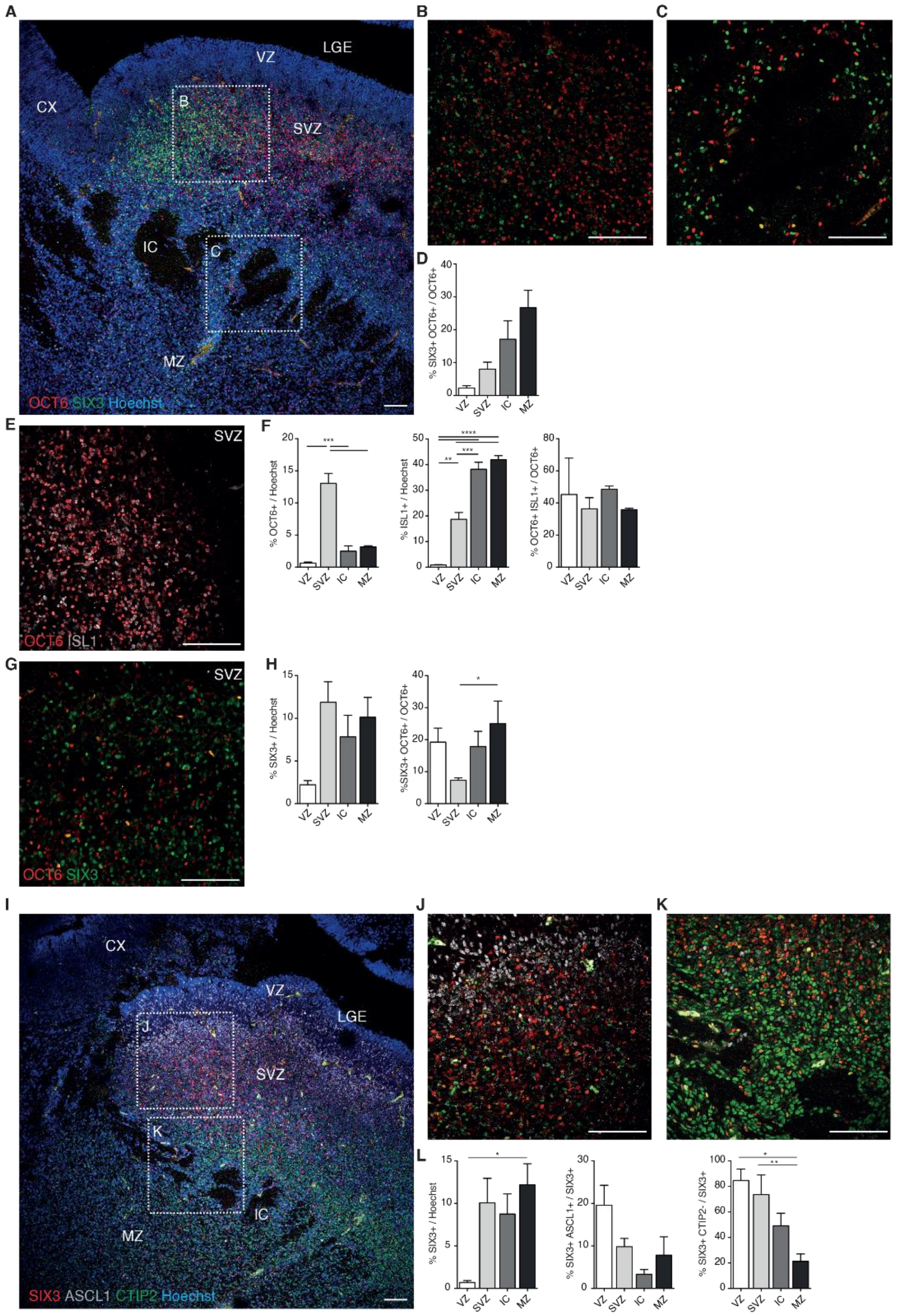

Fig. S8 IHC validation of the pre-D2 MSN state. (**A**) OCT6 and SIX3 staining of a telencephalic coronal hemisection at 9pcw (scale bar 100µm at bottom right) with (**B, C**) 40x magnification of the SVZ (B) and the internal capsule (IC) (C). (**D**) Automatic quantification of the percentage of cells positive for OCT6 and SIX3 with the NIS software on confocal images at 40X magnification. (**E**-**H**) 40x magnification of the SVZ at 11pcw and automatic quantification of the relative markers for OCT6-ISL1 (E, F) and OCT-SIX3 (G, H). (**I**) SIX3, ASCL1 and CTIP2 staining of telencephalic coronal hemisection at 9pcw (scale bar 100µm at bottom right). (**J, K**) 40x magnification of different regions: SVZ (F), IC (G). (**L**) Automatic quantification of the percentage of cells positive for SIX3, ASCL1 and CTIP2 with the NIS software on confocal images at 40X magnification. N = 3-6 fields for each zone (VZ, SVZ, IC and MZ) from 2-3 coronal slices of 1 fetus, 2-3 z stacks of each field are pre-mediated. Statistics were performed with Prism: Anova One Way, Bonferroni post test, * p< 0,05; ** p< 0,01; ***p<0,001;**** p<0,0001.

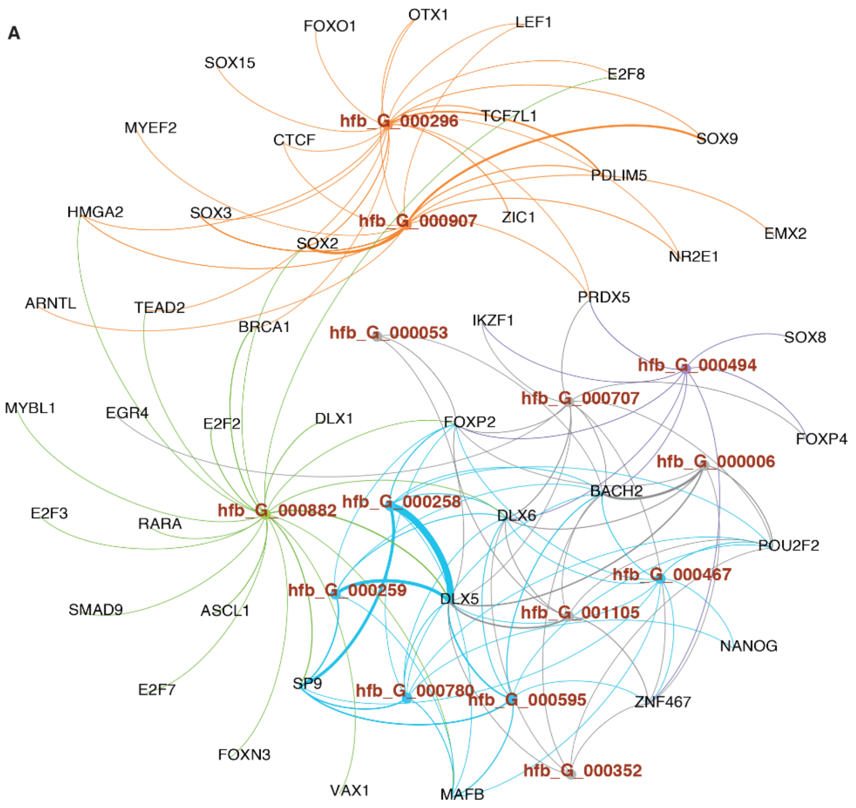

Fig. S9 Predicted GRNs between highly specific lincRNAs and key transcription factors (**A**) Inferred gene regulatory networks between lincRNAs and transcription factors that were highly specific in APs, BPs, D1-MSNs, D2-MSNs and general MSNs. Connections were calculated using GENIE3 in SCENIC. GRN plots of the different interactions were done using the Gephi software package. Edge thickness correlates with the strength of the connection. Colors represent the cell community the lincRNA is specific for: orange, APs; green, BPs; grey, general MSNs; blue, D2-MSNs; D1-MSNs purple.

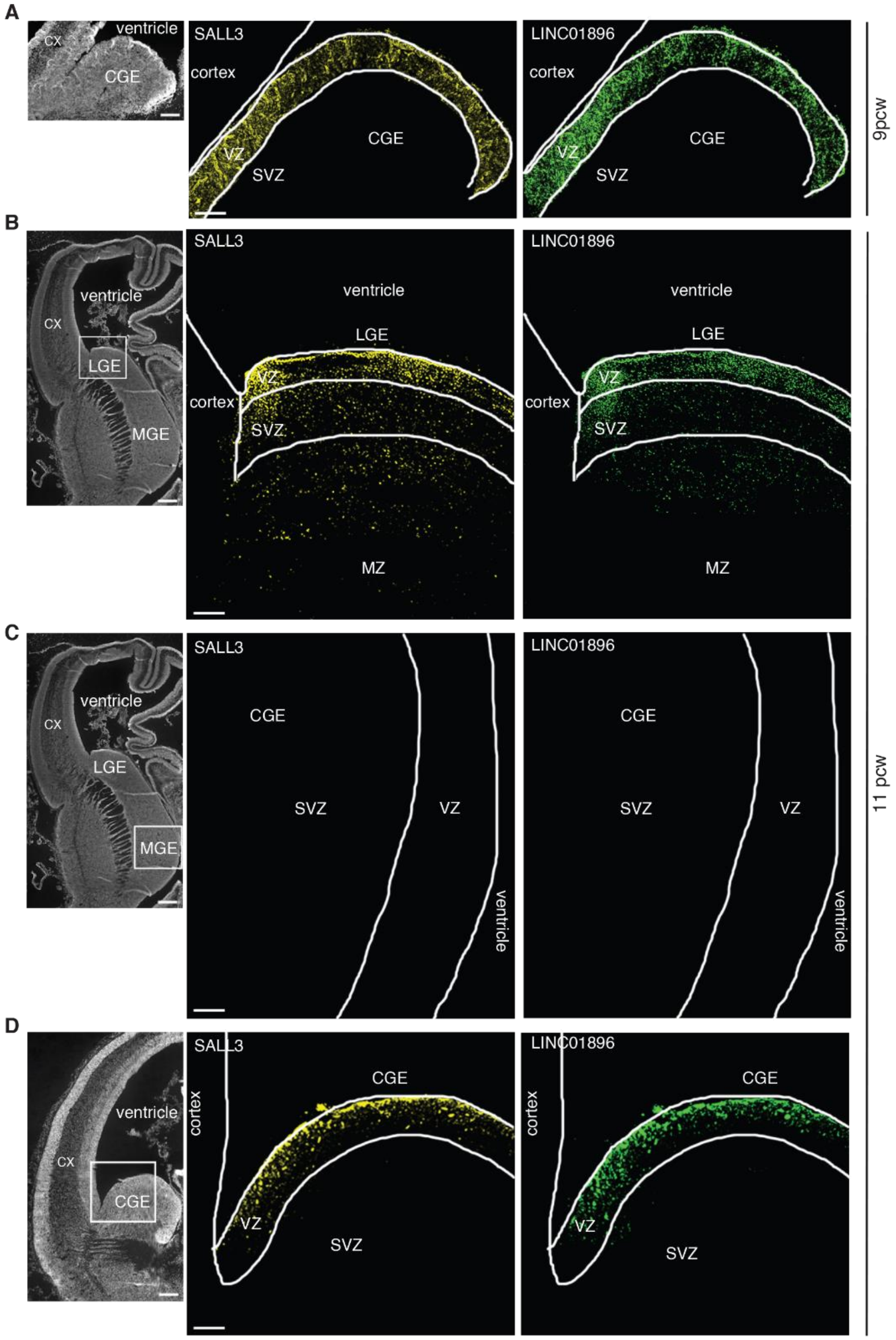

Fig. S10 Validation of specific APs markers in the VZ of the LGE. (**A**) FISH on a telencephalic coronal hemisection showing the presence of *SALL3* and *LINC01896* in the VZ of the CGE at 9pcw (Scale bars: 500 µm, 200 µm). (**B-D**) FISH on a telencephalic coronal hemisection showing the presence of *SALL3* and *LINC01896* in the VZ of the LGE and CGE and absence in the MGE at 11pcw (Scale bars: 500 µm, 200 µm).

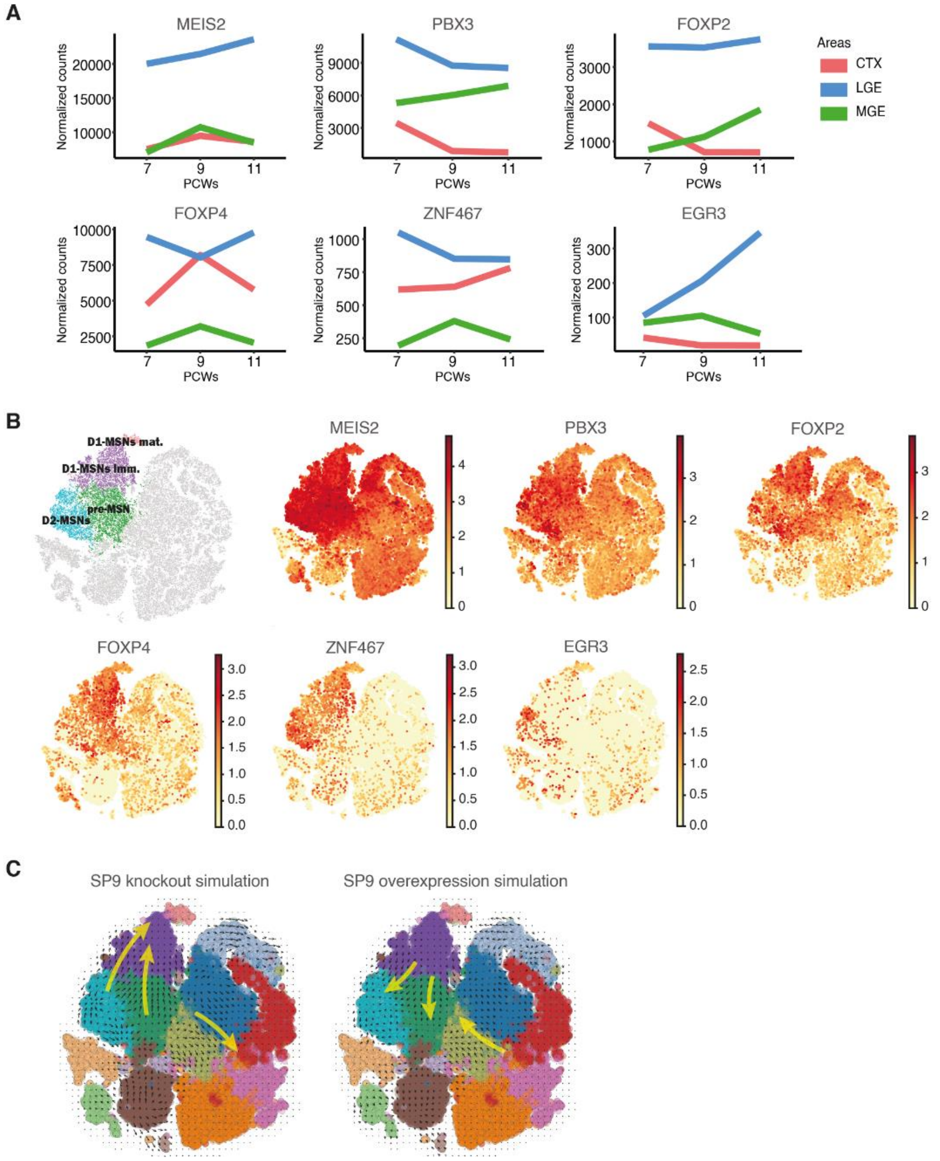

Fig. S11 Core regulons of MSNs. (**A**, **B**) Bulk (A) and single-cell (B) expression levels of LGE and MSNs specific transcription factors. (**C**) CellOracle simulation of knockout and overexpression of *SP9*. The effect of the perturbation in shown on the t-SNE plot with projections of cell state transition vectors for each cell. Yellow arrows were manually added to represent overall directionality.

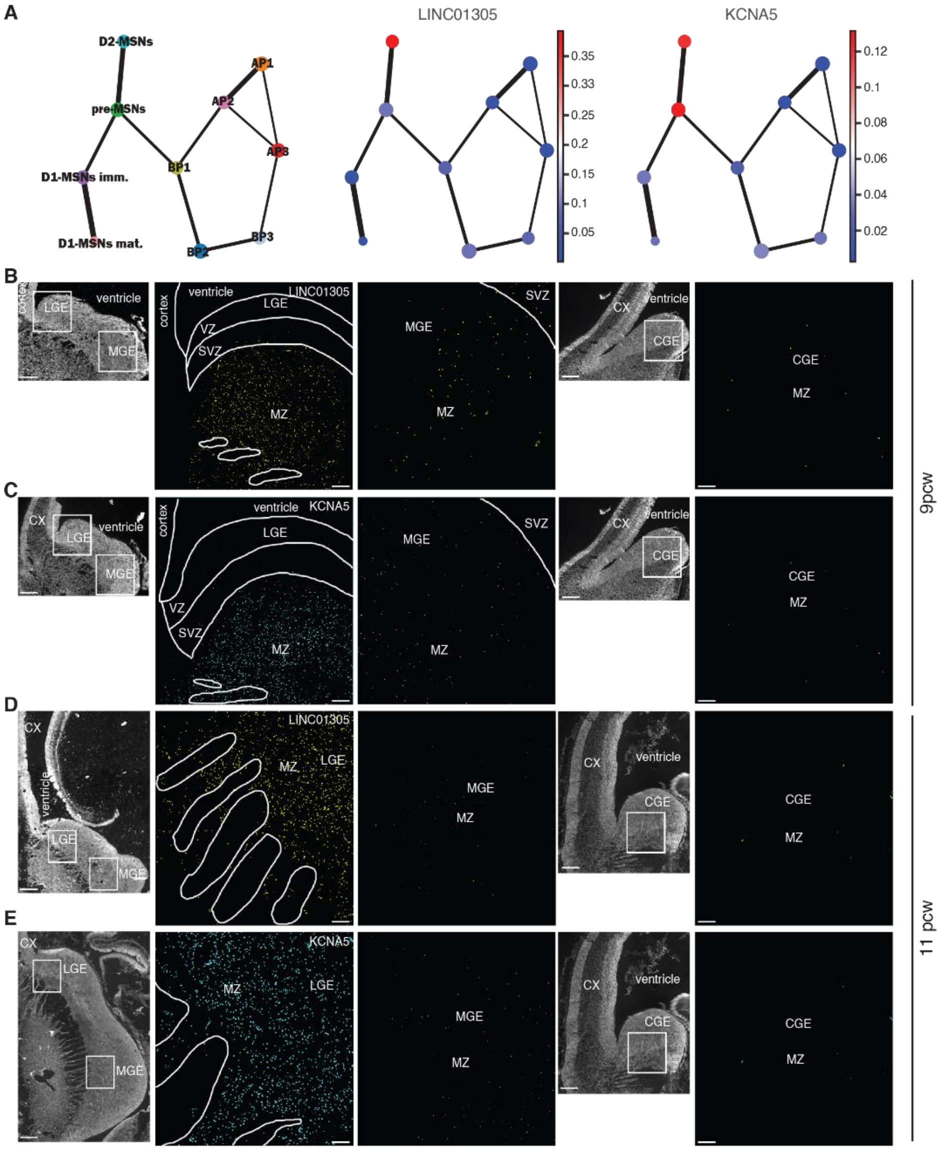

Fig. S12 The molecular signature of D2-MSNs. (**A**) Abstracted graph of the LGE lineage showing average expression of D2-MSN specific genes *LINC01305* and *KCNA5*. (**B-E**) FISH on a telencephalic coronal hemisection at 9pcw (B, C) and 11pcw (D, E) showing exclusive expression in the MZ of the LGE of the D2-MSNs specific markers *LINC01305* (B, D) and *KCNA5* (C, E) (Scale bars: 500 µm, 200 µm).

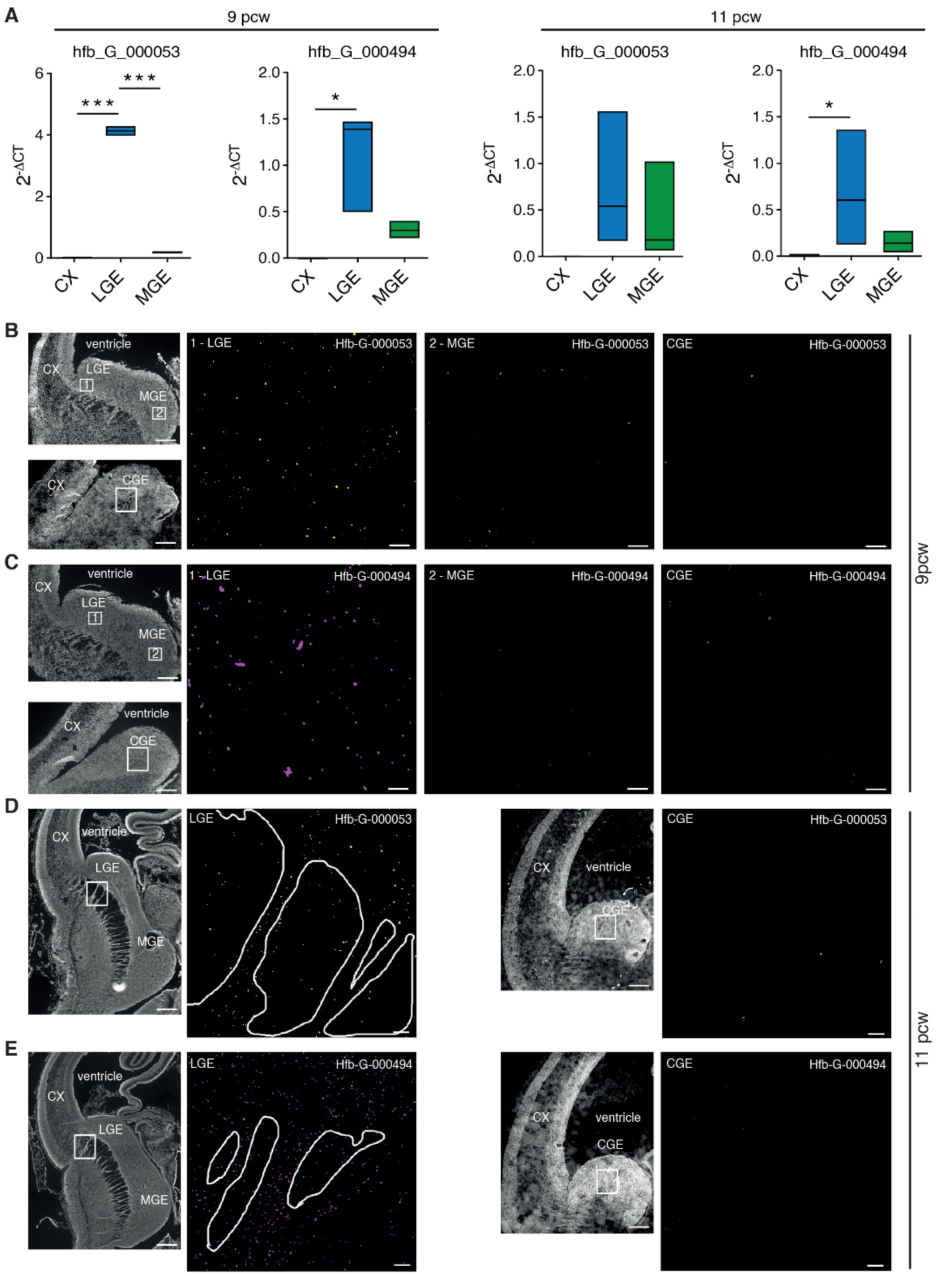

Fig. S13 Validation of human specific lincRNAs of the LGE lineage. (**A**) qPCR validation of *hfb_G_000053* and *hfb_G_000494* expression in bulk tissues of the neocortex, LGE and MGE at 9 and 11pcw. Expression of lincRNAs was normalized to 18S expression. Data are the mean ± SEM (n=3); one tailed Student t-test,*p<0.05, **p<0.01 and ***p<0.001. **(B, C**) FISH on a telencephalic coronal hemisection at 9pcw of the MSNs specific lincRNAs *hfb_G_000053* (B) and *hfb_G_000494* (C) in the mantle zone of the LGE, MGE and CGE (500 µm, 100 µm). (**D, E**) FISH on a telencephalic coronal hemisection at 11pcw of the MSNs specific lincRNAs *hfb_G_000053* (D) and *hfb_G_000494* (E) in the internal capsule of the developing striatum and in the CGE (500 µm, 100 µm).

Tables S1-S12

Table S1

List of the fetal samples used for bulk, scRNA-seq, FISH, IHC and qPCR experiments.

Table S2

List of all novel lincRNAs identified by this study with chromosome number and start and end positions.

Table S3

Lists of differentially expressed genes in the LGE, MGE and neocortex at 7,9,11 and 20pcw.

Table S4

List of the functional GO terms related to the specific LGE transcriptional signatures at 7,9,11pcw and 20pcw.

Table S5

List of the functional upstream regulators of the neocortex, LGE and MGE transcriptional signature at 7,9,11pcw and 20pcw.

Table S6

List of differentially expressed genes in the different cell populations resulting from scRNA-seq.

Table S7

A list of the Functional GO terms related to the specific signature of each cell community of the LGE lineage

Table S8

List of top 200 driver genes of velocity vectors

Table S9

List of regulons for each cell state

Table S10

List of lincRNAs and transcription factors connections

Table S11

List of genes that are cell type specific and LGE specific together with the specific APs and D2-MSNs signatures.

Table S12

Predicted functional terms that correlate and anti-correlate with the human specific lincRNAs hfb_G_000053 and hfb_G_000494.