UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI BIOSCIENZE

DOTTORATO DI RICERCA
DBMC
IN BIOLOGIA MOLECOLARE E CELLULARE

UNIVERSITÀ DEGLI STUDI DI MILANO
Scuola di Dottorato in Biologia molecolare e cellulare
XXXI Ciclo

The coding and non-coding transcriptome of the human fetal striatum from a single-cell perspective

Vittoria Dickinson Bocchi

PhD Thesis

Scientific tutor: Prof. Elena Cattaneo

Academic year: 2017-2018

SSD: BIO/14

Thesis performed at the Laboratory of Stem Cell Biology and Pharmacology of Neurodegenerative Diseases. Department of Biosciences, University of Milan and Istituto Nazionale di Genetica Molecolare.

Table of Contents

Abstract

The human brain is a tissue of vast complexity in terms of the cell types it comprises. Understanding how this complexity arises requires a deep understanding of how neuronal patterning progresses at the single cell level. Previous studies have concentrated on cell fate decisions during cortical development but little is known about how the lateral ganglionic eminences (LGE) develops and gives rise to the different cell populations of the striatum. Conventional approaches to classify cell types in this area have been limited to exploring relatively few markers and therefore have provided a narrow characterization of any given cell type. Furthermore, most studies have bounded their inquiries to protein-coding genes and have not investigated the role of lncRNAs that have been shown to have a high cell and tissue specificity.

Taking these aspects under consideration, here we combined bulk RNA-seq and single-cell RNA-seq to decode an unambiguous gene signature of the striatum and reveal how neural progenitors of this domain are able to differentiate at the single cell level. In particular, we deeply profiled the LGE and the surrounding neocortex and medial ganglionic eminences (MGE), from 7 to 20 postconceptional weeks (pcw), and performed *de novo* lncRNAs analysis that enabled us to define the first dictionary of novel lincRNAs for these areas. Furthermore, this analysis led to the establishment of a unique gene signature for the three different regions. Subsequently, we performed single-cell RNA-seq of the LGE at 7pcw that unravelled a plethora of different cell populations of the LGE. Pseudotemporal ordering of these cells uncovered the first developmental trajectory of striatal neurons and how they transition from early progenitors to mature medium spiny neuron (MSNs) and their coding and non-coding transcriptional signature. This array of cells will shortly be complemented by another batch of single-cell libraries from later time points (9-11pcw) that will be used to full characterize the early steps of neural ramifications that lead to the generation of the human striatum.

The relevance of the approach relies on the availability of extremely rare human fetal samples combined with the most revolutionary RNA sequencing technologies and highly elaborate computational tools that enable the investigation of a brain region, the striatum, whose development is poorly understood and which plays a major role in human brain physiology and pathology.

Abstract (Italian)

Il cervello umano è altamente eterogeneo dal punto di vista cellulare. Questa eterogeneità si realizza durante lo sviluppo fetale attraverso l'attività di reti geniche la cui identità e ruolo sono ancora largamente sconosciute. Molti studi si sono concentrati alla comprensione dello sviluppo della neocorteccia, mentre le conoscenze relative ai meccanismi cellulari e molecolari che sottendono la formazione del nucleo striato umano, sono ancora estremamente limitate. Inoltre, la maggior parte degli studi si sono focalizzati sui geni codificanti proteine con minor attenzione verso il ruolo dei lncRNAs prodotti da geni non codificanti, che sono molto più specifici nel definire stati cellulari rispetto a geni codificanti proteine. Lo scopo del mio progetto di tesi è quello di approfondire le conoscenze sullo sviluppo striatale umano e identificare un panello di geni specifico per questa area cerebrale. Per questo scopo abbiamo raccolto tessuto autoptico proveniente da feti umani di epoca gestazionale compresa tra le 7 e le 20 settimane e dissezionato LGE (area che dà origine allo striato), la neocorteccia (presente dorsalmente al LGE) e MGE (area che dà origine al globo pallido e agli interneuroni corticali, che si trova ventralmente all'LGE). Questi tessuti sono stati processati e l'RNA di ogni area è stato analizzato mediante RNA-seq ad alta profondità. Questa profondità ha permesso di identificare, grazie ad un'analisi computazionale *de novo* del trascrittoma, un pannello di nuovi lncRNA altamente specifici per lo sviluppo striatale. Lo studio effettuato con RNA-seq ha permesso di identificare un codice unico per lo sviluppo dello striato, ma non le diverse sottopopolazioni cellulari. Per questo abbiamo deciso di ampliare l'analisi mediante uno studio pilota di dati single-cell RNA-seq da tessuti fetali tra le 7 e le 8 settimane gestazionali. Grazie ad analisi bioinformatiche di questi dati abbiamo "catturato" la complessità cellulare o dello striato a tempi gestazionali precoci, classificando i diversi sottotipi di cellule e il loro specifico profilo trascrizionale. Inoltre, sono stati applicati particolari algoritmi per decifrare la traiettoria di differenziamento intrapresa da neuroni striatali e i relativi "master regulators" che permettono di indirizzare le cellule verso questo particolare destino cellulare. Questa matrice iniziale di dati verrà a breve integrata dai dati ottenuti dall'analisi delle settimane gestazionali più tardive (9 e 11 settimane) che verranno utilizzate per caratterizzare pienamente i diversi codici trascrizionali e le sottopopolazioni cellulari che portano alla formazione dello striato umano.

La rilevanza e importanza di questo approccio è basata sulla combinazione di tecnologie rivoluzionare di sequenziamento dell'RNA, di metodi computazionali avanzati e sulla disponibilità di materiale fetale autoptico umano per investigare lo sviluppo di un'area cerebrale fortemente implicata nella fisiologia e patologia del cervello umano.

Aim

The full experimental design is shown in Figure 1. The four main goals of this project can be summarized as:

I. Setup a pipeline for *de novo* intergenic lncRNAs (lincRNA) discovery from bulk RNA-seq data.
II. Identify a specific transcriptional blueprint of the neocortex, LGE and MGE.
III. Implement a computational pipeline to analyse single-cell RNA-seq data of the LGE to pinpoint different cell populations in the developing striatum and infer developmental trajectories of MSNs. Initial analysis will be performed on a pilot experiment to determine optimal experimental design.
IV. Look at distribution of protein coding genes and lincRNAs in these subpopulations to identify potential candidate marker genes.

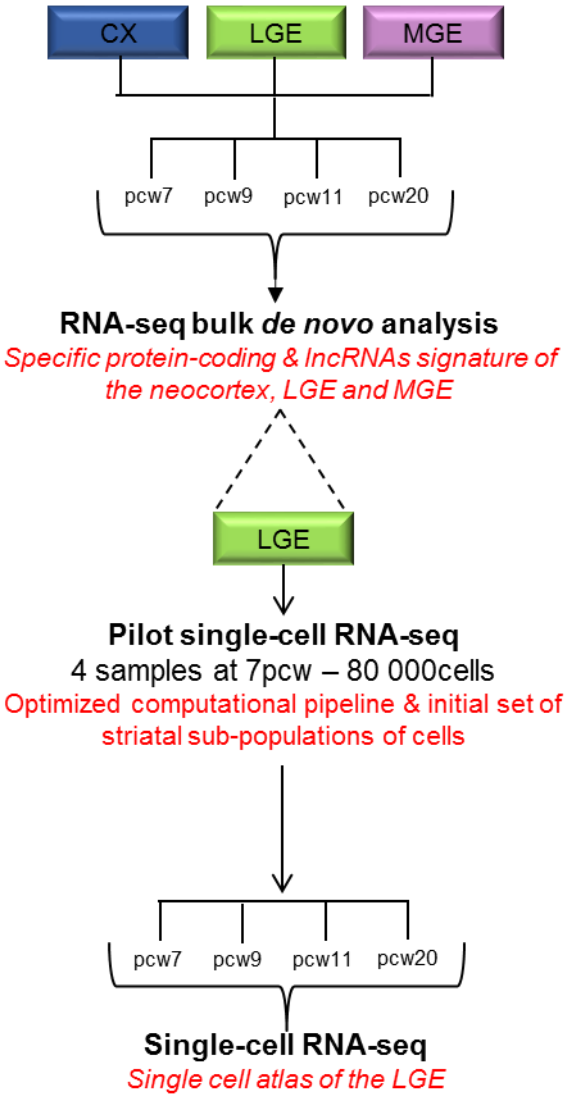


Figure 1 Full experimental design. LGE-Lateral ganglionic eminence; MGE-Medial ganglionic eminence; CGE-Caudal ganglionic eminence. Output of each step is shown in red.

1 | Introduction

1.1 Rationale behind this study

One of the promising treatments for a selected number of neurodegenerative disorders is cell-replacement therapy using clinically relevant neuron types generated *in vitro* from stem cells (Steinbeck and Studer, 2015).

In the past years this field has greatly expanded, especially for Parkinson's disease (PD) and clinical trials using committed dopaminergic neural progenitors from human fetal tissue in PD (Lindvall *et al.*, 1989, 1990) have proved the concept that this may become a promising treatment for neurodegenerative diseases. Based on this proof of principle many groups have concentrated their effort on generating highly reproducible stem cell-derived midbrain dopaminergic neurons *in vitro* to use in cell replacement therapy in PD (Fasano *et al.*, 2010; Kirkeby *et al.*, 2012, 2017; Nolbrant *et al.*, 2017). Thanks to these efforts the field is now very close to moving to clinical trials using *in vitro* generated cells (Barker *et al.*, 2017).

In the case of the work conducted during my PhD thesis, its long term goal is to reach the same standards achieved for PD but for Huntington's disease (HD). The expectation is that by unravelling human striatum development *in vivo*, we will be able to more properly instruct stem cells *in vitro* to turn into a donor neuronal population similar, if not identical, to the one lost in HD, therefore allowing the latter to be replaced with new neurons. As for PD, the early days in the HD field were characterized by clinical trials where donor human fetal tissue was shown to hold some potential for HD patients (Bachoud-Lévi *et al.*, 2000; Reuter *et al.*, 2008). As a consequence, different groups (including ours) have tried to reproduce in a dish neostriatal medium spiny projection neurons (MSNs) that are the specific neural subtype that degenerates in HD (L. Ma *et al.*, 2012; Carri *et al.*, 2013). However, the scientific community working on this protocol has not reached the efficiency of PD protocols. One of the main reasons behind this scarce efficiency is that to fully recapitulate this neuronal subtype *in vitro* requires a deep understanding on how these cells develop and diversify in the fetal striatum. In line, the main gap that sets apart the HD and the PD "world" is the basic knowledge on how specific cells differentiate and develop in the human fetal brain. To this point, the PD community has dedicated much of its efforts to dissect how cells differentiate within the developing human (La Manno *et al.*, 2016) and mouse (Kee *et al.*, 2017) mesencephalon. To learn from early development timepoints, these groups have exploited one of the newest and most revolutionary technologies of the past few years, called

single-cell RNA-seq (scRNA-seq). This approach has inevitably led to a refinement in the understanding of different sub-groups of dopaminergic neurons and their associated genetic code that was then exploited in the lab as a compass to predict the authenticity of mesencephalic dopamine progenitors during differentiation (Kirkeby *et al.*, 2017) .

To fill the gap we have therefore decided to follow in the steps of the PD community and dissect the highly wired and organized transcriptional program that, over time, support the emergence of MSNs. To understand which gene expression patterns determine a specific cell lineage we wanted to combine both bulk RNA-seq and scRNA-seq to fully capture the transcriptional signatures that control striatal developmental, both from a coding and a non-coding perspective. To this point, we decided to add a layer of complexity and also explored the role of lncRNAs in defining cell states as recent evidence suggests that these biotypes are more tissue specific compared to protein-coding genes, especially in the brain (Cabili *et al.*, 2011; Derrien *et al.*, 2012; Liu *et al.*, 2016).

In this thesis work, we therefore decided to first exploit bulk RNA-seq to identify new lncRNAs expressed in the striatum and then to decode a specific coding and non-coding signature for this area that distinguishes it from the surrounding neocortex and medial ganglionic eminence. The single cell data was then used to capture the cell diversity of this area and understand the dynamics underlying MSN differentiation together with the master regulators of fate decision for this specific cell subtype.
This thesis will start by describing what we know of the adult striatum in terms of structure and cell subtypes. I will then illustrate the genetic programs, known up till now that characterize the striatum during development. The second section will elaborate on the nature and function of lncRNAs and why they are an important component of cell function and nervous system development. Finally, in the third section, I will delineate basic concepts of bulk and scRNA-seq and what variables must be considered during experimental design. Specific considerations employed during the construction of the bioinformatics pipeline will be discussed in the "Rationale" section of the Methods part of the thesis.

1.2 The striatum

1.2.1 Cell types and structure

The striatum, that consists in the caudate and putamen, is a structure embedded deep in the brain hemispheres and, together with the globus pallidus, is part of the basal ganglia (Figure 1.1). Together with the related nuclei (subthalamic nucleus, substantia nigra and pedunculopontine nucleus) the basal ganglia form complex circuits that are engaged in motor control, motor learning, behavior and emotions (Jain *et al.*, 2001). The functional organization of the basal ganglia will not be discussed in this thesis but is reviewed here (Lanciego, Luquin and Obeso, 2012).

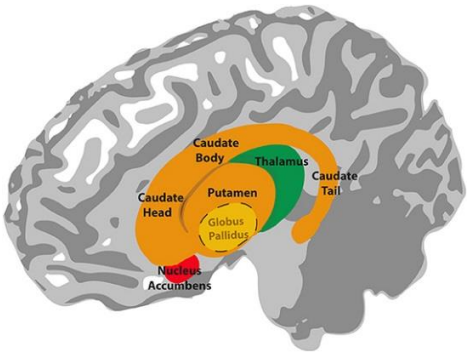


Figure 1.1 The basal ganglia. Schematic illustration of the anatomical location of the the putamen and the caudate that form the striatum, the globus pallidus , the thalamus that is located underneath the basal ganglia and the nucleus accumbens (Lim, Fiez and Holt, 2014)**.**

The complexity of the striatum can already be appreciated in terms of cell types it comprises. In line, striatal neurons can be generally divided into medium spiny projection neurons (MSNs) and aspiny interneurons. MSNs can be further compartmentalized according to their projection targets. In particular, MSNs innervating the external globus pallidus (GPe) are characterized by the dopamine receptor subtype 2 (D2R) and the neuropeptide enkephalin and form the indirect pathway (Schiffmann, Jacobs and Vanderhaeghen, 1991; Gerfen, 1992). While, MSNs projecting directly to the internal globus pallidus (GPi) and substantia nigra pars reticulata (SNr) contain dopamine receptor subtype 1 receptors (D1R) together with neuropeptides substance P and dynorphin and form the direct striatopallidal pathway (Gerfen and Scott Young, 1988; Ince, Ciliax and Levey, 1997) (Figure 1.2).

While MSNs are the major occupants of the striatum, interneurons play a fundamental role in coordinating the microcircuit of this area and compose about 2–3% in rodent (Tepper and Bolam, 2004) and possibly up to 23% in primates (Graveland, Williams and Difiglia, 1985). The first group are GABAergic interneurons (that have an abundance around 2% in rats) that can be divided

into 3 sub-populations neurochemically (Figure 1.2). One expresses the peptides somatostatin and neuropeptide Y (NPY) as well as the enzymes NADPH diaphorase and nitric oxide synthase. The other two express either parvalbumin or calretinin (Kawaguchi et al., 1995). The second group of interneurons (that have an abundance of approximately 0.3% in rats) are called the cholinergic interneurons (Figure 1.2) and are the largest neurons in the neostriatum, and are now known to be cholinergic on the basis of choline acetyltransferase immunolabeling (Bolam, Wainer and Smith, 1984).

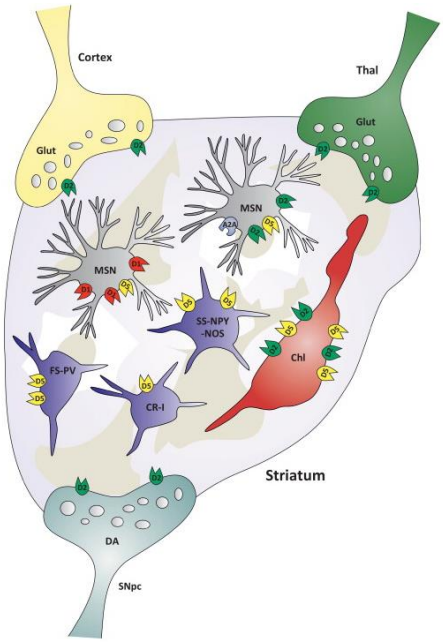


Figure 1.2 Schematic representation of neural subtypes within the stratum. GABAergic medium spiny neurons (MSNs) either express D1DR (D1) or D2DR (D2) receptors. Interneurons can be either fast spiking (FS) parvalbumin (PV), calreticulin (CR-I) or Y (NPY), NOS (nitric oxide synthase) and somatostatin (SS) positive or cetylcholine (ChI) positive (Goodchild, Grundmann and Pisani, 2013).

The diversity of the striatum is not limited to the cell types it comprises, but also from an architectural point of view two different domains can be observed. The first areas are patches rich in μ opioid receptor (Pert, Kuhar and Snyder, 1976; Desban *et al.*, 1995) with strong immunoreactivity against enkephalin, substance P, GABA, and neurotensin (Graybiel *et al.*, 1981; Gerfen, 1984) and are known as striosomes. While the second area is in the background surrounding the striosomes and is enriched in parvalbumin (Prensa, Giménez-Amaya and Parent, 1999) and calbindin (Gerfen, Baimbridge and Miller, 1985) and is known as the matrix.

At present, understanding how this diversity in cell types and distinctiveness in organization is achieved is still unknown. Here I will review what is known on the regional and temporal events that lead to the formation of the striatum from the telencephalon.

1.2.2 Genetic control of striatal development

The striatum is derived from the lateral ganglionic eminence (LGE) of the developing fetal brain, that is found in the most anterior region of the embryo that is called the telencephalon (Stiles and Jernigan, 2010). The telencephalon, gives rise, together with the striatum to different structures of the brain. In particular, the dorsal part of the telencephalon gives rise to the neocortex in the anterior and lateral region, while the posterior and medial area develops into the hippocampus, the cortical hem and the choroid plexus. Instead, the ventral telencephalon gives rise, together with the LGE, to the medial ganglionic eminence (MGE) and the caudal ganglionic eminence (CGE) (Figure 1.3). The MGE produces projection neurons of the globus pallidus, the amygdala and the septum together with migrating parvalbumin (PV) and somatostatin SST interneurons (Xu, Tam and Anderson, 2008) destined to the cortex, striatum or hippocampus. The LGE contributes to the formation of the already discussed striatum and to interneurons that migrate to the olfactory bulb (Deacon, Pakzaban and Isacson, 1994; Jain *et al.*, 2001; Wichterle *et al.*, 2001; Wonders and Anderson, 2006). Instead, the CGE generates the local caudal striatal and pallidal neurons, as well as different subtypes of calretinin (CR), vasoactive intestinal peptide (VIP) or reelin migrating interneurons that migrate to the cortex hippocampus, amygdala and other limbic system nuclei (Laclef and Métin, 2018).

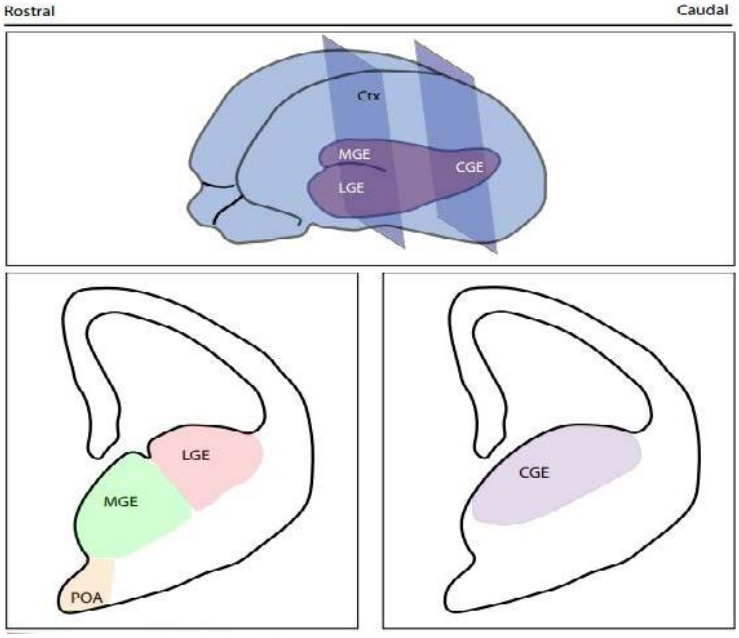


Figure 1.3 Schematic representation of the telencephalon during mouse brain development. LGE-Lateral ganglionic eminence; MGE-Medial ganglionic eminence; CGE-Caudal ganglionic eminence (Greenberg, Ramshaw and Schwarz, 2015).

The key question we are trying to answer with this thesis work is how cells within the telencephalon undertake a specific cell and positional identity to form these different areas. What set of signaling cascades lead to the segregation of the striatum? Here I will briefly summarize what we know on the genetic program that leads to the formation of the striatum.

The LGE, has been the most deceptive areas of the basal ganglia and a unique blueprint for this area, especially within the progenitor domain, is still to be delineated. In line, the available albeit limited information describes the LGE as displaying a blend of gene expression patterns that lies in between the cortical and MGE signatures and is characterized by the expression of *Dlx2* (shared with the MGE), *Gsx2* (shared with the MGE), and *Pax6* (shared with the neocortex), and – in mice - lack of *Nkx2-1* expression (shared with the neocortex) (Stoykova *et al.*, 1996; Szucsik *et al.*, 1997; Sussel *et al.*, 1999).

Within the developing LGE, two domains can be easily observed, the dorsal third of the LGE (dLGE) that is thought to give rise to olfactory bulb interneurons and is characterized by *Er81* expression in the VZ and Sp8 expression in the SVZ and in mature interneurons (Stenman, Toresson and Campbell, 2003; Waclaw *et al.*, 2006). While, the ventral two-thirds of the LGE (vLGE), that gives rise to striatal projection neurons, has low levels of *Er81* expression in the VZ but displays high levels of *Islet1(ISL1)* expression in the SVZ and mantle regions (Stenman, Toresson and Campbell, 2003). The mantle zone is also defined by *Ikzf1* (Tucker *et al.*, 2008), *Meis2* (Toresson *et al.*, 1999), *Ctip2* (Arlotta *et al.*, 2008), *Foxp1*, (Tamura *et al.*, 2004) and *Darpp-32* (Anderson and Reiner, 1991) expression. A more detailed examination of these two areas (Flames *et al.*, 2007) further subdivided the dLGE in 2 subdomains, the first characterized by high levels of *Pax6, Gsx2,* and *Er81*, while the second dLGE territory contains low levels of *Pax6* and high levels of *Gsx2*. Also the vLGE can be divided in another two sub-domains, both characterized by *Er81* and *Isl1* expression, however, the most ventral domain expresses *Nkx6-2* in the progenitor territory, whereas the district between the dLGE and this more ventral region does not.

However, the main challenge in this area, as mentioned above, is deciphering a unique and also functional gene signature for the progenitor domains, especially between LGE and MGE. For example, *Mash1*, in the LGE, is an important regulator of neurogenesis, where it is required both to specify neuronal precursors and to control the timing of their production (Casarosa, Fode and Guillemot, 1999). However, this genes characterizes all ganglionic eminences (Lo *et al.*, 1991) and *Mash1* null mice also suffer loss of neuronal precursors in the MGE (Casarosa, Fode and Guillemot, 1999). Another gene expressed in both the LGE and MGE is *Gsx2*, that although is not a unique marker of the LGE, it appears to have a fundamental function in maintaining the

molecular identity of striatal progenitors during LGE development (Szucsik *et al.*, 1997). In line, mice with a targeted mutation of *Gsx2* show defects in LGE development, with a lack of *Dlx2*, *Mash1* and *Ebf1* expression (Szucsik *et al.*, 1997; Corbin *et al.*, 2000; Toresson, Potter and Campbell, 2000). Another marker expressed in both the LGE and MGE is *Ebf1*. This gene also plays an essential role in the acquisition of the striatal mantle cell molecular identity, as *Ebf1* deficient mice are unable to downregulate SVZ-specific genes and are unable to progress through differentiation and activate mantle-specific genes. However, the MGE remains unaffected in these mutants as tangentially migrating cells remain untouched, suggesting that this gene specifically exerts its function in the LGE territory (Garel *et al.*, 1999). With regards, to MSN specific markers and their role, only a few markers have been investigated. One key MSN marker is CTIP2 (that is also found in the neocortex (Arlotta *et al.*, 2005) that has a fundamental role in MSN differentiation and patch-matrix organization within the striatum (Arlotta *et al.*, 2008). Another key elements is *FoxP1*, that has been shown to have an evolutionarily conserved role in regulating pathways involved in striatal neuron identity (Araujo *et al.*, 2015).

Regarding the development of the human striatum, only one studied performed in 2014 in the laboratory where this thesis was conducted, deeply examined the progression of this area compared to cortical domains (Onorati *et al.*, 2014). The results of this study showed that the VZ is characterized by OTX2, FOXG1, GSX2 and ASCL1, while the SVZ is defined by ASCL1, ISL1 and EBF1. This study also determined a unique signature for early striatal neurons that express ISL1, FOXP1, FOXP2, CTIP2, EBF1, GABA, DARPP-32 and interestingly NKX2-1. NKX2-1 in this region was a peculiar result, as in mouse, *Nkx2-1* is strongly expressed in the MGE but not in the LGE (Nobrega-Pereira *et al.*, 2010). The same was previously described in humans that defined NKX2-1 as an MGE marker (Fertuzinhos *et al.*, 2009; Hansen *et al.*, 2013; Ma *et al.*, 2013). Another difference identified in humans compared to rats, is that MSNs of this area have lower ISL1 expression but maintain DARPP-32 at 20 w, in contrast to rats (Wang and Liu, 2001). Recently, we performed another study on the human striatum (Castiglioni *et al.*, 2018), that pinpointed DACH1 as a key marker of human MSNs adding another key element to help discriminate striatal neurons (for details of my contribution on this work see the Appendix 6.1.2).

1.2.3 Concluding Remarks

Taken together, the literature up till now has given an initial "sketch" on how the striatum develops and diversifies from the surrounding MGE, CGE and neocortex. However, a higher resolution is required to truly partition this area, as a clear cut signature for the LGE is still elusive. Furthermore,

the adult striatum displays a greater diversity than what has been described for early developmental timepoints suggesting that our understanding of neural ramification at early developmental stages is still limited. The first diversity is characterized by the fact that MSNs are not the sole neuronal population within the striatum. In line, as discussed in the introductive remarks of this thesis, at least four different striatal interneurons are present in the striatum (Graveland, Williams and Difiglia, 1985; Tepper and Bolam, 2004). However, a recent study was able to identify 7 different populations of interneurons in mice (Muñoz-Manchado *et al.*, 2018) suggesting that are knowledge on these type of neuron is still limited. Another important source of variety is given by MSNs themselves. In particular, as discussed, two subtypes of MSNs (D1 and D2) have been described both at the functional and molecular level (DeLong and Wichmann, 2009). Even for this population, recent scRNA-seq work on mice (Gokce *et al.*, 2016; Zeisel *et al.*, 2018) suggested that this classification was somewhat reductive. In line, one study showed that D1-MSNs can be partitioned into two discrete populations: *Pchd8*-D1s and *Foxp1*-D1s. Furthermore, the *Foxp1*-D1 population could be separated into *Foxo1*-high or *Dner*-high populations. The D2-MSNs also displays two discrete subpopulations defined as *Htr7*-D2s and *Synpr*-D2s. Even here, the *Synpr*-D2 neurons can be further subdivided into *Calb1*-high or *Cartpt*-high subpopulations. A third population of MSNs instead was characterized by co-expression of *Drd1a* and *Drd2* (Gokce *et al.*, 2016). Another very recent mouse study from the Linnarsson group (Zeisel *et al.*, 2018) showed the presence of two D1-type MSNs one enriched in dorsal and one in ventral striatum, as well as two D2-type MSNs also with a dorsal and ventral distribution. This study showed that a molecular contrast is present that corresponds to the distinct circuits and functions of dorsal MSNs (initiate and control movements) and ventral MSNs (involved in motivation, reward and aversion). Furthermore, the group found putative patch-specific D1-/D2-type neurons (expressing Tshz1) and matrix-specific D2 neurons (expressing Gng2).

In conclusion, what we know about the cell types within the developing striatum and how the complexity described above is achieved is still narrow and unravelling key genes that define specific cell states will greatly aid understanding how we can obtain these neurons *in vitro*.

1.3 The role of lncRNAs in defining cell identity and function

1.3.1 Why lncRNAs

Today, lncRNA transcripts have emerged as a critical layer in the genetic regulatory code and the total number of lncRNAs is surging, catalysed by deeper and more sensitive RNA sequencing technologies combined with improved epigenomic technologies and computational prediction models (Guttman *et al.*, 2009). To date, the ENCODE (the Encyclopedia Of DNA Elements, in humans) project (Derrien *et al.*, 2012) has produced a catalog of 15,779 lncRNA genes that give rise to 28,468 distinct transcripts (GENCODE v28) (Derrien *et al.*, 2012). These can be found overlapping protein-coding genes in both antisense and sense orientations or in gene deserts.

lncRNAs are found in every branch of life and they outnumber protein-coding genes (Djebali *et al.*, 2012). However, one of the most alluring aspects of lncRNAs is that organism complexity is better correlated with the variety of non-coding RNA than with that of protein-coding genes (Mattick, Taft and Faulkner, 2010). Furthermore, 40% of lncRNAs are expressed specifically in the brain (Derrien *et al.*, 2012). This number sparks even more interest in this particular biotype, as protein-coding genes and miRNAs do not show this level of specificity for the brain (Kozomara and Griffiths-Jones, 2014). This collection of data begs the question, what is the functional role (if any) of these lncRNA transcripts in the brain and in particular in the striatum? Can we exploit this biotype to better describe striatal complexity and, subsequently, to promote striatal specification from stem cells?
Here I will briefly summarize the main attributes and roles of lncRNAs before describing how we can study them from a transcriptional point of view.

1.3.2 Defining lncRNAs

lncRNAs are broadly defined as noncoding RNA molecules that have a primary sequence that is longer than 200 nucleotides and they can space from small single-exon loci to large multi-exonic transcripts with several alternative splice forms (Figure 1.4). Although the 200 nucleotides threshold is an arbitrary classification, this size cut-off clearly discriminates lncRNAs from small regulatory RNAs such as miRNAs or piRNAs. Most of them share a set of common attributes with messenger RNAs (mRNAs) including a 5′ 7-methylguanosine cap, a 3′ poly(A) tail and the fact that they are transcribed by RNA polymerase II (Figure 1.4). In contrast, lncRNAs tend to be shorter than mRNAs, have fewer but longer exons and are generally enriched in the nucleus (Derrien *et al.*, 2012). However, the main features that set apart lncRNAs from mRNAs is their

lack in coding capacity (Figure 1.4), their low levels of expression and their poor primary sequence conservation (Cabili *et al.*, 2011; Derrien *et al.*, 2012).

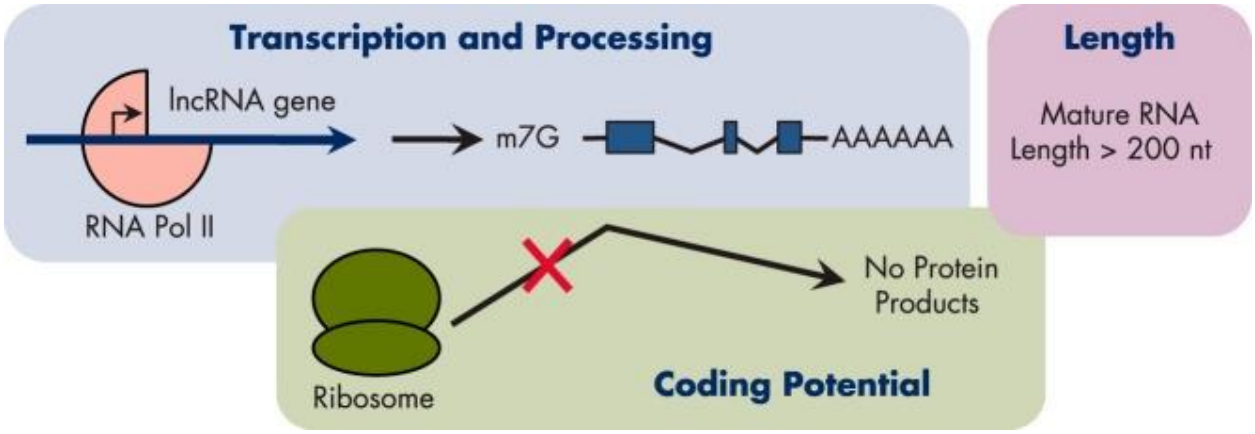


Figure 14 Molecular features of lncRNAs (Sun and Kraus, 2015).

Although lncRNAs exhibit poor primary sequence conservation, lncRNAs have been shown to have promoters that are as conserved as those of mRNAs (Fantom Consortium, 2006; Guttman *et al.*, 2009; Derrien *et al.*, 2012). Furthermore, a subset of lncRNAs have conserved regulatory elements, including repeating RNA domains that act as nuclear localization signals (Hacisuleyman *et al.*, 2016), transcription factor binding sites (Necsulea *et al.*, 2014; Melé *et al.*, 2017) and splicing motifs (Ponjavic, Ponting and Lunter, 2007; Haerty and Ponting, 2015; Melé *et al.*, 2017). These conserved elements and their diverse abundance (Cabili *et al.*, 2011; Derrien *et al.*, 2012), stability (Clark *et al.*, 2012), biogenesis (Ayupe *et al.*, 2015) and evolution (Necsulea *et al.*, 2014; Washietl, Kellis and Garber, 2014; Hezroni *et al.*, 2015; J. Chen *et al.*, 2016) suggests that this biotype is not the result of transcriptional noise resulting from low RNA polymerase fidelity (Struhl, 2007) but is instead, explicitly regulated to exert its diverse set of function. lncRNAs are highly diversified with regards to their mechanistic role. Indeed, whereas some lncRNAs act as functional RNA molecules, other seem to act as local regulators and are independent of their transcripts (Martens, Laprade and Winston, 2004; Ebisuya *et al.*, 2008; Latos *et al.*, 2012; Engreitz *et al.*, 2016; Melé and Rinn, 2016). Functionally, lncRNAs can exert their function on DNA sequences by either coordinating gene expression and chromatin structure in loco, therefore in *cis*, or they can leave the site of transcription and perform cellular functions in a different region from where they are transcribed and therefore act in *trans*. This biotype is involved in diverse biological phenomena such as allosterically regulating enzymatic activity, imprinting genomic loci, shaping chromosome conformation (Ponting, Oliver and Reik, 2009; Rinn and Chang, 2012) and changing cell states during development and differentiation (Batista and Chang, 2013; Flynn and Chang,

2014). One of the most fascinating aspects of lncRNAs is that they exhibit more tissue specific expression profiles than mRNAs (Cabili *et al.*, 2011; Derrien *et al.*, 2012). This tight regulation in expression has suggested that lncRNAs may play a pivotal role in determining cell state throughout differentiation (Guttman *et al.*, 2009; Cabili *et al.*, 2011)

1.3.3 lncRNAs in nervous system development

Throughout the progression from the pluripotent phase to the terminally differentiated specialized neural identity, cell specific gene regulatory networks are activated that guide the cells through neural differentiation. lncRNAs appear to have a pivotal role during this route as their expression is vigorously controlled during development (Mercer *et al.*, 2010; Belgard *et al.*, 2011; Lin *et al.*, 2011; Aprea *et al.*, 2013). In line, a diverse set of lncRNAs have been functionally validated during mouse brain development. Two examples that have been characterized *in vivo* include *Pnky* (Ramos *et al.*, 2015) and *Evf2* (Bond *et al.*, 2009). The lncRNA *Pnky* is a neural-specific lncRNA that controls neurogenesis as it serves as a regulator of NSC turnover in the embryonic and postnatal brain. *Pnky* knockdown leads to loss of the NSC phenotype in the ventricular zone and potentiates neuronal lineage commitment. *Pnky* interacts with the splicing regulator PTBP1 that together, regulate the expression and alternative splicing of a core set of transcripts that allow the cell to maintain the NSC phenotype (Ramos *et al.*, 2015). Instead, *Evf2* is a critical lncRNA for proper formation of GABA-dependent neuronal circuitry in adult brain as it maintains the excitatory to inhibitory neuron balance in the postnatal hippocampus and dentate gyrus. *Evf2* is able to bind the transcription factor DLX and methyl-CpG-binding protein MECP2 and guide these proteins to regulatory regions controlling the expression of interneuron lineage genes, including as *Dlx5*, *Dlx6*, and *Gad1*. Genetic deletion of *Evf2* in mice causes failure of GABAergic interneuron specification and therefore reduced synaptic inhibition (Bond *et al.*, 2009).

The roles of lncRNAs during neural differentiation have also been systematically investigated *in vitro*. In line, loss-of-function studies have identified dozens of lncRNAs that are decisive to establish pluripotency or direct neural lineage entry (Guttman *et al.*, 2011; Ng, Johnson and Stanton, 2012). One example of the action of lncRNAs in fine-tuning neural commitment from pluripotency is the lncRNA *Tuna*. *Tuna* modulates neuronal fate by forming a complex with three RNA-binding proteins, NCL, PTBP1, and hnRNP-K, that then bind to neural gene promoters like in differentiating mouse ESCs. The function of this lncRNA is conserved across relatively distantly related vertebrates as knockdown of *Tuna* is sufficient to inhibit neural differentiation in both mESCs and zebrafish (Lin *et al.*, 2014).

Also the lncRNA *Rmst* executes an integral role in establishing the neural state. This lncRNA is induced by REST during neural differentiation. *Rmst* then interacts with SOX2 and guides this transcription factor to a set of core neurogenesis-promoting genes, such as *Dlx1*, *Ascl1*, *Hey2*, and *Sps* (Ng *et al.*, 2013). Loss of *Rmst* blocks cell in the pluripotent state and inhibits the start of neural differentiation (Ng *et al.*, 2013).

However, the function of lncRNAs during nervous system development is not limited to cell fate commitment. They also play a key part during neurite elaboration and synaptogenesis, making them fundamental players during the convoluted process of neurite elaboration throughout brain development. Examples include: *Malat1* that calibrates synaptic density (Bernard *et al.*, 2010), *Bdnf-AS* that controls neurite elaboration (Modarresi *et al.*, 2012) and *BC1/BC200* that regulates synaptic turnover (Skryabin *et al.*, 2003; Lewejohann *et al.*, 2004; Zhong *et al.*, 2009).

1.3.4 Concluding Remarks

Taken all these evidence into consideration one can easily appreciate the importance of this biotype in defining cell states. For this reason, we opted to add a layer of complexity and investigate how these lncRNAs are distributed within the LGE and how well they define this area and specific cell types. A recent study further supports this idea as it has shown that lncRNAs appear to have a high expression in a few cell types and not a low average expression as described up till now (Liu *et al.*, 2016). Although different atlases of lncRNAs have been produced in the past years (Guttman *et al.*, 2010; Cabili *et al.*, 2011; Ranzani *et al.*, 2015; Liu *et al.*, 2016; Hon *et al.*, 2017) by *ab initio* transcriptome reconstruction, none have deeply catalogued lncRNAs expressed in ventral regions (LGE and MGE) of the developing human brain. In line, only one study evaluated lncRNAs in the human neocortex during fetal development (Liu *et al.*, 2016). Filling this gap and creating a more comprehensive and complete atlas of lncRNAs in these regions will greatly improve our knowledge of lncRNAs in the brain and will help us in defining specific cell states in the developing LGE.

1.4 RNA-seq for transcriptome exploration and reconstruction.

Here I will briefly describe RNA-seq and the specific experimental design required for lncRNA *de novo* discovery. I will then summarize scRNA-seq technologies and the different technologies available and how we decided to opt for the 10xTM (Zheng *et al.*, 2017) methodology. As mentioned, all computational aspects will be described in the Methods section under the "Rational" section of each computational tool.

1.4.1 Why RNA-sequencing

The two technologies used to measure the transcriptome are microarrays and RNA-sequencing (RNA-seq). Microarrays contain hundreds of thousands of short single stranded DNA molecules called probes, which are attached to specific locations on a glass or polymer slide. Samples of cDNA are formed from RNA and each molecule is labelled with a fluorescent dye and allowed to hybridize on the glass slide. Expression can then be estimated by the optical measurement of the amount of fluorescence coming from each probe on the slide (Allison *et al.*, 2006). Since the mid-1990s, DNA microarrays have been the technology of choice for large-scale studies of gene expression levels. However, array technology has several limitations. For example, background levels of hybridization limit the accuracy of expression measurements, particularly for transcripts present in low abundance. Therefore aspecific binding makes the comparison between different transcripts in the same microarray unreliable and the use of microarrays is usually limited to the detection of differential expression of the same probe target between samples (Marioni *et al.*, 2008). Furthermore, arrays are limited to interrogating transcripts that are already known that have relevant probes on the array. Instead, sequencing-based approaches have the potential to overcome these limitations. In an RNA-seq experiment it is possible to investigate not only gene expression, but also alternative splicing (Pan *et al.*, 2008), gene fusion events (Edgren *et al.*, 2011; Chen *et al.*, 2017) and novel transcript expression (Guttman *et al.*, 2010; Cabili *et al.*, 2011). Taking these aspects under consideration, for *de novo* lncRNA discovery, RNA-seq results the only possible choice to reach our objectives. In line, in the past years, high-throughput sequencing coupled with computational pipelines have driven enormous progress in the discovery of novel lncRNAs (Cabili *et al.*, 2011; Alvarez-Dominguez *et al.*, 2014, 2015; Liu *et al.*, 2016). Nonetheless, due to the high tissue-specific expression property of lncRNAs the current annotation is still incomplete and the human fetal brain still remains highly unexplored.

1.4.2 Optimal experimental design

Although the technology enables one to achieve a large amount of information, experimental and methodological biases are still frequently being reported. Therefore, several aspects must be kept under consideration. In fact, even though a sequencing run is capable of producing millions of reads, these represent only a fraction of the RNA actually present in the library. This happens because there is a certain amount of "space" that the actual transcripts in the sequencing library have to "share" in the sequencing instrument (McIntyre *et al.*, 2011). This means that highly expressed transcripts will often make up a large amount of the sequencing library, and in a shallow sequencing experiment less expressed genes may not be represented in the final data even though they were present. There is therefore a built in sampling variance and this must be considered when setting the optimal sequencing depth and during data analysis, especially for lncRNAs that have a low expression. A study (Tarazona *et al.*, 2011), showed that sequencing depth highly depends on the target biotype and its expression. In line, the coding transcriptome reaches saturation earlier than lncRNAs (Figure 1.5). However, increasing depth also causes a non-negligible percentage of off-target RNA species (snRNA, miRNA, snoRNA, misc_RNA, scRNA_pseudogenes) to be identified (Figure 1.5). Therefore, in our experimental design, that has the objective of identifying lncRNAs that were previously unannotated, 80 million reads would enable to reach a good level of representation of lncRNA transcripts. This is shown in Figure 1.5 (red box): increasing sequencing depth over 80 million reads does not allow the curve to reach plateau, however the detection of new genes remains stable at approximately one new gene per million reads up to a sequencing depth of 200 million reads, however this depth reduces detection of transcriptional noise and off-target transcripts (Figure 1.5).

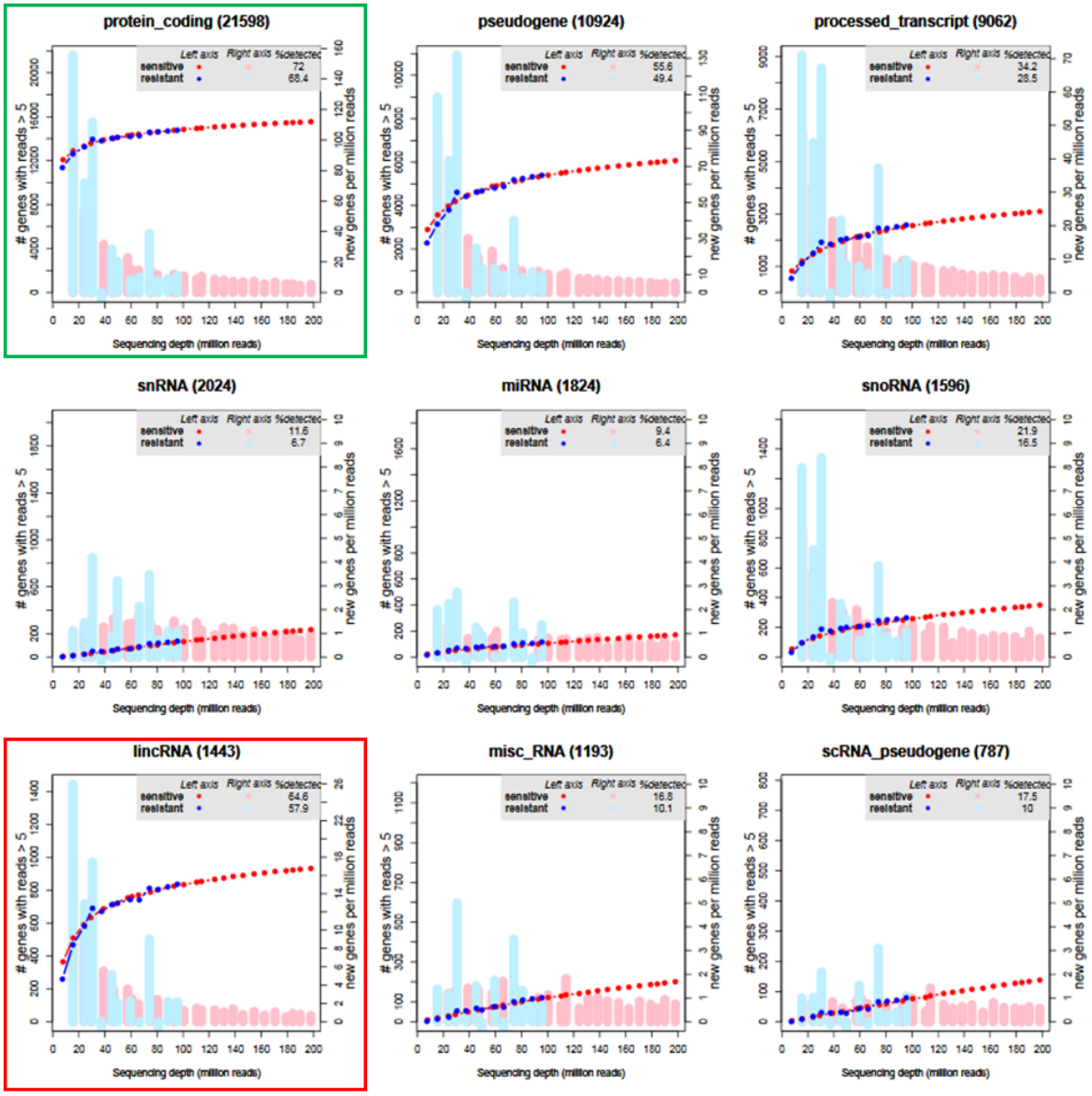


Figure 1.5 Saturation curves per biotype on a set of published data on the transcriptome of two human colorectal cancer cell lines only differing in the fluorouracil resistance phenotype (Griffith *et al.*, 2010). The lefty-axis displays the number of genes detected by more than 5 uniquely mapped reads as a function of the sequencing depth for each experimental condition (red=sensitive; blue=resistant). Vertical bars represent the number of newly detected genes per million additional reads (NDR, righty-axis) for each experimental condition (Tarazona *et al.*, 2011).

Another fundamental aspect that must be considered during experimental design is what is called biological variation that refers to the variability within a biological group. Biological systems are inherently complex and very sensitive to perturbations. Thus, even in the absence of sampling and technical variance biological variance will always exist. Specifically, this type of variance can be described as the natural variance that is present within a sample group (Marioni *et al.*, 2008).

Therefore, a change observed in gene expression between two groups can only be called significant if the difference between the groups is large compared to the variability within the group (Bullard *et al.*, 2010). When biological variation is large relative to technical variation statistical power is gained due to additional biological replicates. In conclusion when analyzing datasets derived from RNA-seq one must always evaluate the biological variability within a group to understand how many samples need to be treated and at the same to time one must insure enough depth to capture the complexity of the transcriptome.

1.4.3 Concluding Remarks

The main aim of utilizing RNA-Seq is *de novo* lncRNA identification and then the characterization of a specific signature for the LGE. In line, we opted to sequence 80 million reads per sample to have an optimal representation of the lncRNA population and have at least 3 samples for each area and pcw.

Although bulk RNA-seq has revolutionized the way we can interrogate the transcriptome and is fundamental to identify *de novo* expressed lncRNAs it can mask fundamental cellular heterogeneity (Wills *et al.*, 2013; de Vargas Roditi and Claassen, 2015) and lead to averaging artifacts (Trapnell *et al.*, 2014). In line, in the context of a specific region of the human brain (although in all tissues to some degree) an immense number of cells can be classified. Therefore, bulk RNA-seq limits our knowledge on the different cell classes present in a specific area and probably restricts our insight to the most abundant cell types that are able to contribute more, from a numerical point of view, to bulk measurements. In the past few years, scRNA-seq has been developed to accurately measure intra-population heterogeneity and capture cell states and transitions at very high resolution. As a consequence, this technology would enable us to expose population structure and cell dynamics that are masked at the group level. In the next paragraph I will summarize the advantages of the technologies and the different tools available to study the transcriptome of single cells.

1.5 scRNA-sequencing to understand cell population dynamics

1.5.1 Why scRNA-sequencing

Single cell offers the scientific community an enormous set of advantages. The first, as discussed, is to partition cells by transcriptomic similarity and detect the underlying population structure in an unsupervised manner, without averaging gene expression through all cells (Figure 1.6). In line, bulk measurements are not able to discriminate changes due to gene regulation from those that arise due to shifts in the ratio of different cell types in a mixed sample (Trapnell, 2015).

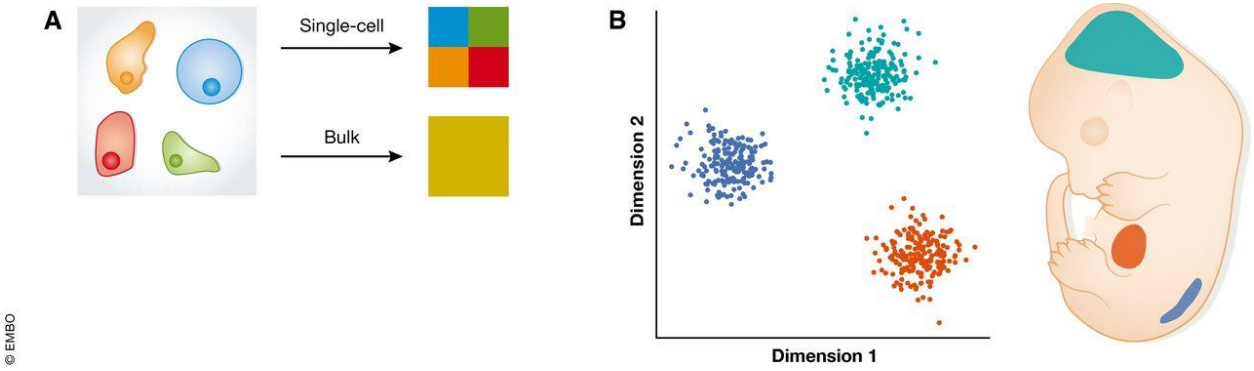


Figure 1.6 scRNA-seq enables the identification of cellular heterogeneity. (**A**) Bulk gene expression RNA-seq provides an average read-out of transcription over many cells, while single-cell RNA-seq allows the assaying of gene expression in individual cells without averaging gene expression levels. (**B**) Single-cell approaches enable the discrimination of different cell types and their transcriptional blueprint. (Griffiths, Scialdone and Marioni, 2018)

Clustering cells to capture consistent cell classes and the "attractors" of these cell states is one of the main aims of this technology and this has enabled the identification of known cell types but also the discovery of novel cell subtypes and their relative transcriptomic signature (Sousa *et al.*, 2017) (Figure 1.7A).

In many systems, however, this type of "discrete" classification is reductive, for example, during cell differentiation. In line, these type of cells exhibit a continuous spectrum of states that enables a cell to differentiate and bifurcate through different phases. scRNA-seq is able to capture these transitions and map cell trajectories through processes like differentiation (Figure 1.7B). Once cells have been ordered along a trajectory, gene signatures that lead to bifurcation events and therefore lineage commitment, can be pinpointed and master regulators within this network can be established (Trapnell, 2015). This cannot be achieved with bulk measurements as a single time point may contain cells from different stages in the process, and averaging expression profiles causes the loss of this signal.

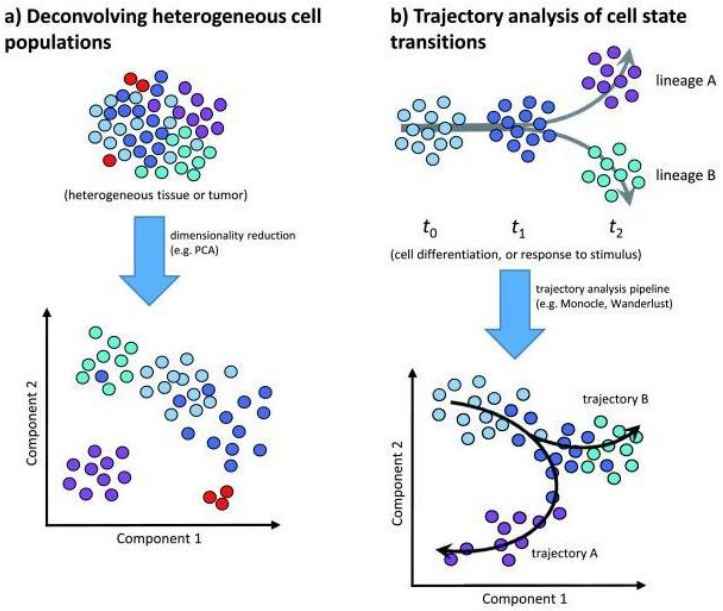


Figure 1.7 Common applications of single-cell RNA sequencing (**a**) Decoding heterogeneous cell populations. Clustering by single-cell transcriptomic profiles can reveal population substructure (**b**) Single-cell RNA sequencing time-series data can be used to untangle cell developmental trajectories (Liu and Trapnell, 2016).

The contribution of single-cell expression data is not limited to identifying unique cell types and developmental trajectories. This type of data can also be exploited to infer gene regulatory networks (Bansal *et al.*, 2007; Padovan-Merhar and Raj, 2013) and to assign specific modules to different cell states (Xue *et al.*, 2013; Patel *et al.*, 2014; Min *et al.*, 2015; Sousa *et al.*, 2017). Furthermore, this data can be employed to determine biases of expression of different alleles (G. Chen *et al.*, 2016; Reinius *et al.*, 2016), lineage tracing (Frieda *et al.*, 2017) and for spatial transcriptomics (Achim *et al.*, 2015; Satija *et al.*, 2015)

In the brain, scRNA-seq has been used to study cellular heterogeneity of different brain regions (Pollen *et al.*, 2014; Achim *et al.*, 2015; Llorens-Bobadilla *et al.*, 2015; Luo *et al.*, 2015; Usoskin *et al.*, 2015; Zeisel *et al.*, 2015; La Manno *et al.*, 2016; Nowakowski *et al.*, 2017; Sousa *et al.*, 2017) and this has led to the identification of new subpopulations of cells specific for the human lineage (Sousa *et al.*, 2017), molecular programs controlling human midbrain development (La Manno *et al.*, 2016) together with insight into the topographical, typological, and temporal hierarchies governing cell-type diversity in the developing human telencephalon (Nowakowski *et al.*, 2017). The striatum has only been looked in two mouse adult studies (Gokce *et al.*, 2016; Zeisel *et al.*, 2018), however a deep investigation of how this area arises and differentiates in human is still to be performed and is, as mentioned, the goal of this PhD project.

In conclusion, scRNA-seq has transformed the way we can interrogate heterogeneous tissues making it possible to answer previously unanswerable questions. However, although a high potential is enclosed in this technology, hazards and difficulties are still present and great care must be taken during the experimental design and the data analysis to avoid confounding factors and technical artefacts to mask true biological insight.

1.5.2 A brief history of single-cell RNA-seq technologies

The first pioneer studies that started to evaluate the whole transcriptome of single cells used microarrays (Tietjen *et al.*, 2003; Kurimoto *et al.*, 2006, 2007; Esumi *et al.*, 2008). However, as mentioned for the bulk measurements of tissues this technology suffers from limited sensitivity, reproducibility and resolution. The first that allowed an unbiased transcriptome-wide investigation of the mRNA in a single cell was conducted by Tang et al. where more than 75% of genes were identified compared with microarrays and also previously unknown splice junctions were pinpointed in a single mouse blastomere (Tang *et al.*, 2009). Then Guo et al. advanced the single cell transcriptomic field by manually collecting single cells from embryos at different developmental stages demonstrating that distinct cell types could be identified without pre-sorting (Guo *et al.*, 2010). The Linnarsson group (Islam *et al.*, 2011) extended this idea and developed single-cell tagged reverse transcription (STRT-seq), a highly multiplexed method that used cell capture plates preloaded with lysis buffer to allow capture and lysis of single cells, then cDNA was generated and sequenced by RNA-seq that allowed to analyze 96 single cells.

Since then, there has been an explosion in the development of protocols for RNA-sequencing of individual cells (Figure 1.8). However, even if great improvements have been made in this field, all protocols suffer a diverse set of challenges. For example, the Fluidigm microfluidic-based approach (https://www.fluidigm.com/about/aboutfluidigm, Pollen et al., 2014) (Figure 1.8) allows the sequential delivery of very small and precise volumes into tiny reaction chambers allowing to capture up to 96 cells in the chambers that are then subjected to the SMART-seq2 protocol (Picelli *et al.*, 2013) to create cDNA libraries. However, this technique suffers from limited throughput, low efficiency and in many cases only a fraction of cells are captured and cell specific biases are present in the capture process.

Another approach recently developed uses plate-based methods like MARS-Seq (Figure 1.8), where single cells are isolated into individual wells using fluorescence-activated cell sorting (FACS) (Jaitin et al., 2014; Soumillon et al., 2014). However, the process of sorting cells is highly time consuming and ultimately may lead to suffering of cells during the process.

In 2015, two revolutionary droplet-based techniques that enable high throughput were published by McCarroll's group (Drop-Seq) (Macosko et al., 2015) and Kirschner's group (inDrop) (Klein et al., 2015) (Figure 1.8). The concept behind these two methods is to encapsulate each cell inside a nanoliter droplet together with a bead. The bead is loaded with the enzymes required to construct the library that will contain a unique barcode which is attached to all of the reads originating from that cell. Oil droplet are created by the combination of two flows, one containing the lysis buffer, the reverse transcriptase and the beads with poly(T) RT primers, the other containing cells. This flow is separated into droplets by the addition of oil at fixed intervals. By adjusting the rate of the two flows and regulating the creation of droplets, one can ensure that most of the times only single cells will be isolated in droplets. After this all the droplets are pooled and sequenced in parallel. However, due to the random nature of the creation of droplets, a large number of cells are required making these techniques improper for samples with scarce availability of cells. Furthermore, both technologies require the generation of custom microfluidic devices and reagents. In line, most groups have opted to use another recent commercially available tool called the 10x Chromium (10xTM GemCodeTM Technolog) (Figure 1.8). This technology consists in a droplet-based system that enables 3′ messenger RNA (mRNA) digital counting of thousands of single cells (Zheng et al., 2017).

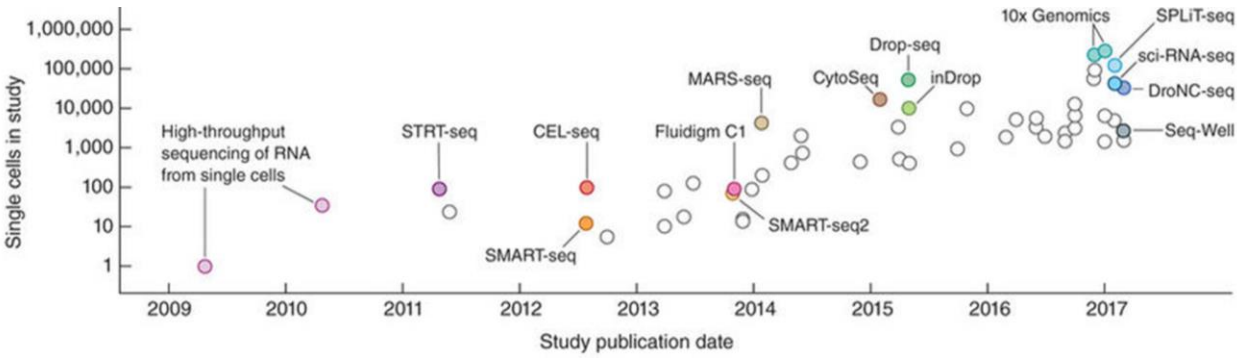


Figure 1.8 Single-cell RNA-seq technologies. Number of single cells reported in representative publications by publication date together with the technology used. (Svensson, Vento-Tormo and Teichmann, 2018)

1.5.3 Comparing platforms for scRNA-seq and understanding the limits of the technology

Deciding on a platform ultimately depends on the question addressed and determining which protocol best suits the needs of the project is a challenging task as they all present limitations in one direction or the other. Svensson et al. (Svensson *et al.*, 2017) evaluated the performance of a large number of published scRNA-seq protocols based on their capacity to quantify the expression of spike-ins of known concentrations. The quality of a protocol was based on the sensitivity and accuracy of each method. In particular, sensitivity was calculated as the minimum number of input

RNA molecules required for a spike-in to be detected as expressed. Therefore, high sensitivity allows the capture of lowly expressed genes. Instead, accuracy measures how close to reality the abundance of a molecule is compared to molecules of known abundance. In line, high accuracy indicates that the variation observed between molecules reflects a biological reality and not technical noise. What was highlighted during this work (Svensson *et al.*, 2017) showed that while these protocols fluctuate in their detection sensitivity, their accuracy in quantification of gene expression is mostly high. (Figure 1.9)

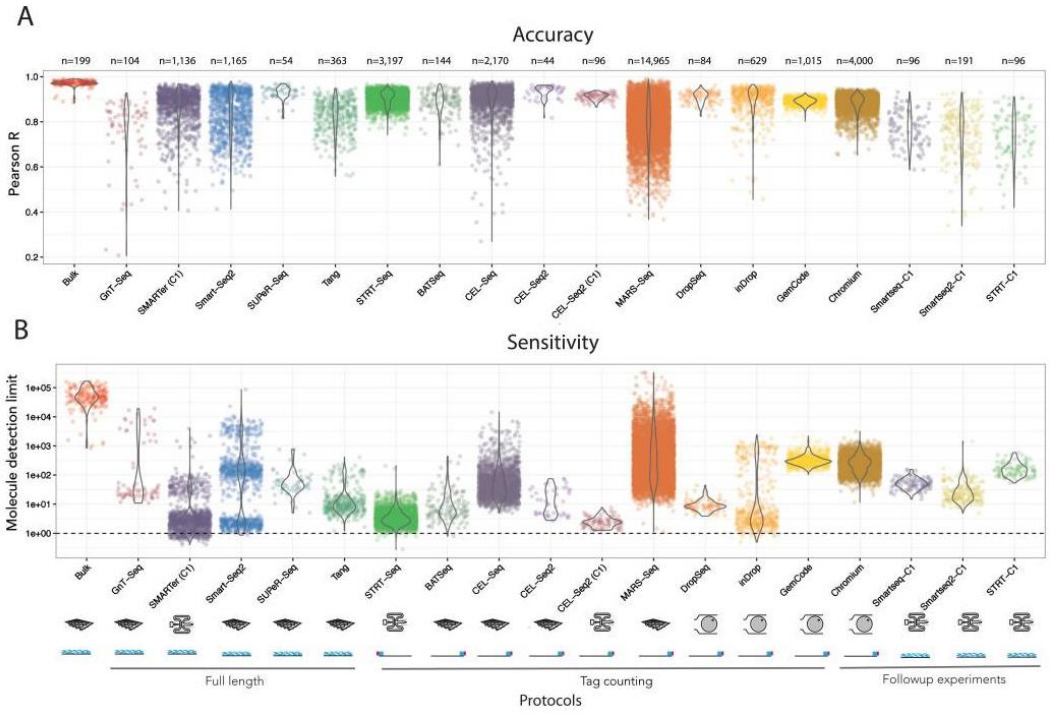


Figure 1.9 Power analysis of scRNA-seq protocols. (A) Graph illustrating the pearson correlation between estimated expression levels and actual input RNA molecule concentration. **(B)** Sensitivity graph showing the distributions of molecule detection limits of different technologies. The number of samples (n) is shown above each protocol (Svensson *et al.*, 2017).

During the experimental planning of our experiment the Fluigm microfluidic-based approach (https://www.fluidigm.com/about/aboutfluidigm, Pollen et al., 2014) and the 10X Chromium (Zheng et al., 2017) were taken into consideration as they were two well described and used protocols throughout the scientific community and each offered different advantages as highlighted in (Svensson *et al.*, 2017). The first major difference between these two techniques is how they generate cDNA. In particular, the Fluidigm platform uses the Smart-seq2 protocol (Picelli *et al.*, 2013) and is able to capture the full-length mRNA, while the 10x protocol is based on a 3′-tag sequencing method and is able to capture only the 3’ end of the transcript. This initial difference already poses one in front of a decision: if alternative isoforms need to be detected and measured the Fluidigm system that allows a full-length capture is needed. If simple quantification

is enough, then the 10x system can be used. Another key difference between the systems is that the 10x protocol enables to capture more cells (100-100 000 cells/run) than the Fluidigm system (96-384 cells/run) and thus is an optimal technique when the tissue under examination is highly heterogeneous and needs a large amount of cells to confidentially call cell subtypes (Baran-Gale, Chandra and Kirschner, 2017). A further important difference between the protocols is the inclusion of unique molecular identifiers (UMIs) in the 10x system that are used to correct for amplification bias (Islam *et al.*, 2014). These UMIs consist of a 10bp barcode that are included in each read. This allows to correct for amplification bias (Kivioja *et al.*, 2012) as each RNA molecule is tagged only once before amplification (Figure 1.10). Furthermore, as the fraction of UMI duplicates depends on the depth of sequencing, these can be used to estimate the levels of sequencing saturation. Adding UMIs to the Smart-Seq2 protocol is difficult, as each 'full length' transcript is fragmented following reverse transcription, and each fragment would need to be linked to the single UMI for that transcript (Baran-Gale, Chandra and Kirschner, 2017).

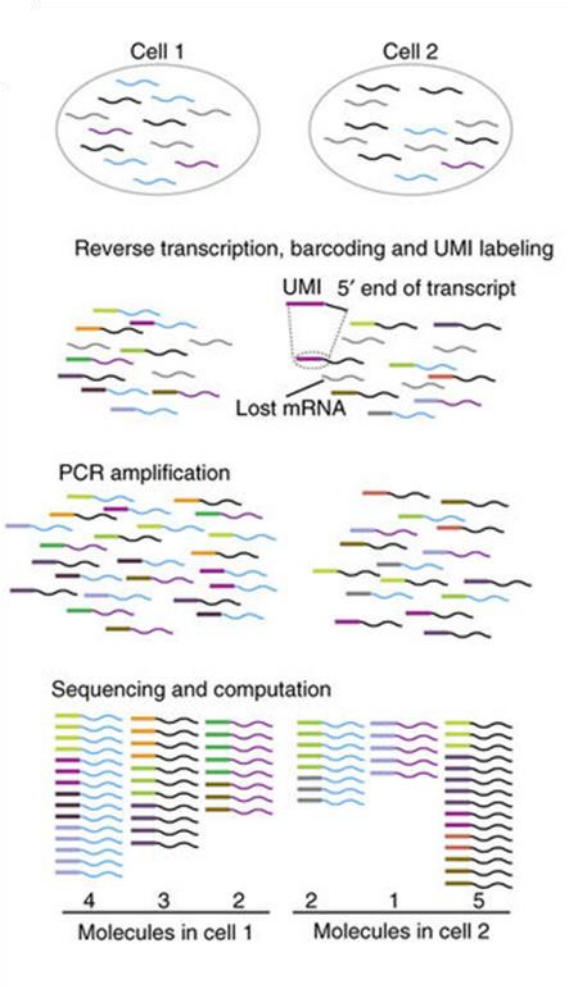


Figure 1.10 Molecule counting using UMIs. Schematic representation of tagging single mRNA molecules with UMIs. UMIs are represented by colored boxes (middle and bottom) (Islam *et al.*, 2014)

Finally, both Fluidigm and the 10x Chromium approaches allow the use of spike-in controls, however, these have been mainly employed with the Smart-Seq2 protocol. This entails adding a constant amount of spike-in RNA from the External RNA Controls Consortium (ERCC) (Baker *et al.*, 2005) to each cell's lysate prior to library preparation that is then quantified by measuring the number of reads mapped to the spike-in reference sequences (Brennecke *et al.*, 2013). These spike-in can be used for quality control, normalization and as a measurements of technical variability (Brennecke *et al.*, 2013; Stegle, Teichmann and Marioni, 2015). The first quality control that can be evaluated using spike-in is the ratio of the number of reads mapped to the endogenous RNA and the number of reads mapped to the extrinsic spike-ins. Low quality runs show a high proportion of reads mapped to the spike-ins and a high percentage of unmapped reads (Figure 1.11), caused by a low amount of RNA that could be attributed to cell lysis. These cells should be removed from downstream analysis. However, this parameter may also be confounding as this ratio can fluctuate from cell to cell for biologically important reasons for example if the cells are captured at different stages of the cell cycle.

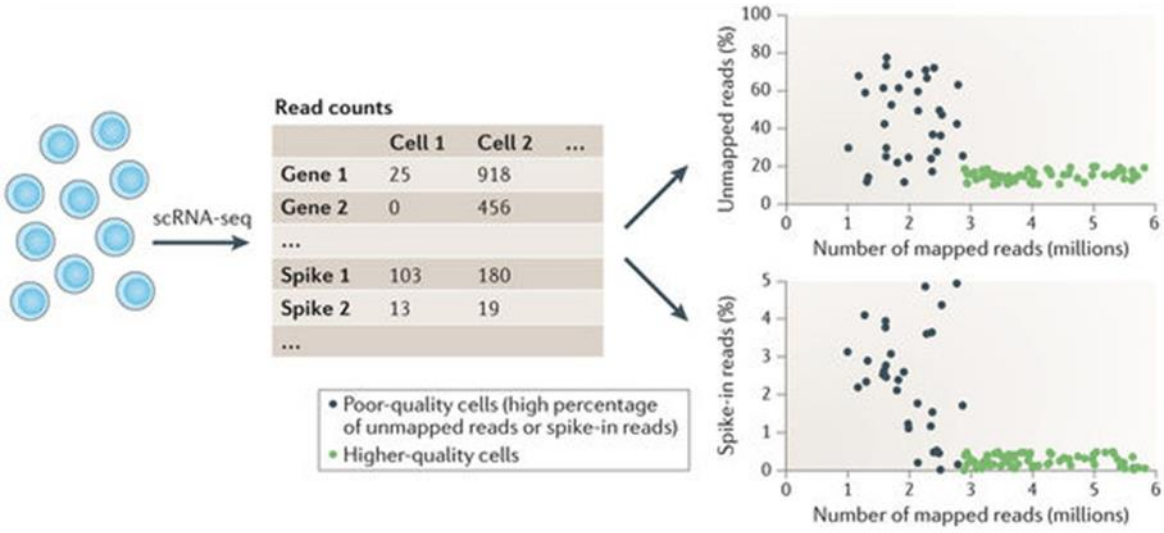


Figure 1.11 Quality control of scRNA-seq data. To determine whether cells have been captured efficiently and the mRNA fraction amplified faithfully two important criteria have to be evaluated. The first is to compare the percentage of unmapped reads (top panel) and the percentage of reads mapped to the external spike-in molecules (bottom panel). Cells that have high values for these two parameters (grey) should be discarded. Only cells having low values for each parameter are high quality (green) and should be kept for further analysis (Stegle, Teichmann and Marioni, 2015).

Spike-ins can also be used to measure reaction efficiency and therefore technical variability within an experiment. Reaction efficiency can be divided in: capture efficiency (percentage of mRNA molecules in the cell lysate that are captured and amplified), amplification bias (non-uniform amplification of transcripts), and sequencing efficiency (rate at which cDNAs in a library are

sequenced). These three steps are critical in defining the amount of noise in an experiment and have the greatest effect on lowly expressed genes (Brennecke *et al.*, 2013; Bhargava *et al.*, 2015; Kim *et al.*, 2015). Since the number of mRNA spike-in molecules is theoretically constant and known across cells and spike-ins are exposed to the same conditions as endogenous genes, one can use this information to measure systematic variability in the number of spike-in genes across cells and use it as a measure of changes in reaction efficiency (Brennecke *et al.*, 2013; Islam *et al.*, 2014). In particular, Brennecke et al. illustrated the presence of a strong non-linear relationship between gene expression and CV^2 (square of the coefficient of variation) for spiked-in genes (Figure 1.12). One can use this measure and set a threshold to distinguish true biological variability from the high levels of technical noise (Figure 1.12). Finally, the counts associated with each gene can be normalized into absolute numbers of mRNA molecules based on the levels of the spike-ins, which have been added at a known concentration. This can be done by calculating a cell-specific factor that adjusts for the differences between the observed and expected expression of spike-ins. Then by using this factor on endogenous genes, normalized expression estimates can be obtained (Brennecke et al., 2013).

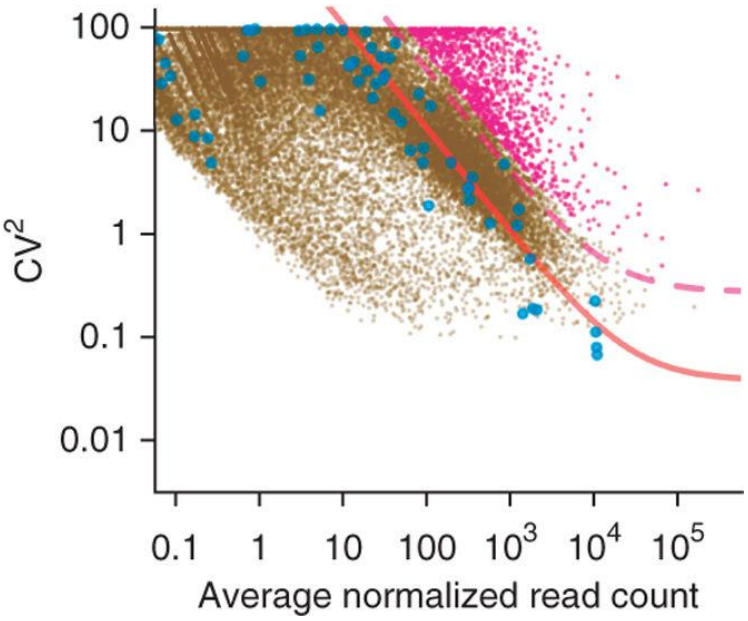


Figure 1.12 Statistical method to fit technical noise and infer highly variable genes. The graph shows the relationship between technical variability and expression strength. The squared coefficients of variation (CV^2) are plotted against the means of normalized read counts for each gene. The solid red line shows the technical noise fit and the dashed magenta line illustrates the expected position of genes with 50% biological CV^2. External RNA Controls Consortium (ERCC) are shown in blue dots; brown dots represents the values for each gene while the magenta dots are the highly variable genes (Brennecke *et al.*, 2013).

The practicality of spike-in controls however, remains highly questioned (Grün, Kester and Van Oudenaarden, 2014; Burns *et al.*, 2015; Stegle, Teichmann and Marioni, 2015). One of the main issues is that the External RNA Controls Consortium (ERCC) set is composed of nucleotides that are shorter than an average human mRNA, have a different GC content, have shorter poly(A) tails and lack 5′ caps, and therefore will be processed differently and with a different efficiency during reverse transcription compared to endogenous RNAs and therefore may depict a different picture than reality. Furthermore, it is extremely challenging to pinpoint the optimal spike in concentration per sample to normalize the data without causing the spike-ins to take up a relatively large proportion of reads (so that if a fixed amount of space for read sequencing is available for each cell, most of the reads would be derived from spike-ins and not from RNAs of each cells) overwhelming the biological signal with a high risking in masking true biological information with technical reads (Stegle, Teichmann and Marioni, 2015).

1.5.4 Concluding Remarks

Since the objective of this thesis is to catalogue all subtypes of cells in the developing human striatum and taking into consideration all the aspects discussed above, we opted to utilize the 10x Chromium technology (Zheng *et al.*, 2017). Although the main disadvantage of this technology is the low sequencing depth and only 3’ information of the transcript, this limit is overcome by the large number of cells sequenced in parallel that will probably capture most of the diversity of this area. Instead, we reasoned that the disadvantage of the Fluidigm system would have led to a few cells sequenced from each fetal sample and probably a bias during the cell capture process that would have caused an underrepresentation of the cell types present and a deficit in cell type classification. Furthermore, considering the disadvantages of spike-in described above we opted to avoid using them in our preparation and used only UMI for correct inference of transcript expression.

2 | Results and Discussion

2.1 Creating a lincRNA catalog of the human fetal telencephalon

To create an atlas of lincRNAs of the developing human telencephalon we set up a computational pipeline that can be summarized in Figure 2.1.

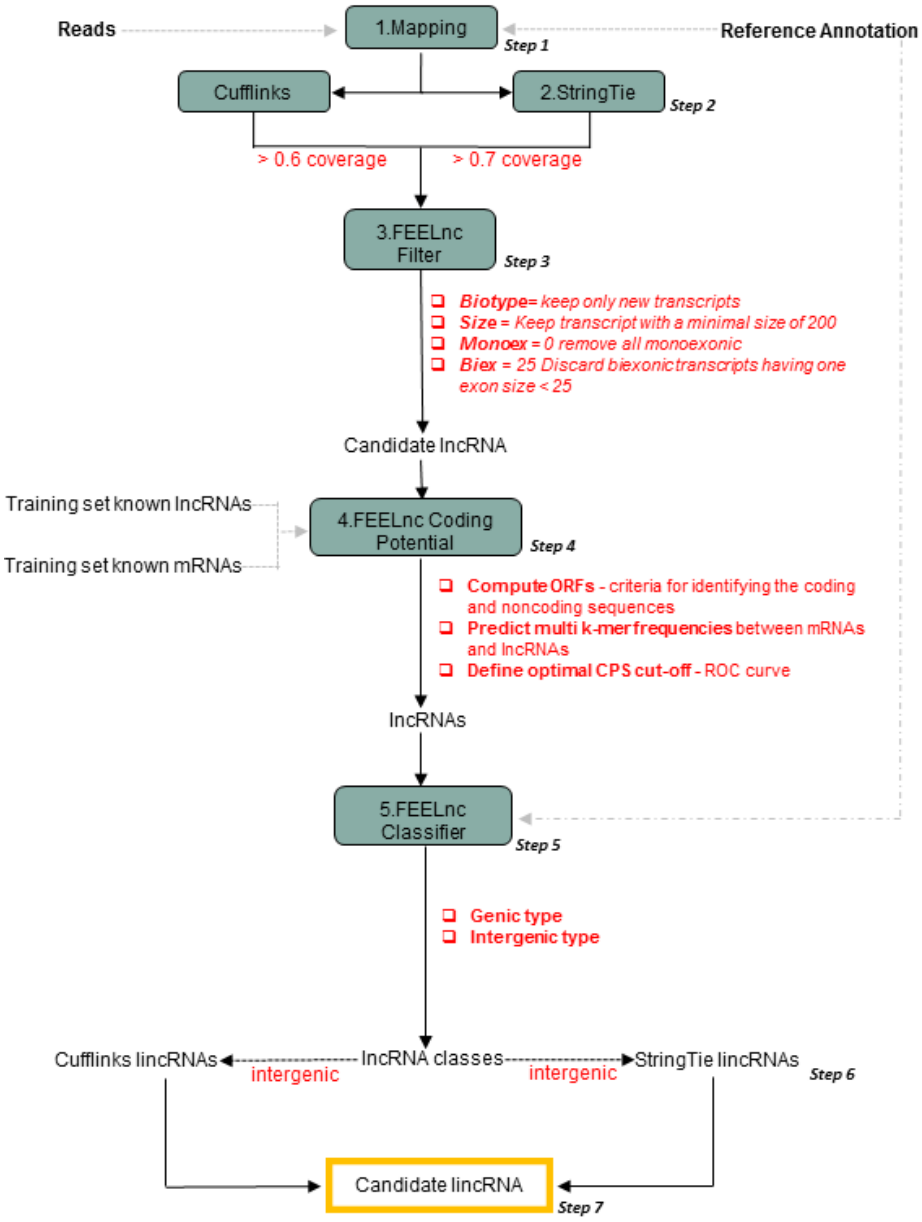


Figure 2.1 ***De novo*** **analysis pipeline**. All filtering steps are shown in red.

2.1.1 Integrating recently identified lincRNAs in the reference annoation

Before starting with the *de novo* analysis we integrated in the reference annotation (Genecode GRCh38) lincRNAs identified in a set of recently published papers (Cabili *et al.*, 2011; Liu *et al.*, 2016; Hon *et al.*, 2017) using cuffcompare. In particular we integrated:

- 3061 lincRNAs identified in a study of 24 human tissues (Cabili *et al.*, 2011).
- 6468 lincRNAs derived from human major cell types and tissues using data from FANTOM5 cap analysis of gene expression analysis (Hon *et al.*, 2017).
- 290 lincRNAs identified in the human neocortex from 13.5 to 23 post conceptional weeks (Liu *et al.*, 2016).

This enabled us to have an updated reference annotation that allowed us to evaluate lincRNA expression more thoroughly and to identify new lincRNAs of the neocortex, LGE and MGE.

2.1.2 Mapping and quality control

We then mapped (Figure 2.1, step 1) the reads derived from each sample (~80 million reads/sample) to the newly integrated reference genome using STAR (Trapnell *et al.*, 2011). We opted for 80 million reads per sample as this enables a good level of saturation for lncRNAs detection and simultaneously reduces transcriptional noise and sequencing of off-target transcripts that would be caused by sequencing deeper (see section 1.4.2). We performed quality control analysis using FastQC (Andrews, 2010) and then summarized the results by using MultiQC (Ewels *et al.*, 2016) (Methods 4.2.1). We observed an overall high quality of the samples with a high percentage (>80%) of uniquely mapped reads to the reference genome (Figure 2.2A). However, a large bias in GC content was detected in the samples at 7pcw (Figure 2.2B). This bias was confirmed after we checked with principal component analysis (PCA) how the different areas divided for each pcw (Figure 2.2C). In line, we observed a separation for GC content/batch at 7pcw (Figure 2.2C) but a good separation between the areas from 9, 11 (with one outlier that was removed from the analysis) and 20pcw (Figure 2.2E-G). This GC bias at 7pcw was corrected using EDASeq (Risso *et al.*, 2011) and then batch effects were removed using ComBat (Johnson, Li and Rabinovic, 2007) before differential expression analysis (Figure 2.2D). (Methods 4.2.6-4.2.7).

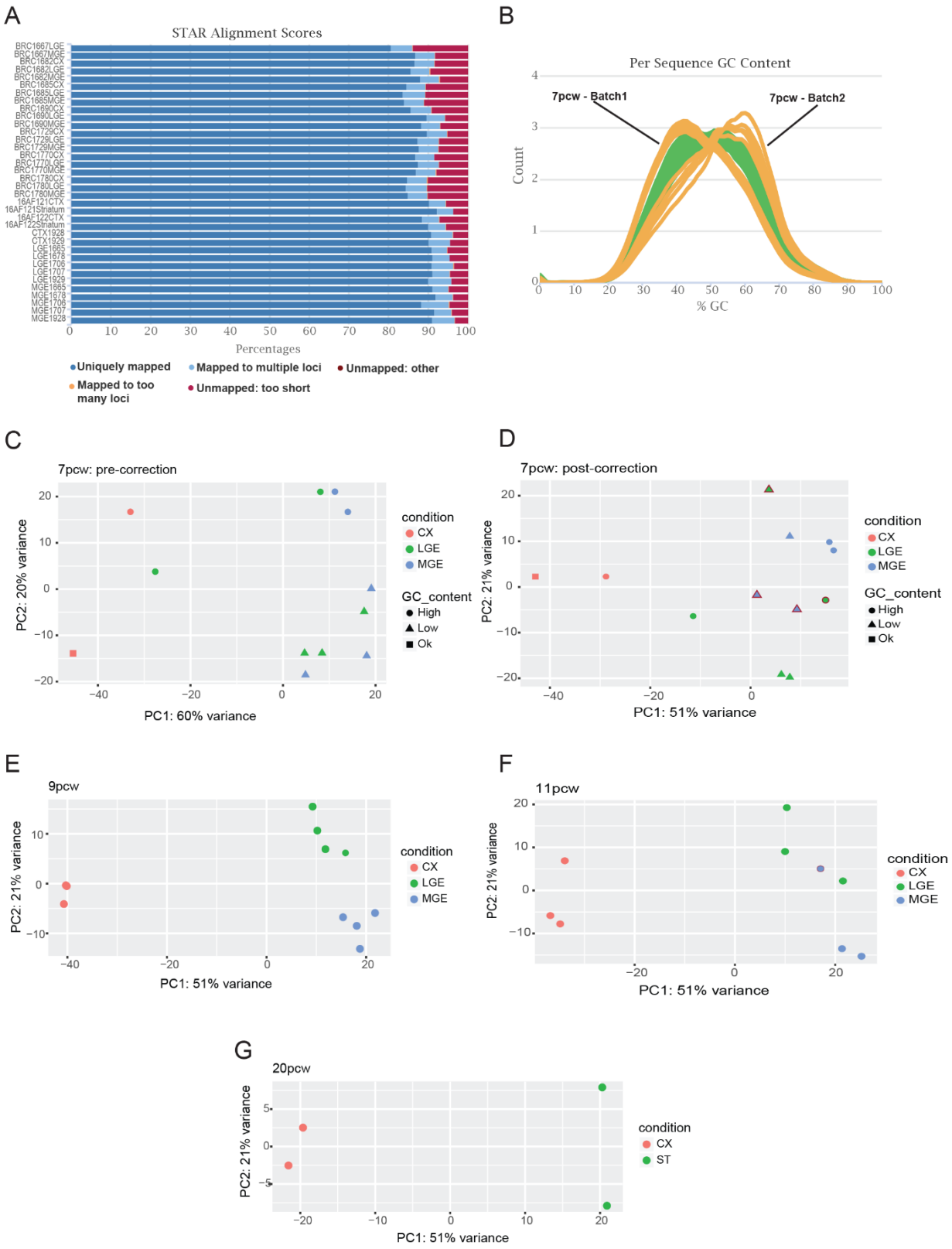


Figure 2.2 Overall QC of samples. (A) STAR alignment scores for all samples. Graph shows the percentage of uniquely mapped reads together with reads mapping to multiple loci or unmapped reads. **(B)** GC content distribution with all samples (green: quality pass; orange: bias in distribution) **(C)** PCA at 7pcw before correction **(D)** PCA at 7pcw after GC content and batch correction **(E-G)** PCA at 9pcw **(E)** at 11pcw **(F)** at 20pcw **(G)**. Red borders around samples in the PCA mark outliers that were removed from the analysis

2.1.2 Transcriptome Reconstruction

After mapping the reads we reconstructed the transcriptome (Figure 2.1, step 2) of each sample using two different genome-guided transcript assemblers: Cufflinks (Trapnell *et al.*, 2011) and StringTie (Pertea *et al.*, 2015) that both use spliced reads to reconstruct the transcriptome (Methods 4.2.2). To distinguish between authentic lowly expressed lincRNAs and technical noise we calculated an optimum read coverage threshold for both assemblers. For Cufflinks we determined this cutoff based on whether each assembler was able to reconstruct, with full read coverage, known protein coding genes and lincRNAs (Methods 4.2.2). We determined a threshold of 0.6 as the minimum read coverage needed to fully reconstruct known lincRNAs with a FPR < 0.05 and therefore filtered all transcripts under this level of expression (Figure 2.3A). For StringTie we did not have the information derived from full red coverage we therefore opted to use the median coverage of lincRNAs identified in the study performed in the neocortex (0.7) (Liu *et al.*, 2016) as we hypothesized that it would reflect a similar expression to lincRNAs in the tissues under study (Figure 2.3B).

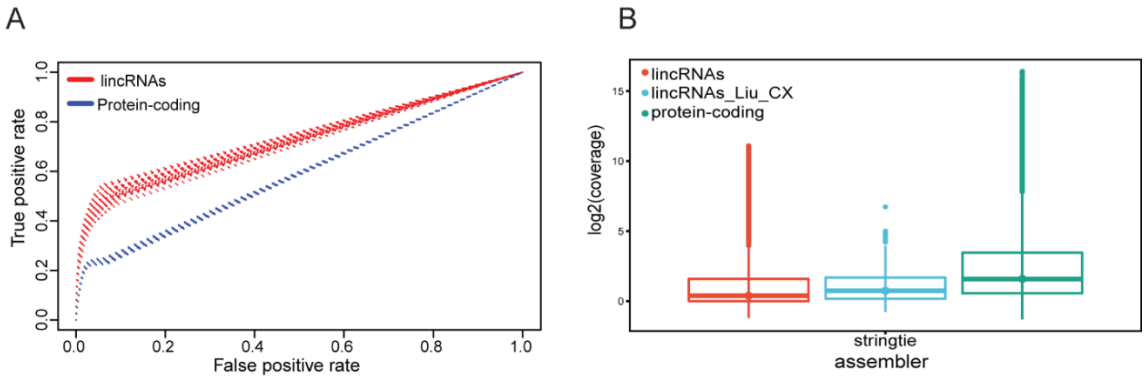


Figure 2.3. Threshold to discriminate transcripts associated with technical noise. (A) ROC analysis of coverage thresholds required to fully annotate known protein-coding and lincRNAs for the assembly created with Cufflinks. **(B)** Coverage distribution of known protein-coding genes, lincRNAs and lincRNAs identified in the human developing neocortex (Liu *et al.*, 2016).

2.1.3 Determination of lincRNA potential of novel transcripts

To identify lincRNAs from protein coding transcripts and unreliable single exon fragments after computational reconstruction of transcripts models we employed FEELnc (FlExible Extraction of LncRNAs) (Methods 4.2.3). For each reconstructed transcriptome (Cufflinks assembly and Stringtie assembly) we used the first module ($FEELnc_{filter}$; Figure 2.1, step 3) to flag and remove from further analysis all assembled transcripts that overlapped a known exon. We then filtered transcripts shorter that 200nt (as defined by Wang & Chang, 2011) and removed mono-exonic transcripts due to the fact that they may be mapping artifacts or transcriptional noise caused by

low polymerase fidelity (Struhl, 2007). We then determined the coding potential of each transcript by using the second module (FEELnc$_{codpot}$; Figure 2.1, step 4) that uses a Random Forest model trained on a set of known protein-coding transcripts and lincRNAs that allowed us to determine a coding potential score (CPS) cut-off equal to 0.438 at a performance of 0.88 (Figure 2.4A). This cutoff enabled us to maximize both sensitivity and specificity and classify transcripts with a value under this threshold as lncRNAs and transcripts above as potential protein-coding transcripts. We observed a higher rate of detection of both protein-coding and lncRNAs by StringTie compared to Cufflinks (Figure 2.4B) and this is expected as StringTie, is highly sensitive and reports as many isoforms as explained by the assembled reads (Pertea *et al.*, 2015), while Cufflinks reports the minimal number of compatible isoforms and therefore allows maximum precision (Trapnell *et al.*, 2011). Finally, the third module (FEELnc$_{classifier}$; Figure 2.1, step 5) was exploited to classify different lncRNAs subclasses based on their relationship with the closest annotated transcripts (Methods 4.2.4). Within our set of lncRNAs a higher number of genic transcripts were detected compared to intergenic transcripts (Figure 2.4C). To reduce potential of overlap with coding transcripts and the uncertainty of strand origin since this is an unstranded protocol (You, Yoon and Nam, 2017) we decided to keep only intergenic lncRNAs (lincRNAs) for both assemblies (Figure 2.1, step 6).

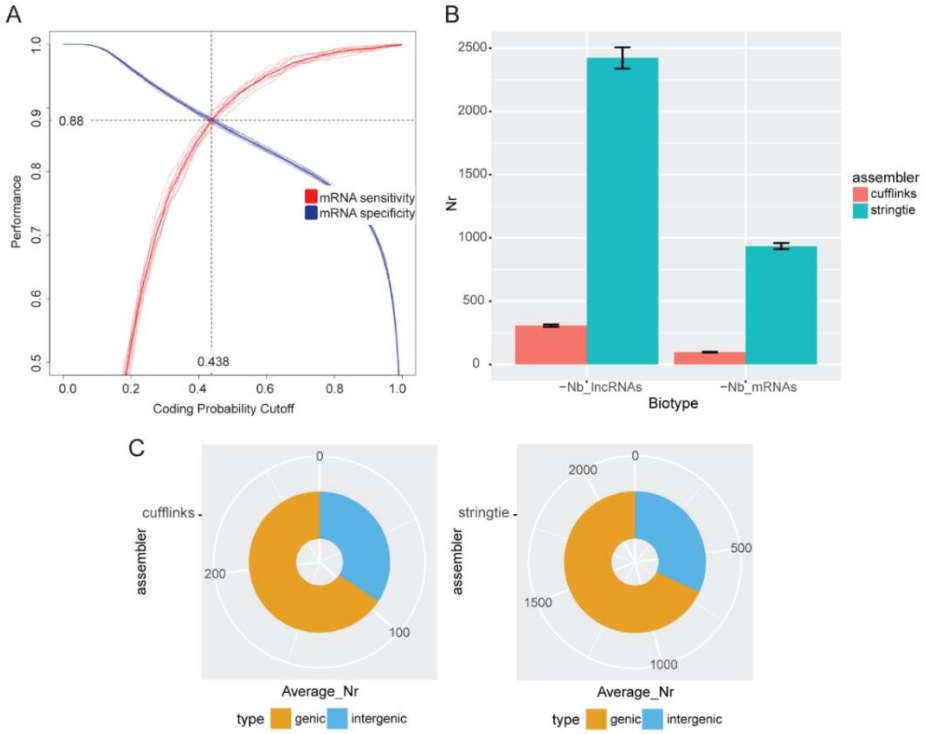


Figure 2.4 Catalogue of potential coding and non-coding genes. (A) Two graph ROC curves for optimized CPS threshold detection using known protein-coding and lncRNAs. **(B)** Number of lncRNAs and mRNAs identified by each assembler by the FEELnc coding potential module **(C)** Sub-classification of intergenic and genic lncRNA/transcripts by the FEELnc classifier module.

2.1.4 Atlas of lincRNAs of the human fetal telencephalon

We created a stringent set of genes by first creating an atlas for each area and pcw that consisted in lincRNAs identified in all biological replicates using Cuffcompare (Trapnell *et al.*, 2011) to create a reliable consensus for each area, pcw and assembler. We then compared each atlas to pinpoint lincRNAs that were identified by both assemblers (Figure 2.1, step 7). This enabled us to keep tract of which sample, area and pcw the transcript was identified and to create a highly reliable "consensus" between both assemblers. We observed, that although the number of lincRNAs that overlapped between the two assemblers was higher compared to that of protein-coding genes (Figure 2.5A), we had a lower percentage of overlap between non-coding transcripts compared to protein-coding transcripts (Figure 2.5B) and this can be attributed to how each assembler reconstructs low-abundance transcripts (Garber *et al.*, 2011). The stringent atlas includes 1116 lincRNA loci (1504 transcripts) of which most can be detected only at 1 pcw for each area (Figure 2.5C). The developing LGE shows the highest number of detected lincRNAs while the neocortex shows the lowest (Figure 2.5D). This result could have been caused by two different reasons. The first is that the neocortex has already been investigated for lincRNAs (Liu *et al.*, 2016) and therefore lincRNA detection is reaching saturation for this area. The second is linked to the distribution of lincRNAs in each cell type. Whether the low abundance of lincRNAs in bulk cell population depends on the fact that they have a low expression in each cell type (Cabili *et al.*, 2015) or that they have a high expression in a small subset of 'jackpot' cells (Liu *et al.*, 2016), the greater diversity of cell types in the neocortex compared to the striatum would result in a lower average expression for each lincRNA in the neocortex bulk RNA-seq data that would in turn cause each lincRNA to have a lower probability of being sequenced and detected compared to the LGE and MGE that are more homogenous. We then evaluated the average number of exons present in newly identified lincRNAs, to see if we were in line with other studies (Figure 2.5E). We found that lincRNAs had an average number of exons equal to 2.27. This pattern was similar for all atlases expect for the Fantom Atlas (Hon *et al.*, 2017) that had an average number of exons lower than 2 (1.77) that may be caused by analysis of gene expression (CAGE) data instead of RNA-seq data. In line, most studies on *de novo* discovery of lincRNAs show that most transcripts of these biotypes have 2 exons suggesting that we are in line with the structure of recently identified lincRNAs (Cabili *et al.*, 2011; Harrow *et al.*, 2012). However, this average number of exons may be an under-estimation since their lower abundance may result in incomplete assembly.

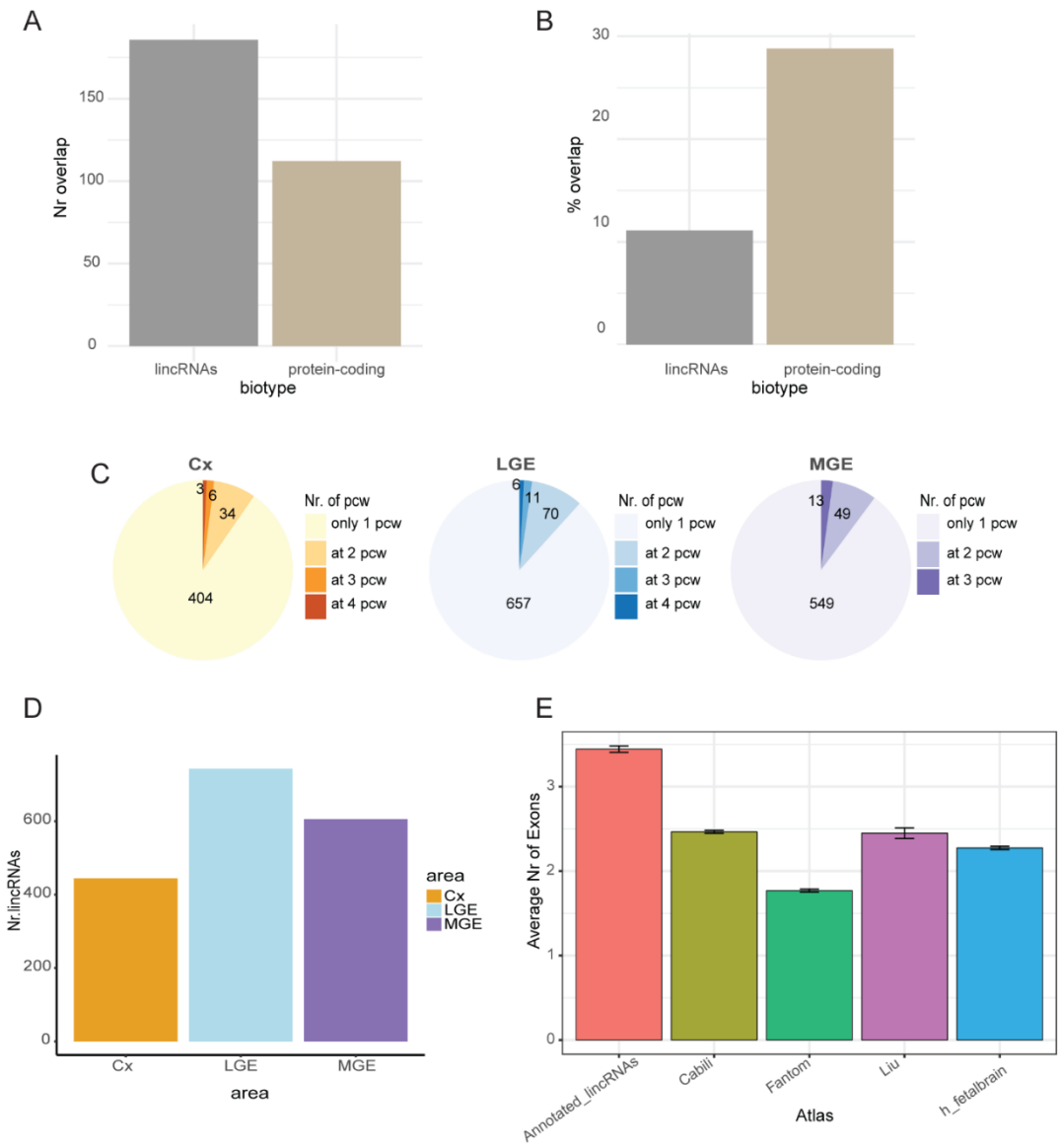


Figure 2.5 LincRNA catalog generation. (A,B) Converge between lincRNAs identified by Cufflinks and StringTie in terms of percentage **(A)** and number **(B).** Distribution of final set of newly identified lincRNAs between regions and pcw C) Final number of lncRNAs identified in each area **(D)** Average number of exons for different atlases of lincRNAs.

2.1.5 Specific attributes of lincRNAs.

Once the final stringent set of lincRNAs was identified, we integrated them into the modified reference annotation described above (see section 2.1.1 and Methods 4.2.5). We subsequently proceeded with re-mapping all reads to the new reference genome with STAR (Dobin *et al.*, 2013). We then evaluated specific characteristics of lincRNAs starting by the expression levels compared to protein-coding genes. As reported in previous manuscripts (Cabili *et al.*, 2011; Derrien *et al.*, 2012), the maximal expression levels of lincRNAs is lower than those of protein coding genes (Figure 2.6A). However, when comparing the different atlases of lincRNAs integrated in this

study, lincRNAs identified in Liu et al., (Liu *et al.*, 2016) in the neocortex and the ones identified in this study (h_fetalbrain) displayed the highest expression compared to the rest of the atlases (Figure 2.6B), probably suggesting that these lincRNAs are explicitly expressed in the tissues they were identified in. In line, when measuring biotype specificity using an entropy-based metric called Jenson Shannon divergence (Cabili *et al.*, 2011) (Methods 4.2.10) we observed an overall broader range of specificity scores (ranging from 0 to 1) compared to protein-coding genes that had a more tight distribution around the median (Figure 2.6C). In line, when applying a specific threshold based on known marker genes (Methods 4.2.10) for each area, we observed a higher percentage of specific lincRNAs (56%) compared to protein-coding genes that displayed only 14% of specific genes (Figure 2.6D).

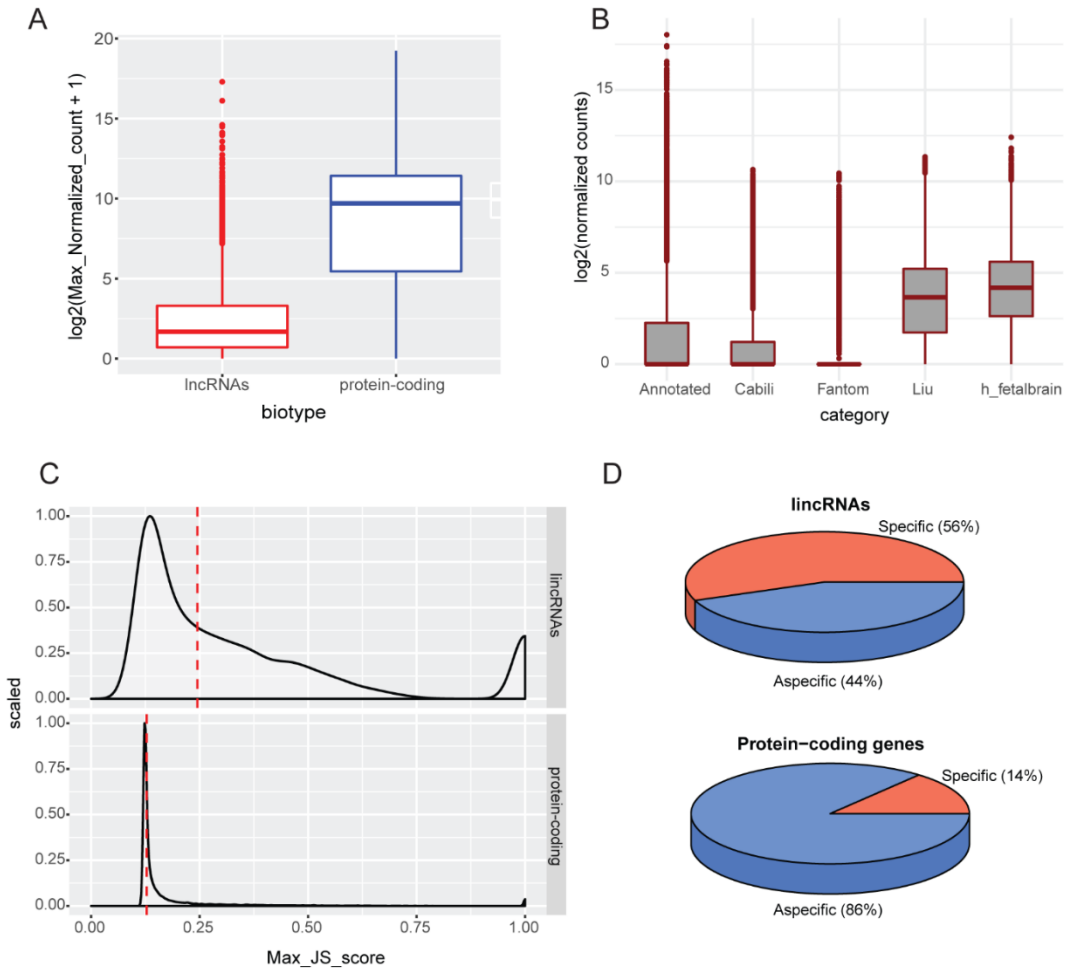


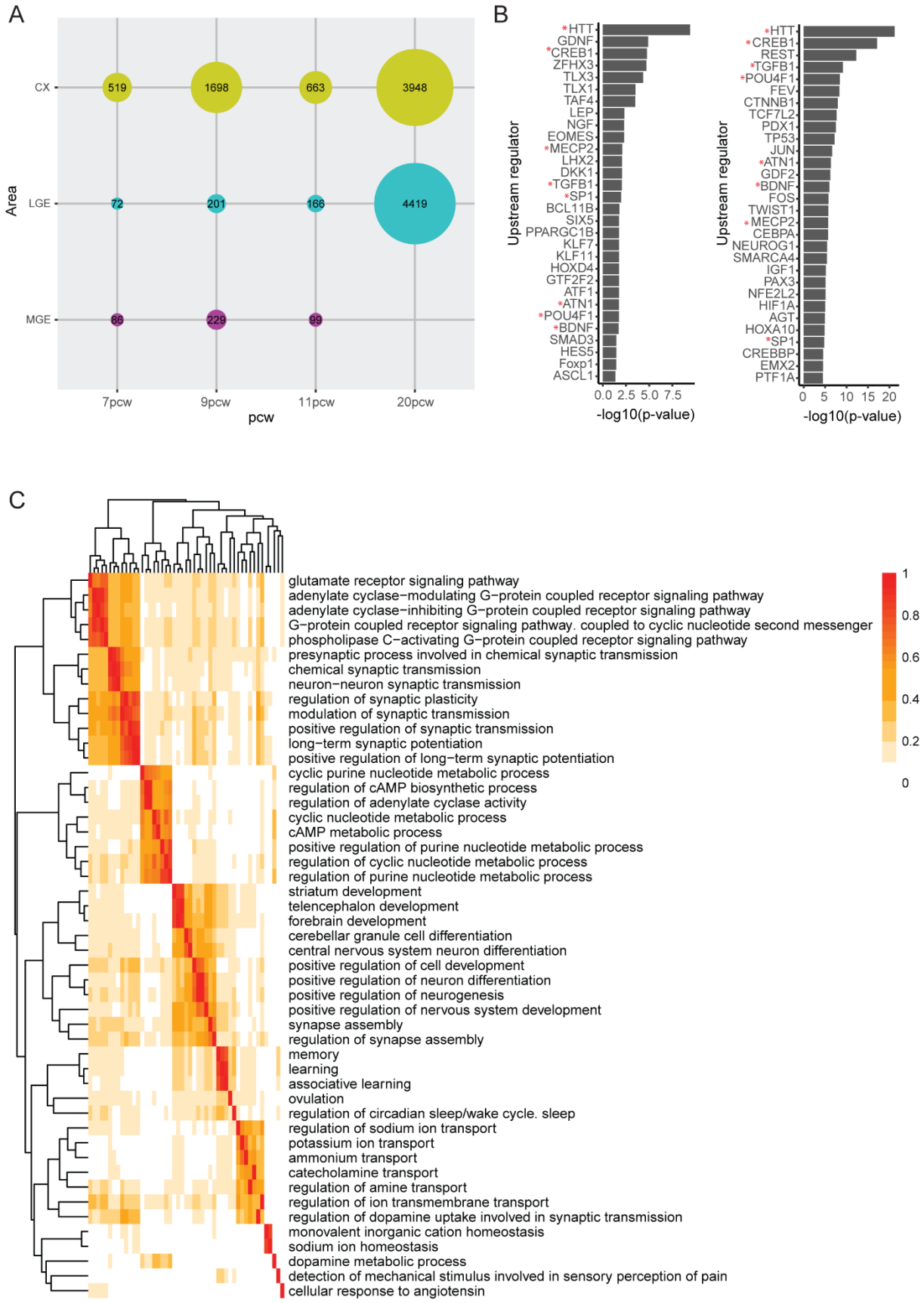
Figure 2.6 Tissue specificity of lincRNAs and coding genes. (A) Normalized expression of lincRNAs and protein-coding-genes **(B)** Normalized expression of different lincRNAs of different atlases. **(C)** Distributions of maximal tissue specificity scores calculated for each transcript across the different areas and pcw (red dashed line = median). **(D)** Percentage of specific lincRNAs and protein-coding genes.

2.1.6 Identifying a unique signature for the LGE

We then wanted to pinpoint a signature of both protein-coding genes and lncRNAs that could be uniquely attributed to the LGE/developing striatum. In particular, for each pcw, we identified a unique set of genes (p-adjusted value < 0.05) that defined each of the three areas (Figure 2.7A).
We then exploited Ingenuity Pathway Analysis (IPA, Ingenuity Systems) to identify upstream transcriptional regulators of genes that were specifically associated with the LGE. Curiously, this led to the identification of HTT as the most significant upstream regulator when looking at both early (7-11pcw; Figure 2.7B left panel) and late developmental time-point (20pcw; Figure 2.7B right panel). This gene did not appear as an upstream regulator of the unique gene signatures of the neocortex or MGE. This sparks great interest, especially for our laboratory, since it has been shown that HTT is necessary for normal brain development (White *et al.*, 1997; Dragatsis, Efstratiadis and Zeitlin, 1998; Tong *et al.*, 2011; Nguyen *et al.*, 2013) and recent evidence have suggested a neurodevelopmental component in HD (Nopoulos *et al.*, 2011; Molero *et al.*, 2016; Lim *et al.*, 2017; Wiatr *et al.*, 2017). Furthermore, a study performed in our laboratory (Conforti *et al.*, 2018) (to see my contribution in this work see the Appendix 6.2.1) showed that during *in vitro* MSN differentiation mHTT leads to defects in early striatal progenitor specification and commitment (Conforti *et al.*, 2018). Therefore having the specific network that is activated during normal human striatal development will help us elucidate what could be the potential targets of HTT and what could be associated to the effects observed in mHTT conditions during MSN *in vitro* differentiation (Conforti *et al.*, 2018). In line, the genes downstream of HTT include *NTRK2*, a receptor for brain-derived neurotrophic factor (BDNF) that has been shown to be required for the development of striatum and survival of striatopallidal MSNs (Baydyuk *et al.*, 2011; Li *et al.*, 2012) and *EBF1* (Garel *et al.*, 1999; Onorati *et al.*, 2014) that is fundamental for the acquisition of mantle cell molecular identity.
Other upstream regulators (Figure 2.7B) that are known to have a role in striatal development are GDNF that is essential for survival and organization of the striatum (Chermenina *et al.*, 2014), ZFH3 that is a known marker of the developing striatum (Onorati *et al.*, 2014), BDNF that is fundamental for the survival of immature MSNs of the indirect and direct pathways before they migrate to their final destination (Baydyuk *et al.*, 2013), HES5 a downstream target of the Notch signalling pathway present in the VZ of the LGE between E12.5-E15.5 (Mason, 2005), FOXP1 that characterizes MSN precursors prior to the expression of DARPP-32 during development (Onorati *et al.*, 2014; Precious *et al.*, 2016) and ASCL1 (MASH1) an important regulator of neurogenesis in the LGE (Casarosa, Fode and Guillemot, 1999; Castro *et al.*, 2011). Other curious upstream regulators include HOXD4 and ATN1. In particular, HOXD4 contains a retinoid-

responsive transcriptional enhancer in the 5' region of the gene (Morrison *et al.*, 1996) and is upstream of *RARβ,* an important gene for MSN development (Rataj-Baniowska *et al.*, 2015). This may be an interesting target to evaluate since retinoid signalling has been shown to have a pivotal role during striatal development (Toresson *et al.*, 1999; Wang and Liu, 2005; Evans *et al.*, 2012) . Instead, ATN1 is highly correlated with HTT expression in the striatum (Keo *et al.*, 2017) and is upstream of *EBF1,* therefore understanding how these signals converge would give insight on the molecular networks that guide striatal development and their relationship with HTT and ATN1.

To further understand the biological relevance of the LGE specific genes, we carried out a GO analysis using the clueGO plugin of Cytoscape (Bindea *et al.*, 2009) and then calculated the semantic similarity scores of all GO-term pairs (*p-adj* < 0.05). Between 7 and 11pcw we found a high correlation between terms for neural differentiation, striatal development and forebrain development (Figure 2.7C). Genes linked specifically to striatal development included retinoic acid receptor β (*RARβ*) that is known to control development of a subpopulation of striatonigral projection neurons (Rataj-Baniowska *et al.*, 2015), *EBF1, ISL1* and *FOXP2* that are known human striatal MSN marker (Onorati *et al.*, 2014) and the MSN markers *DRD1* and *DRD2* (Rataj-Baniowska *et al.*, 2015). The rest of the genes were generally linked to the development of the telencephalon, glutamate receptor signaling or synaptic transmission (Figure 2.7C) but their role in the LGE are still to be determined. At 20pcw terms were linked to neuron projection development, central nervous system myelination, axon development and long−term synaptic potentiation reflecting the more mature signature of the 20pcw striatum (Figure 2.7D). Even at this time-point we could pinpoint SVZ-mantle zone specific markers of the developing human striatum like *ISL1* and *FOXP2* (Onorati *et al.*, 2014). Furthermore, we identified specific markers of sub-population of DRD1 and DRD2 MSNs (Gokce *et al.*, 2016) in the 20pcw striatum that were not present between 7 and 11pcw. This included *FOXO1* (D1 sub-population A), *DNER* (D1 sub-population B), *HTR7* (D2 sub-population A) and *SYNPR* (D2 sub-population B) suggesting that the signature observed at this time point is more linked to genes that characterize the mantle zone and therefore more mature neuronal states.


A
CX
LGE
MGE
Area
519
1698
663
3948
72
201
166
4419
86
229
99
7pcw
9pcw
11pcw
20pcw
pcw
B
Upstream regulator
*HTT
GDNF
*CREB1
ZFHX3
TLX3
TLX1
TAF4
LEP
NGF
EOMES
*MECP2
LHX2
DKK1
*TGFB1
*SP1
BCL11B
SIX5
PPARGC1B
KLF7
KLF11
HOXD4
GTF2F2
ATF1
*ATN1
*POU4F1
*BDNF
SMAD3
HES5
Foxp1
ASCL1
0.0 2.5 5.0 7.5
-log10(p-value)
Upstream regulator
*HTT
*CREB1
REST
*TGFB1
*POU4F1
FEV
CTNNB1
TCF7L2
PDX1
TP53
JUN
*ATN1
GDF2
*BDNF
FOS
TWIST1
*MECP2
CEBPA
NEUROG1
SMARCA4
IGF1
PAX3
NFE2L2
HIF1A
AGT
HOXA10
*SP1
CREBBP
EMX2
PTF1A
0 5 10 15 20
-log10(p-value)
C
glutamate receptor signaling pathway
adenylate cyclase−modulating G−protein coupled receptor signaling pathway
adenylate cyclase−inhibiting G−protein coupled receptor signaling pathway
G−protein coupled receptor signaling pathway. coupled to cyclic nucleotide second messenger
phospholipase C−activating G−protein coupled receptor signaling pathway
presynaptic process involved in chemical synaptic transmission
chemical synaptic transmission
neuron−neuron synaptic transmission
regulation of synaptic plasticity
modulation of synaptic transmission
positive regulation of synaptic transmission
long−term synaptic potentiation
positive regulation of long−term synaptic potentiation
cyclic purine nucleotide metabolic process
regulation of cAMP biosynthetic process
regulation of adenylate cyclase activity
cyclic nucleotide metabolic process
cAMP metabolic process
positive regulation of purine nucleotide metabolic process
regulation of cyclic nucleotide metabolic process
regulation of purine nucleotide metabolic process
striatum development
telencephalon development
forebrain development
cerebellar granule cell differentiation
central nervous system neuron differentiation
positive regulation of cell development
positive regulation of neuron differentiation
positive regulation of neurogenesis
positive regulation of nervous system development
synapse assembly
regulation of synapse assembly
memory
learning
associative learning
ovulation
regulation of circadian sleep/wake cycle. sleep
regulation of sodium ion transport
potassium ion transport
ammonium transport
catecholamine transport
regulation of amine transport
regulation of ion transmembrane transport
regulation of dopamine uptake involved in synaptic transmission
monovalent inorganic cation homeostasis
sodium ion homeostasis
dopamine metabolic process
detection of mechanical stimulus involved in sensory perception of pain
cellular response to angiotensin
1
0.8
0.6
0.4
0.2
0

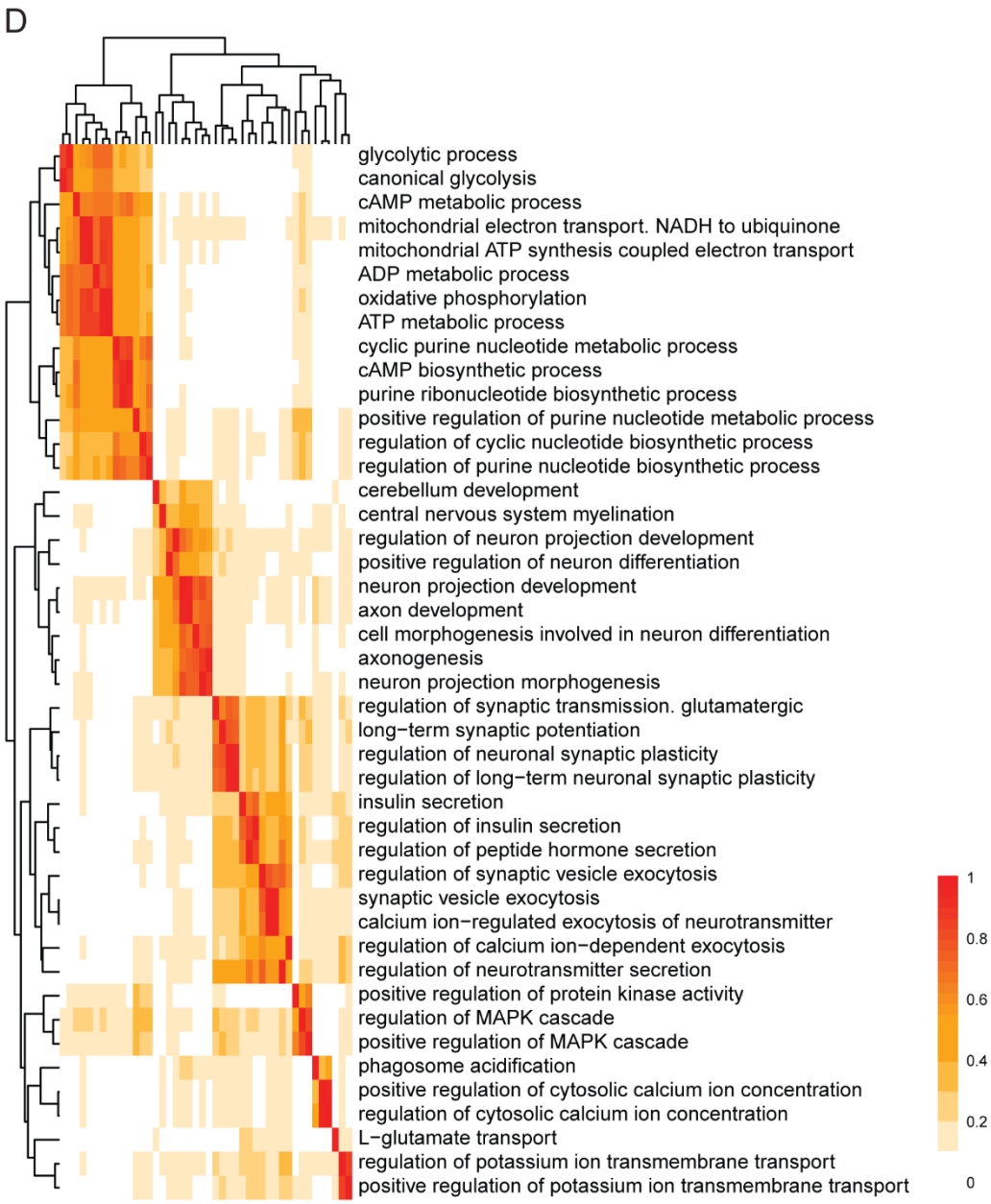


Figure 2.7 Functional Enrichment of the LGE specific gene signature. (A) Number of differentially expressed genes per area and pcw (p-adjusted < 0.05) **(B)** Upstream regulators associated with differentially expressed genes at early pcw (7-9-11pcw, left panel) and late pcw (20pcw, right panel). **(C-D)** Gene Ontology semantic similarity matrix of the LGE specific signature clustered by hierarchical clustering for early pcw (7-9-11pcw, **C**) and late pcw (20pcw, **D**).

2.1.7 Predicting the role of specific lincRNAs of the LGE

Since lncRNAs have been reported to influence the expression of neighbouring genes (Orom *et al.*, 2010; Collier *et al.*, 2012; Gomez *et al.*, 2013; Hu *et al.*, 2013) we retrieved for each specific lincRNA of the LGE, the proximal protein-coding and investigated whether these protein-coding genes were involved in key neural functions. A set of upstream regulators of this array of protein-coding genes were linked to neuronal function (Figure 2.8A). For example, SOX14 is found in

post-mitotic interneurons (Uchikawa, Kamachi and Kondoh, 1999; Kamachi, Uchikawa and Kondoh, 2000), while EMX2 is highly expressed in the early neuronogenic pallium and is involved in the transition between early and late neural progenitors in the dorsal telencephalon (Theil *et al.*, 2002; Gangemi *et al.*, 2006; Falcone and Mallamaci, 2015). Furthermore, ASCL1 (MASH1) and HTT are shared upstream regulators of specific protein-coding genes of the LGE and are important for the development of this area as discussed above.

We then examined the pathways that were enriched in this set of protein-coding genes, and we observed a high significance (p-value < 0.05) for processes linked to development of neuroglia, differentiation of neurons, function of brain and migration of neurons (Figure 2.8B). In line, genes involved in these pathways were *SOX1* that is an important transcription factor involved in the maintenance of a neural progenitor identity (Pevny *et al.*, 1998; Bylund *et al.*, 2003), *SOX11* that is of critical importance for the establishment of pan-neuronal protein expression (Bergsland *et al.*, 2006) and *VGF*, that encodes secretory peptides with established roles in neurogenesis (Sha *et al.*, 2012). In conclusion, all these results point towards a potential symbiotic role of these specific lincRNAs of the LGE and their relative neighbouring protein-coding genes in defining the striatal fate. However, although these results are promising they must be interpreted with care since the co-expression between a lincRNA and its protein-coding neighbour may result from a true *cis* relationship between the two biotypes, or a state of open chromatin that leads to proximal transcriptional activity in the area surrounding the specific gene (Ebisuya *et al.*, 2008).

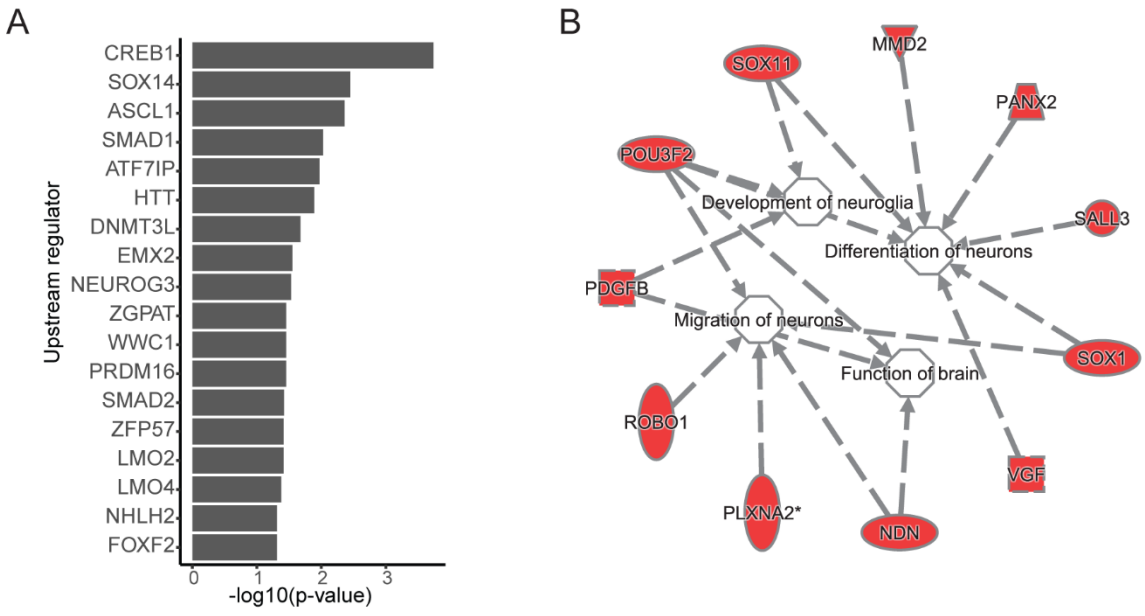


Figure 2.8 Functional Enrichment in neighbors of LGE specific lincRNAs. (**A)** Upstream regulators associated with neighbors of specific lincRNAs of the LGE (**B)** Function Analysis using IPA of neighbors of LGE specific lincRNAs

2.2 Single-cell transcriptional profiling of the developing striatum

2.2.1 Sample processing and quality control

To characterize the different sub-populations present in the developing striatum we designed an initial pilot experiment to evaluate how the 10x Chromium (10xTM GemCodeTM Technolog) captured cell diversity in this area and to grasp the sequencing depth and number of cells required to correctly describe the different cell types of this area. In line, we performed scRNA-seq on an initial set of 80 000 cells at a sequencing depth of 50 000 reads/cells. Initial sample demultiplexing, barcode processing and single-cell 3′ gene counting was performed using the Cell Ranger Single-Cell Software (Methods 4.3.1) and enabled us to retrieve 66080 cells out of the 80000 predicted (82.6%). This suggests that high number of cells can be retrieved with the 10x Chromium settings determined for this pilot experiment. Following gene expression normalization for read counts we evaluated quality of cells by examining the relationship between gene and molecule counts and filtered out potential cell artifacts and multiplets by removing cells with unique gene counts below 500 and over 6000 (Figure 2.9A). We further analyzed the percentage of mitochondrial genes present as the relationship between mtDNA and cell death has been well documented (Detmer and Chan, 2007; Galluzzi, Kepp and Kroemer, 2012; Ilicic *et al.*, 2016) and a high proportion of mitochondrial genes indicates poor-quality cells possibly due to increased apoptosis and/or loss of cytoplasmic RNA from lysed cells (Islam *et al.*, 2014; Ilicic *et al.*, 2016). We observed a low percentage of mitochondrial genes throughout the fetal samples (Figure 2.9B). We decided, however, to remove cells with a percentage of mitochondrial genes above 0.05% as we assume that most cells in the dataset are of high quality and that this percentage captures most of the outliers that can skew downstream analysis. This quality control filtering step allowed us to extract 59962 cells out of the initial 66080 cells (90,7% recovery) suggesting a good technique in terms of fetal tissue sample handling and library preparation for this pilot experiment. After filtering for the quality cells we identified highly variable genes (HVGs) (Methods 4.3.2). HVGs were selected to focus on genes that are driving heterogeneity across the population of cells (Butler *et al.*, 2018a; Yip, Sham and Wang, 2018). This led to the identification of 1797 HVGs that were then used in downstream analysis (Figure 2.9C).

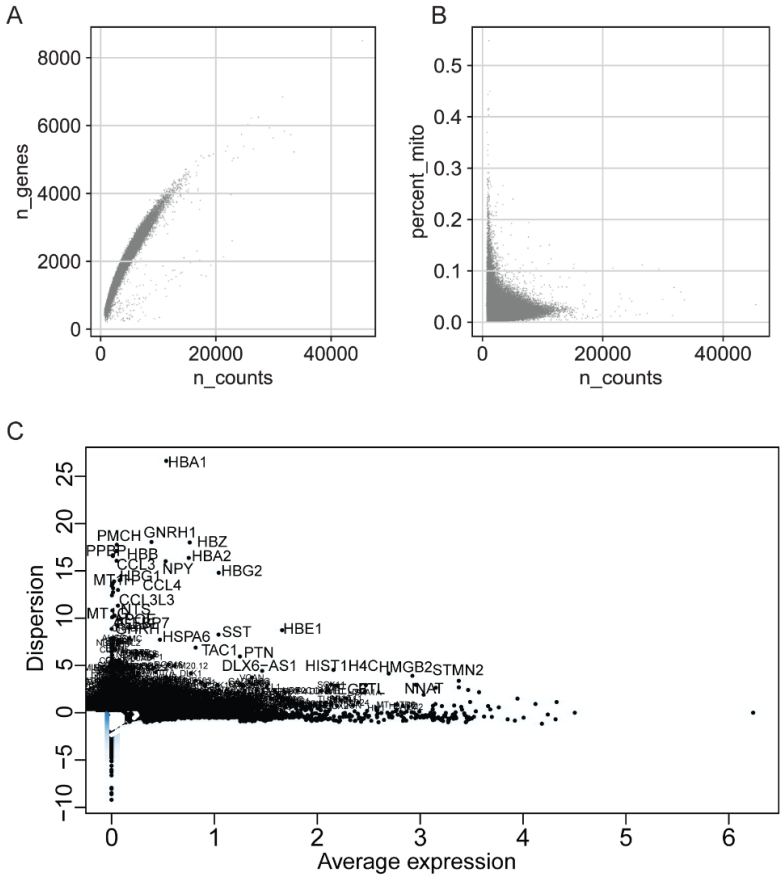


Figure 2.9 Quality Control of single-cell RNA-seq data. (**A)** Plot shows relationship between number of count and genes in each cell, this relationship was used to find cells that were outliers in terms of genes and UMIs. (**B)** Plot shows relationship between number of count and the percent of mitochondrial genes present in each cell, this relationship was used to pinpoint cells that had a high proportion of mitochondrial DNA that may be indicative of low quality cells undergoing apoptosis or cells that have lost cytoplasmic RNA. (**C)** Graph showing the average expression and dispersion for each gene. HVGs are all points (genes) between 0.0125 and 3 and that have a dispersion greater than 0.5.

2.2.2 Cell communities within the developing human striatum

To identify sub-populations in the developing striatum at 7-8pcw we initially applied principle component analysis (PCA) on the cleaned dataset (Methods 4.3.2). This step allowed us to reduce the dimension of the data but also evaluate which PCs are the primary source of heterogeneity within the cells. To select which PCs are the most pronounced in the data we plotted the standard deviations of the principle components and observed that the first 10 PCs had the strongest variability (Figure 2.10A). Although no clear cut was visible, and by testing for significant PCs using the jackStraw procedure (Macosko *et al.*, 2015) (Methods 4.3.2) the first 60 PCs were significant, we decided to take the first 10 as we initially wanted to evaluate the populations of the LGE at a lower resolution and once the identity of these "coarse" clusters have been established we will increase granularity (therefore PCs in the analysis) and identify sub-clusters within the

main established clusters. We therefore used these 10 highly informative principle components to classify cells into groups using graph-based clustering (Methods 4.3.2). This approach identified 15 clusters (Figure 2.10B) that were well covered by cells from different fetal samples (Figure 2.10C). An exception was the sample BRC1990 that clustered separately from the main group of cells (Figure 2.10C). This is a probable dissection error since the sulcus separating the LGE and MGE is still not well formed at these early pcw and also the limit between the rostral MGE-LGE and the caudal CGE is difficult to establish. Also BRC2013 had cells clustering with BRC1990, however it also had cells in other cell communities (Figure 2.10C), suggesting that this dissection covered a broader area of the ganglionic eminences.

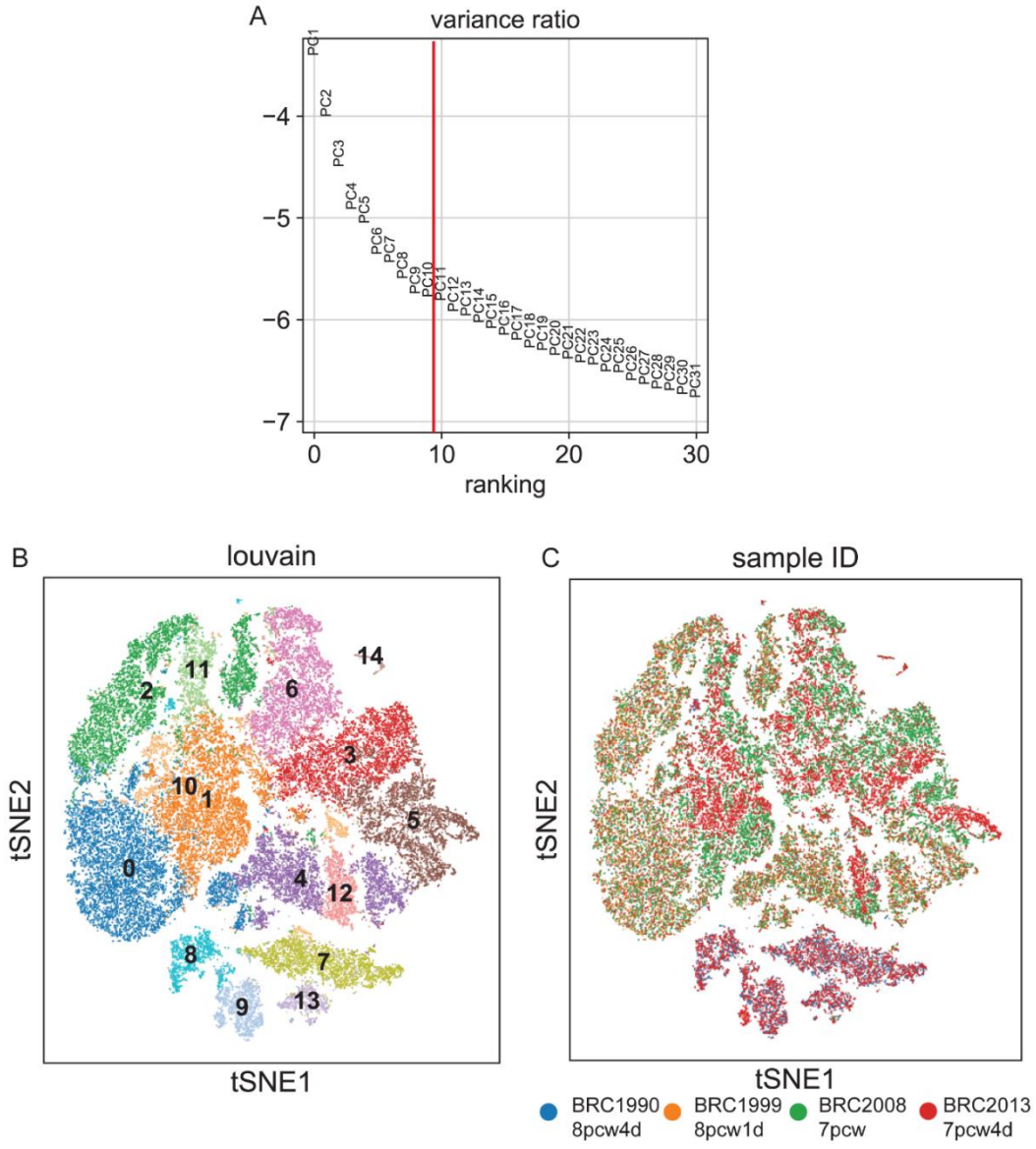


Figure 2.10 The LGE cell type atlas (A) plot showing the standard deviations of the first 31 principle components. The first 10PCs were taken into consideration for downstream analysis (**B)** tSNE representation of the single-cell transcriptomics data with clusters colored according to the partitioning established with the Louvain algorithm. (**C)** tSNE representation of the single-cell transcriptomics data with single cell colored by sample of origin.

We defined the different populations based on canonical markers of the LGE and the areas surrounding it (MGE, CGE and neocortex) and assigned to each cell community a temporary name that will be changed if further analysis on this set of data suggest a different cell origin (Figure 2.11A, left panel). We observed 6 clusters (VZ.Pr.1, VZ.Pr.2, SVZ.Pr.1, SVZ.Pr.2, IZ.Pr, Cx.Pr) of cells that expressed progenitor marker genes like *SOX2* (Ellis *et al.*, 2004; Zhang, 2014) (Figure 2.11A, middle panel). While the other communities were all positive for doublecortin (*DCX*), an immature neuronal marker (Brown *et al.*, 2003), indicating that these cells are early born neurons (Figure 2.11A, right panel). We were able to distinguish the MSN population that was characterized by the known markers *MEIS2* (Ghanem *et al.*, 2008), *EBF1* (Garel *et al.*, 1999; Onorati *et al.*, 2014), *ISL1* (Stenman, Toresson and Campbell, 2003; Onorati *et al.*, 2014) , *FOXP1* (Onorati *et al.*, 2014; Precious *et al.*, 2016), *BCL11B* (*CTIP2*) (Arlotta *et al.*, 2008; Onorati *et al.*, 2014) and *GAD1* (Erlander *et al.*, 1991) (Figure 2.11B). Together with the striatal MSNs we were able to pinpoint a population of interneurons (Int.1) (Figure 2.11C) that expressed *NKX2.1* and *LHX6* (Lavdas *et al.*, 1999; Alifragis, 2004; Flames *et al.*, 2007; Fogarty *et al.*, 2007; Liodis *et al.*, 2007; Flandin, Kimura and Rubenstein, 2010). However, these cells were negative for *PV* (Kelsom and Lu, 2013), *SST* (Nóbrega-Pereira *et al.*, 2008), *TH* (Ibáñez-Sandoval *et al.*, 2010) and *CHAT* (Lim, Kang and McGehee, 2014) suggesting an immature transcriptional program (data not shown). Since the MGE gives rise to interneurons (Lavdas *et al.*, 1999; Sussel *et al.*, 1999; Xu, 2004; Butt *et al.*, 2005) this population of cells may have derived from an erroneous dissection of part of the MGE together with the LGE or they may be a population of migrating interneurons. Another population of interneurons, that we called Int.2, expressed the marker for striatal interneurons *ERBB4* (Villar-Cervino *et al.*, 2015) together with somatostatin (*SST*) (Nóbrega-Pereira *et al.*, 2008) (Figure 2.11C). This population, did not express *NKX2.1*, suggesting that these cells had migrated away from the MGE and were actually found in the LGE at the time of dissection. However, since interneurons that downregulate *Nkx2-1* migrate to the cortex and those that maintain *Nkx2-1* expression localize in the striatum (Sussel *et al.*, 1999; Nóbrega-Pereira *et al.*, 2008), these interneurons were probably destined to the neocortex. This hypothesis was further confirmed by the expression of the MGE derived interneuron markers, *PDE4DIP* (Zhong *et al.*, 2018) and *NXPH1* (Fan *et al.*, 2018), that have been recently identified in the neocortex during human development. However, to fully elucidate the origin and localization of these two types of interneurons further analysis and fluorescent in situ hybridization (FISH) validation are necessary.

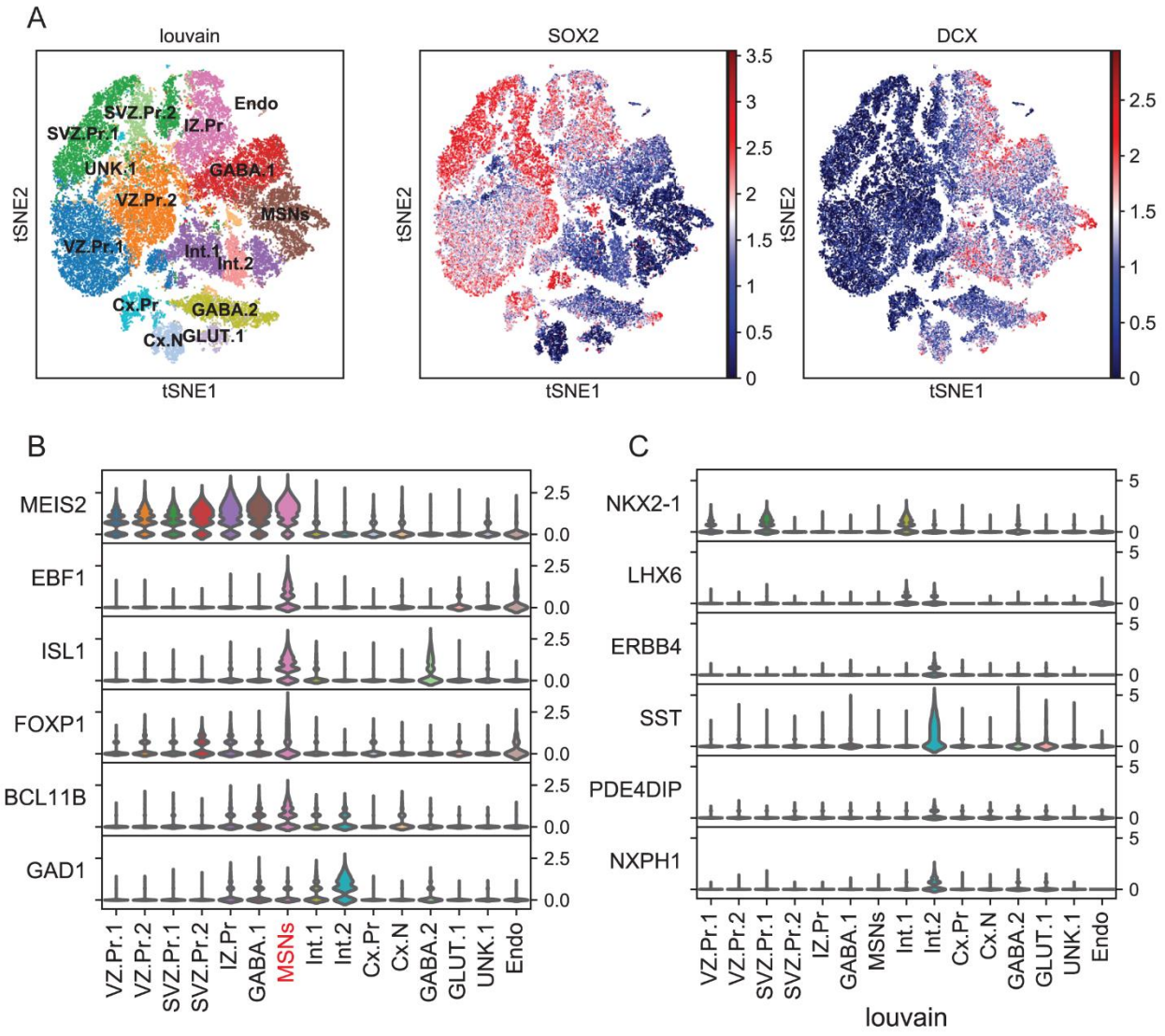


Figure 2.11 Gene signatures of the cell communities of the LGE (A) Left panel: tSNE representation of the single-cell transcriptomics data with clusters colored according to the partitioning established with the Louvain algorithm; middle and right panel: expression of the neural progenitor marker SOX2 and the early new born neuron marker DCX. (**B)** MSN canonical marker expression in the different cell population identified. (**C)** Interneuron markers in the different cell population identified. In red the populations under examination are highlighted. VZ.Pr.1=Ventricular zone progenitors 1; VZ.Pr.2 = Ventricular zone progenitors 2; SVZ.Pr.1 = Subventricular zone progenitors1; SVZ.Pr.2 = Subventricular zone progenitors 2; IZ.Pr = Intermediate zone progenitors; GABA.1 = GABAergic neurons 1; MSNs = Medium spiny neurons; Int1 = Interneurons 1; Int2 = Interneurons 2; Cx.Pr = Cortical progenitors; Cx.N = Cortical projection neurons; GABA.2 = GABAergic neurons 2; GLUT.1 = Glutamatergic neurons 1; UNK1 = Unknown population 1; Endo = Endothelial cells.

In this pilot experiment we also detected a small percentage of cells from derived from a contamination of surrounding areas. This included a population of ventral neocortical progenitors (Cx.Pr) that co-expressed *PAX6* and *COUP-TFII* (*NR2F2*) that are found to be expressed in the progenitor zone of the ventro-temporal cortex and ventro-medial frontal neocortex in humans

(Alzu'bi *et al.*, 2017) and a population of cortical projection neurons (Cx.N) that expressed the cortical markers *NEUROD2* (Bormuth *et al.*, 2013), *BCL11B* (Arlotta *et al.*, 2005; Leone *et al.*, 2008), *TBR1* (Bedogni *et al.*, 2010) and the glutamate transporters *VGLUT2* (*SLC17A6*) (Boulland *et al.*, 2004) (Figure 2.12B). Furthermore, a small percentage of endothelial cells (ENDO) that expressed the markers *IGFBP7* (Akaogi *et al.*, 1996) and *COL4A1* (van Agtmael *et al.*, 2010) was also present (Figure 2.12C). Within the 15 identified clusters we were not able to identify the origin of four clusters. The first two populations, that we temporarily named GABA.1 and GABA.2 (Figure 2.12A), were GABAergic since they expressed *GAD1* (Erlander *et al.*, 1991) but no other known marker tested up till now (Figure 2.11B). The other cluster (GLUT.1) expressed the glutamate transporters *VGLUT2* (*SLC17A6*) (Boulland *et al.*, 2004) (Figure 2.12B) as well as the newly identified neocortical marker *NRN1* (Nowakowski *et al.*, 2017), that was shared with the cell community defined as cortical projection neurons (Cx.N) suggesting a cortical origin also for this cell type (Figure 2.12D). The last populations was temporarily named UNK.1 (Figure 2.12A), but we were not able to find any overlap with known markers. Further analysis are required to understand the origin and relevance of these three clusters. Finally, we did not observe any markers of oligodendrocytes, astrocytes or microglia (Pollen *et al.*, 2015; Zhong *et al.*, 2018).

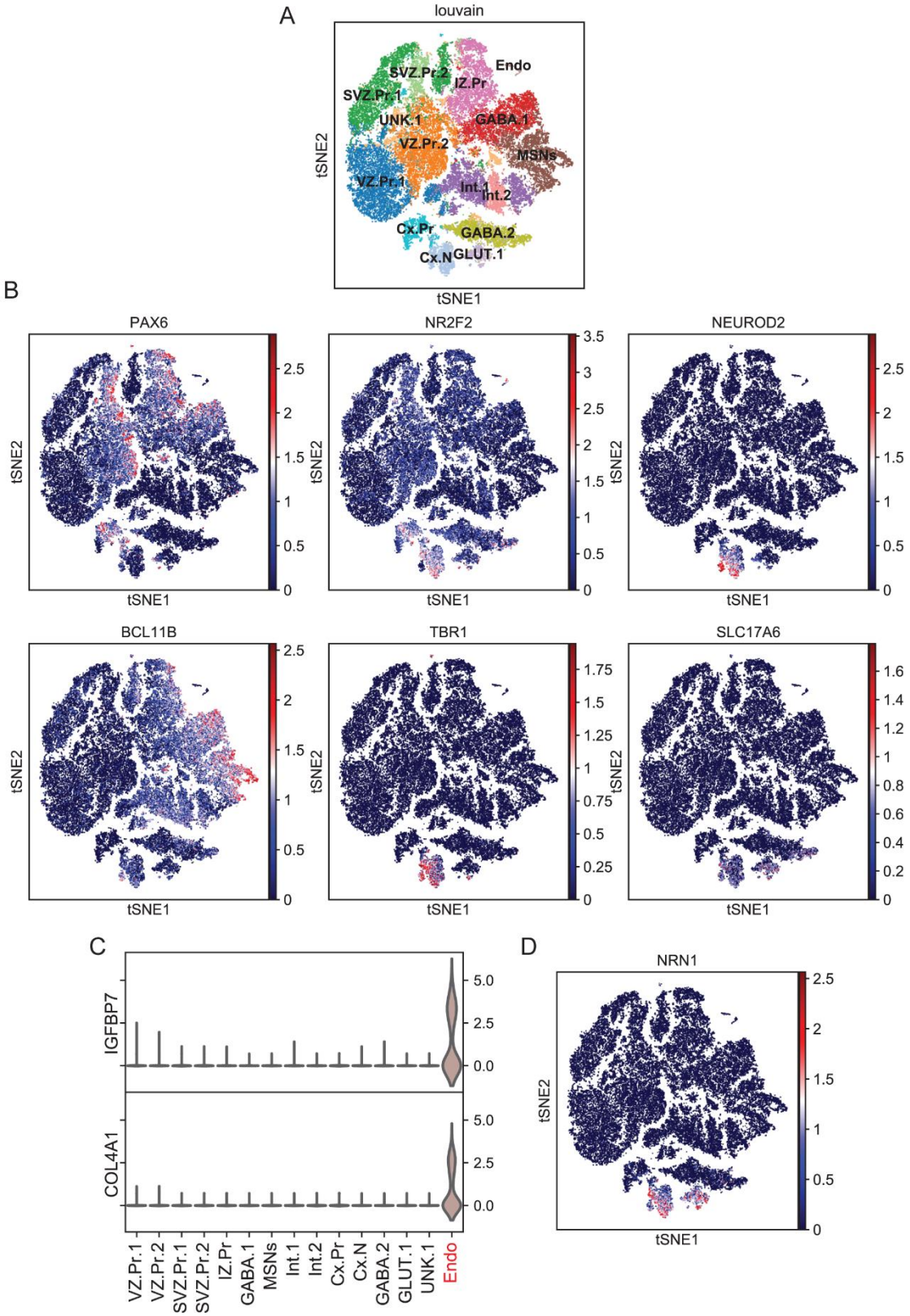


Figure 2.12 Gene signatures of the cell communities of the LGE (A) tSNE representation of the single-cell transcriptomics data with clusters colored according to the partitioning established with the Louvain algorithm (**B)** Expression levels of canonical markers of the neocortex in different cell population. (**C)** Violin plot showing the expression of endothelial markers in different cell population. **(D)** Expression of the neuronal GABAergic marker GAD1 in different cell populations.

2.2.3 Decoding the identity of the progenitor sub-classes of the LGE lineage

To infer cell trajectories and differentiation trees that define the cell populations identified in this pilot experiment we used a method called partition-based graph abstraction (PAGA) (Methods 4.3.4). In particular, we calculated the connectivity of cell clusters and then generated an abstracted graph in which nodes correspond to the previously identified cell communities and edges represent the potential transitions between the communities. The resulting abstracted graph (Figure 2.13A) showed a highly interconnected set of clusters (VZ.Pr.1, VZ.Pr.2, SVZ.Pr.1, SVZ.Pr.2, IZ.Pr, GABA.1 and MSNs) that converge to give rise to MSNs, suggesting that this cell trajectory is specific to the LGE lineage. Additionally, the tree reflects the relationship between different groups of cells that derive from different areas. For example, it predicts the relationship between the two different subtypes of interneurons (Int.1 and Int.2) and the two neocortical populations (Cx.Pr and Cx.N).

To full comprehend the nature of the striatal (LGE) lineage we used a measure of graph distance (diffusion pseudotime, DPT) (Haghverdi *et al.*, 2016) (Methods 4.3.4) to explore the potential temporal connections between cell states in the LGE lineage (Figure 2.13B-C). We observed that the cells within the VZ.Pr.1, VZ.Pr.2, SVZ.Pr.1 and SVZ.Pr.2 clusters had the lowest DPT score suggesting that these are the earliest progenitor types. While the populations that we named IZ.Pr, as we hypothesised these may be derived from an intermediate zone of the striatum, showed an increase in the DPT score that reached a peak in MSNs (Figure 2.13C).

This pseudotemporal state between the progenitor communities is confirmed by the analysis of known markers of the VZ-SVZ (Figure 2.13D). In line, the progenitors VZ.Pr.1 and VZ.Pr.2 exhibit a high expression of *VIM, SOX2* and *NES* and a lower expression of *DLX2* (Figure 2.13D), a known marker of the SVZ of the basal ganglia during differentiation (Liu *et al.*, 1997; Eisenstat *et al.*, 1999). While SVZ.Pr.1 and SVZ.Pr.2 exhibit a higher expression of *DLX2* and also *ASCL1* (a key gene of the SVZ of the developing human striatum (Onorati *et al.*, 2014)) advocating for a more committed precursor state (Figure 2.13D). Curiously, within the VZ (VZ.Pr.1 and VZ.Pr.2) and SVZ (SVZ.Pr.1 and SVZ.Pr.2) progenitors, we observed a mutually exclusive expression pattern of the MGE marker *NKX.2.1* (VZ.Pr.1 and SVZ.Pr.1) (Fertuzinhos *et al.*, 2009; Nobrega-Pereira *et al.*, 2010; Hansen *et al.*, 2013; Ma *et al.*, 2013; Onorati *et al.*, 2014) and the neocortical (Schuurmans and Guillemot, 2002) and dLGE (Stenman, 2003) marker *PAX6* (VZ.Pr.2 and SVZ.Pr.2) (Figure 2.13D). One can postulate that the progenitors positive for *PAX6* are derived from the dLGE and give rise to olfactory bulb interneurons (Stenman, Toresson and Campbell, 2003; Waclaw *et al.*, 2006) or neocortical interneurons (T. Ma *et al.*, 2012). However, these *PAX6* progenitors are negative for the SVZ dLGE *SP8* (Figure 2.13D) suggesting a different origin as

depicted by PAGA analysis that connects *PAX6* progenitors to the main LGE lineage (Figure 2.13A). Also the *NKX2.1* progenitor populations are connected to the MSN trajectory in the lineage tree (Figure 2.13D) advocating for a LGE specific progenitors and not a MGE progenitor as further confirmed by the absence of the MGE marker *LHX8* (Flames *et al.*, 2007), the presence of OTX2, a known marker of the VZ of the human LGE (Onorati *et al.*, 2014) (Figure 2.13D) and the absence of MGE radial glia-like ventral progenitors or intermediate progenitors markers that were recently characterized in a single cell study (Nowakowski *et al.*, 2017). Furthermore, the presence of *NKX-2.1* expression has been shown in the VZ-SVZ of the human developing LGE (Onorati *et al.*, 2014). Finally, the last progenitor sub-population connected to the LGE lineage, IZ.Pr, presents the most mature signature and was characterized by high levels of *HES6* (Figure 2.13D), that is found in cells committing to neural differentiation (Fior and Henrique, 2005) and *DLX5* that defines the SVZ and mantle zone of the ganglionic eminence (Liu *et al.*, 1997; Eisenstat *et al.*, 1999). This suggests that IZ.Pr is the most committed neural progenitor of the LGE lineage as defined by diffusion pseudotime (Figure 2.13C). Immediately, connected to the IZ.Pr population was the community of cells we named GABA.1. These cells are tightly linked to the MSN population but with a lower DPT score. These cells are probably an "immature MSN state" that express the known MSN markers *BCL11B* (*CTIP2*) (Arlotta *et al.*, 2008; Onorati *et al.*, 2014) together with the GABAergic marker *GAD1* (Erlander *et al.*, 1991) but lack expression of the canonical MSN markers *EBF1* (Garel *et al.*, 1999; Onorati *et al.*, 2014), *ISL1* (Stenman, Toresson and Campbell, 2003; Onorati *et al.*, 2014) and *FOXP1* (Onorati *et al.*, 2014; Precious *et al.*, 2016) (Figure 2.11B).

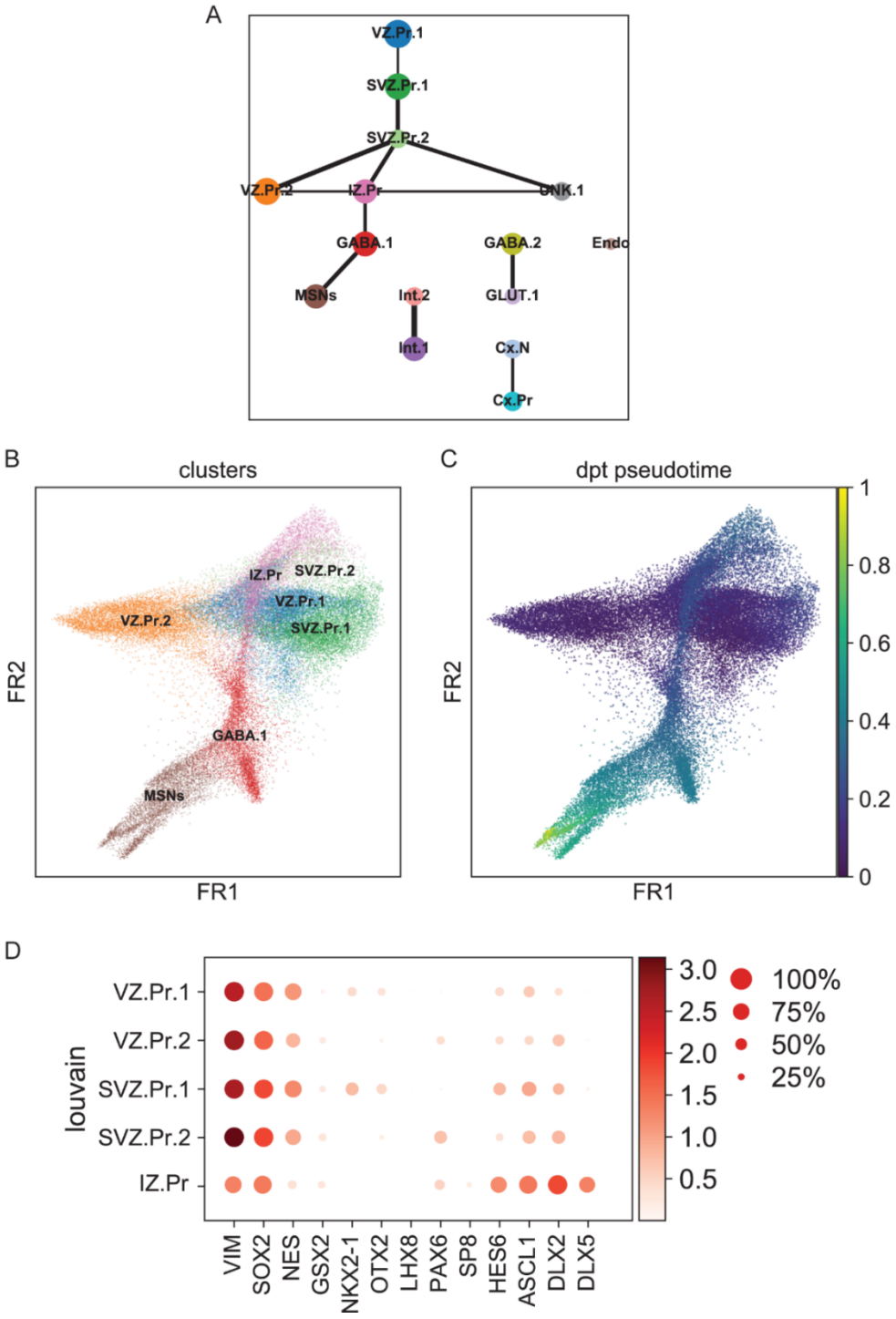


Figure 2.13 Lineage tree and progenitor-fate relationships of the LGE. (A) Abstracted graph showing all the possible edges connecting clusters with a threshold higher than 0.15. Each node corresponds the clusters identified with the Louvain algorithm. The size of nodes is proportional to the amount of cells in the cluster. The thicker edges represent more probable paths. (**B)** Single-cell embedding of the abstracted graph of the LGE lineage (**C)** Single-cell embedding of the abstracted graph of the LGE lineage colored by diffusion pseudotime score **(D)** Dotplot showing expression of different progenitor markers in each progenitor of the LGE lineage. Each dot represents two values: mean expression within each category (visualized by color) and fraction of cells expressing the gene in the category (visualized by size).

Although these preliminary data suggest a route for MSNs differentiation, delineating a solid hypothesis will need further analysis that include integrating the newly identified markers of the neocortex, LGE and MGE identified in the bulk RNA-seq section of the thesis (see section 2.1.6). In line, we have started evaluating lncRNA specificity within single-cell populations and we found that lincRNAs identified in the previous section by *de novo* analysis were ranked between the top specific genes of different cell communities. In particular, the lincRNA *hfb.lincRNA1* was highly specific for the type 1 VZ and SVZ progenitors (VZ.Pr.1, SVZ.Pr.1), *hfb.lincRNA2* was highly enriched in the IZ progenitors (IZ.1) progenitor and in GABA.1 neurons together with LINC01305 that was found as striatal specific in a recent single cell RNA-seq study (Nowakowski *et al.*, 2017), while *lincRNA3* was unique to MSNs (Figure 2.14). These results give an initial draft of how lincRNAs contribute to the different cell lineages. However, further analysis are required to fully comprehend their nature within these cell communities.

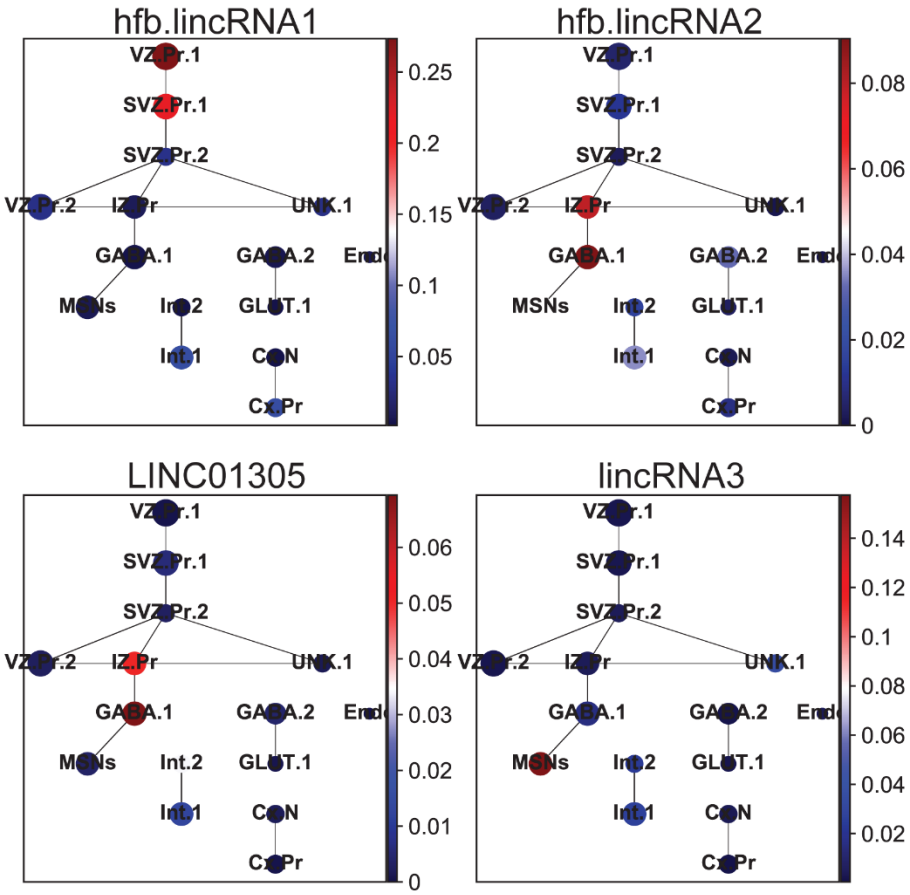


Figure 2.14 lincRNAs in the LGE. Abstracted graph showing all the possible edges connecting clusters with a threshold higher than 0.15. The color scale represents the expression intensity of specific lincRNAs of the LGE lineage. The lincRNAs with "hfb" prefix have been identified in this study by *de novo* analysis.

3 | Conclusions and Future Perspectives

This study reveals the first high resolution snapshot of the early development of the human striatum from a coding and non-coding perspective. In particular, the combination of deep RNA-sequencing with scRNA-sequencing allowed us to identify different cell sub-populations within the LGE and pinpoint cell type specific protein-coding genes and lincRNAs.

De novo analysis of lincRNAs enabled us to create a dictionary of this biotype, not only of the developing striatum, but also of the MGE and neocortex surrounding it. This dictionary is one of the few insights in the expression of lincRNAs in the human fetal brain and hopefully this catalogue will lead to a deeper understanding in the molecular networks that define the complexity of the human brain during development. In line, specificity analysis in this study confirmed that lincRNAs are remarkably tissue specific compared to protein-coding genes and therefore may be fundamental players in specific fine tuning of different cell states.

Transcriptomic analysis of single cells enabled us to create an atlas of the different cell communities present within the striatum at early post conceptional weeks and revealed the diversity of cells present in this area especially within the progenitor domain. By using a graph-based approach we were able to map the developmental landscape of individual cell states transitioning from early progenitors of the LGE to MSNs. Furthermore, we classified how subtype-specific heterogeneity progresses through the expression of cardinal genes including lincRNAs identified *de novo* in the bulk RNA-seq analysis.

Although this gives an initial sketch on the plethora of different cell subtypes present in the developing human striatum, integration with later developmental time-points (9, 11 and 20pcw), will help us reveal the kinetics that drive heterogeneity observed in MSNs in adulthood (Schiffmann, Jacobs and Vanderhaeghen, 1991; Gerfen, 1992; Ince, Ciliax and Levey, 1997; Gokce *et al.*, 2016; Zeisel *et al.*, 2018). In particular, adding later time-point could reveal the difference between early and late born MSNs and how the different progenitors contribute to the emergence of this diversity.

From a computational point of view further analysis are required to independently validate the differentiation trajectories predicted by PAGA. To do this we will use velocyto (La Manno *et al.*, 2017), a method that estimates the cellular expression state to which the cell is moving to. In particular, velocyto measures the relative abundance of nascent (unspliced) and mature (spliced) RNAs. In line, during an active cell state there is an induction of gene expression and one can observe an excess of unspliced RNAs, whereas when the cell transcriptional rate drops one can

measure a rapid reduction in unspliced RNA, followed by a reduction in spliced RNAs. Velocyto captures this balance between unspliced and spliced RNA abundance and uses it to predict the future state of a cell.
Finally, to paint a more complete picture of each cell and to validate the computational predictions made, we will perform RNA FISH for both protein-coding genes and lincRNAs on fetal samples ranging from 7 to 20pcw. FISH has the advantage of preserving the spatial context of the assayed transcripts adding, therefore, another level of fundamental information on the spatial dynamics of the candidate transcripts.

In conclusion, these findings reveal a preliminary blueprint of the transcriptional networks underlying MSN specification and hopefully, in the near future, we will identify genes that have important functional roles in the establishment and maintenance of MSNs that can be used *in vitro* as a compass to guide differentiation towards this specific cell subtype. Our findings mark an initial step towards the goal of ultimately linking specific genes to the development of striatal MSNs.

4 | Materials and Methods

4.1 Fetal sample processing

4.1.1 Human Tissue

Postmortem human brain specimens were obtained from University of Cambridge, UK from patients that requested pregnancy terminations and autopsy diagnostic procedures. 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, with the latest approved November 2017). The ethics were approved in the UK in accordance with the Human Tissue Act 2006. The documents were submitted to the Ethics Committee of the University of Milano and ethical 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 (http://www.wma.net/en/30publications/10policies/b3/index.html).

4.1.2 Human Tissue Collection

This study was conducted with postmortem human fetal brain specimens from tissues collected at the John van Geest Centre for Brain Repair, University of Cambridge, Cambridge, UK. Embryonic and fetal age was extrapolated based on the date of the mother's last menstruation, crown to rump length (CRL) and visual inspection. Depending on the condition and period of the procured specimens, the neocortex, LGE and MGE from different dissection methods were used. Specimens were chilled on ice during dissection and placed onto a chilled plate on ice. The dissected samples were placed in RLT buffer (Qiagen) and immediately frozen in dry ice and stored in −80 °C for later RNA extraction. For scRNA-seq dissected LGEs were maintained in Hibernate-E media (Thermo Fisher) for shipment to Milano.

4.1.3 RNA sample preparation for bulk RNA-seq

This procedure was performed by Andrea Faedo, Paola Conforti and Ira Espuny Camacho. Briefly, total RNA was extracted from LGE/striatal and neocortical human fetal tissues or from 30-DIV differentiated striatal and neocortical hPNs with an RNeasy kit (Qiagen), according to the manufacturer's instructions. 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 were used for cRNA preparations.

4.1.4 RNA sample preparation for single-cell RNA-seq

This procedure was performed by Paola Conforti, Draio Besusso and Ira Espuny Camacho. Briefly, LGE tissues were processed by dissociation using Papain Dissociation System (Worthington) following the manufacturer's recommendations, adjusting incubation time based on tissue piece size, 25-45 min. Briefly, after papain incubation, glass pipettes of increasingly smaller tip diameter (fire-polished) were used to dissociate to single-cell suspension followed by a centrifugation through a BSA single step discontinuous density gradient. Cells were pelleted, resuspended, and stored in N2 media with DNaseI.

4.1.5 Sequencing library construction using the GemCode platform

This procedure was performed by Paola Conforti, and Ira Espuny Camacho. Briefly, for experiments using the 10x Genomics platform, 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) were used according to the manufacturer's instructions in the Chromium Single Cell 3′ Reagents Kits V2 User Guide.

4.2 Bulk RNA-seq Bioinformatics Pipeline

4.2.1 Integrating lincRNAs to the reference annotation.

To integrate into the reference annoation lincRNAs identified in different studied (Cabili *et al.*, 2011; Liu *et al.*, 2016; Hon *et al.*, 2017) Cuffcompare v.2.1.1 (Anders *et al.*, 2014) was used with the -r option. This allows to compare GTF annotations and report transcripts found in only one GTF annotation (either q1 or q2) or the one present in both (both q1 and q2). In particular, Cuffcompare look for overlaps based on the agreements of the coordinates of the transcripts and order of all of their introns, as well as strand. Matching transcripts were allowed to differ on the length of the first and last exons, since these lengths will vary from sample to sample due to the random nature of sequencing. After selecting for unique transcripts in each annoation the transfrag "U" was used to filter and select only intergenic lncRNAs.

4.2.1 Quality Control

Rationale

Quality problems typically originate either in the sequencing itself or in the preceding library preparation. They include low-confidence bases, sequence-specific bias, 3′/5′ positional bias, polymerase chain reaction (PCR) artifacts, untrimmed adapters, and sequence contamination. These problems can seriously affect mapping to reference, assembly, and expression estimates and therefore need to be evaluated and corrected (Li *et al.*, 2015).

Computational details

All reads were tested for QC using FastQC (Andrews, 2010) and then all samples were evaluated simultaneously to examine specific biases in quality using MultiQC (Ewels *et al.*, 2016). To remove any low quality reads or to "cleanup" reads Trimmomatic (Bolger, Lohse and Usadel, 2014) was used. The first preprocessing step involved removing adapters that were still present in the reads. Low quality bases (below quality 3) were removed in the leading and trailing end of the read. Then a sliding window approach was performed where any 4 bases that had an average quality per base of 15 were removed. Finally, all remaining reads shorten than 50 bases long were removed.

4.2.2 Aligning reads to reference genome

Rationale

Powerful computational methods are required for ab-initio transcriptome reconstruction, for de-novo lincRNA discovery, and then transcript quantification. In the next section fundamental aspects of the computational pipeline developed for this work will be described.

RNA-sequencing entails fractioning the polyA transcripts, converting them to cDNA and then adding adapters to both ends of the fragments. cDNA is then amplified and fragments are sequenced starting from both ends of the adapters. This procedure generates short sequence reads that then need to be analyzed computationally to reconstruct the full transcriptome. The initial step consists in exploiting a fast aligner to map the short reads to the reference genome. However, aligning reads to a reference genome is challenging; in fact, reads generated by the sequencer are relatively short while genomes are large and contain non unique sequence such as repeats and pseudogenes lowering the "mappability" of these areas. Furthermore, aligners have to cope with mismatches and indels caused by genomic variation and sequencing errors. Finally, due to the

presence of introns, RNA-seq reads align to genome non-contiguously. In light of these factors, aligners must be flexible when applying mapping criteria (Fonseca *et al.*, 2012). Different alignment tools are available and can be summarized in two main categories: the first comprises tools that map to the genome while the second consists of those mapping to the annotated transcriptome (Conesa *et al.*, 2016). Genome alignment algorithms such as Bowtie and BWA rely on Burrows–Wheeler indexing for very fast genome alignment, but they struggle with transcriptome alignment due to splicing and variations from the reference such as insertions, deletions and substitutions (Burrows and Wheeler, 1994; Langmead *et al.*, 2009; Li and Durbin, 2009). Instead tools that align to the transcriptome, must align reads across splice. A plethora of algorithms are available (Li and Durbin, 2009; Trapnell, Pachter and Salzberg, 2009; Au *et al.*, 2010; Wang *et al.*, 2010; Wu and Nacu, 2010). However, different compromises need to be made in mapping sensitivity, precision and speed. Furthermore, alignment to the transcriptome forces reads to align to known transcripts, some of which could have been better aligned on an unannotated region of the genome (Grant *et al.*, 2011). We therefore opted to use STAR (Spliced Transcripts Alignment to a Reference) that allows for both high mapping accuracy and high speed. STAR initially performs seed searching by looking for the longest sequence that matches one or more locations on the reference genome. These matches are called the Maximal Mappable Prefixes (MMPs) and the first match will become the first seed (*seed1*). STAR will start the process again with the unmapped portion of the read and matches the next MMP that will be defined as *seed2*. When mismatches are found during the alignment process the MMPs is extended to cover a longer region of the genome. However, if the extension does not give a high quality alignment, then the poor quality sequence is removed (Figure 3.1). The mapped sequence seeds are merged by clustering around a selected set of anchor seeds within a defined genomic window and then stitched together. When paired end reads are present the seeds are merged concurrently. The stitching is guided by a scoring scheme that applies scores and penalties for matches, mismatches, insertions, deletions, and junction gaps. The stitched combination with the highest score is reported as the alignment of the read. The sequential searching of only the unmapped portions of reads is the main advantage of STAR as it allows quick and efficient searching along the reference genomes. Instead, other aligners, like TopHat2 (Kim *et al.*, 2013), use algorithms that often search for the entire read sequence before splitting reads and performing iterative rounds of mapping.

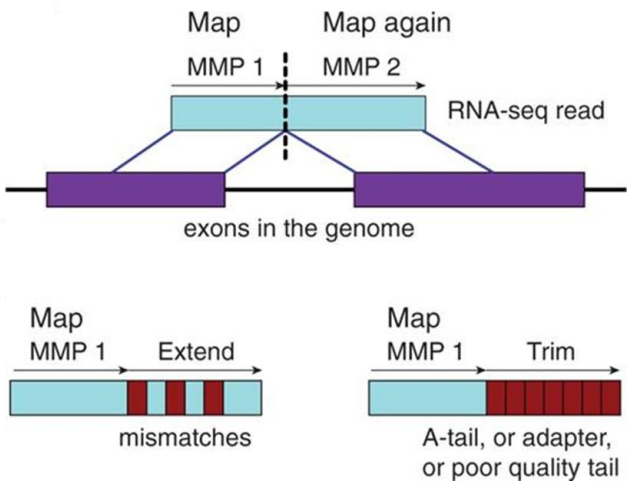


Figure 3.1 The STAR mapping algorithm. Representation of the Maximum Mappable Prefix (MMP) search to detect splice junctions (top panel), mismatches and (bottom, left panel) tails (bottom, left panel) (Dobin *et al.*, 2013).

Computational details

All reads were aligned to the human custom reference annoation created in the previous step using the spliced read aligner STAR (Spliced Transcripts Alignment to a Reference) v2.5.2b (Harrow *et al.*, 2012). We used the reference annotation GTF file to extract splice junctions to improve mapping accuracy by STAR. To preserve +/- strand information in unstranded RNA-seq data, that is needed during trancriptome reconstruction, we used the outSAMstrandField option with "intronMotif" specified that retains this attribute for all alignments that contain splice junctions. In addition, to feed assemblers (StringTie and Cufflinks) only with the highest confidence alignments, non-canonical junctions (that are not GT-AG, GC-AG and AT-AC (Burset, Seledtsov and Solovyev, 2000) were filtered out using the option outFilterIntronMotifs with RemoveNoncanonical specified. We then counted the number of reads per gene using the quantMode GeneCounts option. This modality counts a gene if it overlaps uniquely with one gene. Both ends of the paired end read are checked for overlaps. Using this procedure makes the mapping process more accurate as it excludes from the count reads that map to more than one gene.

4.2.2 lncRNA discovery - assembling transcripts

Rationale

To outline the dictionary of all expressed transcripts in a sample, all reads that have been mapped to the genome need to be assembled into transcripts. Identifying novel transcripts using short reads is one of the most demanding and critical operations for correct lncRNA discovery. Short reads rarely span across several splice junctions and thus make it difficult to directly infer all full-length transcripts and assemble multiple isoforms unambiguously. Different tools are available to

reconstruct the transcriptome and they can either operate by a genome independent (Grabherr *et al.*, 2011) or genome guided methodology (Guttman *et al.*, 2010; Trapnell *et al.*, 2011; Pertea *et al.*, 2015). Genome independent approaches are more challenging due to the presence of fluctuations in expression levels, a high number of alternatively spliced variants and multicopy gene families making this approach less accurate than genome-guided assembly (that use the genome as a map during the assembly process) and is more appropriate with organisms that don't have a reference genome (Zhao *et al.*, 2011). Therefore, for this study, that is performed on the human fetal brain, a genome-guided approach is more computationally efficient as a reference genome is available.

The most efficient assemblers that use genome guided approaches are Cufflinks (Trapnell *et al.*, 2011) and StringTie (Pertea *et al.*, 2015). These methods use spliced reads to reconstruct the transcriptome and build an assembly graph, but they have two different modalities to resolve this graph into transcripts. In particular, Cufflinks connects fragments into a graph if the overlapping fragments agree on their spliced alignment locations, it then identifies incompatible fragments that derive from unique isoforms and defines these incompatible fragments as the minimum number of transcripts that "explain" all fragments. This method allows for maximum precision as it all identifies the minimal number of compatible isoforms (Trapnell *et al.*, 2011). Instead StringTie (Pertea *et al.*, 2015), assembles each locus into as many isoforms as are needed to describe the data. It then uses a network flow algorithm to assemble and quantify simultaneously the most highly-expressed transcript. It then discards the reads connected with that transcript and reiterates the process with the remaining reads. It continues until all the reads are used and therefore assembles all possible isoforms allowing, in contrast to Cufflinks, maximum sensitivity.

The combination of different de-novo assembly approaches provides a clearer and more precise understanding of a newly explored. Therefore, using transcripts identified by both assemblers allows to minimize incorrect or miss-assembled transcripts that introduce ambivalence and complexity which may result in impracticable computation (Garber *et al.*, 2011).

Computational details

The transcriptome of each sample was assembled from the mapped reads separately by both Cufflinks (Trapnell *et al.*, 2011) and StringTie (Pertea *et al.*, 2015). To create the transcriptome assemblies after read alignments, Cufflinks version v2.2.1 was run using a likelihood based approach for fragment bias correction (Roberts *et al.*, 2011). This was done as studies have shown

that RNA-seq fragments can be preferentially located towards either the start or the end of a transcript causing a positional bias (Bohnert and Rätsch, 2010) and that sequence surrounding the beginning or end of fragments affects their probability of being captured for sequencing creating a sequence-specific bias (Hansen, Brenner and Dudoit, 2010; Srivastava and Chen, 2010). Therefore, this correction allows for an increased accuracy in expression estimates.

In addition, to more accurately weight reads mapping to multiple locations in the genome, we applied multi-mapped read correction. In line, using default parameters Cufflinks will evenly divide each multi-mapped read to all of the positions it maps to. For example, if a read maps to 5 positions it will count as 5% of a read at each position. With multi-mapping read correction Cufflinks will initially calculate abundance estimates for all transcripts using the modality mentioned above, however, it will then re-estimate the abundances dividing each multi-mapped read probabilistically based on the initial abundance estimation of the genes it maps to, the inferred fragment length, and the fragment bias (Trapnell *et al.*, 2011).

We then performed a ROC analysis to determine the optimal read coverage thresholds to limit technical noise and leaky expression based on whether Cufflinks classified previously known protein coding and lncRNAs as having full read support. To limit to a FDR of 0.05 we identified the average coverage thresholds of lncRNAs at 0.6. All assembled transcripts below this threshold were filtered out.

Instead StringTie v1.2.3 was run with default parameters with a filter of 0. 7 as the minimum per bp coverage to consider a read for transcript assembly. This threshold was based on the median bp coverage for lncRNAs identified in the human developing cortex (Liu *et al.*, 2016) as we hypothesized that these lncRNAs would have a similar expression to the ones identified in this study as the tissue is the same for the cortex tissue and could correctly predict coverage for the LGE and MGE.

4.2.3 lncRNA discovery - de-novo lncRNA prediction

Rationale

The next step after transcript assembly is to determine whether a novel transcript encodes a small, long non-coding or a protein coding RNA. This task is usually based on a diverse set of principles. The first, that enables discrimination between short and long non coding transcripts is that small RNAs (e.g. microRNAs, PIWI-associated RNAs, and endogenous small interfering RNAs) are smaller than 200nt in length while long ncRNAs are longer than 200nt in length (Kaikkonen, Lam and Glass, 2011). The second is based on the presence of an open reading frame (ORF) as proteins

are encoded in ORFs that consist of sequences of triplet codons beginning with ATG and ending with a stop codon and are usually, larger than 100 codons (Frith, Bailey, *et al.*, 2006). Therefore, the presence of an ORF ≥100 codons is frequently taken as an indication of the nature of the transcript. Since short putative ORFs can be expected to occur by chance within long noncoding sequences, minimum ORF cutoffs are usually enforced to reduce the likelihood of falsely categorizing lncRNAs as mRNAs (Dinger *et al.*, 2008). However, using ORF length alone, is limiting for different reasons. The first is that some lncRNAs have been shown to contain putative ORFs. This is the case for *H19*, *Xist*, *Mirg*, *Gtl2*, and *KcnqOT1* that all have putative ORFs >100 codons, but have been characterized as functional ncRNAs (Prasanth and Spector, 2007). Second, when a cutoff of 300nt is applied, proteins <100 aa in size may also be incorrectly classified as ncRNAs since ~3,700 proteins in the mammalian proteome have been shown to have ORF below this size (Frith, Forrest, *et al.*, 2006). Lastly, ~10% of protein-coding transcripts lack a start codon and ~25% lack a stop codon in the human EnsEMBL (v83) annotation (Yates *et al.*, 2016) creating false classification of protein-coding genes as lncRNAs as they don't fulfill the requirement of a start and stop codon. Given the problems of relying solely upon ORF size another fundamental principle is added to classify transcripts. This involves looking at codon usage and structure which is non-random in functional ORFs of protein-coding RNAs together with examining the nature of third position of codons, as evolutionary conserved ORFs are characterized by a more relaxed constrain in this position. (Sharp *et al.*, 1988).

To this aim, a series of bioinformatic tools are available in order to calculate a coding potential score (CPS) used to differentiate the coding status of the RNA gene. There are two main approaches to calculate CPS of RNAs: one uses sequence alignments, either between species (Lin, Jungreis and Kellis, 2011; Washietl *et al.*, 2011) or alignments to protein databases (Kong *et al.*, 2007), the other is based on an alignment-free approach (Wang *et al.*, 2013; Li, Zhang and Zhou, 2014). An example of an alignment based approach is PhyloCSF (Lin, Jungreis and Kellis, 2011) that requires a multispecies sequence alignment to exploit as a genome-wide training data to create phylogenetic codon models of known coding and non-coding regions. It then classifies transcripts based on whether the alignment has a higher probability under the coding or non-coding model. However, these alignment-dependent methods, have the disadvantages of depending on the quality of the input alignments (Schloss, 2010) and may also fail in classifying non-coding RNAs as protein-coding due to high conservation of some lncRNAs. In line, some lncRNAs show signs of purifying selection at the sequence level instead others show selection only for transcription (J.

Chen *et al.*, 2016). Therefore, lncRNAs like *TUG1*, *MALAT1*, and *XIST* are misclassified due to their high conservation (J. Chen *et al.*, 2016).

In contrast, alignment-free programs compute a CPS only depending on intrinsic characteristics of the input RNA sequences. One of the main traits that is used during the classification process is the length of the longest ORF (Blanco, Parra and Guigó, 2007; Brent, 2007) since a transcript comprising a long ORF will most likely be translated into a protein. Another complementary feature to segregate coding and non-coding RNAs is to examine biases in nucleotide frequencies and composition together with codon usage within different gene biotypes that has already been shown as being powerful discriminative features between coding and non-coding RNAs (Blanco, Parra and Guigó, 2007; Bussotti *et al.*, 2011). To measure this compositional bias within transcripts the relative frequency of k-mers (that refer to all the possible subsequences, of length k, from a transcript) can be measured. The longer the *k* the higher the resolution when discriminating lncRNAs and mRNAs (Li, Zhang and Zhou, 2014). Some programs already use k-mer frequencies but are limited to a small number of k-mers ($k \leq 6$), whereas longer stretches of k-mers could help resolve uncertainty by taking into account lncRNA-specific domains, repeats or spatial information (Johnson and Guigo, 2014; Zucchelli *et al.*, 2015).

To best characterize lncRNAs a recent tool has been developed called FlExible Extraction of LncRNAs (FEELnc) (Wucher *et al.*, 2017) that addresses the limits mentioned above and enables high classification performance compared to PhyloCSF (Lin, Jungreis and Kellis, 2011), CPC (Kong *et al.*, 2007), CPAT (Wang *et al.*, 2013), PLEK (Li, Zhang and Zhou, 2014) and CNCI (Sun *et al.*, 2013) that are five state-of-the-art tools.
FEELnc implements an alignment-free strategy using Random Forests (Breiman, 2001) to classify lncRNAs and mRNAs based on a relaxed ORF and a range of *k*-mer frequencies (from $k = 1$ to 12). In particular, to circumvent the difficulties of the ORF annotations described above, FEELnc computes all ORFs and discriminates five ORF types from the stringent ‘type 0’ that equals to the longest ORF having both a start and a stop codon, to the more loose ‘type 4’ that is, the whole RNA sequence, without either the start or stop codon (Figure 3.2).

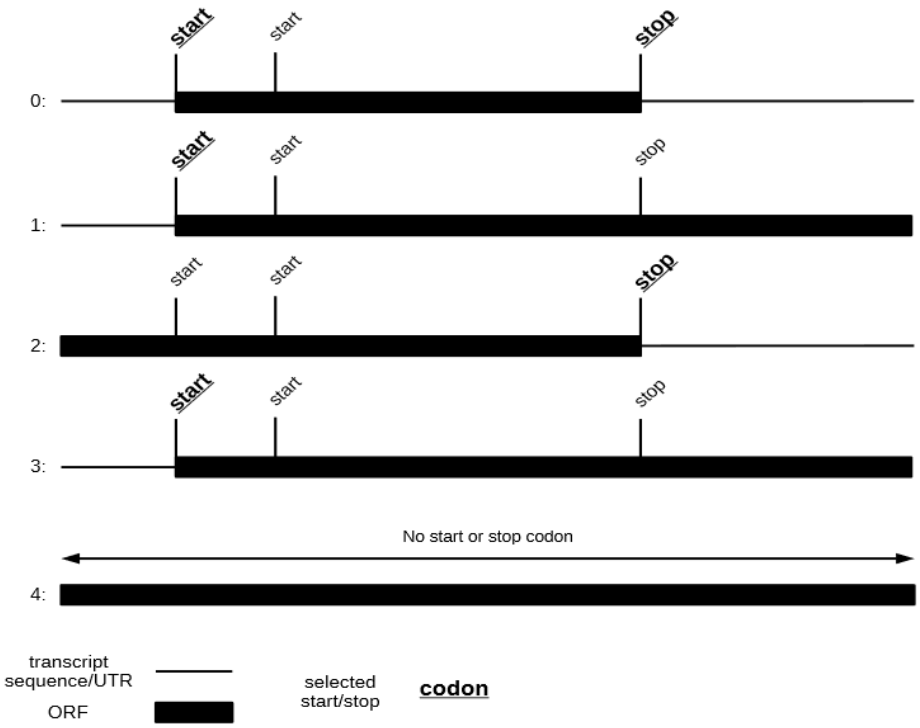


Figure 3.2 Annotation of ORFs by FEELnc. The start and/or stop codons selected by FEELnc are underlined and in bold characters. (Wucher *et al.*, 2017).

The second feature FEELnc uses to correctly interpret ORFs data is the proportion of the transcript size covered by an ORF (ORF coverage) as the size of the protein-coding ORF is generally correlated with the length of the input RNA sequence. This feature also has good discriminative power as some large bona fide ncRNAs may contain putative long ORFs by chance (Cabili *et al.*, 2011), however these large ncRNAs usually have a lower ORF coverage than protein-coding RNAs (Wang *et al.*, 2013). To measure the underlying structure of the transcript and the biases in nucleotide frequencies and codon usage FEELnc calculates each k-mer frequency in known mRNAs and lncRNAs by using a scoring method that enables a *k*-mer score to be associated to each sequence for each *k*-mer size. Finally, FEELnc uses total RNA sequence length as an additional predictor of the model since mRNAs have been shown to be significantly longer than lncRNAs (Cabili *et al.*, 2011; Harrow *et al.*, 2012). All these predictors are then integrated into a machine learning method (Random Forest) that then measures a CPS for each input training transcripts. It then classifies the *de novo* input sequence to be either coding or non-coding based on an optimal CPS cut-off that maximizes both sensitivity and specificity.

Computational details

To classify newly assembled transcripts into either protein-coding or lncRNAs, the alignment-free tool called FEElnc v0.1.1 was used. The first module (FEELnc_filter) was used to remove transcripts that were shorter that 200nt and then filter out unwanted/spurious transcripts that were either monoexonic or biexonic with one exon that was shorter that 25nt transcripts. Then all transcripts that overlapped exons of the reference annotation were removed. The FEELnc_codpot module was then used to compute the CPS (between 0 and 1) for each of the candidate transcripts that passed the previous module. This score was calculated based on the combination of a set of parameters described above. These predictor scores were then incorporated into a Random Forest model that computes a CPS for each input training transcripts. To calculate the CPS optimal cutoff, a 10 fold cross-validation on the input training set of known mRNAs and lncRNAs was used to extract the CPS that maximized both sensitivity and specificity and enabled classification of the newly assembled trancripts.

4.2.4 lncRNA classification

Rationale

To coherently predict lncRNA function in the plethora of newly identified transcripts one can classify these transcripts with respect to the localization and the direction of transcription of proximal protein-coding transcripts as shown in Figure 3.3.

Therefore, according to the GENCODE annotation (Harrow *et al.*, 2012) lncRNAs can be either:

- Sense lncRNAs - that are transcripts transcribed from the sense strand of protein-coding genes and therefore contain exons from protein-coding genes. They may overlap with part of protein-coding genes, or cover the entire sequence of a protein-coding gene.
- Intronic lncRNAs - that are lncRNA that are transcribed entirely from introns within protein-coding genes
- Antisense lncRNAs - have transcripts that intersect exons of a protein-coding but are transcribed from the antisense strand of protein-coding genes.
- Intergenic lncRNAs - that are those transcripts transcribed from intergenic regions that don’t overlap known genes.

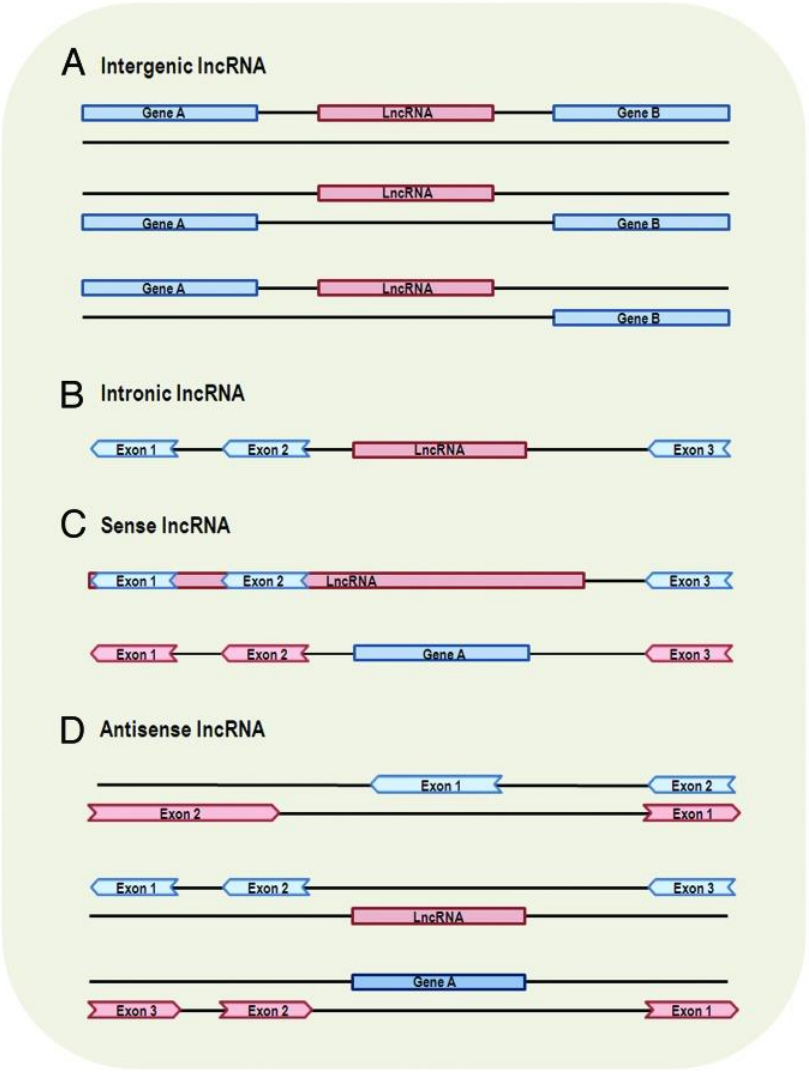


Figure 3.3 lncRNA classification. Protein-coding genes and their exons are shown in blue, while lncRNAs and their exons in red. (**A**) Intergenic lncRNA, from both strands. (**B**) Intronic lncRNA, are transcribed from introns of protein-coding genes. (**C**) Sense lncRNA, transcribed from the sense strand of protein-coding genes. (**D**) Antisense lncRNA, transcribed from the antisense strand of protein-coding genes (Ma, Bajic and Zhang, 2013).

Computational details

The last step of the pipeline consisted in classifying the newly lncRNAs with respect to the localization and the direction of transcription of proximal RNA transcripts. To classify these lncRNAs a sliding window approach was used to scan 10000nt around the lncRNA to define a possible overlap with the nearest transcripts from the reference annotation. A first level of classification discriminates between lncRNA that overlap an RNA gene from the reference annotation file (genic lncRNAs) and the ones that don't overlap a reference gene (intergenic lncRNAs (lincRNAs). Then, subtypes and locations were defined according to the orientation of

the interactions and the localisation of the interactions. In particular, the following sub classifications were made:

GENIC lncRNAs:

- overlapping subtype: the lncRNA partially overlaps the RNA partner transcript
- containing subtype: the lncRNA contains the RNA partner transcript
- nested subtype: the lncRNA is contained in the RNA partner transcript

then each of these is classified as either exonic or intronic locations.

INTERGENIC lincRNAs :

- divergent subtype: the lncRNA is transcribed in head to head orientation with RNA_partner
- convergent subtype: the lncRNA is oriented in tail to tail with orientation with RNA_partner
- same_strandsubtype: the lncRNA is transcribed in the same orientation orientation with RNA_partner

then for each of these we then defined whether the transcript was upstream or downstream of the known gene. Finally, for each lncRNA interaction, a best lncRNA:RNA partner interaction was identified. For lincRNAs the best RNA partner is the closest to the lincRNA. Instead for genic lncRNAs the best RNA partner is first the exonic, then the ones that overlap a fraction of an exon, then the intronic and finally the containing. For this study only intergenic lncRNAs (lincRNAs) were selected.

4.2.5 Integrating lincRNAs in the modified reference annoation

After lincRNAs were defined for both the Cufflinks and the StringTie assembled transcripts, only lincRNAs that were identified by both assembles and that passed the FEELnc were integrated in the modified reference annotation created previously (see section 2.1.1). This new atlas has therefore, the standard reference genes in the GENCODE atlas, together with the lincRNAs identified in our study, lincRNAs identified by Cabili et al., (Cabili *et al.*, 2011), lincRNAs identified by Liu et al., (Liu *et al.*, 2016) and the ones identified in the Fantom project recently (Hon *et al.*, 2017). This new reference annotation was then used to remap read and for all analysis of differentiation expression and specificity described below as well as in the scRNA-seq analysis.

4.2.6 GC content bias removal

Rationale

GC content biases have been frequently reported in RNA-seq data (Boeva *et al.*, 2011; Risso *et al.*, 2011; Benjamini and Speed, 2012). These biases are created due to the different sequencing

efficiency of genomic regions (Risso *et al.*, 2011) and cause an underrepresentation of GC-rich and GC-poor fragments making counts not directly comparable between genes. These biases must be corrected as they confound differential expression analysis and therefore proper downstream interpretation. One of the few tools to do this type of normalization is called EDASeq (Risso *et al.*, 2011) that stratifies genes in equally-sized bins based on GC-content and then matches parameters of the count distributions across bins.

Computational details

For GC content normalization we initial retrieved the length and GC content of each gene of the reference annotation using the nuc tool from bedtools v2.27.0 (Quinlan and Hall, 2010). We then stratified the genes according to GC-content and performed full quantile normalization using EDASeq (Risso *et al.*, 2011).

4.2.7 Batch effect removal

Rationale

After GC content normalization we still had samples separating by batch and not be area. We therefore opted to perform batch effect removal using ComBat (Johnson, Li and Rabinovic, 2007) that uses empirical Bayes frameworks for adjusting data for batch effects, is robust to outliers and has been shown to work well with a small number of samples like ours (Johnson, Li and Rabinovic, 2007).

Computational details

To perform batch effect correction we used the *sva* package (Leek *et al.*, 2012). Since we knew our batch effect, the ComBat (Johnson, Li and Rabinovic, 2007) function was employed by using a parametric empirical Bayesian adjustment (par.prior = true) based on the 2 covariates that formed the batch (+/- glycogen in RNA sample preparation).

4.2.8 Differential expression analysis

Rationale

To determine the highly specific genes for each area analyzed in this study (Cx, LGE and MGE) a number of different tools can be used, however, the number of biological replicates determines the types of differential expression analysis. With a fair number of biological replicates (at least 5–10 biological replicates per group), a nonparametric method is more ideal as it does not make

assumptions about the form of the statistical distribution of the observed data. With few biological replicates (like in our experimental design), however, nonparametric methods are typically underpowered and in these cases, parametric methods, that assume a certain form of the distribution based on empirical data, are used. The most common package that uses this type of distribution is and DESeq2 (Love, Huber and Anders, 2014). This package is based on the concept of generalized linear models (GLM). Its principle is to model the expression of each gene as a linear combination of different variables. In particular, a GLM is a more flexible version of a standard linear model that allows the distribution of the response variable to be different from the normal distribution used in standard linear regression. The GLMs used in DESeq2 assumes that the read counts are distributed according to the negative binomial distribution. It then uses a median of ratios method, which accounts for sequencing depth and RNA composition to normalize the data. DESeq2 then proceeds with fitting the negative binomial model and then estimating the variance associated with biological variation. This variation is the coefficient of variation (CV) with which the (unknown) true abundance of the gene varies between replicate RNA samples. It represents the CV that would remain between biological replicates if sequencing depth could be increased indefinitely. The technical CV decreases as the size of the counts increases. BCV on the other hand does not, and for this reason, it is considered the dominant source of uncertainty, therefore reliable estimation of this parameter is crucial for realistic assessment of differential expression in RNA-seq experiments. To estimate the variance DESeq2 assumes that genes of similar average expression strength have similar dispersion. First every gene is treated separately and the gene-wise dispersion estimates are calculated (using maximum likelihood), this relies only on the data of each individual gene. Then, a curve is fitted to capture the dependence of these estimates on average expression strength. This provides an accurate estimate for the expected dispersion value for genes of a given expression strength but cannot represent deviations of individual genes from this overall trend. Finally, the gene-wise dispersion estimates are shrunk towards the values predicted by the curve to obtain final dispersion values (Anders and Huber, 2010). Therefore, this approach calculates gene-specific variation to the extent that the data provides this information, while the fitted curve helps to estimate variation in less information rich areas. The strength of shrinkage does not depend simply on the mean count, but rather on the amount of information available for the fold change estimation. Two genes with equal expression strength but different dispersions will experience different amount of shrinkage.

Once negative binomial models are fitted and dispersion estimates are obtained, DESeq2 tests for each model coefficient whether it differs significantly from zero. For significance testing, DESeq2 uses Wald tests: the shrunken estimate of LFC is divided by its standard error, resulting

in a z statistic that can be compared to a standard normal. Then the Wald test p-values are adjusted for multiple testing using the procedure of Benjamini and Hochberg to find the FDR. DESeq2 also detects automatically count outliers using Cooks's distance and removes these genes from analysis (Anders and Huber, 2010). It also automatically removes genes whose mean of normalized counts is below a threshold determined by an optimization procedure. Removing these genes with low counts improves the detection power by making the multiple testing adjustments of the p-values less severe.

Computational details

To determine a specific linRNA/protein-coding signature for each area and pcw DESeq2 was exploited. In particular, for each pcw every combination of comparisons (CTX vs LGE, CTX vs MGE, LGE vs MGE, LGE vs CTX, MGE vs CTX and MGE vs LGE) was tested. For each of these comparisons only significantly overexpressed or downregulated genes with a p-adjusted value < 0.05 were kept. To pinpoint a set of genes that were significant only for a certain region the list of genes from each set of comparisons (e.g. LGE vs MGE and LGE vs CTX) were compared and only the ones that were significant in both comparisons for the LGE were considered "LGE specific". To avoid excessive shrinking of genes with low expression (that may remove a significant number of lncRNAs) by DESeq2 (Love, Huber and Anders, 2014) we separated lncRNAs and tested them independently from the rest of the gene set.

4.2.9 Pathway Analysis

Computational details

To identify Gene Ontology (GO) terms related to LGE specific genes we used the ClueGo plugin of Cytoscape (Bindea *et al.*, 2009). The GO "Biological component" was used to query these genes against the background of all genes expressed in the fetal brain tissue. Only GO with a *P-adj* < 0.05 were considered (κ-score threshold = 0.4). We then calculated the semantic similarity score between GO terms using the the R package GOSemSim (Yu *et al.*, 2010). We then subjected the resulting matrix to hierarchical clustering to find the most represented GO terms. Functional annotation of the genes that were LGE specific was performed using Ingenuity Pathway Analysis (IPA; Ingenuity Systems, https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/). For upstream regulators the first 30 significant genes are shown for both the set of specific genes between 7 and 11pcw and the ones at 20pcw.

4.2.10 Tissue-specificity analysis

Computational details

To quantify tissue specificity of lncRNAs and protein-coding genes a probability distribution distance metric called Jensen-Shannon divergence (JSD) was employed (Cabili *et al.*, 2011). The metric quantifies the similarity between expression pattern in a given area/pcw and an extreme pattern that represents that a transcript is expressed in only one area/pcw. We calculated a JS score for each transcript and selected in which area/pcw it had the max JS score to represent where that particular transcript was specific. We then determined a specificity threshold that was based on the JS scores of genes that resulted highly specific (p-adj < 0.001 and $\log_2$FC > 2) in the DESeq2 analysis. We selected a specificity threshold for the JS score of 0.6 that represented the value for the 3rd quartile of specific genes.

4.3 Single-cell RNA-seq Bioinformatics Pipeline

4.3.1 Sample demultiplexing, barcode processing, and single cell 3' gene counting

To perform sample demultiplexing, barcode processing and single-cell 3′ gene counting the Cell Ranger Single-Cell Software was employed (http://software.10xgenomics.com/single-cell/overview/welcome). To identify reads deriving from each sample, the first step involved demultiplexing each sample based on the sample index read to generate FASTQs. Read2, which contains the cDNA fragment was then aligned to the reference annotation created in the previous steps (see section 2.1.1) using STAR (Dobin *et al.*, 2013). During this alignment reads were classified as either exonic, intronic or intergenic. A read was considered exonic if at least 50% of it intersected an exon, intronic if it was non-exonic and intersecting an intron, or intergenic otherwise. For reads that align to a single exonic locus but also align to 1 or more non-exonic loci, the exonic locus was prioritized and the read was considered to be confidently mapped to the exonic locus with MAPQ 255. Cell Ranger further aligns exonic reads to annotated transcripts. A read that was compatible with the exons of an annotated transcript, and aligned to the same strand, was considered mapped to the transcriptome. If the read was compatible with a single gene annotation, it was considered uniquely (confidently) mapped to the transcriptome. Only reads that were confidently mapped to the transcriptome were used for UMI counting.

Next, GemCode barcodes and UMIs were filtered. All cell barcodes that are 1 mismatch from an observed barcode were corrected to the candidate barcode. UMIs that were not homopolymers and with a sequencing quality score >10 (90% base call accuracy) were discareded. UMIs that were

1-nucleotide mismatch away from a higher-count UMI were corrected to the UMI with more reads, if they shared a cell barcode and gene.

Then if two sets of read pairs shared a barcode sequence, a UMI tag, and a gene ID they were marked as PCR duplicates and were removed from counts. Finally, GEMs that likely contained cells were selected. This was done by summing the UMI counts for each barcode/cell and then selecting the barcodes with a total UMI count $\geq$ 10% of the 99th percentile of the expected recovered cells. This allowed us to recover cells with counts with the same order of magnitude. Barcodes that had a lower values of cell counts were considered backround. This step was done as we assume that RNA content varies by roughly an order of magnitude among cells in a sample, which is based on an empirical observation.

4.3.2 Quality control and data normalization

Rationale

One of the first steps is to check for the quality of the cells. Different variables can influence the integrity and quality of the transcriptome, for example incomplete cell lysis or failures during library preparation can cause altered transcriptional profiles that confounds analyses. In line, one must always evaluate the total number of transcripts detected, the total number of genes found to be expressed and the fraction of expression contributed by mitochondrial genes (Griffiths, Scialdone and Marioni, 2018). Any cells that display an inconsistent profile in any of these three domain must be removed from the analysis. However, the clear cut-off for the first two parameters is not easily established as cells with a divergent number of genes/transcripts may be found in a highly heterogeneous population (Ilicic *et al.*, 2016). Another crucial element to consider is that with droplet technologies cells may be captured in doublets and not alone. To remove such cells one can either remove cells expressing sets of biologically mutually exclusive markers (Ibarra-Soria *et al.*, 2018) or identify cells with an abnormally large library size (Bach *et al.*, 2017).

Another aspect that must be considered is whether confounding factors are present in the data which can contribute to biases in data distribution. These include: batch effects (Lun and Marioni, 2017), sequencing depth which can be controlled by normalization (Vallejos *et al.*, 2017) and cell-cycle stage than can confound underlying transcriptional architecture (Buettner *et al.*, 2015).

With regards to normalization different approaches are available, however all have their pitfalls and advantages and no true decision for the best practice has been reached. One approach to normalize data is to use bulk RNA-seq methods that then account for artifacts specific to scRNA-seq in downstream analysis. This includes methods like, SCDE that uses reads per million (RPM) (Kharchenko, Silberstein and Scadden, 2014) and then uses a particular model to capture dropout

events to reduce technical variation, and MAST that uses Transcripts Per Kilobase Million (TPM) (Finak *et al.*, 2015) and then uses a fraction of genes that are detectably expressed in each cell to model technical and biological sources of variation. Another way to normalize data is to use methods that were built around single-cell data. Examples include Scran (Lun, McCarthy and Marioni, 2016) that initially groups cells into clusters of similar expression and then computes size factors that are used to scale the counts in each cell within each cluster and then the factors are rescaled by normalization between clusters. This method allows cell-specific size factors to be estimated more robustly in the presence of zero inflation and unbalanced differential expression of genes across groups of cells. Another example is BASiCS (Vallejos, Marioni and Richardson, 2015) that is an integrated Bayesian hierarchical model that also estimates cell-specific normalization constants. Recently, the newly developed Seurat pipeline (Butler *et al.*, 2018b) implements a global-scaling normalization method that normalizes the gene expression measurements for each cell by the total expression, multiplies this by a scale factor and log-transforms the result. Understanding which method is the most appropriate is still a work in progress. A study (Vallejos *et al.*, 2017) showed that methods like Scran and BASICS lead to similar results in terms of scaling factor estimation and highly variable gene selection and are more robust to features of scRNA-seq (like a zero-inflated matrix) data compared to bulk-based approaches (Vallejos *et al.*, 2017). However, since a plethora of studies have being following the Seurat Pipeline (Butler *et al.*, 2018b) we opted to follow the standard "workflow" of this pipeline to initially characterize the data and then evaluate potential other normalization methods and their effects on the results.

Finally, to complete the "cleaning" step of the data, the last step is to detect genes that vary more between cells than would be expected by chance, that are called highly variable genes (HVGs) (Brennecke *et al.*, 2013). This principle avoids prioritizing low-abundance genes that have large variances due to technical noise and enables true biological variation to surface. However, since heteroscedasticity (high dispersion) is observed in this type of data (Law *et al.*, 2014) different methods have been developed that evaluate the variance taking into account the mean expression of genes. For example, Seurat (Satija *et al.*, 2015) calculates the variance divided by mean, scVEGs (H.-I. H. Chen *et al.*, 2016) uses the coefficient of variation while Brennecke (Brennecke *et al.*, 2013) uses the squared coefficient of variation. A study (Yip, Sham and Wang, 2018) compared different tools to detect HVGs and showed that although advantages and disadvantages are found in each method, Seurat (Satija *et al.*, 2015) performs well in terms of runtime and shows stable performance on different types of real datasets making us opt to follow this method also for HVGs detection.

Computational Details

Raw gene expression matrices generated per sample using CellRanger, as described above, were converted to an AnnData object using Scanpy (Wolf, Angerer and Theis, 2018) in Python (v1.2.0 of Scanpy and v0.22.0 Pandas). From this object, all genes expressed in less than 3 cells and all cells with less than 200 detected genes were removed. We then did a second round of filtering and removed cells with unique gene counts under 500 or over 6000. We then normalized gene expression for each cell by the total expression, then multipled by a scale factor of 10,000 and log-transformed the results. To determine highly variable genes, we selected those that had an average normalized expression between 0.0125 and 3 and a dispersion greater than 0.5. Finally, we regressed out variation in gene expression driven by the number of detected molecules and mitochondrial gene expression to avoid these variations to dominate downstream analysis and bias the final interpretation of the data.

4.3.3 Clustering to determine major cell types

Rationale

Different groups have implemented a range of unsupervised clustering methods. Their performance, however varies depending on high-depth versus low-depth sequencing data (Menon, 2017). For example the Seurat package (Satija *et al.*, 2015; Shekhar *et al.*, 2016) first reduces dimensionality using PCA, and then clusters cells in PCA space using the Jaccard overlap to compute a cell–cell distance and the Louvain algorithm to identify communities of cells and this method performs better on low read depth and high cell number data (Menon, 2017). Instead the iterative Weighted Gene Co-Expression Network Analysis (iWGCNA) approach (Tasic *et al.*, 2016; Yao *et al.*, 2017) initially clusters genes into modules (Langfelder and Horvath, 2008) and then groups cells into groups in eigengene coordinate space. This approach is similar to PCA, as they both use linear combinations of genes to group cells and works more efficiently on cells that have a deeper sequencing but are smaller in number (Menon, 2017). Another method that works better on deeper sequing is BackSPIN algorithm (Zeisel *et al.*, 2015) that is a biclustering-based approach that iteratively divides cells and genes into subgroups so as to maximize their separation (Zeisel *et al.*, 2015).

Computational Details

To reduce dimensionality of this dataset, cells that passed QC and the resulting 1797 variably expressed genes were subjected to principle component analysis. To select only PCs that explain

a statistically significant proportion of the variance we performed a permutation test (Macosko *et al.*, 2015) and identified 60 statistically significant PCs. The top 10 significant PCs were further used to construct a K-nearest neighbor (KNN) graph and then the Louvain algorithm was used to cluster cells together and define highly interconnected cell communities (Blondel *et al.*, 2008; Macosko *et al.*, 2015). We then performed t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van Der Maaten and Hinton, 2008) on the 10PCs 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 gene for each cluster were found by ranking genes with a t test.

4.3.4 Trajectory inference and pseudotime estimation

Rationale

Although discrete cell states are highly informative, this technology also enables, as mentioned above, to find transition paths between stable states, reconstruct cellular trajectories and pinpoint genes triggering state transitions during development. The methods to infer such transition, known as *pseudotime* methods, interpret an ordering of cells according to some characteristic in the data. From a computational point of view, there are different ways to approach the problem and a plethora of different tools have been developed to try and infer cell trajectories based on the assumption that cells with similar expression profiles arise from the same lineage. All methods are based on two common steps: the first is to reduce the dimension of the data, the second is to find a smooth progression through the low dimensional data, assuming that cells that are nearby one another in the low dimensional space have similar expression patterns. However, all these tools are based on different principles, for example if the topology of the trajectory is inferred computationally, or is imposed by the design. Furthermore, some methodologies need prior information on where the trajectory originates. This obviously implies that the output trajectories will highly depend on the model imposed by the user (Saelens *et al.*, 2018).

The first group that introduced such a tool was a team led by John Rinn that developed a package called Monocle (Trapnell *et al.*, 2014) that is based on using independent-component analysis (ICA) to project cells in a low dimensional space and then constructing of a minimum spanning tree (MST) on the cells based of their transcriptomic profiles (Trapnell *et al.*, 2014). Initially the main limitation of this algorithm was that it did not allow for bifurcations in the lineage prediction. Monocle2 (Qiu *et al.*, 2017) has now resolved this issue, however, it still forces cells into a tree-like topology and does not provide a statistical measure for the fitting. Similar algorithms to Monocle are SLICE (Guo *et al.*, 2017), TSCAN (Ji and Ji, 2016), and Waterfall (Shin *et al.*, 2015).

In contrast to Monocle2, these methods share the advantage of using predefined clusters of cells rather than single cells to determine trajectories and therefore are less sensitive to outliers, but they still suffer from a lack of a statistical model to infer significant trajectories. Another class of algorithms recently developed include Wanderlust (Bendall *et al.*, 2014) and Wishbone (Setty *et al.*, 2016) that are all based on similar concepts as above but use k-nearest neighbor graphs (k-NNGs) to link similar cells to each other. However, the main limitation of the methods described above is that they assume that the data is structured in a tree-lie topology with a beginning and end and don't allow disconnected group of cells but, instead try to "force" cells to be connected. This is limiting since it is difficult to evaluate how data should be structured, especially if the tissue of interest is not well documented. Recently a new method named partition-based graph abstraction (PAGA) (Wolf *et al.*, 2017) was developed that enables trajectory estimates for both continuous and discontinuous structured data. In particular, PAGA determines a trajectory by treating predefined group of cells as nodes and quantifies the connectivity between groups of cells enabling the creation of different paths and by statistically weighing the strength of each path (Figure 3.4).

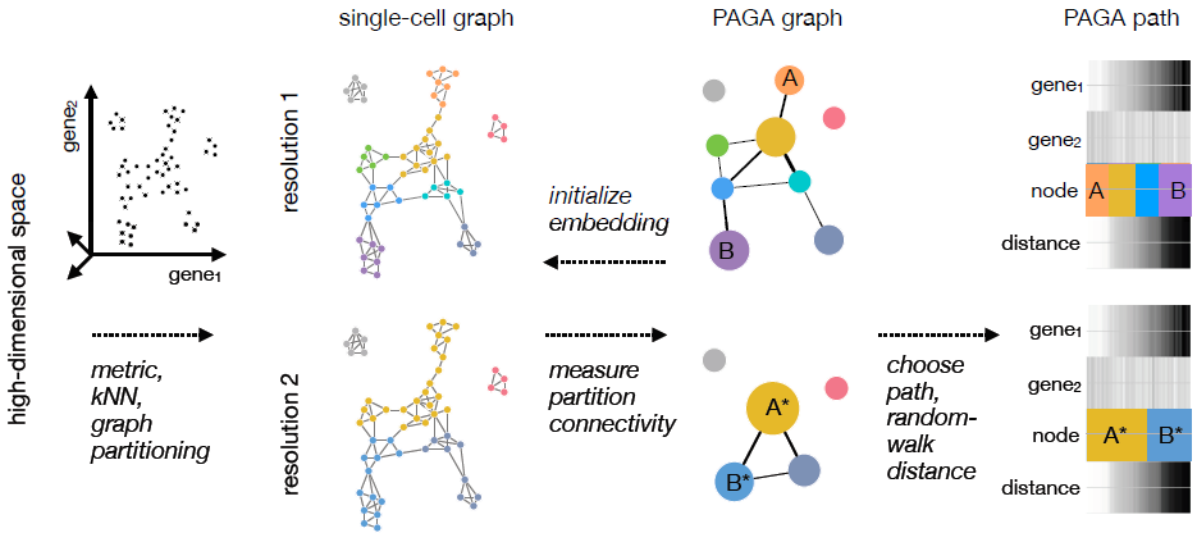


Figure 3.4 A topology preserving map of single cells Left graph illustrates a sketch of high-dimensional gene expression data and a partitioned k-nearest-neighbor graph with different resolutions (middle, left graph). Nodes represent single cells and colors represent groups of similar cells. Middle, left graph shows abstracted graph whose nodes correspond to groups of similar cells and whose edges represent statistically estimated connectivity between groups of cells. Right heatmap shows a measure (distance) according to their distance from a predefined root cell (Wolf *et al.*, 2017).

In light of the fact that we dissect an area of the developing telencephalon that is not easily recognizable and that may include some cells from the neighboring neocortex, MGE and CGE we

opted to start are trajectory reconstruction with PAGA since we expect to find cluster of cells from these surrounding areas to be detached from the main lineage tree of the developing LGE.

Computational Details

We used the PAGA (Wolf *et al.*, 2017) method for inferring the lineage tree of the developing LGE. PAGA analysis was used to measure the connectivity between the clusters that were predefined using the Louvain algorithm described above. We used a resolution equal to 0.15 to evaluate different potential relationships between the main cell communities identified. We then visualized the resulting abstracted graphs with Reingold-Tilford algorithm (layout=rt) to enable a hypothetical hierarchical structure to emerge between cell types. We then calculated the diffusion pseudotime (DPT) (Haghverdi *et al.*, 2016) on the LGE lineage by defining the "NES_NKX2.1" as the route (earliest) progenitor based on known markers.

5 | References

Achim, K. *et al.* (2015) 'High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin', *Nature Biotechnology*, 33(5), pp. 503–509. doi: 10.1038/nbt.3209.

van Agtmael, T. *et al.* (2010) 'Col4a1 mutation in mice causes defects in vascular function and low blood pressure associated with reduced red blood cell volume', *Human Molecular Genetics*. doi: 10.1093/hmg/ddp584.

Akaogi, K. *et al.* (1996) 'Specific accumulation of tumor-derived adhesion factor in tumor blood vessels and in capillary tube-like structures of cultured vascular endothelial cells.', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.93.16.8384.

Alifragis, P. (2004) 'Lhx6 Regulates the Migration of Cortical Interneurons from the Ventral Telencephalon But Does Not Specify their GABA Phenotype', *Journal of Neuroscience*, 24(24), pp. 5643–5648. doi: 10.1523/JNEUROSCI.1245-04.2004.

Allison, D. B. *et al.* (2006) 'Microarray data analysis: From disarray to consolidation and consensus', *Nature Reviews Genetics*, pp. 55–65. doi: 10.1038/nrg1749.

Alvarez-Dominguez, J. R. *et al.* (2014) 'Global discovery of erythroid long noncoding RNAs reveals novel regulators of red cell maturation', *Blood*, 123(4), pp. 570–581. doi: 10.1182/blood-2013-10-530683.

Alvarez-Dominguez, J. R. *et al.* (2015) 'De Novo Reconstruction of Adipose Tissue Transcriptomes Reveals Long Non-coding RNA Regulators of Brown Adipocyte Development', *Cell Metabolism*, 21(5), pp. 764–776. doi: 10.1016/j.cmet.2015.04.003.

Alzu'bi, A. *et al.* (2017) 'Distinct cortical and sub-cortical neurogenic domains for GABAergic interneuron precursor transcription factors NKX2.1, OLIG2 and COUP-TFII in early fetal human telencephalon', *Brain Structure and Function*, 222(5), pp. 2309–2328. doi: 10.1007/s00429-016-1343-5.

Anders, S. *et al.* (2014) 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.', *PLoS ONE*, 8(1), pp. 1–13. doi: 10.1038/nbt.1621.

Anders, S. and Huber, W. (2010) 'Differential expression analysis for sequence count data', *Genome Biology*, 11(10). doi: 10.1186/gb-2010-11-10-r106.

Anderson, K. D. and Reiner, A. (1991) 'Immunohistochemical localization of DARPP-32 in striatal projection neurons and striatal interneurons: implications for the localization of D1-like dopamine receptors on different types of striatal neurons', *Brain Research*, 568(1–2), pp. 235–243. doi: 10.1016/0006-8993(91)91403-N.

Andrews, S. (2010) *FastQC: A quality control tool for high throughput sequence data.*, *Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/*. doi: citeulike-article-id:11583827.

Aprea, J. *et al.* (2013) 'Transcriptome sequencing during mouse brain development identifies long non-coding RNAs functionally involved in neurogenic commitment', *EMBO Journal*, 32(24), pp. 3145–3160. doi: 10.1038/emboj.2013.245.

Araujo, D. J. *et al.* (2015) 'FoxP1 orchestration of ASD-relevant signaling pathways in the striatum', *Genes and Development*, 29(20), pp. 2081–2096. doi: 10.1101/gad.267989.115.

Arlotta, P. *et al.* (2005) 'Neuronal subtype-specific genes that control corticospinal motor neuron development in vivo', *Neuron*, 45(2), pp. 207–221. doi: 10.1016/j.neuron.2004.12.036.

Arlotta, P. *et al.* (2008) 'Ctip2 Controls the Differentiation of Medium Spiny Neurons and the Establishment of the Cellular Architecture of the Striatum', *Journal of Neuroscience*, 28(3), pp. 622–632. doi: 10.1523/JNEUROSCI.2986-07.2008.

Au, K. F. *et al.* (2010) 'Detection of splice junctions from paired-end RNA-seq data by SpliceMap', *Nucleic Acids Research*, 38(14), pp. 4570–4578. doi: 10.1093/nar/gkq211.

Ayupe, A. C. *et al.* (2015) 'Global analysis of biogenesis, stability and sub-cellular localization of lncRNAs mapping to intragenic regions of the human genome', *RNA Biology*, 12(8), pp. 877–892. doi: 10.1080/15476286.2015.1062960.

Bach, K. *et al.* (2017) 'Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing', *Nature Communications*, 8(1). doi: 10.1038/s41467-017-02001-5.

Bachoud-Lévi, A. C. *et al.* (2000) 'Motor and cognitive improvements in patients with Huntington's disease after neural transplantation.', *Lancet*, 356(9246), pp. 1975–9. doi: 10.1016/S0140-6736(00)03310-9.

Baker, S. C. *et al.* (2005) 'The external RNA controls consortium: A progress report', *Nature Methods*, 2(10), pp. 731–734. doi: 10.1038/nmeth1005-731.

Bansal, M. *et al.* (2007) 'How to infer gene networks from expression profiles', *Molecular Systems Biology*, 3. doi: 10.1038/msb4100120.

Baran-Gale, J., Chandra, T. and Kirschner, K. (2017) 'Experimental design for single-cell RNA sequencing', *Briefings in Functional Genomics*. doi: 10.1093/bfgp/elx035.

Barker, R. A. *et al.* (2017) 'Human Trials of Stem Cell-Derived Dopamine Neurons for Parkinson's Disease: Dawn of a New Era', *Cell Stem Cell*, pp. 569–573. doi: 10.1016/j.stem.2017.09.014.

Batista, P. J. and Chang, H. Y. (2013) 'Long noncoding RNAs: Cellular address codes in development and disease', *Cell*, pp. 1298–1307. doi: 10.1016/j.cell.2013.02.012.

Baydyuk, M. *et al.* (2011) 'TrkB receptor controls striatal formation by regulating the number of newborn striatal neurons', *Proceedings of the National Academy of Sciences*, 108(4), pp. 1669–1674. doi: 10.1073/pnas.1004744108.

Baydyuk, M. *et al.* (2013) 'Midbrain-Derived Neurotrophins Support Survival of Immature Striatal Projection Neurons', *Journal of Neuroscience*, 33(8), pp. 3363–3369. doi: 10.1523/JNEUROSCI.3687-12.2013.

Bedogni, F. *et al.* (2010) 'Tbr1 regulates regional and laminar identity of postmitotic neurons in developing neocortex', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1002285107.

Belgard, T. G. *et al.* (2011) 'A transcriptomic atlas of mouse neocortical layers', *Neuron*, 71(4), pp. 605–616. doi: 10.1016/j.neuron.2011.06.039.

Bendall, S. C. *et al.* (2014) 'Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development', *Cell*, 157(3), pp. 714–725. doi: 10.1016/j.cell.2014.04.005.

Benjamini, Y. and Speed, T. P. (2012) 'Estimation and correction for GC-content bias in high throughput sequencing', *Nucleic Acid Research*. Available at: http://stat-www.berkeley.edu/tech-reports/804.pdf.

Bergsland, M. *et al.* (2006) 'The establishment of neuronal properties is controlled by Sox4 and Sox11', *Genes and Development*, 20(24), pp. 3475–3486. doi: 10.1101/gad.403406.

Bernard, D. *et al.* (2010) 'A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression', *EMBO Journal*, 29(18), pp. 3082–3093. doi: 10.1038/emboj.2010.199.

Bhargava, V. *et al.* (2015) 'Technical Variations in Low-Input RNA-seq Methodologies', *Scientific Reports*, 4. doi: 10.1038/srep03678.

Bindea, G. *et al.* (2009) 'ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks', *Bioinformatics*, 25(8), pp. 1091–1093. doi: 10.1093/bioinformatics/btp101.

Blanco, E., Parra, G. and Guigó, R. (2007) 'Using geneid to Identify Genes', in *Current Protocols in Bioinformatics*. doi: 10.1002/0471250953.bi0403s18.

Blondel, V. D. *et al.* (2008) 'Fast unfolding of communities in large networks', *Journal of Statistical Mechanics: Theory and Experiment*. doi: 10.1088/1742-5468/2008/10/P10008.

Boeva, V. *et al.* (2011) 'Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization', *Bioinformatics*, 27(2), pp. 268–269. doi: 10.1093/bioinformatics/btq635.

Bohnert, R. and Rätsch, G. (2010) 'rQuant.web: A tool for RNA-Seq-based transcript quantitation', *Nucleic Acids Research*, 38(SUPPL. 2). doi: 10.1093/nar/gkq448.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: A flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

Bond, A. M. *et al.* (2009) 'Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry', *Nature Neuroscience*, 12(8), pp. 1020–1027. doi: 10.1038/nn.2371.

Bormuth, I. *et al.* (2013) 'Neuronal Basic Helix-Loop-Helix Proteins Neurod2/6 Regulate Cortical Commissure Formation before Midline Interactions', *Journal of Neuroscience*, 33(2), pp. 641–651. doi: 10.1523/JNEUROSCI.0899-12.2013.

Boulland, J. L. *et al.* (2004) 'Expression of the vesicular glutamate transporters during development indicates the widespread corelease of multiple neurotransmitters', *Journal of Comparative Neurology*. doi: 10.1002/cne.20354.

Breiman, L. (2001) 'Random Forests', *Machine Learning*, 45(1), pp. 5–32. doi: 10.1023/A:1010933404324.

Brennecke, P. *et al.* (2013) 'Accounting for technical noise in single-cell RNA-seq experiments', *Nature Methods*, 10(11), pp. 1093–1098. doi: 10.1038/nmeth.2645.

Brent, M. R. (2007) 'How does eukaryotic gene prediction work?', *Nature Biotechnology*, pp. 883–885. doi: 10.1038/nbt0807-883.

Brown, J. P. *et al.* (2003) 'Transient Expression of Doublecortin during Adult Neurogenesis', *Journal of Comparative Neurology*. doi: 10.1002/cne.10874.

Buettner, F. *et al.* (2015) 'Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells', *Nature Biotechnology*, 33(2), pp. 155–160. doi: 10.1038/nbt.3102.

Bullard, J. H. *et al.* (2010) 'Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments', *BMC Bioinformatics*, 11. doi: 10.1186/1471-2105-11-94.

Burns, J. C. *et al.* (2015) 'Single-cell RNA-Seq resolves cellular complexity in sensory organs from the neonatal inner ear', *Nature Communications*, 6. doi: 10.1038/ncomms9557.

Burrows, M. and Wheeler, D. (1994) 'A block-sorting lossless data compression algorithm', *Algorithm, Data Compression*, (124), p. 18. doi: 10.1.1.37.6774.

Burset, M., Seledtsov, I. A. and Solovyev, V. V (2000) 'Analysis of canonical and non-canonical splice sites in mammalian genomes.', *Nucleic Acids Research*, 28(21), pp. 4364–4375. doi: 10.1093/nar/28.21.4364.

Bussotti, G. *et al.* (2011) 'BlastR-fast and accurate database searches for non-coding RNAs', *Nucleic Acids Research*, 39(16), pp. 6886–6895. doi: 10.1093/nar/gkr335.

Butler, A. *et al.* (2018a) 'Integrating single-cell transcriptomic data across different conditions, technologies, and species', *Nature Biotechnology*, 36(5), pp. 411–420. doi: 10.1038/nbt.4096.

Butler, A. *et al.* (2018b) 'Integrating single-cell transcriptomic data across different conditions, technologies, and species', *Nature Biotechnology*. doi: 10.1038/nbt.4096.

Butt, S. J. B. *et al.* (2005) 'The temporal and spatial origins of cortical interneurons predict their physiological subtype', *Neuron*, 48(4), pp. 591–604. doi: 10.1016/j.neuron.2005.09.034.

Bylund, M. *et al.* (2003) 'Vertebrate neurogenesis is counteracted by Sox1-3 activity', *Nature Neuroscience*, 6(11), pp. 1162–1168. doi: 10.1038/nn1131.

Cabili, M. *et al.* (2011) 'Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses', *Genes and Development*, 25(18), pp. 1915–1927. doi: 10.1101/gad.17446611.

Cabili, M. N. *et al.* (2015) 'Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution', *Genome Biology*. doi: 10.1186/s13059-015-0586-4.

Carri, A. D. *et al.* (2013) 'Developmentally coordinated extrinsic signals drive human pluripotent stem cell differentiation toward authentic DARPP-32+ medium-sized spiny neurons', *Development*, 140(2), pp. 301–312. doi: 10.1242/dev.084608.

Casarosa, S., Fode, C. and Guillemot, F. (1999) 'Mash1 regulates neurogenesis in the ventral telencephalon.', *Development (Cambridge, England)*, 126, pp. 525–534. doi: 10.1371/journal.pcbi.0020117.

Castiglioni, V. *et al.* (2018) 'Dynamic and Cell-Specific DACH1 Expression in Human Neocortical and Striatal Development', *Cerebral Cortex*. doi: 10.1093/cercor/bhy092.

Castro, D. S. *et al.* (2011) 'A novel function of the proneural factor Ascl1 in progenitor proliferation identified by genome-wide characterization of its targets', *Genes and Development*, 25(9), pp. 930–945. doi: 10.1101/gad.627811.

Chen, B. *et al.* (2017) 'Identification of fusion genes and characterization of transcriptome features in T-cell acute lymphoblastic leukemia', *Proceedings of the National Academy of Sciences*, p. 201717125. doi: 10.1073/pnas.1717125115.

Chen, G. *et al.* (2016) 'Single-cell analyses of X Chromosome inactivation dynamics and pluripotency during differentiation', *Genome Research*, 26(10), pp. 1342–1354. doi: 10.1101/gr.201954.115.

Chen, H.-I. H. *et al.* (2016) 'Detection of high variability in gene expression from single-cell RNA-seq profiling', *BMC Genomics*. BMC Genomics, 17(S7), p. 508. doi: 10.1186/s12864-016-2897-6.

Chen, J. *et al.* (2016) 'Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs', *Genome Biology*, 17(1). doi: 10.1186/s13059-016-0880-9.

Chermenina, M. *et al.* (2014) 'GDNF is important for striatal organization and maintenance of dopamine neurons grown in the presence of the striatum', *Neuroscience*, 270, pp. 1–11. doi: 10.1016/j.neuroscience.2014.04.008.

Clark, M. B. *et al.* (2012) 'Genome-wide analysis of long noncoding RNA stability', *Genome Research*, 22(5), pp. 885–898. doi: 10.1101/gr.131037.111.

Collier, S. P. *et al.* (2012) 'Cutting Edge: Influence of Tmevpg1, a Long Intergenic Noncoding RNA, on the Expression of Ifng by Th1 Cells', *The Journal of Immunology*, 189(5), pp. 2084–2088. doi: 10.4049/jimmunol.1200774.

Conesa, A. *et al.* (2016) 'A survey of best practices for RNA-seq data analysis', *Genome Biology*. doi: 10.1186/s13059-016-0881-8.

Conforti, P. *et al.* (2018) 'Faulty neuronal determination and cell polarization are reverted by modulating HD early phenotypes', *Proceedings of the National Academy of Sciences*, 115(4), pp. E762–E771. doi: 10.1073/pnas.1715865115.

Corbin, J. G. *et al.* (2000) 'The Gsh2 homeodomain gene controls multiple aspects of telencephalic development.', *Development (Cambridge, England)*, 127(23), pp. 5007–5020. Available at: papers2://publication/uuid/70AD126B-B1FA-4677-A38F-

96E67803BFEB%5Cnhttp://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=11060228&retmode=ref&cmd=prlinks%5Cnpapers2://publication/uuid/E207D484-477F-4A07-B4FA-88CD6DECA6C8.

Deacon, T. W., Pakzaban, P. and Isacson, O. (1994) 'The lateral ganglionic eminence is the origin of cells committed to striatal phenotypes: neural transplantation and developmental evidence', *Brain Research*, 668(1–2), pp. 211–219. doi: 10.1016/0006-8993(94)90526-6.

DeLong, M. and Wichmann, T. (2009) 'Update on models of basal ganglia function and dysfunction', *Parkinsonism and Related Disorders*, 15(SUPPL. 3). doi: 10.1016/S1353-8020(09)70822-3.

Derrien, T. *et al.* (2012) 'The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression', *Genome Research*, 22(9), pp. 1775–1789. doi: 10.1101/gr.132159.111.

Desban, M. *et al.* (1995) 'Heterogeneous topographical distribution of the striatonigral and striatopallidal neurons in the matrix compartment of the cat caudate nucleus', *Journal of Comparative Neurology*, 352(1), pp. 117–133. doi: 10.1002/cne.903520109.

Detmer, S. A. and Chan, D. C. (2007) 'Functions and dysfunctions of mitochondrial dynamics', *Nature Reviews Molecular Cell Biology*. doi: 10.1038/nrm2275.

Dinger, M. E. *et al.* (2008) 'Differentiating protein-coding and noncoding RNA: Challenges and ambiguities', *PLoS Computational Biology*. doi: 10.1371/journal.pcbi.1000176.

Djebali, S. *et al.* (2012) 'Landscape of transcription in human cells', *Nature*, 489(7414), pp. 101–108. doi: 10.1038/nature11233.

Dobin, A. *et al.* (2013) 'STAR: Ultrafast universal RNA-seq aligner', *Bioinformatics*, 29(1), pp. 15–21. doi: 10.1093/bioinformatics/bts635.

Dragatsis, I., Efstratiadis, A. and Zeitlin, S. (1998) 'Mouse mutant embryos lacking huntingtin are rescued from lethality by wild-type extraembryonic tissues.', *Development (Cambridge, England)*, 125(8), pp. 1529–1539.

Ebisuya, M. *et al.* (2008) 'Ripples from neighbouring transcription', *Nature Cell Biology*, 10(9), pp. 1106–1113. doi: 10.1038/ncb1771.

Edgren, H. *et al.* (2011) 'Identification of fusion genes in breast cancer by paired-end RNA-sequencing', *Genome Biology*, 12(1). doi: 10.1186/gb-2011-12-1-r6.

Eisenstat, D. D. *et al.* (1999) 'DLX-1, DLX-2, and DLX-5 expression define distinct stages of basal forebrain differentiation', *Journal of Comparative Neurology*, 414(2), pp. 217–237. doi: 10.1002/(SICI)1096-9861(19991115)414:2<217::AID-CNE6>3.0.CO;2-I.

Ellis, P. *et al.* (2004) 'SOX2, a persistent marker for multipotential neural stem cells derived from embryonic stem cells, the embryo or the adult', *Developmental Neuroscience*. doi: 10.1159/000082134.

Engreitz, J. M. *et al.* (2016) 'Local regulation of gene expression by lncRNA promoters, transcription and splicing', *Nature*, 539(7629), pp. 452–455. doi: 10.1038/nature20149.

Erlander, M. G. *et al.* (1991) 'Two genes encode distinct glutamate decarboxylases', *Neuron*. doi: 10.1016/0896-6273(91)90077-D.

Esumi, S. *et al.* (2008) 'Method for single-cell microarray analysis and application to gene-expression profiling of GABAergic neuron progenitors', *Neuroscience Research*, 60(4), pp. 439–451. doi: 10.1016/j.neures.2007.12.011.

Evans, A. E. *et al.* (2012) 'Molecular Regulation of Striatal Development: A Review', *Anatomy Research International*, 2012, pp. 1–14. doi: 10.1155/2012/106529.

Ewels, P. *et al.* (2016) 'MultiQC: Summarize analysis results for multiple tools and samples in a single report', *Bioinformatics*, 32(19), pp. 3047–3048. doi: 10.1093/bioinformatics/btw354.

Falcone, C. and Mallamaci, A. (2015) 'Tuning of neocortical astrogenesis rates by emx2 in neural stem cells', *Neural Regeneration Research*, pp. 550–551. doi: 10.4103/1673-5374.155418.

Fan, X. *et al.* (2018) 'Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell RNA-seq analysis', *Cell Research*. Springer US, (February). doi: 10.1038/s41422-018-0053-3.

Fantom Consortium, T. (2006) 'The Transcriptional Landscape of the Mammalian Genome', *Science*, 309(5740), pp. 1559–1563. doi: 10.1126/science.1112014.

Fasano, C. A. *et al.* (2010) 'Efficient Derivation of Functional Floor Plate Tissue from Human Embryonic Stem Cells', *Cell Stem Cell*, 6(4), pp. 336–347. doi: 10.1016/j.stem.2010.03.001.

Fertuzinhos, S. *et al.* (2009) 'Selective depletion of molecularly defined cortical interneurons in human holoprosencephaly with severe striatal hypoplasia', *Cerebral Cortex*, 19(9), pp. 2196–2207. doi: 10.1093/cercor/bhp009.

Finak, G. *et al.* (2015) 'MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data', *Genome Biology*, 16(1). doi: 10.1186/s13059-015-0844-5.

Fior, R. and Henrique, D. (2005) 'A novel hes5/hes6 circuitry of negative regulation controls Notch activity during neurogenesis', *Developmental Biology*, 281(2), pp. 318–333. doi: 10.1016/j.ydbio.2005.03.017.

Flames, N. *et al.* (2007) 'Delineation of Multiple Subpallial Progenitor Domains by the Combinatorial Expression of Transcriptional Codes', *Journal of Neuroscience*, 27(36), pp. 9682–9695. doi: 10.1523/JNEUROSCI.2750-07.2007.

Flandin, P., Kimura, S. and Rubenstein, J. L. R. (2010) 'The Progenitor Zone of the Ventral Medial Ganglionic Eminence Requires Nkx2-1 to Generate Most of the Globus Pallidus But Few Neocortical Interneurons', *Journal of Neuroscience*, 30(8), pp. 2812–2823. doi: 10.1523/JNEUROSCI.4228-09.2010.

Flynn, R. A. and Chang, H. Y. (2014) 'Long noncoding RNAs in cell-fate programming and reprogramming', *Cell Stem Cell*, 14(6), pp. 752–761. doi: 10.1016/j.stem.2014.05.014.

Fogarty, M. *et al.* (2007) 'Spatial Genetic Patterning of the Embryonic Neuroepithelium Generates GABAergic Interneuron Diversity in the Adult Cortex', *Journal of Neuroscience*, 27(41), pp. 10935–10946. doi: 10.1523/JNEUROSCI.1629-07.2007.

Fonseca, N. A. *et al.* (2012) 'Tools for mapping high-throughput sequencing data', *Bioinformatics*, pp. 3169–3177. doi: 10.1093/bioinformatics/bts605.

Frieda, K. L. *et al.* (2017) 'Synthetic recording and in situ readout of lineage information in single cells', *Nature*, 541(7635), pp. 107–111. doi: 10.1038/nature20777.

Frith, M. C., Bailey, T. L., *et al.* (2006) 'Discrimination of non-protein-coding transcripts from protein-coding mRNA', *RNA Biology*, 3(1), pp. 40–48. doi: 10.4161/rna.3.1.2789.

Frith, M. C., Forrest, A. R., *et al.* (2006) 'The abundance of short proteins in the mammalian proteome', *PLoS Genetics*, 2(4), pp. 515–528. doi: 10.1371/journal.pgen.0020052.

Galluzzi, L., Kepp, O. and Kroemer, G. (2012) 'Mitochondria: Master regulators of danger signalling', *Nature Reviews Molecular Cell Biology*. doi: 10.1038/nrm3479.

Gangemi, R. M. R. *et al.* (2006) 'Effects of Emx2 inactivation on the gene expression profile of neural precursors', *European Journal of Neuroscience*, 23(2), pp. 325–334. doi: 10.1111/j.1460-9568.2005.04559.x.

Garber, M. *et al.* (2011) 'Computational methods for transcriptome annotation and quantification using RNA-seq', *Nature Methods*, pp. 469–477. doi: 10.1038/nmeth.1613.

Garel, S. *et al.* (1999) 'Ebf1 controls early cell differentiation in the embryonic striatum.', *Development (Cambridge, England)*, 126, pp. 5285–5294. doi: 10.1242/dev.00416.

Gerfen, C. R. (1984) 'The neostriatal mosaic: Compartmentalization of corticostriatal input and striatonigral output systems', *Nature*, 311(5985), pp. 461–464. doi: 10.1038/311461a0.

Gerfen, C. R. (1992) 'The neostriatal mosaic: multiple levels of compartmental organization', *Trends in Neurosciences*, pp. 133–139. doi: 10.1016/0166-2236(92)90355-C.

Gerfen, C. R., Baimbridge, K. G. and Miller, J. J. (1985) 'The neostriatal mosaic: compartmental distribution of calcium-binding protein and parvalbumin in the basal ganglia of the rat and monkey.', *Proceedings of the National Academy of Sciences*, 82(24), pp. 8780–8784. doi: 10.1073/pnas.82.24.8780.

Gerfen, C. R. and Scott Young, W. (1988) 'Distribution of striatonigral and striatopallidal peptidergic neurons in both patch and matrix compartments: an in situ hybridization histochemistry and fluorescent retrograde tracing study', *Brain Research*, 460(1), pp. 161–167. doi: 10.1016/0006-8993(88)91217-6.

Ghanem, N. *et al.* (2008) 'Characterization of a distinct subpopulation of striatal projection neurons expressing the Dlx genes in the basal ganglia through the activity of the I56ii enhancer', *Developmental Biology*. doi: 10.1016/j.ydbio.2008.07.029.

Gokce, O. *et al.* (2016) 'Cellular Taxonomy of the Mouse Striatum as Revealed by Single-Cell RNA-Seq', *Cell Reports*, 16(4), pp. 1126–1137. doi: 10.1016/j.celrep.2016.06.059.

Gomez, J. A. *et al.* (2013) 'The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-γ locus', *Cell*, 152(4), pp. 743–754. doi: 10.1016/j.cell.2013.01.015.

Goodchild, R. E., Grundmann, K. and Pisani, A. (2013) 'New genetic insights highlight "old" ideas on motor dysfunction in dystonia', *Trends in Neurosciences*, pp. 717–725. doi: 10.1016/j.tins.2013.09.003.

Grabherr, M. G. *et al.* (2011) 'Full-length transcriptome assembly from RNA-Seq data without a reference genome', *Nature Biotechnology*, 29(7), pp. 644–652. doi: 10.1038/nbt.1883.

Grant, G. R. *et al.* (2011) 'Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM)', *Bioinformatics*, 27(18), pp. 2518–2528. doi: 10.1093/bioinformatics/btr427.

Graveland, G. A., Williams, R. S. and Difiglia, M. (1985) 'A Golgi study of the human neostriatum: Neurons and afferent fibers', *Journal of Comparative Neurology*, 234(3), pp. 317–333. doi: 10.1002/cne.902340304.

Graybiel, A. M. *et al.* (1981) 'An immunohistochemical study of enkephalins and other neuropeptides in the striatum of the cat with evidence that the opiate peptides are arranged to form mosaic patterns in register with the striosomal compartments visible by acetylcholinesterase stainin', *Neuroscience*, 6(3). doi: 10.1016/0306-4522(81)90131-7.

Greenberg, Z., Ramshaw, H. and Schwarz, Q. (2015) 'Time Windows of Interneuron Development: Implications to Our Understanding of the Aetiology and Treatment of Schizophrenia', *neuroscience 2015, Vol. 2, Pages 294-321*, 2(September), pp. 294–321. doi: 10.3934/Neuroscience.2015.4.294.

Griffith, M. *et al.* (2010) 'Alternative expression analysis by RNA sequencing', *Nature Methods*. doi: 10.1038/nmeth.1503.

Griffiths, J. A., Scialdone, A. and Marioni, J. C. (2018) 'Using single-cell genomics to understand developmental processes and cell fate decisions', *Molecular Systems Biology*, 14(4), p. e8046. doi: 10.15252/msb.20178046.

Grün, D., Kester, L. and Van Oudenaarden, A. (2014) 'Validation of noise models for single-cell transcriptomics', *Nature Methods*, 11(6), pp. 637–640. doi: 10.1038/nmeth.2930.

Guo, G. *et al.* (2010) 'Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst', *Developmental Cell*, 18(4), pp. 675–685. doi: 10.1016/j.devcel.2010.02.012.

Guo, M. *et al.* (2017) 'SLICE: Determining cell differentiation and lineage based on single cell entropy', *Nucleic Acids Research*, 45(7). doi: 10.1093/nar/gkw1278.

Guttman, M. *et al.* (2009) 'Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals', *Nature*, 458(7235), pp. 223–227. doi: 10.1038/nature07672.

Guttman, M. *et al.* (2010) 'Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs', *Nature Biotechnology*, 28(5), pp. 503–510. doi: 10.1038/nbt.1633.

Guttman, M. *et al.* (2011) 'LincRNAs act in the circuitry controlling pluripotency and differentiation', *Nature*, 477(7364), pp. 295–300. doi: 10.1038/nature10398.

Hacisuleyman, E. *et al.* (2016) 'Function and evolution of local repeats in the Firre locus', *Nature Communications*, 7. doi: 10.1038/ncomms11021.

Haerty, W. and Ponting, C. P. (2015) 'Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci', *RNA*, 21(3), pp. 320–332. doi: 10.1261/rna.047324.114.

Haghverdi, L. *et al.* (2016) 'Diffusion pseudotime robustly reconstructs lineage branching', *Nature Methods*. doi: 10.1038/nmeth.3971.

Hansen, D. V. *et al.* (2013) 'Non-epithelial stem cells and cortical interneuron production in the human ganglionic eminences', *Nature Neuroscience*, 16(11), pp. 1576–1587. doi: 10.1038/nn.3541.

Hansen, K. D., Brenner, S. E. and Dudoit, S. (2010) 'Biases in Illumina transcriptome sequencing caused by random hexamer priming', *Nucleic Acids Research*, 38(12). doi: 10.1093/nar/gkq224.

Harrow, J. *et al.* (2012) 'GENCODE: The reference human genome annotation for the ENCODE project', *Genome Research*, 22(9), pp. 1760–1774. doi: 10.1101/gr.135350.111.

Hezroni, H. *et al.* (2015) 'Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species', *Cell Reports*, 11(7), pp. 1110–1122. doi: 10.1016/j.celrep.2015.04.023.

Hon, C. C. *et al.* (2017) 'An atlas of human long non-coding RNAs with accurate 5′ ends', *Nature*, 543(7644), pp. 199–204. doi: 10.1038/nature21374.

Hu, G. *et al.* (2013) 'Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation', *Nat Immunol*, 14(11), pp. 1190–1198. doi: doi: 10.1038/ni.2712.

Ibáñez-Sandoval, O. *et al.* (2010) 'Electrophysiological and morphological characteristics and synaptic connectivity of tyrosine hydroxylase-expressing neurons in adult mouse striatum.', *The Journal of neuroscience : the official journal of the Society for Neuroscience*. doi: 10.1523/JNEUROSCI.5996-09.2010.

Ibarra-Soria, X. *et al.* (2018) 'Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation', *Nature Cell Biology*, 20(2), pp. 127–134. doi: 10.1038/s41556-017-0013-z.

Ilicic, T. *et al.* (2016) 'Classification of low quality cells from single-cell RNA-seq data', *Genome Biology*, 17(1). doi: 10.1186/s13059-016-0888-1.

Ince, E., Ciliax, B. J. and Levey, a I. (1997) 'Differential expression of D1 and D2 dopamine and m4 muscarinic acetylcholine receptor proteins in identified striatonigral neurons.', *Synapse (New York, N.Y.)*, 27(4), pp. 357–66. doi: 10.1002/(SICI)1098-2396(199712)27:4<357::AID-SYN9>3.0.CO;2-B.

Islam, S. *et al.* (2011) 'Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq', *Genome Research*, 21(7), pp. 1160–1167. doi: 10.1101/gr.110882.110.

Islam, S. *et al.* (2014) 'Quantitative single-cell RNA-seq with unique molecular identifiers', *Nature Methods*, 11(2), pp. 163–166. doi: 10.1038/nmeth.2772.

Jain, M. *et al.* (2001) 'Cellular and molecular aspects of striatal development', *Brain Research Bulletin*, pp. 533–540. doi: 10.1016/S0361-9230(01)00555-X.

Ji, Z. and Ji, H. (2016) 'TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis', *Nucleic Acids Research*, 44(13), p. e117. doi: 10.1093/nar/gkw430.

Johnson, R. and Guigo, R. (2014) 'The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs', *RNA*, 20(7), pp. 959–976. doi: 10.1261/rna.044560.114.

Johnson, W. E., Li, C. and Rabinovic, A. (2007) 'Adjusting batch effects in microarray expression data using empirical Bayes methods', *Biostatistics*, 8(1), pp. 118–127. doi: 10.1093/biostatistics/kxj037.

Kaikkonen, M. U., Lam, M. T. Y. and Glass, C. K. (2011) 'Non-coding RNAs as regulators of gene expression and epigenetics', *Cardiovascular Research*, pp. 430–440. doi: 10.1093/cvr/cvr097.

Kamachi, Y., Uchikawa, M. and Kondoh, H. (2000) 'Pairing SOX off: With partners in the regulation of embryonic development', *Trends in Genetics*, pp. 182–187. doi: 10.1016/S0168-9525(99)01955-1.

Kee, N. *et al.* (2017) 'Single-Cell Analysis Reveals a Close Relationship between Differentiating Dopamine and Subthalamic Nucleus Neuronal Lineages', *Cell Stem Cell*, 20(1), pp. 29–40. doi: 10.1016/j.stem.2016.10.003.

Kelsom, C. and Lu, W. (2013) 'Development and specification of GABAergic cortical interneurons', *Cell and Bioscience*. doi: 10.1186/2045-3701-3-19.

Keo, A. *et al.* (2017) 'Co-expression Patterns between ATN1 and ATXN2 Coincide with Brain Regions Affected in Huntington's Disease', *Frontiers in Molecular Neuroscience*, 10. doi: 10.3389/fnmol.2017.00399.

Kharchenko, P. V., Silberstein, L. and Scadden, D. T. (2014) 'Bayesian approach to single-cell differential expression analysis', *Nature Methods*, 11(7), pp. 740–742. doi: 10.1038/nmeth.2967.

Kim, D. *et al.* (2013) 'TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions', *Genome Biology*, 14(4). doi: 10.1186/gb-2013-14-4-r36.

Kim, J. K. *et al.* (2015) 'Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression', *Nature Communications*, 6. doi: 10.1038/ncomms9687.

Kirkeby, A. *et al.* (2012) 'Generation of Regionally Specified Neural Progenitors and Functional Neurons from Human Embryonic Stem Cells under Defined Conditions', *Cell Reports*, 1(6), pp. 703–714. doi: 10.1016/j.celrep.2012.04.009.

Kirkeby, A. *et al.* (2017) 'Predictive Markers Guide Differentiation to Improve Graft Outcome in Clinical Translation of hESC-Based Therapy for Parkinson's Disease', *Cell Stem Cell*, 20(1), pp. 135–148. doi: 10.1016/j.stem.2016.09.004.

Kivioja, T. *et al.* (2012) 'Counting absolute numbers of molecules using unique molecular identifiers', *Nature Methods*, 9(1), pp. 72–74. doi: 10.1038/nmeth.1778.

Kong, L. *et al.* (2007) 'CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine', *Nucleic Acids Research*, 35(SUPPL.2). doi: 10.1093/nar/gkm391.

Kozomara, A. and Griffiths-Jones, S. (2014) 'MiRBase: Annotating high confidence microRNAs using deep sequencing data', *Nucleic Acids Research*, 42(D1). doi: 10.1093/nar/gkt1181.

Kurimoto, K. *et al.* (2006) 'An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis', *Nucleic Acids Research*, 34(5). doi: 10.1093/nar/gkl050.

Kurimoto, K. *et al.* (2007) 'Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis', *Nature Protocols*, 2(3), pp. 739–752. doi: 10.1038/nprot.2007.79.

Laclef, C. and Métin, C. (2018) 'Conserved rules in embryonic development of cortical interneurons', *Seminars in Cell and Developmental Biology*, pp. 86–100. doi: 10.1016/j.semcdb.2017.09.017.

Lanciego, J. L., Luquin, N. and Obeso, J. A. (2012) 'Functional neuroanatomy of the basal ganglia', *Cold Spring Harbor Perspectives in Medicine*, 2(12). doi: 10.1101/cshperspect.a009621.

Langfelder, P. and Horvath, S. (2008) 'WGCNA: An R package for weighted correlation network analysis', *BMC Bioinformatics*, 9. doi: 10.1186/1471-2105-9-559.

Langmead, B. *et al.* (2009) 'Ultrafast and memory-efficient alignment of short DNA sequences to the human genome', *Genome biology*, 10(3), p. R25. doi: 10.1186/gb-2009-10-3-r25.

Latos, P. A. *et al.* (2012) 'Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing', *Science*, 338(6113), pp. 1469–1472. doi: 10.1126/science.1228110.

Lavdas, a a *et al.* (1999) 'The medial ganglionic eminence gives rise to a population of early neurons in the developing cerebral cortex.', *The Journal of neuroscience*, 19(19), pp. 7881–7888. doi: 10.1523/JNEUROSCI.19-18-07881.1999.

Law, C. W. *et al.* (2014) 'Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts', *Genome Biology*. doi: 10.1186/gb-2014-15-2-r29.

Leek, J. T. *et al.* (2012) 'The SVA package for removing batch effects and other unwanted variation in high-throughput experiments', *Bioinformatics*, 28(6), pp. 882–883. doi: 10.1093/bioinformatics/bts034.

Leone, D. P. *et al.* (2008) 'The determination of projection neuron identity in the developing cerebral cortex', *Current Opinion in Neurobiology*. doi: 10.1016/j.conb.2008.05.006.

Lewejohann, L. *et al.* (2004) 'Role of a neuronal small non-messenger RNA: Behavioural alterations in BC1 RNA-deleted mice', *Behavioural Brain Research*, 154(1), pp. 273–289. doi: 10.1016/j.bbr.2004.02.015.

Li, A., Zhang, J. and Zhou, Z. (2014) 'PLEK: A tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme', *BMC Bioinformatics*, 15(1). doi: 10.1186/1471-2105-15-311.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.

Li, X. *et al.* (2015) 'Quality control of RNA-seq experiments', in *RNA Bioinformatics*. doi: 10.1007/978-1-4939-2291-8_8.

Li, Y. *et al.* (2012) 'Conditional ablation of brain-derived neurotrophic factor-TrkB signaling impairs striatal neuron development', *Proceedings of the National Academy of Sciences*, 109(38), pp. 15491–15496. doi: 10.1073/pnas.1212899109.

Lim, R. G. *et al.* (2017) 'Developmental alterations in Huntington's disease neural cells and pharmacological rescue in cells and mice', *Nature Neuroscience*, 20(5), pp. 648–660. doi: 10.1038/nn.4532.

Lim, S.-J., Fiez, J. A. and Holt, L. L. (2014) 'How may the basal ganglia contribute to auditory categorization and speech perception?', *Frontiers in Neuroscience*. doi: 10.3389/fnins.2014.00230.

Lim, S. A. O., Kang, U. J. and McGehee, D. S. (2014) 'Striatal cholinergic interneuron regulation and circuit effects', *Frontiers in Synaptic Neuroscience*. doi: 10.3389/fnsyn.2014.00022.

Lin, M. *et al.* (2011) 'RNA-Seq of human neurons derived from iPS cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders', *PLoS ONE*, 6(9). doi: 10.1371/journal.pone.0023356.

Lin, M. F., Jungreis, I. and Kellis, M. (2011) 'PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions', *Bioinformatics*, 27(13). doi: 10.1093/bioinformatics/btr209.

Lin, N. *et al.* (2014) 'An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment', *Molecular Cell*, 53(6), pp. 1005–1019. doi: 10.1016/j.molcel.2014.01.021.

Lindvall, O. *et al.* (1989) 'Human Fetal Dopamine Neurons Grafted Into the Striatum in Two Patients With Severe Parkinson's Disease', *Archives of Neurology*, 46(6), p. 615. doi: 10.1001/archneur.1989.00520420033021.

Lindvall, O. *et al.* (1990) 'Grafts of fetal dopamine neurons survive and improve motor function in Parkinson's disease.', *Science (New York, N.Y.)*, 247(4942), pp. 574–577. doi: 10.1126/science.2105529.

Liodis, P. *et al.* (2007) 'Lhx6 Activity Is Required for the Normal Migration and Specification of Cortical Interneuron Subtypes', *Journal of Neuroscience*, 27(12), pp. 3078–3089. doi: 10.1523/JNEUROSCI.3055-06.2007.

Liu, J. K. *et al.* (1997) 'Dlx genes encode DNA-binding proteins that are expressed in an overlapping and sequential pattern during basal ganglia differentiation', *Developmental Dynamics*, 210(4), pp. 498–512. doi: 10.1002/(SICI)1097-0177(199712)210:4<498::AID-AJA12>3.0.CO;2-3.

Liu, S. J. *et al.* (2016) 'Single-cell analysis of long non-coding RNAs in the developing human neocortex', *Genome Biology*, 17(1). doi: 10.1186/s13059-016-0932-1.

Liu, S. and Trapnell, C. (2016) 'Single-cell transcriptome sequencing: recent advances and remaining challenges', *F1000Research*. doi: 10.12688/f1000research.7223.1.

Llorens-Bobadilla, E. *et al.* (2015) 'Single-Cell Transcriptomics Reveals a Population of Dormant Neural Stem Cells that Become Activated upon Brain Injury', *Cell Stem Cell*, 17(3), pp. 329–340. doi: 10.1016/j.stem.2015.07.002.

Lo, L. C. *et al.* (1991) 'Mammalian achaete-scute homolog 1 is transiently expressed by spatially restricted subsets of early neuroepithelial and neural crest cells', *Genes and Development*, 5(9), pp. 1524–1537. doi: 10.1101/gad.5.9.1524.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), pp. 1–21. doi: 10.1186/s13059-014-0550-8.

Lun, A. T. L. and Marioni, J. C. (2017) 'Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data.', *Biostatistics (Oxford, England)*, pp. 1–14. doi: 10.1093/biostatistics/kxw055.

Lun, A. T. L., McCarthy, D. J. and Marioni, J. C. (2016) 'A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor', *F1000Research*, 5, p. 2122. doi: 10.12688/f1000research.9501.2.

Luo, Y. *et al.* (2015) 'Single-cell transcriptome analyses reveal signals to activate dormant neural stem cells', *Cell*, 161(5), pp. 1175–1188. doi: 10.1016/j.cell.2015.04.001.

Ma, L. *et al.* (2012) 'Human embryonic stem cell-derived GABA neurons correct locomotion deficits in quinolinic acid-lesioned mice', *Cell Stem Cell*, 10(4), pp. 455–464. doi: 10.1016/j.stem.2012.01.021.

Ma, L., Bajic, V. B. and Zhang, Z. (2013) 'On the classification of long non-coding RNAs', *RNA Biology*. doi: 10.4161/rna.24604.

Ma, T. *et al.* (2012) 'A subpopulation of dorsal lateral/caudal ganglionic eminence-derived neocortical interneurons expresses the transcription factor Sp8', *Cerebral Cortex*, 22(9), pp. 2120–2130. doi: 10.1093/cercor/bhr296.

Ma, T. *et al.* (2013) 'Subcortical origins of human and monkey neocortical interneurons', *Nature Neuroscience*, 16(11), pp. 1588–1597. doi: 10.1038/nn.3536.

Van Der Maaten, L. and Hinton, G. (2008) 'Visualizing Data using t-SNE', *Journal of Machine Learning Research*. doi: 10.1007/s10479-011-0841-3.

Macosko, E. Z. *et al.* (2015) 'Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets', *Cell*, 161(5), pp. 1202–1214. doi: 10.1016/j.cell.2015.05.002.

La Manno, G. *et al.* (2016) 'Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells', *Cell*, 167(2), p. 566–580.e19. doi: 10.1016/j.cell.2016.09.027.

La Manno, G. *et al.* (2017) 'RNA velocity in single cells', *bioRxiv*. Available at: http://biorxiv.org/content/early/2017/10/19/206052.abstract.

Marioni, J. C. *et al.* (2008) 'RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays', *Genome Research*, 18(9), pp. 1509–1517. doi: 10.1101/gr.079558.108.

Martens, J. A., Laprade, L. and Winston, F. (2004) 'Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene', *Nature*, 429(6991), pp. 571–574. doi: 10.1038/nature02538.

Mason, H. A. (2005) 'Notch signaling coordinates the patterning of striatal compartments', *Development*, 132(19), pp. 4247–4258. doi: 10.1242/dev.02008.

Mattick, J. S., Taft, R. J. and Faulkner, G. J. (2010) 'A global view of genomic information - moving beyond the gene and the master regulator', *Trends in Genetics*, pp. 21–28. doi: 10.1016/j.tig.2009.11.002.

McIntyre, L. M. *et al.* (2011) 'RNA-seq: Technical variability and sampling', *BMC Genomics*, 12. doi: 10.1186/1471-2164-12-293.

Melé, M. *et al.* (2017) 'Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs', *Genome Research*, 27(1), pp. 27–37. doi: 10.1101/gr.214205.116.

Melé, M. and Rinn, J. L. (2016) '"Cat's Cradling" the 3D Genome by the Act of LncRNA Transcription', *Molecular Cell*, pp. 657–664. doi: 10.1016/j.molcel.2016.05.011.

Menon, V. (2017) 'Clustering single cells: a review of approaches on high-and low-depth single-cell RNA-seq data', *Briefings in Functional Genomics*. doi: 10.1093/bfgp/elx044.

Mercer, T. R. *et al.* (2010) 'Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation', *BMC Neuroscience*, 11. doi: 10.1186/1471-2202-11-14.

Min, J. W. *et al.* (2015) 'Identification of distinct tumor subpopulations in lung adenocarcinoma via single-cell RNA-seq', *PLoS ONE*, 10(8). doi: 10.1371/journal.pone.0135817.

Modarresi, F. *et al.* (2012) 'Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation', *Nature Biotechnology*, 30(5), pp. 453–459. doi: 10.1038/nbt.2158.

Molero, A. E. *et al.* (2016) 'Selective expression of mutant huntingtin during development recapitulates characteristic features of Huntington's disease', *Proceedings of the National Academy of Sciences*, 113(20), pp. 5736–5741. doi: 10.1073/pnas.1603871113.

Morrison, A. *et al.* (1996) 'In vitro and transgenic analysis of a human HOXD4 retinoid-responsive enhancer', *Development*, 122(6), pp. 1895–1907. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8674428.

Muñoz-Manchado, A. B. *et al.* (2018) 'Diversity of Interneurons in the Dorsal Striatum Revealed by Single-Cell RNA Sequencing and PatchSeq', *Cell Reports*. doi: 10.1016/j.celrep.2018.07.053.

Necsulea, A. *et al.* (2014) 'The evolution of lncRNA repertoires and expression patterns in tetrapods', *Nature*, 505(7485), pp. 635–640. doi: 10.1038/nature12943.

Ng, S. Y. *et al.* (2013) 'The long non coding RNA RMST interacts with SOX2 to regulate neurogenesis', *Molecular Cell*. Elsevier Inc., 51(3), pp. 349–359. doi: 10.1016/j.molcel.2013.07.017.

Ng, S. Y., Johnson, R. and Stanton, L. W. (2012) 'Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors', *EMBO Journal*, 31(3), pp. 522–533. doi: 10.1038/emboj.2011.459.

Nguyen, G. D. *et al.* (2013) 'Functions of Huntingtin in Germ Layer Specification and Organogenesis', *PLoS ONE*, 8(8). doi: 10.1371/journal.pone.0072698.

Nobrega-Pereira, S. *et al.* (2010) 'Origin and Molecular Specification of Globus Pallidus Neurons', *Journal of Neuroscience*, 30(8), pp. 2824–2834. doi: 10.1523/JNEUROSCI.4023-09.2010.

Nóbrega-Pereira, S. *et al.* (2008) 'Postmitotic Nkx2-1 Controls the Migration of Telencephalic Interneurons by Direct Repression of Guidance Receptors', *Neuron*. doi: 10.1016/j.neuron.2008.07.024.

Nolbrant, S. *et al.* (2017) 'Generation of high-purity human ventral midbrain dopaminergic progenitors for in vitro maturation and intracerebral transplantation', *Nature protocols*, 12(9), pp. 1962–1979. doi: 10.1038/nprot.2017.078.

Nopoulos, P. C. *et al.* (2011) 'Smaller intracranial volume in prodromal Huntington's disease: Evidence for abnormal neurodevelopment', *Brain*, 134(1), pp. 137–142. doi: 10.1093/brain/awq280.

Nowakowski, T. J. *et al.* (2017) 'Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex', *Science*, 358(6368), pp. 1318–1323. doi: 10.1126/science.aap8809.

Onorati, M. *et al.* (2014) 'Molecular and functional definition of the developing human striatum', *Nature Neuroscience*, 17(12), pp. 1804–1815. doi: 10.1038/nn.3860.

Orom, U. A. *et al.* (2010) 'Long noncoding RNAs with enhancer-like function in human cells', *Cell*, 143(1), pp. 46–58. doi: 10.1016/j.cell.2010.09.001.

Padovan-Merhar, O. and Raj, A. (2013) 'Using variability in gene expression as a tool for studying gene regulation', *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 5(6), pp. 751–759. doi: 10.1002/wsbm.1243.

Pan, Q. *et al.* (2008) 'Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing', *Nature Genetics*, 40(12), pp. 1413–1415. doi: 10.1038/ng.259.

Patel, A. P. *et al.* (2014) 'Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma', *Science*, 344(6190), pp. 1396–1401. doi: 10.1126/science.1254257.

Pert, C. B., Kuhar, M. J. and Snyder, S. H. (1976) 'Opiate receptor: autoradiographic localization in rat brain.', *Proceedings of the National Academy of Sciences of the United States of America*, 73(10), pp. 3729–33. doi: 10.1073/pnas.73.10.3729.

Pertea, M. *et al.* (2015) 'StringTie enables improved reconstruction of a transcriptome from RNA-seq reads', *Nature Biotechnology*, 33(3), pp. 290–295. doi: 10.1038/nbt.3122.

Pevny, L. H. *et al.* (1998) 'A role for SOX1 in neural determination.', *Development (Cambridge, England)*, 125(10), pp. 1967–78. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9550729.

Picelli, S. *et al.* (2013) 'Smart-seq2 for sensitive full-length transcriptome profiling in single cells', *Nature Methods*, 10(11), pp. 1096–1100. doi: 10.1038/nmeth.2639.

Pollen, A. A. *et al.* (2014) 'Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex', *Nature Biotechnology*, 32(10), pp. 1053–1058. doi: 10.1038/nbt.2967.

Pollen, A. A. *et al.* (2015) 'Molecular Identity of Human Outer Radial Glia during Cortical Development', *Cell*. doi: 10.1016/j.cell.2015.09.004.

Ponjavic, J., Ponting, C. P. and Lunter, G. (2007) 'Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs', *Genome Research*, 17(5), pp. 556–565. doi: 10.1101/gr.6036807.

Ponting, C. P., Oliver, P. L. and Reik, W. (2009) 'Evolution and Functions of Long Noncoding RNAs', *Cell*, pp. 629–641. doi: 10.1016/j.cell.2009.02.006.

Prasanth, K. V. and Spector, D. L. (2007) 'Eukaryotic regulatory RNAs: An answer to the "genome complexity" conundrum', *Genes and Development*, pp. 11–42. doi: 10.1101/gad.1484207.

Precious, S. V. *et al.* (2016) 'FoxP1 marks medium spiny neurons from precursors to maturity and is required for their differentiation', *Experimental Neurology*, 282, pp. 9–18. doi: 10.1016/j.expneurol.2016.05.002.

Prensa, L., Giménez-Amaya, J. M. and Parent, A. (1999) 'Chemical heterogeneity of the striosomal compartment in the human striatum', *Journal of Comparative Neurology*, 413(4), pp. 603–618. doi: 10.1002/(SICI)1096-9861(19991101)413:4<603::AID-CNE9>3.0.CO;2-K.

Qiu, X. *et al.* (2017) 'Reversed graph embedding resolves complex single-cell trajectories', *Nature Methods*, 14(10), pp. 979–982. doi: 10.1038/nmeth.4402.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: A flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26(6), pp. 841–842. doi: 10.1093/bioinformatics/btq033.

Ramos, A. D. *et al.* (2015) 'The long noncoding RNA Pnky regulates neuronal differentiation of embryonic and postnatal neural stem cells', *Cell Stem Cell*, 16(4), pp. 439–447. doi: 10.1016/j.stem.2015.02.007.

Ranzani, V. *et al.* (2015) 'The long intergenic noncoding RNA landscape of human lymphocytes highlights the regulation of T cell differentiation by linc-MAF-4', *Nature Immunology*, 16(3), pp. 318–325. doi: 10.1038/ni.3093.

Rataj-Baniowska, M. *et al.* (2015) 'Retinoic Acid Receptor Controls Development of Striatonigral Projection Neurons through FGF-Dependent and Meis1-Dependent Mechanisms', *Journal of Neuroscience*, 35(43), pp. 14467–14475. doi: 10.1523/JNEUROSCI.1278-15.2015.

Reinius, B. *et al.* (2016) 'Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq', *Nature Genetics*, 48(11), pp. 1430–1435. doi: 10.1038/ng.3678.

Reuter, I. *et al.* (2008) 'Long-term clinical and positron emission tomography outcome of fetal striatal transplantation in Huntington's disease', *Journal of Neurology, Neurosurgery and Psychiatry*, 79(8), pp. 948–951. doi: 10.1136/jnnp.2007.142380.

Rinn, J. L. and Chang, H. Y. (2012) 'Genome Regulation by Long Noncoding RNAs', *Annual Review of Biochemistry*, 81(1), pp. 145–166. doi: 10.1146/annurev-biochem-051410-092902.

Risso, D. *et al.* (2011) 'GC-Content Normalization for RNA-Seq Data', *BMC Bioinformatics*. BioMed Central Ltd, 12(1), p. 480. doi: 10.1186/1471-2105-12-480.

Roberts, A. *et al.* (2011) 'Improving RNA-Seq expression estimates by correcting for fragment bias', *Genome Biology*, 12(3). doi: 10.1186/gb-2011-12-3-r22.

Saelens, W. *et al.* (2018) 'A comparison of single-cell trajectory inference methods: towards more accurate and robust tools', *bioRxiv*, p. 276907. doi: 10.1101/276907.

Satija, R. *et al.* (2015) 'Spatial reconstruction of single-cell gene expression data', *Nature Biotechnology*, 33(5), pp. 495–502. doi: 10.1038/nbt.3192.

Schiffmann, S. N., Jacobs, O. and Vanderhaeghen, J. -J (1991) 'Striatal Restricted Adenosine A2 Receptor (RDC8) Is Expressed by Enkephalin but Not by Substance P Neurons: An In Situ Hybridization Histochemistry Study', *Journal of Neurochemistry*, 57(3), pp. 1062–1067. doi: 10.1111/j.1471-4159.1991.tb08257.x.

Schloss, P. D. (2010) 'The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies', *PLoS Computational Biology*, 6(7), p. 19. doi: 10.1371/journal.pcbi.1000844.

Setty, M. *et al.* (2016) 'Wishbone identifies bifurcating developmental trajectories from single-cell data', *Nature Biotechnology*, 34(6), pp. 637–645. doi: 10.1038/nbt.3569.

Sha, L. *et al.* (2012) 'Transcriptional regulation of neurodevelopmental and metabolic pathways by NPAS3', *Molecular Psychiatry*, 17(3), pp. 267–279. doi: 10.1038/mp.2011.73.

Sharp, P. M. *et al.* (1988) 'Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity', *Nucleic Acids Research*, 16(17), pp. 8207–8211. doi: 10.1093/nar/16.17.8207.

Shekhar, K. *et al.* (2016) 'Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics', *Cell*, 166(5), p. 1308–1323.e30. doi: 10.1016/j.cell.2016.07.054.

Shin, J. *et al.* (2015) 'Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis Resource Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis', *Cell Stem Cell*, 17(3), pp. 360–372. doi: 10.1016/j.stem.2015.07.013.

Skryabin, B. V *et al.* (2003) 'Neuronal untranslated BC1 RNA: targeted gene elimination in mice.', *Molecular and cellular biology*, 23(18), pp. 6435–6441. doi: 10.1128/MCB.23.18.6435-6441.2003.

Sousa, A. M. M. *et al.* (2017) 'Molecular and cellular reorganization of neural circuits in the human lineage', *Science*, 358(6366), pp. 1027–1032. doi: 10.1126/science.aan3456.

Srivastava, S. and Chen, L. (2010) 'A two-parameter generalized Poisson model to improve the analysis of RNA-seq data', *Nucleic acids research*, 38(17), p. e170. doi: 10.1093/nar/gkq670.

Stegle, O., Teichmann, S. A. and Marioni, J. C. (2015) 'Computational and analytical challenges in single-cell transcriptomics', *Nature Reviews Genetics*, pp. 133–145. doi: 10.1038/nrg3833.

Steinbeck, J. A. and Studer, L. (2015) 'Moving stem cells to the clinic: Potential and limitations for brain repair', *Neuron*, pp. 187–206. doi: 10.1016/j.neuron.2015.03.002.

Stenman, J. (2003) 'Tlx and Pax6 co-operate genetically to establish the pallio-subpallial boundary in the embryonic mouse telencephalon', *Development*. doi: 10.1242/dev.00328.

Stenman, J., Toresson, H. and Campbell, K. (2003) 'Identification of two distinct progenitor populations in the lateral ganglionic eminence: implications for striatal and olfactory bulb neurogenesis.', *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 23(1), pp. 167–174. doi: 23/1/167 [pii].

Stiles, J. and Jernigan, T. L. (2010) 'The basics of brain development', *Neuropsychology Review*, pp. 327–348. doi: 10.1007/s11065-010-9148-4.

Stoykova, a *et al.* (1996) 'Forebrain patterning defects in Small eye mutant mice.', *Development (Cambridge, England)*, 122(11), pp. 3453–3465. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8951061.

Struhl, K. (2007) 'Transcriptional noise and the fidelity of initiation by RNA polymerase II', *Nature Structural and Molecular Biology*, pp. 103–105. doi: 10.1038/nsmb0207-103.

Sun, L. *et al.* (2013) 'Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts', *Nucleic Acids Research*, 41(17). doi: 10.1093/nar/gkt646.

Sun, M. and Kraus, W. L. (2015) 'From discovery to function: The expanding roles of long noncoding RNAs in physiology and disease', *Endocrine Reviews*, pp. 25–64. doi: 10.1210/er.2014-1034.

Sussel, L. *et al.* (1999) 'Loss of Nkx2.1 homeobox gene function results in a ventral to dorsal molecular respecification within the basal telencephalon: evidence for a transformation of the pallidum into the striatum', *Development*, 126(15), pp. 3359–3370. doi: 10393115.

Svensson, V. *et al.* (2017) 'Power analysis of single-cell rnA-sequencing experiments', *Nature Methods*, 14(4), pp. 381–387. doi: 10.1038/nmeth.4220.

Svensson, V., Vento-Tormo, R. and Teichmann, S. A. (2018) 'Exponential scaling of single-cell RNA-seq in the past decade', *Nature Protocols*. doi: 10.1038/nprot.2017.149.

Szucsik, J. C. *et al.* (1997) 'Altered forebrain and hindbrain development in mice mutant for the Gsh- 2 homeobox gene', *Developmental Biology*, 191(2), pp. 230–242. doi: 10.1006/dbio.1997.8733.

Tamura, S. *et al.* (2004) 'Foxp1 gene expression in projection neurons of the mouse striatum', *Neuroscience*, 124(2), pp. 261–267. doi: 10.1016/j.neuroscience.2003.11.036.

Tang, F. *et al.* (2009) 'mRNA-Seq whole-transcriptome analysis of a single cell', *Nature Methods*, 6(5), pp. 377–382. doi: 10.1038/nmeth.1315.

Tarazona, S. *et al.* (2011) 'Differential expression in RNA-seq: A matter of depth', *Genome Research*, 21(12), pp. 2213–2223. doi: 10.1101/gr.124321.111.

Tasic, B. *et al.* (2016) 'Adult mouse cortical cell taxonomy revealed by single cell transcriptomics', *Nature Neuroscience*, advance on(January), pp. 1–37. doi: 10.1038/nn.4216.

Tepper, J. M. and Bolam, J. P. (2004) 'Functional diversity and specificity of neostriatal interneurons', *Current Opinion in Neurobiology*, pp. 685–692. doi: 10.1016/j.conb.2004.10.003.

Theil, T. *et al.* (2002) 'Wnt and Bmp signalling cooperatively regulate graded Emx2 expression in the dorsal telencephalon.', *Development (Cambridge, England)*, 129(13), pp. 3045–54. doi: 10.1093/emboj/16.13.3797.

Tietjen, I. *et al.* (2003) 'Single-cell transcriptional analysis of neuronal progenitors', *Neuron*, 38(2), pp. 161–175. doi: 10.1016/S0896-6273(03)00229-0.

Tong, Y. *et al.* (2011) 'Spatial and Temporal Requirements for huntingtin (Htt) in Neuronal Migration and Survival during Brain Development', *Journal of Neuroscience*, 31(41), pp. 14794–14799. doi: 10.1523/JNEUROSCI.2774-11.2011.

Toresson, H. *et al.* (1999) 'Retinoids are produced by glia in the lateral ganglionic eminence and regulate striatal neuron differentiation.', *Development (Cambridge, England)*, 126(6), pp. 1317–1326.

Toresson, H., Potter, S. S. and Campbell, K. (2000) 'Genetic control of dorsal-ventral identity in the telencephalon: opposing roles for Pax6 and Gsh2.', *Development (Cambridge, England)*, 127(20), pp. 4361–4371.

Trapnell, C. *et al.* (2011) 'Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms', *Nature Biotechnology*, 28(5), pp. 511–515. doi: 10.1038/nbt.1621.Transcript.

Trapnell, C. *et al.* (2014) 'The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells', *Nature Biotechnology*, 32(4), pp. 381–386. doi: 10.1038/nbt.2859.

Trapnell, C. (2015) 'Defining cell types and states with single-cell genomics', *Genome Research*, pp. 1491–1498. doi: 10.1101/gr.190595.115.

Trapnell, C., Pachter, L. and Salzberg, S. L. (2009) 'TopHat: Discovering splice junctions with RNA-Seq', *Bioinformatics*, 25(9), pp. 1105–1111. doi: 10.1093/bioinformatics/btp120.

Tucker, E. S. *et al.* (2008) 'Molecular Specification and Patterning of Progenitor Cells in the Lateral and Medial Ganglionic Eminences', *Journal of Neuroscience*, 28(38), pp. 9504–9518. doi: 10.1523/JNEUROSCI.2341-08.2008.

Uchikawa, M., Kamachi, Y. and Kondoh, H. (1999) 'Two distinct subgroups of Group B Sox genes for transcriptional activators and repressors: Their expression during embryonic organogenesis of the chicken', *Mechanisms of Development*, 84(1–2), pp. 103–120. doi: 10.1016/S0925-4773(99)00083-0.

Usoskin, D. *et al.* (2015) 'Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing', *Nature Neuroscience*, 18(1), pp. 145–153. doi: 10.1038/nn.3881.

Vallejos, C. A. *et al.* (2017) 'Normalizing single-cell RNA sequencing data: Challenges and opportunities', *Nature Methods*, pp. 565–571. doi: 10.1038/nmeth.4292.

Vallejos, C. A., Marioni, J. C. and Richardson, S. (2015) 'BASiCS: Bayesian Analysis of Single-Cell Sequencing Data', *PLoS Computational Biology*, 11(6). doi: 10.1371/journal.pcbi.1004333.

de Vargas Roditi, L. and Claassen, M. (2015) 'Computational and experimental single cell biology techniques for the definition of cell type heterogeneity, interplay and intracellular dynamics', *Current Opinion in Biotechnology*, pp. 9–15. doi: 10.1016/j.copbio.2014.10.010.

Villar-Cervino, V. *et al.* (2015) 'Molecular Mechanisms Controlling the Migration of Striatal Interneurons', *Journal of Neuroscience*. doi: 10.1523/JNEUROSCI.4317-14.2015.

Waclaw, R. R. *et al.* (2006) 'The zinc finger transcription factor Sp8 regulates the generation and diversity of olfactory bulb interneurons', *Neuron*, 49(4), pp. 503–516. doi: 10.1016/j.neuron.2006.01.018.

Wang, H. F. and Liu, F. C. (2001) 'Developmental restriction of the LIM homeodomain transcription factor Islet-1 expression to cholinergic neurons in the rat striatum', *Neuroscience*, 103(4), pp. 999–1016. doi: 10.1016/S0306-4522(00)00590-X.

Wang, H. F. and Liu, F. C. (2005) 'Regulation of multiple dopamine signal transduction molecules by retinoids in the developing striatum', *Neuroscience*, 134(1), pp. 97–105. doi: 10.1016/j.neuroscience.2005.04.008.

Wang, K. *et al.* (2010) 'MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery', *Nucleic Acids Research*, 38(18). doi: 10.1093/nar/gkq622.

Wang, K. C. and Chang, H. Y. (2011) 'Molecular Mechanisms of Long Noncoding RNAs', *Molecular Cell*, pp. 904–914. doi: 10.1016/j.molcel.2011.08.018.

Wang, L. *et al.* (2013) 'CPAT: Coding-potential assessment tool using an alignment-free logistic regression model', *Nucleic Acids Research*, 41(6). doi: 10.1093/nar/gkt006.

Washietl, S. *et al.* (2011) 'RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data', *RNA*, 17(4), pp. 578–594. doi: 10.1261/rna.2536111.

Washietl, S., Kellis, M. and Garber, M. (2014) 'Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals', *Genome Research*, 24(4), pp. 616–628. doi: 10.1101/gr.165035.113.

White, J. K. *et al.* (1997) 'Huntington is required for neurogenesis and is not impaired by the Huntington's disease CAG expansion', *Nature Genetics*, 17(4), pp. 404–410. doi: 10.1038/ng1297-404.

Wiatr, K. *et al.* (2017) 'Huntington Disease as a Neurodevelopmental Disorder and Early Signs of the Disease in Stem Cells', *Molecular Neurobiology*, pp. 1–21. doi: 10.1007/s12035-017-0477-7.

Wichterle, H. *et al.* (2001) 'In utero fate mapping reveals distinct migratory pathways and fates of neurons born in the mammalian basal forebrain.', *Development (Cambridge, England)*, 128(19), pp. 3759–3771.

Wills, Q. F. *et al.* (2013) 'Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments', *Nature Biotechnology*, 31(8), pp. 748–752. doi: 10.1038/nbt.2642.

Wolf, F. A. *et al.* (2017) 'Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells', *Doi.Org*, p. 208819. doi: 10.1101/208819.

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

Wonders, C. P. and Anderson, S. A. (2006) 'The origin and specification of cortical interneurons', *Nature Reviews Neuroscience*, pp. 687–696. doi: 10.1038/nrn1954.

Wu, T. D. and Nacu, S. (2010) 'Fast and SNP-tolerant detection of complex variants and splicing in short reads', *Bioinformatics*, 26(7), pp. 873–881. doi: 10.1093/bioinformatics/btq057.

Wucher, V. *et al.* (2017) 'FEELnc: A tool for long non-coding RNA annotation and its application to the dog transcriptome', *Nucleic Acids Research*, 45(8), pp. 1–12. doi: 10.1093/nar/gkw1306.

Xu, Q. (2004) 'Origins of Cortical Interneuron Subtypes', *Journal of Neuroscience*, 24(11), pp. 2612–2622. doi: 10.1523/JNEUROSCI.5667-03.2004.

Xu, Q., Tam, M. and Anderson, S. A. (2008) 'Fate mapping Nkx2.1-lineage cells in the mouse telencephalon', *Journal of Comparative Neurology*, 506(1), pp. 16–29. doi: 10.1002/cne.21529.

Xue, Z. *et al.* (2013) 'Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing', *Nature*, 500(7464), pp. 593–597. doi: 10.1038/nature12364.

Yao, Z. *et al.* (2017) 'A Single-Cell Roadmap of Lineage Bifurcation in Human ESC Models of Embryonic Brain Development', *Cell Stem Cell*, 20(1), pp. 120–134. doi: 10.1016/j.stem.2016.09.011.

Yates, A. *et al.* (2016) 'Ensembl 2016', *Nucleic Acids Research*, 44(D1), pp. D710–D716. doi: 10.1093/nar/gkv1157.

Yip, S. H., Sham, P. C. and Wang, J. (2018) 'Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data', *Briefings in Bioinformatics*. doi: 10.1093/bib/bby011.

You, B. H., Yoon, S. H. and Nam, J. W. (2017) 'High-confidence coding and noncoding transcriptome maps', *Genome Research*, 27(6), pp. 1050–1062. doi: 10.1101/gr.214288.116.

Yu, G. *et al.* (2010) 'GOSemSim: An R package for measuring semantic similarity among GO terms and gene products', *Bioinformatics*, 26(7), pp. 976–978. doi: 10.1093/bioinformatics/btq064.

Zeisel, A. *et al.* (2015) 'Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq.', *Science (New York, N.Y.)*, 347(6226), pp. 1138–42. doi: 10.1126/science.aaa1934.

Zeisel, A. *et al.* (2018) *Molecular architecture of the mouse nervous system*, *bioRxiv*. doi: 10.1101/294918.

Zhang, S. (2014) 'Sox2, a key factor in the regulation of pluripotency and neural differentiation', *World Journal of Stem Cells*. doi: 10.4252/wjsc.v6.i3.305.

Zhao, Q. Y. *et al.* (2011) 'Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study', *BMC Bioinformatics*, 12 Suppl 1, p. S2. doi: 10.1186/1471-2105-12-S14-S2.

Zheng, G. X. Y. *et al.* (2017) 'Massively parallel digital transcriptional profiling of single cells', *Nature Communications*, 8. doi: 10.1038/ncomms14049.

Zhong, J. *et al.* (2009) 'BC1 Regulation of Metabotropic Glutamate Receptor-Mediated Neuronal Excitability', *Journal of Neuroscience*, 29(32), pp. 9977–9986. doi: 10.1523/JNEUROSCI.3893-08.2009.

Zhong, S. *et al.* (2018) 'A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex', *Nature*. doi: 10.1038/nature25980.

Zucchelli, S. *et al.* (2015) 'SINEUPs are modular antisense long non-coding RNAs that increase synthesis of target proteins in cells', *Frontiers in Cellular Neuroscience*, 9. doi: 10.3389/fncel.2015.00174.

6 | Appendix

6.1 Contributions to published articles

6.1.1 Faulty neuronal determination and cell polarization are reverted by modulating HD early phenotypes.

*P. Conforti, D. Besusso[*], **V. D. Bocchi**[*], A. Faedo[*], E. Cesana, G. Rossetti, V. Ranzani, C. N. Svendsen, L. M. Thompson, M. Toselli, G. Biella, M. Pagani, and E. Cattaneo.*

[*]contributed equally to this work.

Proceedings of the National Academy of Sciences 115(4): E762–71. 2018

The aim of this work was to establish the in vitro effects of huntingtin mutations at the level of neuronal progenitor specification. My specific goal in this project was to interrogate the cortical component of the disease by determining transcriptional differences in control and HD cortical organoids by microarray data analysis. All results and analysis are under the paragraph "HD Cerebral Organoids Show an Immature Transcriptional Bluep rint". The figures produced by this bioinformatics analysis are Figure 5 and Supplementary Figure 4F-H.

I specifically contributed to the retrieval, computational analysis and interpretation of the microarray data together with the experimental design of the transcriptomic study. I wrote the section of the bioinformatics analysis in terms of results and methods and was involved in the final revision of the article.

6.1.2 Dynamic and Cell-Specific DACH1 Expression in Human Neocortical and Striatal Development.

*V. Castiglioni[*], A. Faedo[*], M. Onorati[*], **V.D Bocchi**[*], Z. Li, R. Iennaco, R.Vuono, G.P. Bulfamante, L. Muzio, G. Martino, N. Sestan, R.A. Barker, E. Cattaneo.*

[*]Co-first authors

Cerebral Cortex, bhy092. 2018

The aim of this study was to systematically investigate DACH1 expression patterns during human neurodevelopment, from 5 to 21 postconceptional weeks (pcw). My goal in this project was to identify specific sub-populations of DACH1 expressing cells in the developing neocortex by analysing previously published single-cell RNA-seq data on the human developing neocortex.

All results and analysis are under the paragraph “Single-Cell Transcriptional Profiling Reveals DACH1 Expression in Neuroepithelial Cells and vRGCs”. The figures produced by this bioinformatics analysis are Figure 3 and Supplementary Figure 2 and 6.

I specifically contributed to the computational analysis and interpretation of the scRNA-seq data. I was involved, together with the other co-first authors, in drafting the article and then in the revision and final approval of the manuscript before publishing.