UNIVERSITY OF CALIFORNIA

Santa Barbara

The Role of Heterogeneity During Development and Stem Cell Differentiation

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Mechanical Engineering

by

Mahdi Golkaram

Committee in charge:

Professor Linda R. Petzold, Chair

Professor Kenneth S. Kosik

Professor Otger Campas

Professor Megan T. Valentine

September 2018

The dissertation of Mahdi Golkaram is approved.

_____

Kenneth S. Kosik

_____

Otger Campas

_____

Megan T. Valentine

_____

Linda R. Petzold, Committee Chair

September 2018

The Role of Heterogeneity During Development and Stem Cell Differentiation

by

Mahdi Golkaram

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my committee chair, Professor Linda Petzold, without whose help and persistent advice this work would not have been possible. Every step of the way, not only she trusted on my research ideas, she offered me her invaluable guidance.

In addition, I would like to thank my committee members, Professor Kenneth Kosik whose knowledge, experience and support provided me the opportunity to explore new realms of science, Professors Otger Campas and Megan Valentine for generously offering their time, support, guidance and good will throughout the preparation and review of this dissertation.

Last but not least, I would like to dedicate my dissertation to my parents for their love and support throughout my life. Thank you both for giving me strength to reach for the stars and chase my dreams.

Contribution statement: The following had significant contribution to this dissertation:

- Jiwon Jang: performed all the experiments in Chapter 3.
- Neha Rani, Tomasz Nowakowski: performed all the experiments in Chapter 4.
- Tomasz Nowakowski, Zhou Hongjun: helped with the data analysis in Chapter 4.

VITA OF MAHDI GOLKARAM
September 2018

EDUCATION

University of California, Santa Barbara (UCSB) – 2014-2018

Ph.D in Mechanical Engineering – GPA: 3.92/4.00 (Thesis: *The Role of Heterogeneity During Development and Stem Cell Differentiation.*)

Pennsylvania State University (PSU), University Park – 2012-2014

Master of Science in Mechanical Engineering – GPA: 3.94/4.00 (Thesis: *Application of Reactive Molecular Dynamics in Biomaterial Science and Computational Biology.*)

Amirkabir University (Tehran Polytechnic), Tehran, Iran – 2008-2012

Bachelor of Science in Mechanical Engineering - GPA: 4.00/4.00

SKILLS

Bioinformatics & Computational Biology:

- Machine learning (deep learning, RNN/LSTM), data mining and analysis of noisy biological data (regression, clustering, classifications, etc.)
- Statistical inference (linear models, generalized linear models, etc.)
- Bayesian inference
- Next generation sequencing data analysis (scRNA-seq, ChIP-seq, ChIA-PET, Hi-C, etc.)
- Statistical programming languages (R/Bioconductor/Python/MATLAB)
- Familiarity with public bioinformatics resources (NCBI, UCSC, Ensembl, UniProt)

Software Development:

- Version control software (GitHub)
- Programming languages (C/C++)
    - Linux (Redhat and Ubuntu) and Microsoft Windows operating systems
    - High-performance computing (HPC), cloud computing (EC2)

Molecular Biology:

- Developmental biology / Stem cell biology / Neuroscience
- Cancer research / Systems biology
- Immunology / Cancer immunotherapy

EXPERIENCE

Research Assistantships

- Bioinformatics & computational systems biology, University of California, Santa Barbara - 2014-2018:
    1. Analysis of gene expression heterogeneity of embryonic stem cell development using next generation sequencing (scRNA-seq, ChIP-seq, ChIA-PET, Hi-C, etc.) (using R)
    2. Statistical modeling, data manipulation and visualizing of high-throughput next generation sequencing data (using R)
    3. Stochastic modeling of complex biological networks
    4. Statistical and Bayesian inference of clinical cancer data
- Machine learning – Deep learning, University of California, Santa Barbara – 2016-2018:
    5. Predicting *ex vivo* neuronal spiking patterns from Multi-Electrode Array (MEA) using *Recurrent Neural Network* - Long-Short-Term-Memory (LSTM) (using Python *TensorFlow*).
        (i) Analysis of learning using Neural Network (*in silico*) vs. Neuronal Network (*ex vivo*)
        (ii) Detecting biomarkers for memory formation and learning
- Software development and cloud computation, University of California, Santa Barbara – 2014-2018:
    6. Modified pyURDME© software for crowded biochemical systems (using C/Python/MATLAB) on AWS platform
- Computational biochemistry/drug design, Pennsylvania State University – 2014
    7. Predictive modeling of chromosomal translocation using Ewing sarcoma cancer patients' data
    8. Computational modeling of deformed human red blood cells infected with Plasmodium falciparum malaria parasite.

Internships and Industry Experience

- Bioinformatics intern at Genentech, South San Francisco – summer 2017: Deciphering cellular heterogeneity in metastatic colorectal adenocarcinoma using scRNA-seq.
    1. Development of scRNA-seq data analysis workflow in R: including gene and cell quality controls, normalization (eight different methods), clustering (six different methods), differential expression and pseudo-time analysis.
    2. Studying the influence of tumor environment on cancer stem cell differentiation.

- Computational biology at Kressworks © & The Pennylvania State Univerity – summer 2014: Determining the mechanism of chromosomal translocations in Ewing sarcoma cancer.

PUBLICATIONS

Journal Articles

[1] Nowakowski TJ*, Rani N*, **Golkaram M***, Zhou H*, Alvarado B, Huch K, West JA, Leyrat A, Pollen AA, Kriegstein AA, Petzold LR, Kosik KS. Regulation of Cell-Type-Specific Transcriptomes by miRNA Networks During Human Brain Development, *Nature Neuroscience* (accepted).

** These authors contributed equally to this work.*

[2] Jang J, **Golkaram M**, Audouard1 M, Bridges D, Hellander S, Petzold LR, Kosik KS. WNT-driven single cell G1 lengths establish a probability distribution for differentiation outcomes of stem cell populations, *PLOS Biology* (under review).

[3] De Sousa EMF, Piskol R, **Golkaram M**, De Sauvage F, (2016), Deciphering cellular heterogeneity in metastatic colorectal adenocarcinoma using scRNA-seq. (under preparation).

[4] **Golkaram M**, Jang J, Hellander S, Kosik KS, Petzold LR. The Role of Chromatin Density in Cell Population Heterogeneity during Stem Cell Differentiation. *Scientific Reports*. 2017 Oct 17;7(1)13307.

[5] **Golkaram M,** Hellander S, Drawert B, Petzold LR. Macromolecular Crowding Regulates the Gene Expression Profile by Limiting Diffusion. *PLOS Computational Biology*. 2016 Nov 28;12(11):e1005122.

[6] **Golkaram M,** van Duin AC. Revealing graphene oxide toxicity mechanisms: A reactive molecular dynamics study. *Materials Discovery*. 2015 Jan 31;1:54-62.

[7] **Golkaram M**, Shin YK, van Duin AC. Reactive Molecular Dynamics Study of the pH-Dependent Dynamic Structure of α-Helix. *The Journal of Physical Chemistry B*. 2014 Nov 17;118(47):13498-504.

[8] Zhang Y, Huang C, Kim S, **Golkaram M**, Dixon MW, Tilley L, Li J, Zhang S, Suresh S. Multiple stiffening effects of nanoscale knobs on human red blood cells infected with Plasmodium falciparum malaria parasite. *Proceedings of the National Academy of Sciences*. 2015 May 12;112(19):6068-73.

[9] Verlackt CC, Neyts EC, Jacob T, Fantauzzi D, **Golkaram M**, Shin YK, van Duin AC, Bogaerts A. Atomic-scale insight into the interactions between hydroxyl radicals and DNA in solution using the ReaxFF reactive force field. *New Journal of Physics*. 2015 Oct 2;17(10):103005.

[10] **Golkaram M.** Application of Reactive Molecular Dynamics in Biomaterial Science and Computational Biology.

[11] **Golkaram M,** Aghdam M. Free transverse vibration analysis of thin rectangular plates locally suspended on elastic beam. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*. 2012 Oct 25; 227(7) 1515–1524.

[12] Amereh M, Aghdam MM, **Golkaram M**. Design and modeling of a novel translational and angular micro-electromechanical accelerometer. *Aerospace Science and Technology*. 2016 Mar 31;50:15-24.

Conference Proceedings and Presentations

[1] **Golkaram M, "**The analysis of gene expression pattern in a crowded environment and its phenotypes using stochastic simulation methods.", UCSB second Mechanical Engineering grand slam and convocation, Oct. 2015.

[2] **Golkaram M,** "The analysis of gene expression pattern in a crowded environment and its phenotypes using stochastic simulation methods.", Southern California Simulations in Science conference, Oct. 2015.

[3] **Golkaram M**, Zhang S. "Phase Field Modeling of Active Stress Generation in Spreading Cells during Actin Polymerization." ASME 2013 International Mechanical Engineering Congress & Exposition.

[4] Zhang Y, Huang C, **Golkaram M**, Zhang S (2014). Molecular origins of the loss of deformability in Plasmodium-falciparum infected erythrocytes: a coarse-grained modeling, Purdue University.

ABSTRACT

The Role of Heterogeneity During Development and Stem Cell Differentiation

by

Mahdi Golkaram

Heterogeneity is an integral property of many biological responses from molecular scales to members of a population of one species. While several recent studies have demonstrated the extent of the underlying heterogeneity on different scales, its origin, regulation, and consequences have not been thoroughly understood. In this dissertation, first we draw a general picture of how heterogeneity accompanies most biological responses from molecular to tissue scale, then we uncover the factors that can cause or contribute to non-uniformity. We explore the regulation of variation of biological systems as well as its consequences and illustrate the pivotal role that molecular, cellular and tissue heterogeneity plays in survival of an organism.

First, we seek to elucidate the role of macromolecular crowding in transcription and translation. It is well known that stochasticity in gene expression can lead to differential gene expression and heterogeneity in a cell population. Recent experimental observations by Tan *et al*. (*Nature nanotechnology*. 2013 Aug;8(8):602) have improved our understanding of the functional role of macromolecular crowding. It can be inferred from their observations that macromolecular crowding can lead to robustness in gene expression, resulting in a more homogeneous cell population.

We introduce a spatial stochastic model to provide insight into this process. Our results show that macromolecular crowding reduces noise (as measured by the kurtosis of the mRNA distribution) in a cell population by limiting the diffusion of transcription factors (i.e. removing the unstable intermediate states), and that crowding by large molecules reduces noise more efficiently than crowding by small molecules. Finally, our simulation results provide evidence that the local variation in chromatin density as well as the total volume exclusion of the chromatin in the nucleus can induce a homogenous cell population.

Next we incorporate three-dimensional (3D) conformation of chromosome (Hi-C) and single-cell RNA sequencing data together with discrete stochastic simulation, to explore the role of chromatin reorganization in determining gene expression heterogeneity during development. While previous research has emphasized the importance of chromatin architecture on activation and suppression of certain regulatory genes and gene networks, our study demonstrates how chromatin remodeling can dictate gene expression distribution by folding into distinct topological domains. We hypothesize that the local DNA density during differentiation accentuates transcriptional bursting due to the crowding effect of chromatin. This phenomenon yields a heterogeneous cell population, thereby increasing the potential of differentiation of the stem cells.

Finally, we depict the interplay between microRNAs and mRNAs and how this network can regulate human fetal brain development. microRNAs (miRNAs) regulate many cellular events during brain development by interacting with hundreds of mRNA transcripts. However, miRNAs operate non-uniformly upon the transcriptional profile with an as yet unknown logic. Shortcomings in defining miRNA-mRNA networks are limited knowledge of in vivo miRNA targets, and their abundance in single cells. By combining multiple complementary approaches: AGO2-HITS-CLIP, single-cell profiling, and innovative

computational analyses using bipartite and co-expression networks, we show that miRNA-mRNA interactions operate as functional modules that often correspond to cell-type identities and undergo dynamic transitions during brain development. These networks are highly dynamic during development and over the course of evolution. One such interaction is between radial glia-enriched ORC4 and miR-2115, a great ape specific miRNA, which appears to control radial glia proliferation rates during human brain development.

CHAPTER 1

Overview

For decades, biologists have been studying living organisms by identifying the average behavior of a population in different conditions. Often, sampling from a population several times was a convincing approach to describe the average changes between conditions and thus to neglect the inherent variations present in different biological systems. For many reasons, namely a lack of interest in the intrinsic variations across the individual members of a biological system, the absence of quantification tools with single cell resolution, the void of rigorous statistical methods to decipher noisy information in order to render a clear picture from one side, and the vast amount of information encoded in the population scale measurements discouraged researchers from thoroughly investigating the heterogeneity in the biological systems.

By successful completion of the Human Genome Project on 2003, researchers took the first step toward determining the sequence of nucleotide base pairs that make up human DNA, identifying and mapping all of the genes of the human genome. This was proceeded with the 1000 Genome Project, launched on 2008 with a global interest to establish by far the most comprehensive catalogue of human genetic variation. In this project, scientists aimed to sequence the genomes of at least one thousand anonymous participants from a

variety of ethnic groups from all over the world. This can be considered to be the first pivotal study that attracted the attention of the scientific community toward the variations embedded in human biology. Namely, discovering thousands of nucleotide variations between different individuals (also known as single nucleotide polymorphism or SNPs), enabled scientists to map the genetic information of human population to different traits such as eyes and skin color, bone structure, height, as well as variable propensity to certain diseases such as Alzheimer`s disease, certain types of cancer, diabetes, etc.

Understanding the structural variants such as gene copy number variation, gene inversion, short and long repeats, etc. and scrutinizing databases such as 1000 Genome Project have had a dramatic contribution to our understanding on why certain ethnicities are more likely to acquire certain disease. Furthermore, investigating short tandem repeats (STR) – short sequences consisting of a unit of two to thirteen nucleotides repeated hundreds of times in a row on the DNA strand - has been referred to as a must in forensic analysis. Forensic science takes advantage of the population`s variability in STR lengths, enabling us to distinguish one DNA from another.

Aside from the encoded variability in our genome, how genes are expressed can entail drastic variability which is referred to as cellular noise. The central dogma of molecular biology states that genes are transcribed into RNAs and RNAs are translated into proteins. Today our knowledge of molecular systems biology stipulates that the path from genotype to phenotype is much more complex. Indeed, due to several evolutionary devised mechanisms, cells utilize several layers of regulation, so that the final response can significantly diverge from this path. For instance, not all genes are transcribed simultaneously but presence of cis-(distant regulatory sites on one DNA molecules that are brought together to activate or

inactivate the expression of a gene) and trans- (regulation by binding and unbinding of regulatory elements to the DNA strand) regulatory elements can dictate gene expression.

One might date the interest in heterogeneity in biology back to late 20th century, but it is fair to call the seminal study by Elowitz et al. [1] to be the hallmark of scientific curiosity in single cell variability. Elowitz and his colleagues, in a bottom-up approach, demonstrated that due to the underlying stochasticity of gene expression and protein translation, an isogenic (having the same or similar genotype) population of cells exhibit high level of variability that can propagate through out the tissue or in some cases the organism. As it will be discussed in Chapter 2, cellular noise is often investigated in the framework of intrinsic and extrinsic noise. While the precise definition of intrinsic versus extrinsic noise can depend on the context, generally intrinsic noise refers to variation in regulation of a certain gene within one cell and extrinsic noise refers to differences in the regulation of the same gene between cells. The factors that can contribute to intrinsic noise include low copy number, particle diffusion, noise propagation in a reaction network e.g. noise amplification in signaling networks. Extrinsic noise arises from cellular age, cell cycle (mitosis and meiosis), inter cellular signaling gradient, local micro-environment, organelle distribution, etc. Therefore, Chapter 2 describes the causes and the contributing factors to cell-to-cell variability. We will show how both cis- and trans- regulatory elements can regulate gene expression variation by limiting the diffusion of transcription factors (molecules, mainly proteins, that can activate one or more genes). We also will demonstrate that crowdedness of the cellular environment can dramatically influence the diffusion of particles and thus transcriptional variation.

Recent technological advancement has enabled us to study biological systems at single cell resolution. Rapid improvements of RNA and protein quantification methods within the

last two decades provided researches the opportunity to study the behavior of several genes in parallel. The introduction of high throughput messenger RNA (mRNA) sequencing tools such as microarray, and next generation sequencing methods [2] such as RNA-seq, ChIP-seq, ATAC-seq, methylation-seq, etc. facilitated massively parallel identification of several layers of regulation on a population scale.

Besides the traditional staining methods that have been used to show the variability of the expression of one gene in different cells in both mRNA and protein levels, methods such as single molecule fluorescent in situ hybridization (smFISH), and single cell quantitative polymerase chain reaction (sc-qPCR) are two example methods that are designed to accurately quantify gene expression at single cell resolution. Despite the applicability and accuracy of these methods, the introduction of single cell RNA-seq (scRNA-seq) by Tang et al. on 2009 [3] revolutionized our understanding of the biology of single cells. Within the last decade, different researchers have been able to modify scRNA-seq such that today we are able to analyze the transcriptome of single cells of a population of thousands of cells in parallel fast and with relatively low cost (Fig. 1.1).

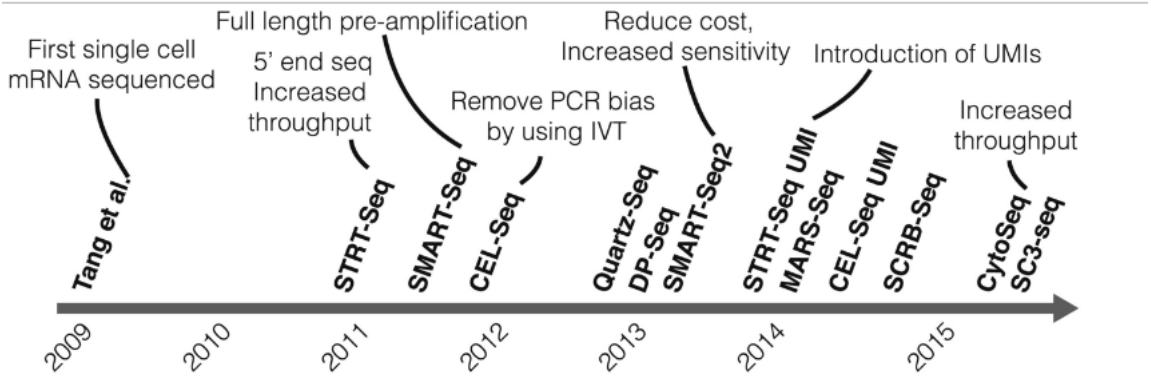


Figure 1.1. The evolution of single cell RNA sequencing

In Chapter 3, we explore the consequences of population heterogeneity and explain the context in which heterogeneity can be fatal, essential, or provide a selection advantage. The main focus of this chapter is to address the role of heterogeneity during early embryonic development. Embryonic stem cells (ESCs) are pluripotent stem cells (i.e. they can divide, differentiate and give rise to almost all cell types of a species) derived from the inner cell mass of blastocyst, an early-stage of pre-implantation embryo. Embryonic stem cells of the inner cell mass can differentiate into three primary germ layers: ectoderm, endoderm, and mesoderm. Ectoderm forms the exoskeleton, mesoderm develops into organs, and endoderm forms the inner lining of organs. For example, ectoderm (outer layer) includes central nervous system and neurons of brain, and epidermal cells of skin, endoderm (middle layer) includes bone tissues, blood cells, and kidney cells, and finally endoderm (internal layer) can give rise to lung, muscle, and stomach cells (Fig. 1.2). We will show how the local density of chromatin (the material of which the chromosomes of organisms other than bacteria i.e. eukaryotes are composed which includes DNA, RNA, proteins and histones) alters during early development and how this can contribute to population heterogeneity. We will show the factors that can potentiate the differentiation capabilities to one cell type over another.

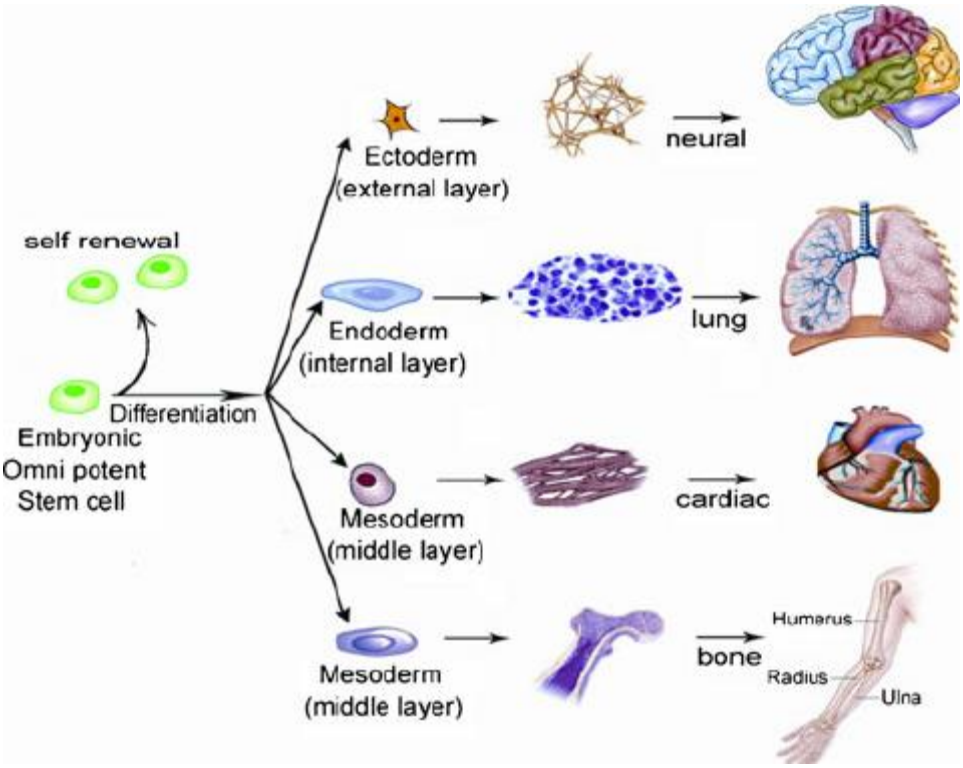


Figure 1.2. Formation of the three primary germ layers: Ectoderm, Endoderm, and Mesoderm.

In Chapter 4, we focus on one of the three primary germ layers, ectoderm layer and mainly brain cells. We will discuss the development of brain from fetal to adult stages and what roles heterogeneity plays in this context. In week 3 of human development the neuro-ectoderm appears and forms the neural plate along the dorsal side of the embryo. Two distinct cell types stem from neural plate: neurons and glial cells of the central nervous system (CNS). While glia cells can proliferate and replenish their peers, neurons once fully matured will no longer divide (i.e. they are post-mitotic). The brain is a diverse collection of cell types. Even though most brain cells can be categorized as either neuron or glia, each category can be further divided into a dozen distinct cell types. For example, glia cells comprise oligodendrocytes, astrocytes, ependymal cells, Schwann cells, microglia, and

satellite cells each of which has a critical function whose absence can be, if not fatal, leading to certain disease. As a matter of fact, the distribution of these cell types across different brain regions is highly regulated. Hence this chapter focuses on the consequences of cell type variability and heterogeneity in an organism and depicts how proper function of a species depends on the underlying variability.

In order to thoroughly understand the regulation of the distribution of each cell type in early brain development, we will introduce a new layer of regulation by small non-coding RNAs. As highlighted earlier, the central dogma of molecular biology can no longer fully capture the complexity of living organisms in that it refuses to acknowledge the other regulatory elements such as post transcriptional and translational modifications (adding or removing small molecules to mRNAs or peptides which can alter the cellular response), alternative splicing (tissue dependent mRNA post processing by which segments of the transcripts are spliced out), epigenetic markers (methylation, chromatin modification, etc.) and non-coding RNAs. Later constructs a large family of RNAs that even though the corresponding gene is encoded in our genome and can be transcribed into RNAs under certain conditions, there won`t be any subsequent translation of these RNAs into proteins. Namely, small interfering RNAs (siRNAs) – evolutionarily acquired by some virus to inhibit the expression of certain genes in the host cell, small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and miRNA (microRNA) (Fig. 1.3).

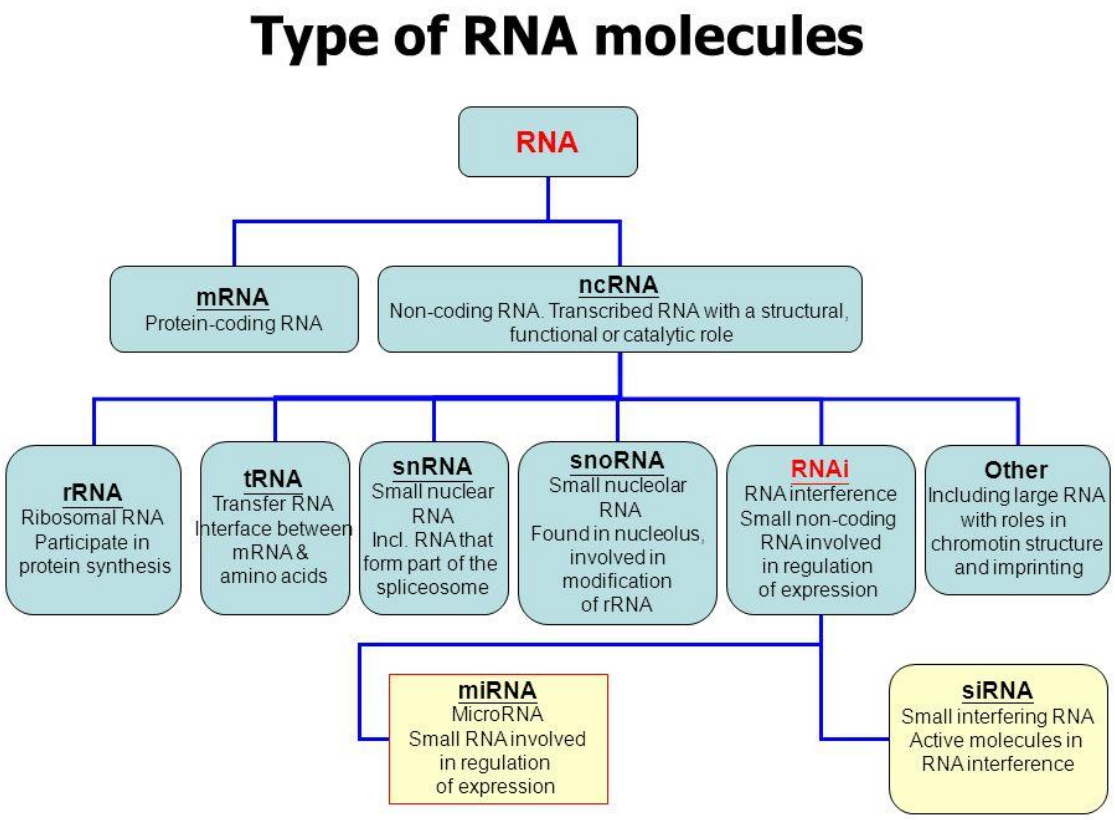


Figure 1.3. RNAs divide into two major classes of messenger RNAs and non-coding RNAs. Several non-coding RNAs have been discovered and demonstrated to participate in gene regulation.

miRNAs have been shown to be involved in several biological functions and diseases such as regulating several developmental processes, cell cycle, metabolism, which explains their involvement in cancer, mental disorders, etc. miRNAs are short (~22 nucleotides) sequences that target mRNA transcripts by recognizing regions that match a segment of their own (a segment of 7-8 nucleotides i.e. miRNA response element or MRE) and hybridizing to target. This can lead to inactivation of translation of the target mRNAs into proteins. In Chapter 3 we elaborate on the functions of miRNAs and illustrate how miRNAs contribute to tissue

heterogeneity. We will construct networks of mRNAs and miRNAs and by recruiting state of the art computational tools will reveal the information encoded in the miRNA-mRNA regulatory network that can shed light on modulation of cell-type-specificity of developing human brain.

References:

1. M. B. Elowitz, A. J. Levine, E. D. Siggia, P. S. Swain, Stochastic gene expression in a single cell. *Science* **297**, 5584 (2002).

2. T. van der Straaten. Next-Generation Sequencing: Current Technologies and Applications. Edited by J. Xu. *ChemMedChem*. **10**, 2 (2015).

3. F. Tang, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods* **6.5**, 377 (2009).

CHAPTER 2

Macromolecular crowding regulates the gene expression profile by limiting diffusion

Introduction:

Even in an isogenic cell population under constant environmental conditions, significant variability in molecular content can be observed. This variability plays an important role in stem cell differentiation [1], cellular adaptation to a fluctuating environment [2], variations in cellular response to sudden stress [3], and evolutionary adaptations [4]. However, it can also be detrimental to cellular function and has been implicated as a factor leading to dangerous diseases such as haploinsufficiency [5], cancer [6], age-related cellular degeneration, and death in tissues of multicellular organisms [7]. The variability stems both from stochasticity inherent in the biochemical process of gene expression (intrinsic noise) and fluctuations in other cellular components (extrinsic noise), namely, stochastic promoter activation, promoter deactivation, mRNA, and protein production and decay, as well as cell-to-cell differences in, for example, number of ribosomes [8-13]. One consequence of biological noise in gene expression is transcriptional bursting, which is observed in both prokaryotes [14] and eukaryotes [12-15]. Transcriptional bursting can bring about a bimodal distribution of mRNA abundance in an isogenic cell population [16-18]. Therefore, understanding critical factors that

influence noise in gene expression can provide us with a new tool to tune cellular variability [19-24].

The cellular environment is packed with proteins, RNA, DNA, and other macromolecules. It is estimated that 30-40% of the cell volume is occupied by proteins and RNA [25]. Macromolecular crowding has been studied extensively in the last few decades [26, 27] and has been ingeniously utilized for numerous medical purposes [28-30]. It is well established that macromolecular crowding can reduce diffusion rates and enhance the binding rates of macromolecules [31], which can change the optimal number of transcription factors [32], the nuclear architecture [33], and the dynamical order of metabolic pathways [34].

It is known that manipulating the binding and unbinding rates (k_{on} and k_{off}) can affect the likelihood of observing transcriptional bursting [42, 43]. Higher values of k_{on} and k_{off} lead to a bimodal distribution and transcriptional bursting, while keeping the basal (i.e. in the absence of the bursts) protein abundance constant. It is also known that macromolecular crowding can alter diffusion and reaction rates [44]. Together, it is implied that macromolecular crowding can have an impact on protein production in a cellular environment.

In a previous study [35], crowding has been modeled by the direct manipulations of reaction rates using experimentally fitted relations. In contrast, we model macromolecular crowding explicitly by altering the effective diffusion rate of transcription factors. This approach is similar to recent studies performed by Isaacson *et al.* [46] and Cianci *et al.* [51]; however, we also consider the effects of the artificial crowding agents, in order to capture analogous experimental conditions performed by Tan *et al.* [35].

It has been observed experimentally [35] that macromolecular crowding can influence cell population homogeneity and gene expression robustness. In this experiment [35], the influence

of the diffusion of macromolecules on transcriptional activity is studied by synthesizing artificial cells in which inert dextran polymers (Dex) assume the role of the artificial crowding agent in the system. To capture the impact of the size of the crowding agent, the experiments are performed on two different sizes of Dex molecules, here referred to as Dex-Big and Dex-Small. It can be inferred from this experiment that a highly crowded environment results in a narrow distribution of fold gene-expression perturbation, suggesting that molecular crowding decreases the fluctuation of gene expression rates due to the perturbation of gene environmental factors.

However, the mechanism by which cellular crowding can control gene expression has not been elucidated. We demonstrate through modeling that macromolecular crowding reduces the noise (kurtosis of the mRNA distribution) in gene expression by limiting the diffusion of the transcription factors. This increases the residence time of the transcription factor on its promoter, thereby reducing the transcriptional noise. As a consequence, unstable intermediate states of gene expression pattern will diminish. Furthermore, our model reveals that small crowding agents reduce noise less than large crowding agents do, which is in agreement with the experimental observations [35]. Finally, our simulation results provide evidence that local variation in the chromatin density, in addition to the total volume exclusion of the chromatin in the nucleus, can alter gene expression patterns.

Results:

A simple and well-studied model was employed to simulate transcription and translation. The model includes: a) one transcription factor (TF) placed randomly in the simulation domain, b) TF diffusion in order to find the gene locus, c) binding and unbinding of TF to its promoter, d) mRNA production, and destruction and e) protein production and

destruction. This model and its corresponding parameters were adopted from Kaeren *et al.* [8] for the sake of comparison. (see Methods.)

We assume that the initial concentrations of mRNA and the target protein are zero, and use spatial stochastic simulation to investigate the gene expression pattern, a model that has been widely used and verified by both theoretical [36-39] and experimental [39, 40] observations. To account for crowding, we developed a modified next subvolume method (NSM) to approximately solve the reaction-diffusion master equation (RDME) [41] capable of explicitly treating the crowding agent amount, distribution, and interactions (see Methods).

The NSM method was modified so that the mesoscopic diffusion coefficient is linearly dependent on the crowding density in the destination voxel. In our model, the macromolecular crowding stems from two primary sources: chromatin structure and artificial crowding agents, akin to the Tan *et al*. experiment [35]. We utilized the 3-dimensional structured illumination microscopy data from [50] to model the chromatin structure. To account for chromatin structure, the crowding density in each voxel was assumed to be proportional to DAPI (4',6-diamidino-2-phenylindole) intensities in that voxel, similar to the method introduced by Isaacson *et al*. [46]. To account for different levels of crowding, we added artificial crowding agents distributed randomly in our simulation domain. We define the crowdedness parameter θ as the probability for each voxel to be occupied by an artificial crowding agent. Thus, we are able to explicitly account for different amounts of crowding in our simulation domain by changing θ. To interpret θ correctly, let`s consider the extreme case where $\theta = 1$. In this case all voxels would be occupied by one and only one crowding agent. Then, crowding reduces the diffusion coefficient depending on the size of the crowding agent (90% reduction for a large crowder vs. 40% reduction for a small crowder.). Note that under no condition would any voxel be completely blocked (i.e. 100%

crowded). For any other θ, approximately θ×N crowding molecules are randomly distributed in θ×N voxels, where N is the total number of the voxels. A convergence study demonstrates that our conclusions are independent of voxel size for a sufficiently small mesh, see (Supplementary Fig. 2.4).

To validate the model, the simulation was run for 1000 minutes with the same parameters as in [8] while θ = 0, i.e. with no artificial crowding agent or chromatin present. As in [8], this resulted in transcriptional bursts. A direct quantitative comparison is not trivial due to the fact that our model is spatially inhomogeneous (Supplementary Fig. 2.2).

Next, we included the artificial crowding agent and the chromatin in our model and investigated mRNA abundance in our simulation domain for low and high θ values (θ = 0% vs. θ = 100%). It can be seen that the system switches more frequently between active and inactive states for low θ values than it does for high θ values (Figs. 2.1a and b). We hypothesized that adding the artificial crowding agent limited the diffusion of the TF. Thus, the TF tends to stay in either of the two stable states (active or inactive states) for a longer period of time. This increase in the residence time of the TF on the promoter results in reduced transcriptional bursting.

Next we studied the effect of the artificial crowding agent on biochemical rates, by comparing the distributions of active state duration (t_{on}) for 3200 trajectories of 1000 min simulations (Figs. 2.1c and d). Fig. 2.1c and d show a significant (p-value < 0.001) decrease in t_{on} for θ = 100%. Likewise, given $t_{on} + t_{off} = 1000$ min, we observe a significant increase in t_{off} for θ = 100%. Therefore, using gene activation rate constant $k_+ \sim < t_{off} >^{-1}$ (<.> denotes the mean), our simulation results suggest a 12% decrease in k_+ (in agreement with [65]) and a 23% increase in the deactivating rate constant (k_-). Our model predicts a smaller reduction in k_- compared to [65], and thus, predicts a 29% decrease in equilibrium constant ($K_{eq}=k_+/k_-$)

whereas [65] predicts an increase in K_{eq}. This discrepancy might be due to the assumption in [65] that the association rates are always diffusion limited. It would be interesting to repeat

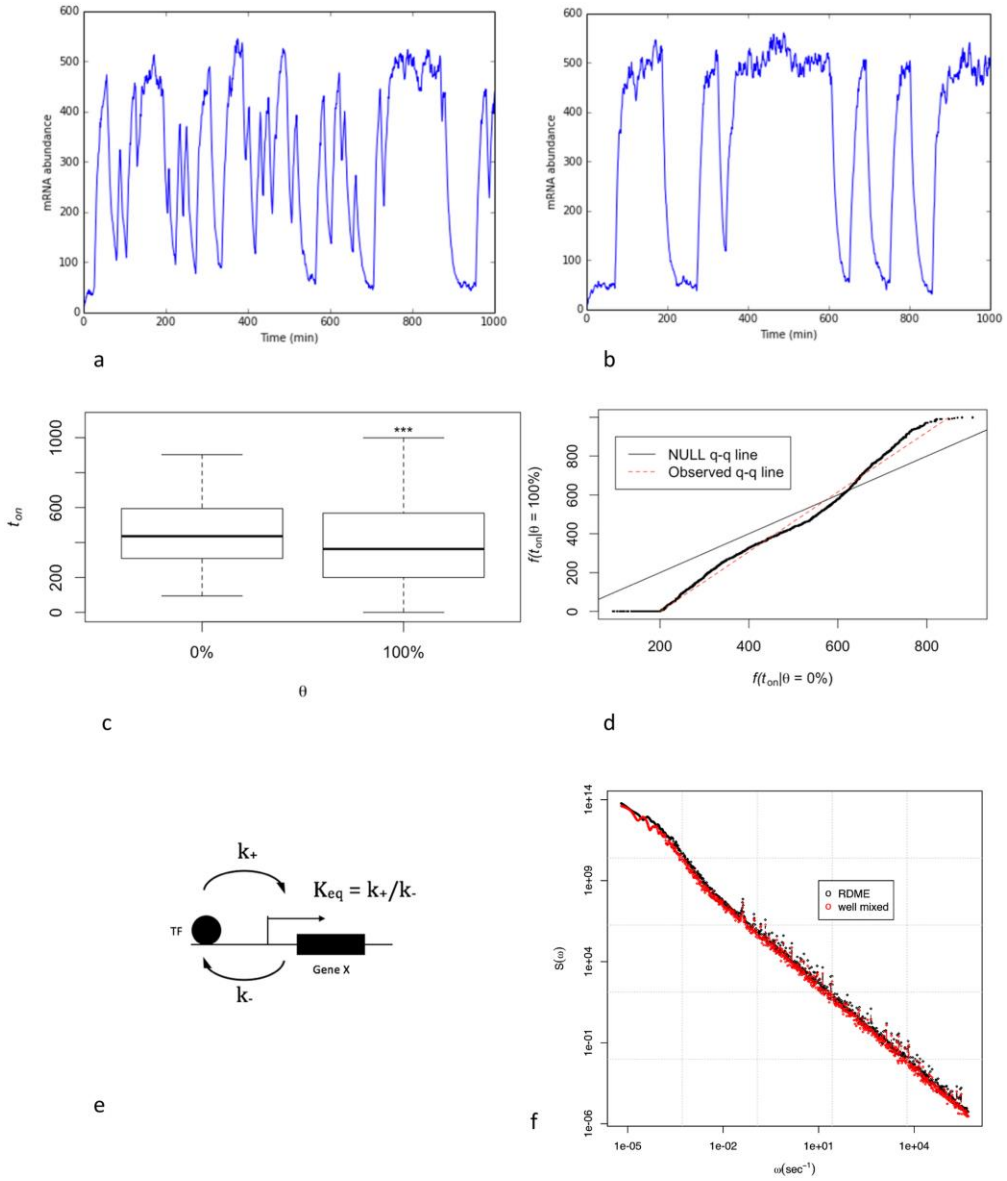


Figure 2.1. Macromolecular crowding can increase the transcriptional bursting by limiting the diffusion and increasing the residence time of TF on the promoter. a) gene expression dynamics in the presence of chromatin as the only volume exclusion factor. b) gene expression dynamics when a large crowding agent (e.g. Dex) is added. c) comparison between the distributions of active state duration (t_{on}) for θ = 0% vs. θ = 100% (*** p-value < 0.001). d) qq-plot of distributions of t_{on} for θ = 0% vs. θ = 100%. Significant deviation between quantiles of t_{on} distributions (fitted red line) and the Null distribution implies a reduction in average t_{on}.

e) effective two states well-mixed gene model. c and d are obtained by 3200 trajectories of 1000 min simulations. f) the power spectrum of mRNA obtained by the spatial model is in agreement with a well-mixed model using the effective rate constants.

similar simulations using particle level methods such as molecular dynamics to obtain a more precise estimate of the change in the equilibrium constant. Our finding is in qualitative agreement with the experimental observation that a crowded condition of heterochromatin can repress gene expression [49] (Fig. 2.1e).

Van Paijmans and Ten Wolde [60] showed that in general the abovementioned biochemical system can be reduced to a well-mixed model if there is a clear separation of time scales between rebinding and binding of molecules from the bulk, which can be deduced from the power spectrum of the mRNA expression. Briefly, a characteristic knee in the low-frequency regime (corresponding to Markovian switching at long times), which is well separated from the regime corresponding to the rebindings at higher frequencies renders it possible for a spatially resolved biochemical system to be reduced to a well-mixed system. To explore whether our system can be reduced into a well-mixed system, we used the effective biochemical rate constants obtained by measuring the transcriptional activity (Fig. 2.1e). By comparing the power spectrum of the spatial model for the special case when θ = 100% with the corresponding well-mixed model, we conclude that once the effective biochemical rate constants are measured using our spatial model for a given configuration (i.e. distinct crowding size and distribution), spatial model can be reduced into a well-mixed model (Fig. 2.1f). Note, however, as shown later in this study, these biochemical rate constants depend strongly on the size and the distribution of the crowding agent molecules, and the local chromatin density. Hence, a spatial model is required to measure these constants.

To analyze the consequences of macromolecular crowding on a cell population, we simulated 16000 isogenic cells in an analogous situation for different values of θ. We observed (Fig. 2.2a) that while low θ values can diversify the cell population and result in

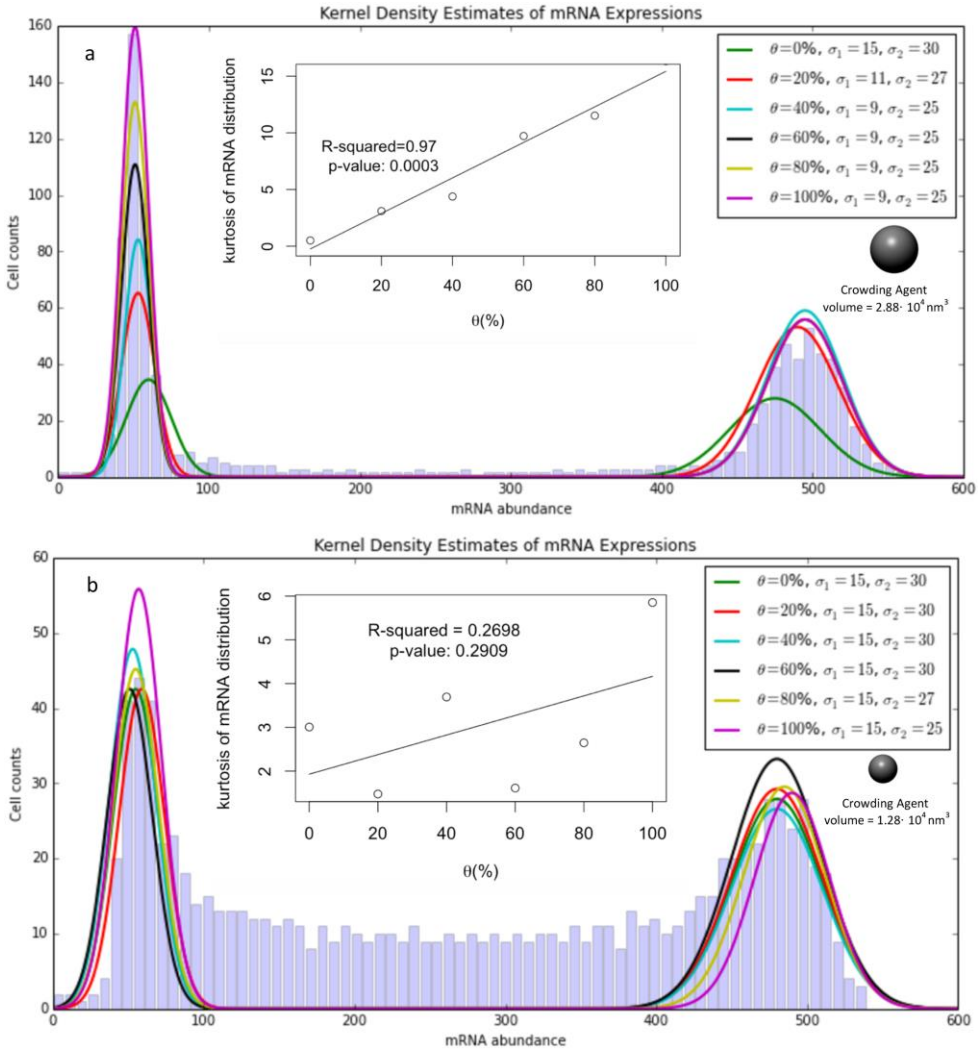


Figure 2.2. a) A large crowding agent can homogenize a cell population effectively. Adding a crowding agent diminishes the probability of the intermediate states (i.e. 100<mRNA<400) and results in a more uniform cell population that lies in either of the two stable states (p-value < 0.01). The results illustrate a strong correlation between the concentration of the crowding agent and the kurtosis of the distribution of mRNA abundance. Each distribution represents 16000 data points and was obtained using kernel-density estimate (KDE) [52]. b) A small crowding agent is incapable of producing a uniform cell population (p-value > 0.01). Intermediate states remain intact after crowding the cells by a small crowding agent. Each distribution represents 16000 data points and was obtained using kernel-density estimate (KDE). The histograms show the total number of the cells with certain expression levels. Here only histograms for θ = 100%, large crowding agent (a) and θ = 0% (b) which provide the least and the most intermediate states are shown. σ_1 and σ_2 are the approximate standard deviation corresponding to the first and the second mode, respectively.

intermediate states (two peaks correspond to two stable states, i.e. active and inactive states), with higher values of θ we observed a more homogeneous population (no intermediate states). This observation is in agreement with recent experimental results [35]. In this situation, the average number of mRNA is close to the number of mRNA obtained when noise is removed from gene expression (deterministic models). Our simulation results show that adding the crowding agent to the simulation domain replaces the intermediate states by two more stable states. The two stable modes (mRNA abundance = 50 and 500) are intact after crowding the simulation domain (Fig. 2.2a). It can be inferred from our linear model that there is a statistically significant correlation between kurtosis of the mRNA distribution and the amount of the crowding agent (p-value < 0.01).

We should stipulate that the kurtosis values define the noise in our system. Low kurtosis values correspond to a cell population in which mRNA expression in each cell is near either the first or the second peak (i.e. ~50 and 500). Conversely, high kurtosis corresponds to a cell population in which certain cells have mRNA expression levels that lay between the peaks (i.e. intermediate states). Likewise, a more homogenous cell population can be obtained by removing the intermediate states (i.e. higher kurtosis value and narrower distributions or lower noise).

It has been observed experimentally [35] that the larger crowding agents (Dex-Big) can contribute robustness to the gene expression pattern more effectively than the smaller crowding agents (Dex-Small). To examine whether our model would reproduce this observation, we repeated the previous simulations using smaller crowding agents (~2 times smaller by volume fraction). Larger crowding agents occupy more volume in a voxel and reduce the diffusion coefficient more effectively than smaller crowding agents (90%

reduction in the diffusion coefficient for larger crowding agents compared to 40% reduction for smaller crowding agents). However, by occupying more voxels (~2 times as many voxels as in the larger crowding agent case), a similar level of volume exclusion can be achieved by smaller crowding agents. Note that in order to assess the effect of the artificial crowding agent size, one should compare the kurtosis of mRNA distributions for θ values that correspond to similar total volume exclusion for Dex-Big vs. Dex-Small (e.g. Dex-Big and θ = 60% vs. Dex-Small and θ = 100%).

Our diffusion-limited gene expression model is capable of reproducing the same experimental observations (Fig. 2.2b). Our simulation results suggest that the intermediate states do not vanish, despite adding a substantial amount of small crowding agents. Our linear regression model illustrates a small correlation between the kurtosis of the mRNA distribution and the amount of the crowding agent (p-value > 0.01). Therefore, we can conclude that, in agreement with experimental observations, our model shows that the smaller crowding agents cannot homogenize the cell population effectively. This is not surprising since small molecules exist in the cellular environment in high concentrations but their impact on gene expression is negligible compared to histones, mRNAs and regulatory proteins.

Next, we analyzed the impact of chromatin reorganization, to understand how the local volume exclusion of chromatin can influence the gene expression patterns of specific genes. Three different genes were selected (Genes 1-3) to account for super dense (Gene 1), dense (Gene 2) and sparse chromatin area (Gene 3). Identical model and simulation parameters were used for all three genes to control for other effects except the volume exclusion of chromatin. By comparing the mRNA distributions of cell populations consisting of 16000 cells, our simulation results suggest that diffusion-limited gene expression can alter mRNA

production in a cell population (Fig. 2.3). Here, the two-sample (all compared to Gene3) Kolmogorov-Smirnov (KS) test (Bonferroni-adjusted) was used to compare different mRNA distributions and a statistically significant difference was obtained (p-value < 0.01).

To demonstrate that macromolecular crowding reduces the gene expression noise primarily by volume exclusion, thus limiting the diffusion, we repeated the simulations in the absence of the crowding agents but using different diffusion coefficients. This was implemented by replacing the diffusion coefficients (D) with the effective diffusion coefficient (D*) (see Methods). Each data point (X, Y) in Fig. 2.4 was found by running the simulation for different D values (X) and evaluating the kurtosis of mRNA distributions. Then the corresponding D* values (Y) were obtained by Eq 7. We hypothesized that if macromolecular crowding is capable of reducing the noise of gene expression primarily by slowing down the diffusion of TF, we should expect to see a linear fit in our data points with the hypothetical line (Fig. 2.4, red dotted line). As shown in Fig. 2.4 our simulation results support this hypothesis for a physical range of θ values (0-100%), for a large crowding agent.

As previously discussed, the size of the crowding agent plays a vital role in obtaining a homogeneous cell population. By comparing the kurtosis values of the mRNA distributions obtained using a large crowding agent (θ = 60%) vs. a small crowding agent from Fig. 2.2 (θ = 100%), where the total volume exclusion is similar, different phenotypes can be observed (kurtosis value of ~10 vs. ~4). Furthermore, the position of the gene within the chromatin matters. It can be inferred that although the overall volume exclusion effect is similar for all three genes, the local chromatin density can alter the time a TF requires to reach its target. In sum, our study suggests that macromolecular crowding can influence the gene expression noise significantly, both locally and globally (Fig. 2.3, yellow curve, p-values < 0.01).

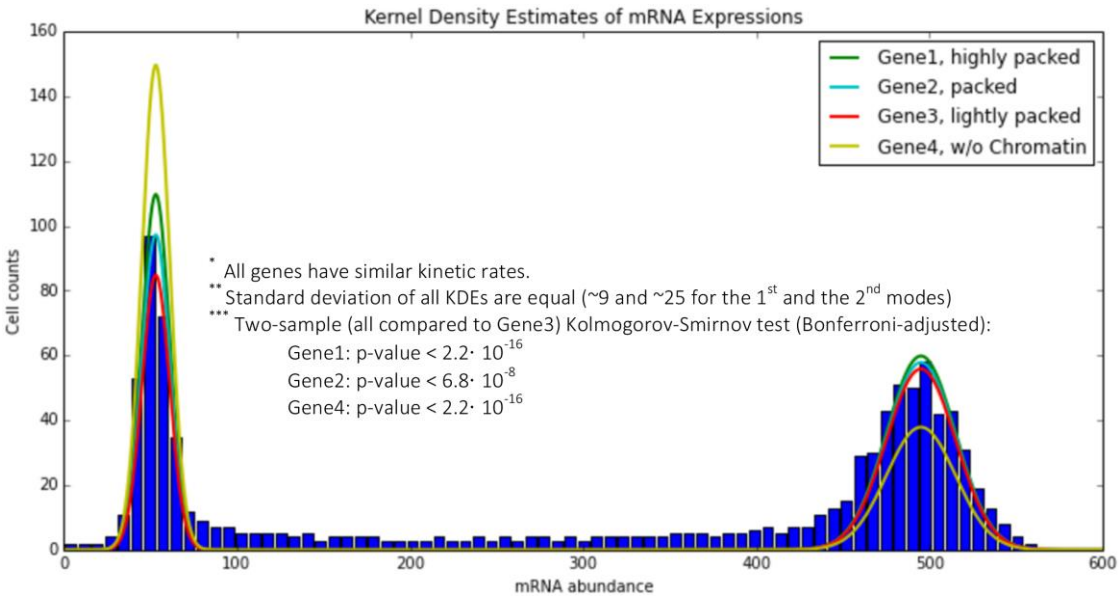


Figure 2.3. Effect of the chromatin structure on the gene expression pattern. Not only does the chromatin structure reduce the diffusion rate due to the macromolecular crowding effect, it also can determine the transcription pattern of different genes due to their location. Each distribution is compared to gene 3 (red curve) using Kolmogorov-Smirnov test (Bonferroni-adjusted for multiple comparisons). p-values show significant difference between distributions (each distribution is obtained by 16000 data points). Histograms show the total number of the cells with certain Gene2 expression levels, indicating the intermediate states for this gene.

Discussion:

A significant portion of cell volume is occupied by proteins, RNAs and other macromolecules. To obtain a complete understanding of the pattern of gene expression, a comprehensive understanding of the impacts of macromolecular crowding is essential. In this study, we have proposed a simple model similar to that of [46] to account for macromolecular crowding in the cellular environment. We utilized the NSM method for simulation of the reaction-diffusion master equation, to include macromolecular crowding. We have avoided any direct manipulation of reaction rates to account for macromolecular

crowding [35]. In addition, our method facilitates an explicit treatment of macromolecular crowding, in that geometric dependency of chromatin structure on gene expression is addressed, and interactions between the crowding agent and different molecules can be considered. This provides a platform to assess how the chromatin structure impacts gene expression. Our model accounts for the addition of the artificial crowding agent and its size, and demonstrates that macromolecular crowding can homogenize a cell population by limiting the diffusion of TFs. Therefore, it improves our understanding of the underlying sources of gene expression noise from that of the earlier models [35, 46].

Our model predicts that a large crowding agent (Dex-big), reduces the diffusion coefficient of TF more effectively than a small crowding agent (Dex-small), in agreement with the experimental observations by Tan *et al*. Likewise, it can be inferred from other experimental observations by Phillies *et al.* [69] that the molecular weight and concentration of crowding molecules can change the diffusion coefficient considerably, whereas the size of a TF has insignificant impact. Finally, although Muramatsu and Minton [68] observed an inverse correlation between the size of the crowder and that of the diffusion coefficient, Phillies *et al*. [69] has shown the opposite (this controversy is discussed in [68] as well).

It is worth noting that Isaacson *et al.* [46] used spatial stochastic simulation to show that the first passage time (the time required for TF to find the gene locus) decreases to a minimum at first, and then increases again as the volume exclusion due to chromatin increases further. That study suggests that crowding can accelerate or decelerate the diffusion depending on the density of the crowding agent, leading to faster or slower chemical kinetics, respectively. Our study, on the other hand, demonstrates the mechanism by which crowding can reduce the transcriptional noise of gene expression. For an intuitive understanding of the gene expression noise reduction mechanism, first note that as shown by

van Zon *et al.* [47], TF diffusion is the dominant source of gene expression noise. Also, macromolecular crowding can effectively partition the available space into smaller compartments, which not only linearizes the input–output relation, but also reduces the noise in the total concentration of the output. In fact, by partitioning the space, macromolecular crowding isolates molecules, as a result of which the molecules in the different compartments are activated independently, thereby reducing the correlations in the gene expression switch. Consequently, this removal of correlations can lower the output noise [48]. We suggest the following function for the macromolecular crowding, by which a uniform cell population can be obtained. By comparing Figs 2.1a and b, it can be inferred that macromolecular crowding can increase the average residence time of TF on its promoter. As a consequence, transcriptional bursts are attenuated which leads to elimination of the intermediate states in the mRNA distributions.

Our findings demonstrate the importance of spatial simulations to fully capture several experimental observations. Morelli *et al.* [65] studied the effect of macromolecular crowding on a gene network by rescaling the association and dissociation constants into a well-mixed model. Here, on the other hand, we provide strong evidence (Figs. 2.2,3 and Supplementary Fig. 2.4) that the impact of crowding structure and distribution cannot be fully understood using well-mixed models.

Furthermore, our model sheds light on how to develop engineered cells to achieve advantages in gene expression, cellular computing and metabolic pathways [35]. Investigations of other epigenetic factors show that DNA methylation and chromatin structure may be linked to transcriptional activity, both in single cells and across populations. Gene silencing by histone modification or formation of repressed chromatin states (heterochromatin) are good examples of how nature has exploited macromolecular crowding and inherent

stochasticity in gene expression to display new traits [49]. Our methodology can be utilized to further assess heterochromatin and euchromatin functional differences at a reasonable resolution.

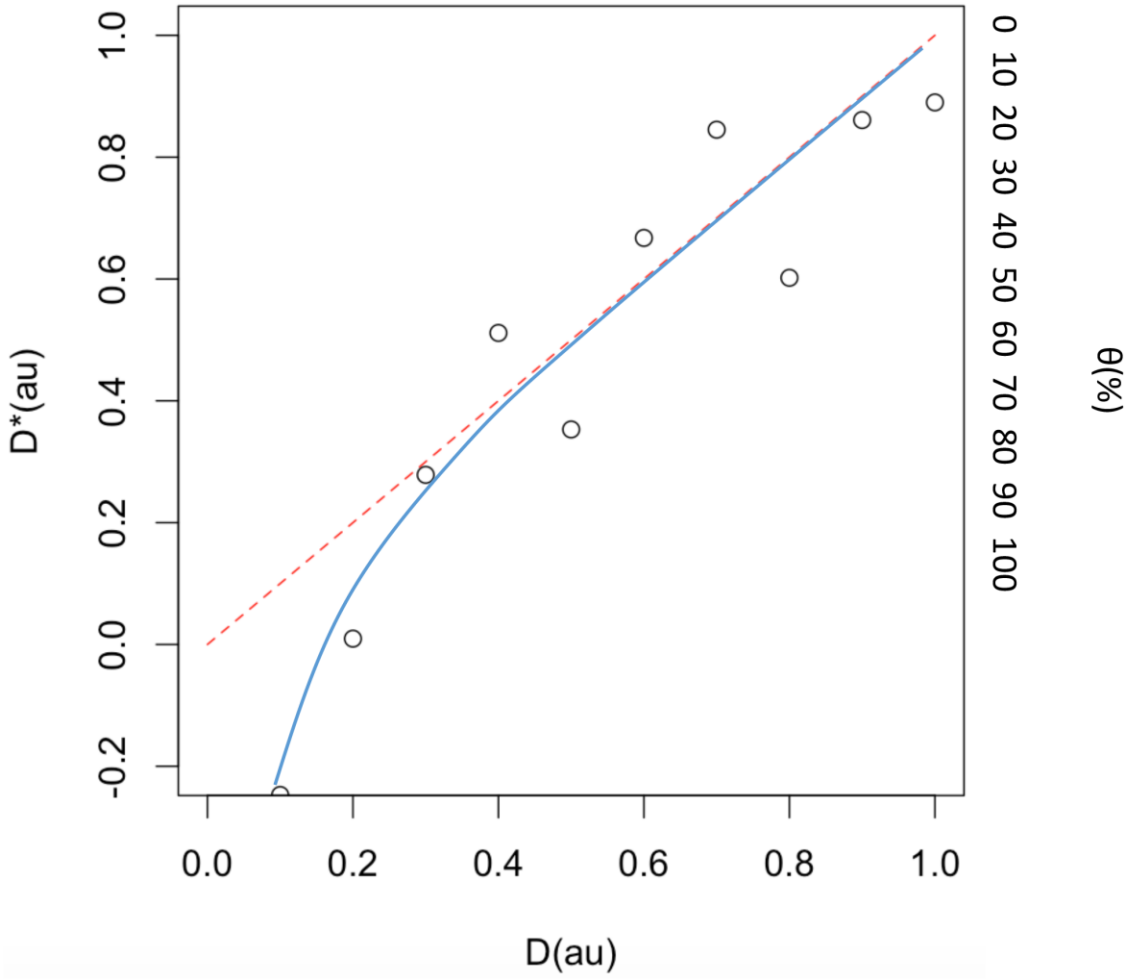


Figure 2.4. Impact of the diffusion-limited gene expression on the cell population diversity (D* is the effective diffusion coefficient). The red dotted line shows the hypothetical line that explains the variability in a cell population completely as a result of diffusion-limited gene expression. The blue line is a logarithmic fit to our simulation results and supports the hypothesis of robust gene expression as a result of diffusion-limited gene expression (a large crowding agent was used).

Methods:

We used a well-known, simple model to describe transcription and translation [8]. Transcription factor (TF) was added to that model to account for spatial effects of TF diffusion in a crowded environment. Given a cubic domain in which protein production takes place, gene expression begins by TF diffusion and finding the locus of the gene of interest. Upon binding/unbinding of TF to/from its promoter, the gene switches between active and inactive states. Without loss of generality, the gene of interest is placed in the center of a box with a characteristic length L. One and only one TF can activate the promoter. Thus, the chemical system of protein production can be written as:

$$TF + Promoter_{Repressed} \overset{k_{on},\, k_{off}}{\rightleftharpoons} Promoter_{Active} \overset{s_A}{\rightarrow} M + Promoter_{Active} \quad \text{(R1)}$$

$$Promoter_{Repressed} \overset{s_R}{\rightarrow} M + Promoter_{Repressed} \quad (R2)$$

$$M \overset{s_P}{\rightarrow} P + M \quad (R3)$$

$$M \overset{\delta_M}{\rightarrow} \emptyset \quad (R4)$$

$$P \overset{\delta_P}{\rightarrow} \emptyset \quad (R5)$$

The simulation parameters were adopted from [8] for the sake of comparison with non-spatial methods (Table 2.1).

Table 2.1. Model Parameters

Parameters	Values	Reference	Species	Initial Conditions
D_{TF}	10 ($\mu m^2.min^{-1}$)	[53]	Transcription Factor (TF)	1
$D_{Chromatin}$	~ 0	[54]	mRNA (M)	0
k_{on}	0.1 (min^{-1})	[55]	Protein (P)	0
k_{off}	0.1 (min^{-1})	[56]	Simulation Parameters	Values
s_A	50 (min^{-1})	[8]	Box Dimensions [35]	1μm by 1μm by 1μm
s_R	5 (min^{-1})	[8]	Number of voxels	50 by 50 by 11
s_P	0.2 (min^{-1})	[8]	Simulation Time	1000 min
$_P$	0.05 (min^{-1})	[8]		
δ_M	0.1 (min^{-1})	[8]		

doi:10.1371/journal.pcbi.1005122.t001

Simulation Algorithm

The inherent stochastic characteristics of gene expression, along with the failure of deterministic models to produce transcriptional bursting, lead us to consider a spatial stochastic model. A modified next subvolume method (NSM) [41] was used to simulate the stochastic reaction-diffusion system, using the implementation in PyURDME on the MOLNS software platform [57]. We developed the following modifications to account for crowding (for access to the software implementation, see URL in [58]).

Inside the cell, the chromatin, histones, etc., are crowding the nucleus. Note that we are ignoring dynamic addition and reduction of newly synthetized proteins (*P*) and mRNAs (*M*) since they are negligible when compared to the chromatin. Given a domain which is discretized into N uniform voxels, each voxel is occupied with the artificial crowding agent with a probability θ. The diffusion between two adjacent voxels is linearly dependent on the crowding density of the destination voxel, consisting of the chromatin and the artificial crowding agent. This model assumption is analyzed in detail and compared with the available experimental data in Supporting Information (Supplementary Fig. 2.1). This model does not explicitly take into account lock-in effects, that crowding in the origin voxel may affect the diffusion rate to adjacent voxels, or that the effective reaction rate in a voxel may

depend on the local crowding and configuration of the chromatin and crowders. For instance, Friedman [66] showed that hydrodynamic effects cause a 15% reduction in the computed rate constant for neutral species or ions in water. The impacts of the electrostatic forces have been widely studied and considered primarily in molecular level simulations [67]. Specific chromatin configurations can affect the hopping rate of particles differently. Namely, even in low chromatin concentrations, distinct configurations might be able to fully trap the particle and reduce the hopping rates significantly. However, we believe that our model is sufficiently accurate to study the qualitative effects of crowding.

It is worth mentioning that our model ignores any non-specific interaction between DNA and TF. Paijmans and ten Wolde [60] showed quantitatively that even in the presence of 1D sliding along the DNA, which makes rebinding events not only more frequent but also longer, the effect of diffusion can still be captured in a well-stirred model by renormalizing the rate constants. However, renormalization does not account for the architecture of chromatin and how it can influence the rate constants. Although several studies suggest that such non-specific interactions can help TF to slide on the DNA strand (facilitated diffusion) to find the target faster [61, 62], recent work by Wang F *et al.* [63] provides evidence that the promoter-search mechanism of E. coli RNAP is dominated by 3D diffusion. Moreover, in another work [64] the sliding length of TF on DNA is measured to be ~30-900 bps. In our simulation, on the other hand, each voxel contains ~Mbps and therefore, on the length scale of our model, facilitated diffusion is insignificant.

The size of the crowding agent is modeled by the parameter δ_i. We assume that smaller crowders reduce the diffusion less than large crowders. In our simulations we let $\delta_i = 0.6$ for

smaller crowding agents, δ_i = 0.1 for larger crowding agents, and δ_i = 1 when no crowding agent is present. Thus, the diffusion rate into voxel *i* is computed as

$$D = D_0 \times (1 - c_i) \times \delta_i \qquad (5)$$

where c_i models the crowding due to the chromatin in voxel *i*. It is unknown exactly how the concentration of chromatin affects the effective diffusion, but as a simple model we assume that

$$c_i = \frac{DAPI\ intensity\ in\ cell\ i}{max_j\ DAPI\ intensity\ in\ voxel\ j} \qquad (6)$$

The diffusion rate thus depends linearly on the DAPI intensity, and we assume that the voxel with the highest intensity of DAPI is fully blocked. For simplicity we assume that neither the chromatin nor the crowding agent diffuses between voxels.

The TF molecule is initially placed randomly in the domain. During the simulation it will diffuse to the gene locus and activate transcription. Recent studies [46, 51] have proposed more complicated relations to obtain the effective diffusion coefficient in the presence of macromolecular crowding. Here, we use a linear relation to calculate the TF diffusion coefficient as a function of the total crowdedness (i.e. the effects of both chromatin structure and artificial crowding agents included). This simple relation can capture physiologically relevant trends and suffices for the purpose of our simulations.

Effective Diffusion Rate Calculation:

Considering the total effect of the artificial crowding agent as

$$D^* = \frac{\sum_i \delta_i}{N} D, \qquad (7)$$

where *i* is the voxel index and *N* is the total number of voxels in the domain. For a large crowding agent, Eq 7 leads to D* = [θ×0.1 + (1- θ) ×1]D = (1-0.9 θ)D. Using the linear model presented in Fig. 2.2a (Kurtosis(θ) = 15 θ), we obtain (for D = 1)

$$D^* = 1 - 0.06 \times \text{Kurtosis} \qquad (8)$$

to calculate the effective diffusion rate. Each data point (X, Y) in Fig. 2.4 is found by running the simulation for different D values (X) and evaluating the kurtosis of the mRNA distributions. Then the corresponding D* values (Y) are obtained by Eq 8.

In summary, in order to obtain the effective diffusion as illustrated in Fig. 2.4, the following procedure has been followed:

1. We performed different simulations by varying the diffusion coefficient *D* in the absence of any crowding, and the kurtosis values of mRNA distributions were calculated.
2. Next, in order to pinpoint the corresponding effective diffusion D^* that leads to the same kurtosis value as *D* in the presence of a crowder, we need to determine the crowding parameter (θ). θ can be estimated using a linear regression model as shown in Fig. 2.2a, Kurtosis(θ) = 15 θ.
3. Finally, by substituting θ into D* = [θ×0.1 + (1- θ) ×1]D = (1-0.9 θ)D from Eq 2.7, D^* can be obtained using Eq 8.
4. 1-3 should be applied to all D values to obtain a set of (D, D*) in order to produce Fig. 2.4.

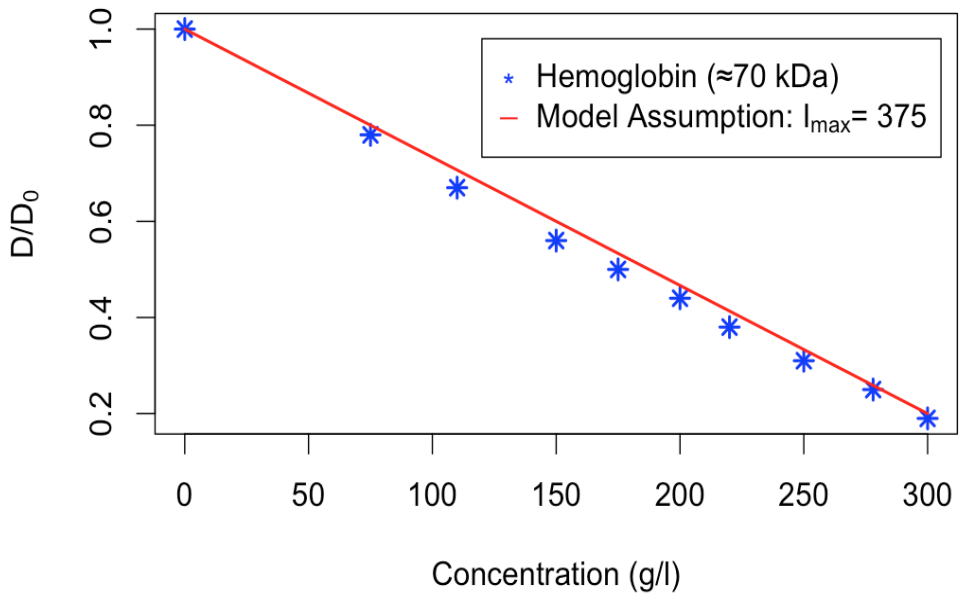
Supporting Information:

Impact of macromolecular crowding on the diffusion coefficient

To model the impact of crowding on the diffusion coefficient, we assume a linear relation between the diffusion coefficient and the concentration of the crowding. Given Eq 5, we have

$$\frac{D}{D_0} = (1 - c_i) = 1 - \frac{I_i}{I_{max}} = 1 + (-\frac{1}{I_{max}})I_i \qquad (S1)$$

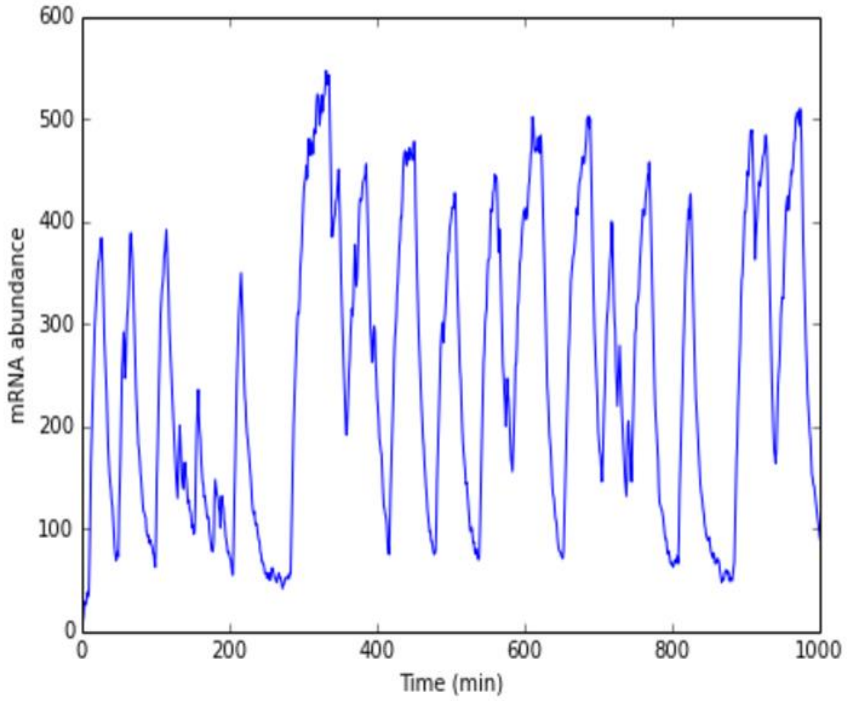
where I_i is the DAPI concentration in voxel i. Comparison between Eq S1 for the right choice of the normalization factor I_{max} and the available experimental data for Hemoglobin [68] over physiological crowding ranges shows a good agreement between the model assumption and the experimental observations (Supplementary Fig. 2.1).



Supplementary Figure 2.1. The validity of a linear relation between the diffusion coefficient and the crowder concentration over physiological crowding ranges.

Transcriptional bursting in the absence of macromolecular crowding

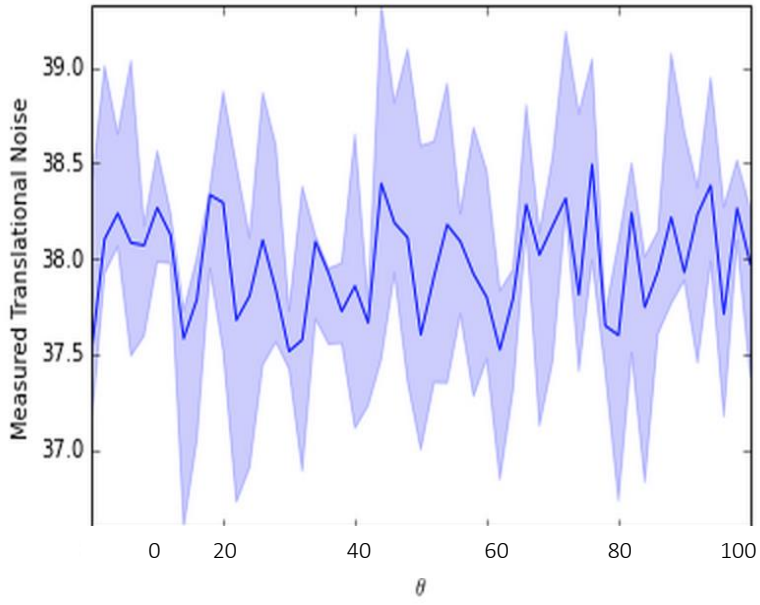
To verify our model with the previous non-spatial stochastic simulations [8], gene expression in the absence of the crowding agent and the chromatin structure (i.e. $D_i = D_0$) was studied. It can be seen (Supplementary Fig. 2.2) that our model is capable of producing transcriptional bursting similar to that reported for well-mixed stochastic simulation results [8].



Supplementary Figure 2.2. Transcriptional bursting in the absence of macromolecular crowding.

Impact of Macromolecular Crowding on Translational Bursting

To assess how translational bursting can be affected by macromolecular crowding, we present a quantitative description of noise for different values of crowdedness (Supplementary Fig. 2.3). Our analysis suggests that macromolecular crowding has an insignificant impact on translational bursting (standard deviation (SD) is used to compare the gene expression noises).

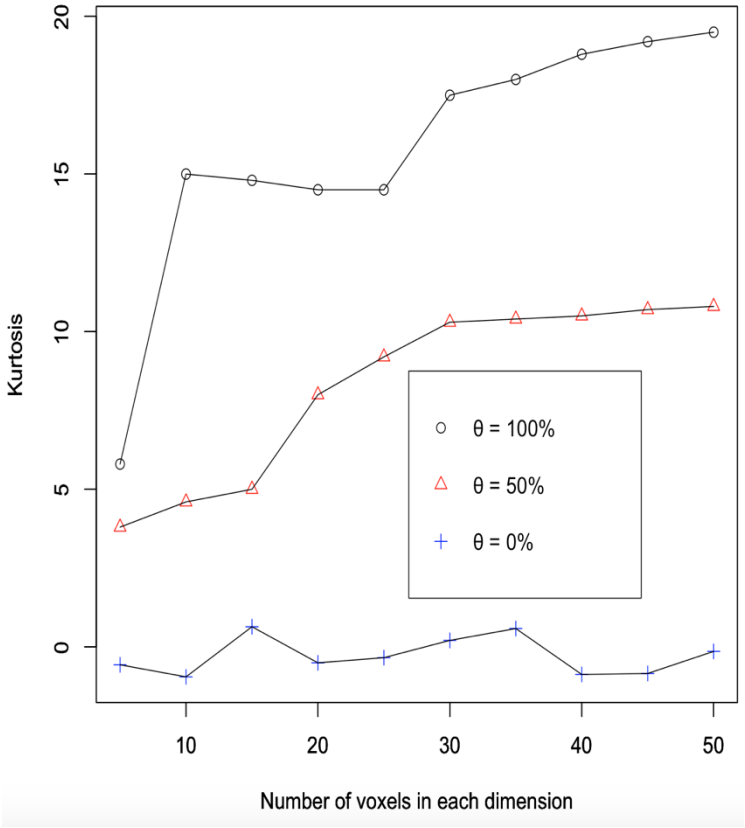


Supplementary Figure 2.3. Comparison between translational bursting and gene expression noise for various amounts of crowding agents. Macromolecular crowding has an insignificant impact on translational bursting and gene expression noise.

Study of voxel size effects

To assess the effects of voxel size on the results of our model, we performed a mesh refinement study. The results are shown in Supplementary Fig. 2.4. We ran our model

for three different crowdedness parameters. For each θ value, we ran our model for different mesh sizes to determine the extent of phenotype dependency on mesh resolution. As can be seen in Supplementary Fig. 2.4, the kurtosis of the mRNA distributions converges as we refine the mesh. The kurtosis tends to near zero as the number of voxels tends to 1, as expected since 1 voxel is a well-mixed simulation. This demonstrates the importance of the use of inhomogeneous stochastic simulation to capture the noise reduction by macromolecular crowding.



Supplementary Figure 2.4. Convergence study of modified NSM method. The kurtosis of the mRNA distributions converge as we refine the mesh; however, a coarser mesh is incapable of showing diffusion-limited gene expression noise reduction (δ_i = 0.1).

References:

1. Yamanaka S. Elite and stochastic models for induced pluripotent stem cell generation. Nature. 2009 Jul 2;460(7251):49.

2. Acar M, Mettetal JT, van Oudenaarden A. Stochastic switching as a survival strategy in fluctuating environments. Nature Genetics. 2008 Apr 1;40(4):471-5.

3. Singh A, Weinberger LS. Stochastic gene expression as a molecular switch for viral latency. Current Opinion in Microbiology. 2009 Aug 31;12(4):460-6.

4. Feinberg AP, Irizarry RA. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proceedings of the National Academy of Sciences. 2010 Jan 26;107(suppl 1):1757-64.

5. Cook DL, Gerber AN, Tapscott SJ. Modeling stochastic gene expression: implications for haploinsufficiency. Proceedings of the National Academy of Sciences. 1998 Dec 22;95(26):15641-6.

6. Capp JP. Stochastic gene expression, disruption of tissue averaging effects and cancer as a disease of development. Bioessays. 2005 Dec 1;27(12):1277-85.

7. Bahar R, Hartmann CH, Rodriguez KA, Denny AD, Busuttil RA, Dollé ME. Increased cell-to-cell variation in gene expression in aging mouse heart. Nature. 2006 Jun 22;441(7096):1011-4.

8. Kaern M, Elston TC, Blake WJ, Collins JJ. Stochasticity in gene expression: from theories to phenotypes. Nature Reviews Genetics. 2005 Jun 1;6(6):451-64.

9. Swain PS, Elowitz MB, Siggia ED. Intrinsic and extrinsic contributions to stochasticity in gene expression. Proceedings of the National Academy of Sciences. 2002 Oct 1;99(20):12795-800.

10. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science. 2002 Aug 16;297(5584):1183-6.

11. Raser JM, O'Shea EK. Noise in gene expression: origins, consequences, and control. Science. 2005 Sep 23;309(5743):2010-3.

12. Raser JM, O'Shea EK. Control of stochasticity in eukaryotic gene expression. Science. 2004 Jun 18;304(5678):1811-4.

13. Pedraza JM, van Oudenaarden A. Noise propagation in gene networks. Science. 2005 Mar 25;307(5717):1965-9.

14. Chong S, Chen C, Ge H, Xie XS. Mechanism of transcriptional bursting in bacteria. Cell. 2014 Jul 17;158(2):314-26.

15. Blake WJ, Kærn M, Cantor CR, Collins JJ. Noise in eukaryotic gene expression. Nature. 2003 Apr 10;422(6932):633-7.

16. Eldar A, Elowitz MB. Functional roles for noise in genetic circuits. Nature. 2010 Sep 9;467(7312):167-73.

17. Chubb JR, Liverpool TB. Bursts and pulses: insights from single cell studies into transcriptional mechanisms. Current opinion in genetics & development. 2010 Oct 31;20(5):478-84.

18. Friedman N, Cai L, Xie XS. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. Physical Review Letters. 2006 Oct 19;97(16):168302.

19. Korobkova E, Emonet T, Vilar JM, Shimizu TS, Cluzel P. From molecular noise to behavioural variability in a single bacterium. Nature. 2004 Apr 1;428(6982):574-8.

20. Spencer SL, Gaudet S, Albeck JG, Burke JM, Sorger PK. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. Nature. 2009 May 21;459(7245):428-32.

21. Di Talia S, Skotheim JM, Bean JM, Siggia ED, Cross FR. The effects of molecular noise and size control on variability in the budding yeast cell cycle. Nature. 2007 Aug 23;448(7156):947-51.

22. Süel GM, Kulkarni RP, Dworkin J, Garcia-Ojalvo J, Elowitz MB. Tunability and noise dependence in differentiation dynamics. Science. 2007 Mar 23;315(5819):1716-9.

23. Niepel M, Spencer SL, Sorger PK. Non-genetic cell-to-cell variability and the consequences for pharmacology. Current Opinion in Chemical Biology. 2009 Dec 31;13(5):556-61.

24. Rao CV, Wolf DM, Arkin AP. Control, exploitation and tolerance of intracellular noise. Nature. 2002 Nov 14;420(6912):231-7.

25. Sharp KA. Analysis of the size dependence of macromolecular crowding shows that smaller is better. Proceedings of the National Academy of Sciences. 2015 Jun 30;112(26):7990-5.

26. Ellis RJ. Macromolecular crowding: obvious but underappreciated. Trends in Biochemical Sciences. 2001 Oct 1;26(10):597-604.

27. Dobson CM. Chemical space and biology. Nature. 2004 Dec 16;432(7019):824-8.

28. Zaki A, Dave N, Liu J. Amplifying the macromolecular crowding effect using nanoparticles. Journal of the American Chemical Society. 2011 Dec 19;134(1):35-8.

29. Dominak LM, Keating CD. Macromolecular crowding improves polymer encapsulation within giant lipid vesicles. Langmuir. 2008 Nov 4;24(23):13565-71.

30. Jones JJ, van der Maarel JR, Doyle PS. Effect of nanochannel geometry on DNA structure in the presence of macromolecular crowding agent. Nano Letters. 2011 Oct 14;11(11):5047-53.

31. Minton AP. The effect of volume occupancy upon the thermodynamic activity of proteins: some biochemical consequences. Molecular and Cellular Biochemistry. 1983 Sep 1;55(2):119-40.

32. Li GW, Berg OG, Elf J. Effects of macromolecular crowding and DNA looping on gene regulation kinetics. Nature Physics. 2009 Apr 1;5(4):294-7.

33. Richter K, Nessling M, Lichter P. Experimental evidence for the influence of molecular crowding on nuclear architecture. Journal of Cell Science. 2007 May 1;120(9):1673-80.

34. Beg QK, Vazquez A, Ernst J, de Menezes MA, Bar-Joseph Z, Barabási AL, et al. Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. Proceedings of the National Academy of Sciences. 2007 Jul 31;104(31):12663-8.

35. Tan C, Saurabh S, Bruchez MP, Schwartz R, LeDuc P. Molecular crowding shapes gene expression in synthetic cellular nanosystems. Nature Nanotechnology. 2013 Aug 1;8(8):602-8.

36. Stundzia AB, Lumsden CJ. Stochastic simulation of coupled reaction–diffusion processes. Journal of Computational Physics. 1996 Aug 31;127(1):196-207.

37. Lis M, Artyomov MN, Devadas S, Chakraborty AK. Efficient stochastic simulation of reaction–diffusion processes via direct compilation. Bioinformatics. 2009 Sep 1;25(17):2289-91.

38. Dobrzyński M, Rodríguez JV, Kaandorp JA, Blom JG. Computational methods for diffusion-influenced biochemical reactions. Bioinformatics. 2007 Aug 1;23(15):1969-77.

39. Lawson MJ, Drawert B, Khammash M, Petzold L, Yi TM. Spatial stochastic dynamics enable robust cell polarization. PLoS Comput Biol. 2013 Jul 25;9(7):e1003139.

40. Holloway DM, Lopes FJ, da Fontoura Costa L, Travençolo BA, Golyandina N, Usevich K, et al. Gene expression noise in spatial patterning: hunchback promoter structure affects noise amplitude and distribution in Drosophila segmentation. PLoS Comput Biol. 2011 Feb 3;7(2):e1001069.

41. Gillespie DT, Hellander A, Petzold LR. Perspective: Stochastic algorithms for chemical kinetics. The Journal of Chemical Physics. 2013 May 7;138(17):170901.

42. Pirone JR, Elston TC. Fluctuations in transcription factor binding can explain the graded and binary responses observed in inducible gene expression. Journal of theoretical biology. 2004 Jan 7;226(1):111-21.

43. Karmakar R, Bose I. Graded and binary responses in stochastic gene expression. Physical Biology. 2004 Nov 16;1(4):197.

44. Minton AP. The influence of macromolecular crowding and macromolecular confinement on biochemical reactions in physiological media. Journal of biological Chemistry. 2001 Apr 6;276(14):10577-80.

45. Huet S, Lavelle C, Ranchon H, Carrivain P, Victor JM, Bancaud A. Relevance and limitations of crowding, fractal, and polymer models to describe nuclear architecture. International Review of Cell and Molecular Bio. 2014 Jan 1;307:443-79.

46. Isaacson SA, McQueen DM, Peskin CS. The influence of volume exclusion by chromatin on the time required to find specific DNA binding sites by diffusion. Proceedings of the National Academy of Sciences. 2011 Mar 1;108(9):3815-20.

47. van Zon JS, Morelli MJ, Tănase-Nicola S, ten Wolde PR. Diffusion of transcription factors can drastically enhance the noise in gene expression. Biophysical Journal. 2006 Dec 15;91(12):4350-67.

48. ten Wolde PR, Mugler A. Importance of crowding in signaling, genetic, and metabolic networks. Int. Rev. Cell Mol. Biol. 2013 Dec 27;307:419-42.

49. Cheutin T, McNairn AJ, Jenuwein T, Gilbert DM, Singh PB, Misteli T. Maintenance of stable heterochromatin domains by dynamic HP1 binding. Science. 2003 Jan 31;299(5607):721-5.

50. Schermelleh L, Carlton PM, Haase S, Shao L, Winoto L, Kner P, Burke B, et al. Subdiffraction multicolor imaging of the nuclear periphery with 3D structured illumination microscopy. Science. 2008 Jun 6;320(5881):1332-6.

51. Cianci C, Smith S, Grima R. Molecular finite-size effects in stochastic models of equilibrium chemical systems. J. Chem. Phys. 2016 Feb 23; 144: 084101.

52. Parzen E. On estimation of a probability density function and mode. The Annals of Mathematical Statistics. 1962 Sep 1;33(3):1065-76.

53. Hager GL, McNally JG, Misteli T. Transcription dynamics. Molecular Cell. 2009 Sep 24;35(6):741-53.

54. Chubb JR, Boyle S, Perry P, Bickmore WA. Chromatin motion is constrained by association with nuclear compartments in human cells. Current Biology. 2002 Mar 19;12(6):439-45.

55. Elf J, Li GW, Xie XS. Probing transcription factor dynamics at the single-molecule level in a living cell. Science. 2007 May 25;316(5828):1191-4.

56. Wong OK, Guthold M, Erie DA, Gelles J. Interconvertible Lac repressor–DNA loops revealed by single-molecule experiments. PLoS Biol. 2008 Sep 30;6(9):e232.

57. Drawert B, Trogdon M, Toor S, Petzold L, Hellander A. MOLNs: A Cloud Platform for Interactive, Reproducible, and Scalable Spatial Stochastic Computational Experiments in Systems Biology Using PyURDME. SIAM Journal on Scientific Computing. 2016 Jun 1;38(3):C179-202.

58. https://github.com/mgolkaram/pyurdme/tree/crowding

59. Abràmoff MD, Magalhães PJ, Ram SJ. Image processing with ImageJ. Biophotonics International. 2004;11(7):36-42.

60. Paijmans J, ten Wolde PR. Lower bound on the precision of transcriptional regulation and why facilitated diffusion can reduce noise in gene expression. Physical Review E. 2014 Sep 25;90(3):032708.

61. Mirny L, Slutsky M, Wunderlich Z, Tafvizi A, Leith J, Kosmrlj A. How a protein searches for its site on DNA: the mechanism of facilitated diffusion. Journal of Physics A: Mathematical and Theoretical. 2009 Oct 13;42(43):434013.

62. Klenin KV, Merlitz H, Langowski J, Wu CX. Facilitated diffusion of DNA-binding proteins. Physical Review Letters. 2006 Jan 9;96(1):018104.

63. Wang F, Redding S, Finkelstein IJ, Gorman J, Reichman DR, Greene EC. The promoter-search mechanism of Escherichia coli RNA polymerase is dominated by three-dimensional diffusion. Nature Structural & Molecular Biology. 2013 Feb 1;20(2):174-81.

64. Wunderlich Z, Mirny LA. Spatial effects on the speed and reliability of protein–DNA search. Nucleic Acids Research. 2008 Jun 1;36(11):3570-8

65. Morelli MJ, Allen RJ, Ten Wolde PR. Effects of macromolecular crowding on genetic networks. Biophysical Journal. 2011 Dec 21;101(12):2882-91.

66. Friedman HL. A Hydrodynamic Effect in the Rates of Diffusion-Controlled Reactions1. The Journal of Physical Chemistry. 1966 Dec;70(12):3931-3.

67. Davis ME, McCammon JA. Electrostatics in biomolecular structure and dynamics. Chemical Reviews. 1990 May;90(3):509-21.

68. Muramatsu N, Minton AP. Tracer diffusion of globular proteins in concentrated protein solutions. Proceedings of the National Academy of Sciences. 1988 May 1;85(9):2984-8.

69. Phillies GD, Ullmann GS, Ullmann K, Lin TH. Phenomenological scaling laws for ‘‘semidilute’’macromolecule solutions from light scattering by optical probe particles. The Journal of Chemical Physics. 1985 Jun 1;82(11):5242-6.

CHAPTER 3

The Role of Chromatin Density in Cell Population Heterogeneity during Stem Cell Differentiation

Introduction:

Recent advances have enabled mapping of the local structure of chromatin by analyzing the complement of DNA-associated proteins and their modifications along chromosomes. The introduction of the chromosome conformation capture (3C) methods by Dekker et al. [1] opened new windows toward a better understanding of chromatin structure. Using spatially constrained ligation followed by locus-specific polymerase chain reaction (PCR), 3C provides information regarding long-range interactions between specific pairs of loci. Other researchers extended the 3C technique to develop chromosome conformation capture (3C)-on-chip (4C) [2, 3], and 3C-Carbon Copy (5C) [4] as fast and unbiased technologies to further study the nuclear architecture. The introduction of the Hi-C method by Lieberman et al. [5] facilitated obtaining the frequency of interactions between all genomic loci in a single experiment. Finally, Fullwood et al. [6] presented chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) for direct analysis of chromatin interactions exclusively to those formed between sites bound by a given DNA- or chromatin-interacting protein.

Recently, Hi-C data has been extensively used to explore chromatin reorganization, by primarily focusing on topologically associating domains (TAD). Such studies revealed the underlying principles of chromatin organization [5] and its variability across single cells of the same population [7], among distinct human cell lineages [8, 9], and between different species [10, 11]. The importance of chromatin remodeling in gene regulation [12-15] has provided compelling evidence that chromatin remodeling can activate or suppress certain genes and can control gene expression profiles, in particular during embryonic stem cell differentiation [16-18]. However, the role of chromatin in inducing heterogeneity in a cellular population has been underappreciated. Here we introduce local chromatin (DNA) density to describe how chromatin condenses differentially in the vicinity of certain gene loci and how this can influence the gene expression profile.

It is well established that phenotypic cell-to-cell variability observed in a clonal cell population is due to gene expression fluctuations [19-22]. Substantial evidence [22-26] supports fast binding and unbinding of RNA polymerase II and transcription factors in transcriptional bursting, which contributes to such fluctuations (noise) at the population level. However, these fluctuations can serve regulatory roles [27-29]. For instance, the experimental reduction of protein expression noise in *Bacillus subtilis* ComK, the regulator for competence for DNA uptake, leads to a decrease in the number of competent cells [27]. Moreover, stochastic activation of both, either one or none of the two alleles of a gene in a heterozygote organism can contribute to the phenomenon of hybrid vigor [28], or as demonstrated by many reports, gene expression noise can lead to differentiation of unicellular organisms into two distinct states of gene expression such as lysis-lysogeny decision in lambda phage-infected *E. coli* as well as the lac operon in *E. coli* [28].

Chang et al. [30] showed that in clonal populations of mouse haematopoietic progenitor cells, those with higher and lower Sca-1 expression are capable of reconstituting the whole population distribution of Sca-1 expression, suggesting hidden multi-stability within one cell type. In particular, several studies have highlighted the vital function of gene expression heterogeneity in cell fate decision [31]. However, despite the convincing evidence underpinning the necessity of heterogeneity in a cell population for stem cells to differentiate, the direct mechanism that governs this heterogeneity and its association with chromatin reorganization during development has not been determined.

Results:

In [32] we illustrated that macromolecular crowding can control gene expression fluctuations. Here we present a model based on the macromolecular crowding effects of chromatin structure to reveal the underlying principles of population heterogeneity and its regulation. By employing Hi-C interaction data [8], we identify the DNA density as a novel element playing a vital role in modulation of gene expression heterogeneity. We verify our model predictions using several new experiments as well as recent single cell RNA-seq data [33]. Our model predicts an increase in population heterogeneity during development, an increased chromatin density of the majority of genes, and an increase in bimodal behavior. Contrary to the current paradigm of epigenetic regulation, the role of chromatin is beyond a mere on and off switch for certain genes: it contains information, even in the unoccupied space of the nucleus, that serves regulatory purposes to control population heterogeneity. In this work we mainly focus on genes with differential heterogeneity but similar mean expression.

Recent discoveries that analyze TADs and chromatin looping interactions [34], as well as recent tools to facilitate building a three-dimensional model of chromatin, allow an unprecedented accuracy [35-39]. Here we first introduce the idea of chromatin density as a regulatory component of gene expression in a cell population. Interphase chromatin is compartmentalized into tightly packed (heterochromatin) and less compact domains (euchromatin) which are localized primarily to the peripheral and inner nucleus, respectively. Gene repression and activation are highly associated with this packaging, and controlled by histone modifications. Trimethylation at H3K4, H3K36, or H3K79 can result in an open chromatin configuration and is, therefore, characteristic of euchromatin, whereas condensed heterochromatin is enriched in trimethylation of H3K9, K3K27, and H4K20 [40]. Furthermore, DNA methylation, histone modification and variants contribute strongly to gene activation and repression during early embryonic development [41]. Aside from this paradigm, the importance of local variation of the chromatin density even within the same chromatin state – heterochromatin or euchromatin – and its regulatory role have not been fully addressed.

We introduce Γ as a parameter for quantification of the local chromatin density (which we will henceforth simply refer to as density) as follows. For any gene X at a certain genomic coordinate j, we imagine a small sphere with a radius R centered at j, and count the number of base pairs (bps) that lie inside this sphere. To be more precise, suppose that x coincides with the transcription start site (TSS) of a gene X. Due to the relatively low resolution of the current Hi-C experiment (~40 Kbps) – which will be discussed later – the center of the density sphere (DS) will be approximated as the start or the end of the genomic coordinates of gene X, depending on whether transcription occurs on the positive or the

negative strand. The second assumption in the evaluation of the density is that any inter-chromosomal interactions are neglected. This assumption is of course valid only due to the lack of entanglement and insignificant inter-chromosomal interactions, as shown by the fractal globule model [5]. This can also be verified using Markov Chain Monte Carlo (MCMC) sapling simulation results [38] (Fig. 3.1a shows the lack of inter-chromosomal entanglement). Given that bps in proximity tend to show stronger interactions, it has been proposed that the physical distance of two bps is proportional to $1/IF_{ij}^{e}$, where e is a constant [38]. IF_{ij} is the (i, j)th element of the interaction frequency matrix of the chromosome in which gene X is located as obtained by Hi-C data. Rousseau et al. [38] used leave-one-out cross-validation and MCMC to show that e ~ 1-3 is the best range for e (e = 1 is used in this study for simplicity). The density Γ_j can then be computed as:

$$\Gamma_j = \sum_i U\left(\frac{1}{IF_{ij}^{e}} - R\right) \qquad (1)$$

where U(t) is the characteristic function defined as 1 for t > 0 and 0 otherwise. One might simply discard those elements of the interaction frequency matrix with zero interaction (insignificant interactions are also often set to zero during the quality control step of the Hi-C data preparations). Note that the density $\Gamma = d_1$ therefore corresponds to d_1 * 40,000 base pairs in the DS. The drawback of the above definition of the density is its dependence on the radius of the density sphere R. While larger radii account for more bps around the gene of interest, they do not properly reflect the local variation in the chromatin compactness. We define the cumulative radial distribution function c(R) as Γ/V, to assess the sensitivity of Γ to R (V is the volume of the density sphere).

In this study, publicly available, normalized intra-chromosomal Hi-C data by Dixon et al. [8] for five cell lines (H1 human embryonic stem cell (hESC) and H1-derived cells including neural progenitor cell (NPC), mesenchymal stem cell (MSC), mesendoderm cell (MES), and trophoblastic cells (TRO)) were used to obtain the DNA densities. Figure 3.1b illustrates a significant change in c(R) at the SOX2 promoter site for hESC vs NPC cell

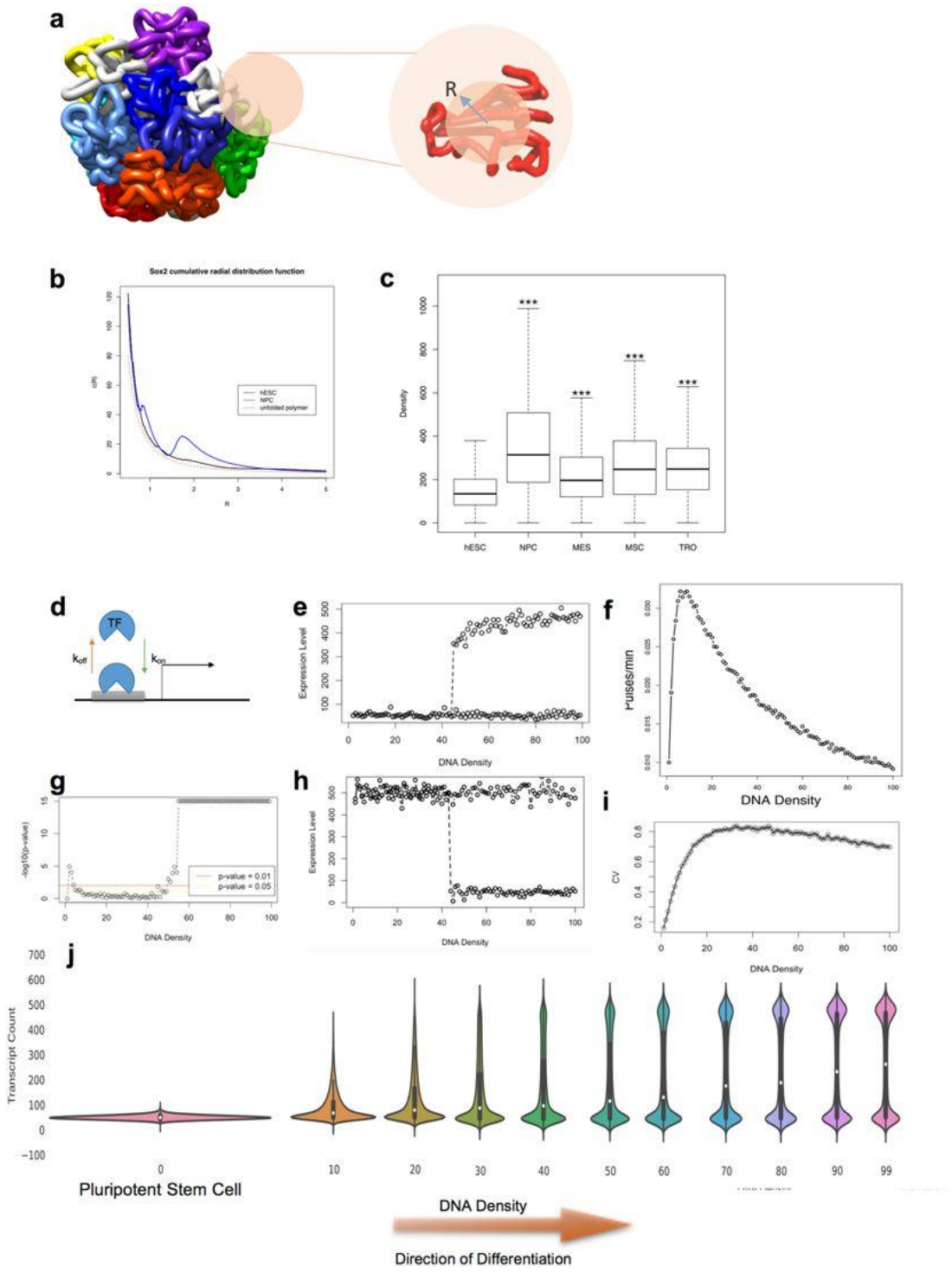


Figure 3.1. Macromolecular crowding model of gene expression. (**a**) MCMC simulations show a lack of inter-chromosomal entanglement. Schematic illustration of the density sphere around an arbitrary gene locus is provided on the right. (**b**) cumulative radial distribution function of DNA bps around the SOX2 gene locus

demonstrates the local density variation between hESC and NPC. (**c**) density increases significantly during hESC differentiation (*** p-value < 0.001). (**d**) binary model of gene expression for gene activating and repressing transcription factors. (**e** and **f**) bifurcation diagrams of gene expression for transcriptionally activated (**e**) and repressed (**h**) genes. (**f**) pulse frequency analysis of transcriptional bursting. (**g**) bimodality inspection during chromatin condensation by Hartigan's Dip test. (**i**) population heterogeneity induced by transcriptional bursting and chromatin condensation. (**j**) steady state gene expression distribution variation during stem cell differentiation aroused from chromatin condensation.

lines. It can be seen from Fig. 3.1b and [16] that chromatin is condensing in the vicinity of the SOX2 gene during differentiation from hESC to NPC, which is particularly interesting because hESC and NPC show similar expression levels of SOX2 [42]. A relatively decondensed chromatin state in hESC can also be inferred by comparing c(R) at the SOX2 gene locus and unfolder polymer (it can be shown that $c(R) \sim 1/R^2$ for an unfolder polymer). We then investigated the density around all genes using R = 2 (recommended by [43]) (Fig. 3.1c), suggesting a significant global increase in the density among all of the four differentiated cell lines studied (p-value << 0.001) (Table S3.1). Further analysis showed that this conclusion is insensitive to R for R ~ 1-3 (Fig. 3.1b shows a similar trend in the density of the SOX2 gene during NPC differentiation regardless of R i.e. the model is robust with respect to the choice of R in this range). Note that given the experimental value of $0.0123 bp/nm^3$ for the density of human nuclei, R=1 can be scaled to a physical distance of ~250nm (similar scale as an average TAD). One might argue that single-cell variation in chromatin interactions can affect this study and therefore, this mode should be trained using single-cell Hi-C data. However, as suggested by [7] individual chromosomes maintain domain organization at the megabase scale (size of a DS) and cell-to-cell variations in chromosome

structure occur at larger scales. We conclude that chromatin experiences a global condensation during differentiation, measured by the density parameter Γ. In addition, we should mention that by analyzing the high resolution promoter capture Hi-C data from the recently published research by Freire-Pritchett et al.[63], we observed that the measured densities are significantly correlated (p-value of correlation test < 2.2e-16) with our previously measured densities from 40kb resolution Hi-C [8], demonstrating that our conclusions are robust to the resolution of Hi-C data.

Higher levels of chromatin looping in differentiated cells is not surprising. For example, it has been shown that regions of condensed heterochromatin form during pluripotent embryonic stem cell differentiation, and silencing histone marks accumulate [16], which result in differential expression in daughter cells. However, we also found density changes in highly expressed genes (i.e. SOX2) and in genes whose mean expression levels remain unchanged, as shown below. These results suggest that, beyond our understanding of the role of the chromatin reorganization limited to gene silencing and activation, chromatin conformation might be involved in other types of gene regulation such as gene expression heterogeneity.

The cellular environment is packed with DNA, RNA, proteins and other macromolecules hindering the diffusion of other molecules such as transcription factors and RNA polymerases. Several studies [32, 44, 45] have demonstrated the impact of macromolecular crowding on the rebinding time of DNA binding proteins (DBP), leading to heterogeneity in gene expression. In particular, *in vitro* experimental studies by Muramatsu et al. [46] show that macromolecular crowding can alter the diffusion and the biochemical rates of globular proteins in concentrated protein solutions. It has been suggested by Morelli et al. [47] that the

effect of crowding can be taken into account by merely scaling the association (k_{on}) and dissociation rate constants (k_{off}) of DBPs. Thus,

$$\frac{k_{on}}{k_{on}^{0}} = \Gamma^{-\gamma} \quad \text{and} \quad \frac{k_{off}}{k_{off}^{0}} = \Gamma^{-\gamma-1}, \qquad (2)$$

where k_{on}^{0} and k_{off}^{0} are the basal association and dissociation constants of DBP and where $\gamma \approx 0.36$ (Table S3.2). Note that k_{on}^{0} and k_{off}^{0} are chosen such that the obtained k_{on} and k_{off} using Eq 2 for a gene with an average local DNA density lies within the range of the reported experimental data for association and dissociation rate constants of DNA binding protein (Table S3.2). Equal k_{on}^{0} and k_{off}^{0} are considered for all genes which can be explained due to the diffusion-limited kinetics of DBPs. In other words, while association and dissociation rates of DBPs to DNA can vary due to factors such as promoter sequence, DNA methylation pattern, etc., the influence of these factors is negligible compared to the effect of the variation the diffusion rates of DBPs which is explained by parameter Γ. The following major assumptions have been made in our macromolecular crowding (MMC) model: 1) *macromolecular crowding is directly correlated to the DNA density*. DNAs are wrapped around histone complexes and are bound by many proteins such as RNA polymerases, transcription factors, mediators, and cofactors. Therefore, we assume that the DNA density is correlated with local molecular crowding. Hence, we assume that high density regions of chromatin are enriched in other proteins, histones and other molecules. The DNA density alone, therefore, can simply represent the concentrations of other bound molecules as well. 2) *RNA polymerases and/or TFs will only diffuse locally during any rebinding events*. In other words, we assume that the diffusive particles are less likely to exit the density sphere with radius R, due to the highly packed chromatin structure which restricts the diffusive

particle to the vicinity of the promoter [43]. 3) *basal association and dissociation biochemical rates do not vary significantly among different genes*. 4) *1D sliding of TF on DNA is ignored.* The sliding length of TF on DNA has been measured to be ~30-900 bps [48]. Given the current resolution of Hi-C data (40 Kbps), precise analysis of the impact of the density alteration on 1D sliding of TF on a DNA strand is not feasible.

We utilized the described simple model to study the influence of the chromatin density on the gene expression patterns by simulating a two state (binary) system (Fig. 3.1d). The Gillespie stochastic simulation algorithm (SSA) [49] was used (See Methods) to simulate the constructed biochemical system (note that here we assume a time-homogeneous Markov process in order to use SSA). We then inspected the transcriptional bursting behavior for different DNA densities by exploring the pulsing frequency, the coefficient of variation (CV) of RNA distributions, and the bimodality of gene expression in a cell population (Fig. 3.1e-j). Our simulation results show a general decrease in pulse frequency, suggesting slower expression kinetics, which could be expected due to the limited diffusion of TF (or RNA polymerase) by crowding. Figure 3.1j depicts steady state RNA distribution variation caused by the increase in DNA density. At very low densities, the MMC model predicts a unimodal normally distributed cell population with a relatively low variation induced by gene intrinsic noise. During stem cell differentiation, chromatin condenses and the DNA density around several promoters increases. Chromatin condensation is followed by transcriptional bursting and heterogeneity in a cell population generating a long tail RNA distribution with high variation. Ultimately, a second cell state emerges due to high RNA variation (measured by CV), resulting in a bimodal RNA distribution. Further increase in the DNA density gradually eliminates the intermediate states [32], indicated by a decrease in the CV (Fig. 3.1i). While Fig.

3.1i initially illustrates an increase in the CV, upon emergence of the second cell state (determined by Hartigan's Dip test for unimodality [50]), the CV tends to decrease. Our simulation results suggest an analogous trait whether the gene is transcriptionally turned on or off (Fig. 3.1e and h). While both undergo a bifurcation, repressive TFs endow a second state with a lower expression level. The bifurcation predicted by our model is consistent with the previous reports on several developmental genes, namely, GATA-1 and PU.1 in a myeloid progenitor cell [51]. All in all, it can be inferred from the MMC model that chromatin reorganization can alter the excluded volume around certain gene loci and thereby can modify the rebinding kinetics of TFs, leading to distinct gene expression patterns. This explains how unoccupied space of the nucleus encodes vital information regarding gene expression heterogeneity. While only one state is accessible at lower densities, increasing the densities can increase the accessibility of other states by rendering wider RNA distributions. Further increase in the densities imposes an all or nothing response which entails a bifurcation toward a second stable state with lower or higher expression levels depending on whether the gene is transcriptionally activated or repressed.

Based on our simulation results, we hypothesize the following:

1. During differentiation the macromolecules accumulate around the promoters of different genes, intensifying the effect of macromolecular crowding.

2. An increase in the chromatin density increases transcriptional bursting due to macromolecular crowding.

3. Transcriptional bursting increases the CV of RNA expression in a cell population. Hence, the CV increases during differentiation and several genes undergo a bifurcation

stochastically, allowing a portion of cells to enter a new cell state while the remainder retain their current cell state.

4. This process creates the potential for pluripotent cells to differentiate.

To test how density changes are related to gene expression variation during stem cell differentiation, we focused on NPC differentiation. We used scRNA-seq data by Chu et al. [33] to assess gene expression variation in H1 hESC and H1-derived NPC because they used the same dual SMAD inhibition method [52] to differentiate hESC to NPC with Dixon et al., which enables direct comparison between DNA density and gene expression variation. First, raw scRNA-seq data of H1 hESC (212 cells) and NPC (173 cells) were counts per million (CPM) normalized. A significant dependency of the CV of RNA expression distributions on their mean expression can be seen, which is due to inherent high technical noise of scRNA-seq experiments (Supplementary Fig. 3.1a and b). To account for any bias, we first discretized the expression levels of all genes within each cell line into 50 logarithmically-spaced intervals, assuming that the mean expression level of those genes that lie inside each interval is similar. For each interval, the biological variation of the genes with low CV is strongly confounded by their technical variation. Therefore, we exclude any genes with CV less than 3 standard deviations (CV < 3stdev), generating a list of genes with high biological variation. Note that since the technical variations of the genes with similar mean expression levels can be assumed to be similar, the obtained list of genes includes only those genes whose biological variations are significant compared to their technical noise. Finally, to account for the technical variation among cell lines, we analyzed only genes whose mean expression is not statistically different (q-value > 0.01) between two cell lines (hESC and NPC).

Figure 3.2a demonstrates that the majority of the genes are more heterogeneous during differentiation, resulting in higher CV in NPC compared to hESC, as predicted by our model. This is not due to non-NPC contamination because Chu et al. sorted out SOX2-positive NPCs for scRNA-seq [33]. Note that all of the genes remaining in the obtained gene

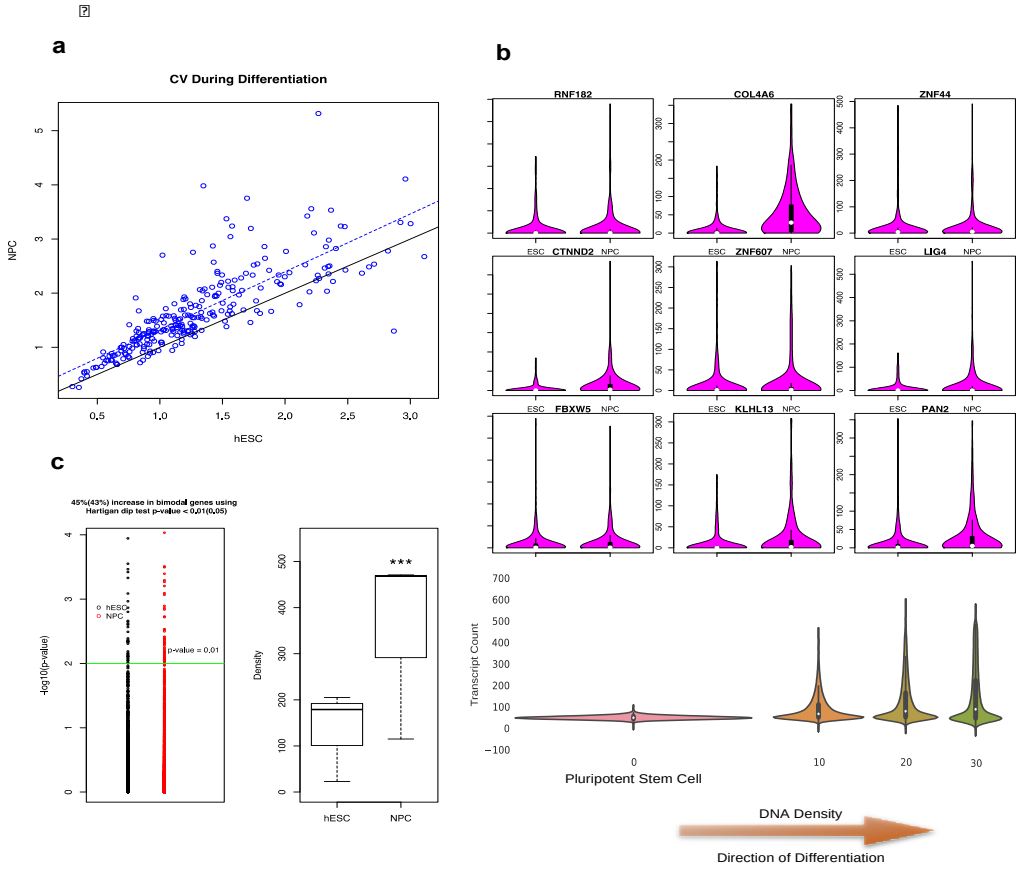


Figure 3.2. Analysis of gene expression heterogeneity during development. (**a**) CV increases during NPC differentiation for the majority of genes. (**b**) long tailed steady state gene expression distributions appear as a result of chromatin condensation, as predicted by the MMC model. (**c**) unimodal to bimodal switch during NPC differentiation due to significant density increase (*** p-value < 0.001).

list have a unimodal RNA distribution, thus the model predicts a steady state long tail unimodal RNA distribution with increasing variation from hESC to NPC. This is illustrated in Fig. 3.2b for 9 sample genes with high gene expression variations.

Another prediction of the MMC model of gene expression is an increased number of genes with a bimodal distribution. We argue that although as previously discussed, the CV values are confounded by the technical noise, RNA expression bimodality is less likely to be affected by that. Thus to study the bimodality we dropped out the CV>3stdev criterion for the gene list selection and compared the number of genes which are bimodal in each cell line (p-value of Hartigan's Dip test for unimodality < 0.01). Interestingly, as predicted by our model, not only do all the bimodal genes in hESC remain bimodal in NPC, but also a 45% increase in the number of bimodal genes was observed. We then analyzed the density values of the genes that undergo a bimodal distribution switch during NPC differentiation. Again, as predicted by our MMC model, there is a significant increase in the density of the genes of interest (Fig. 3.2c).

We introduce density-CV coordinates to facilitate an efficient assessment of the model predictions. A vital prediction of the MMC model is the direction of differentiation in this coordinate plane (Fig. 3.3a). Each arrow represents one gene pointing from hESC to NPC (toward differentiation). The validity of the model can be tested by incorporating Hi-C data (to obtain the densities) and scRNA-seq data (to obtain the CVs). As predicted by our model, 83% of the genes experience an increase in both the density and CV during NPC differentiation. 13% of the genes show an increase in their density while their CVs are slightly reduced (Fig. 3.3b). This 13% discrepancy can be due to the assumptions in the

MMC model and ignoring post transcriptional modifications, cell cycle dependent genes, etc. Another explanation for the decrease in the CV of 13% of the genes is the emergence of

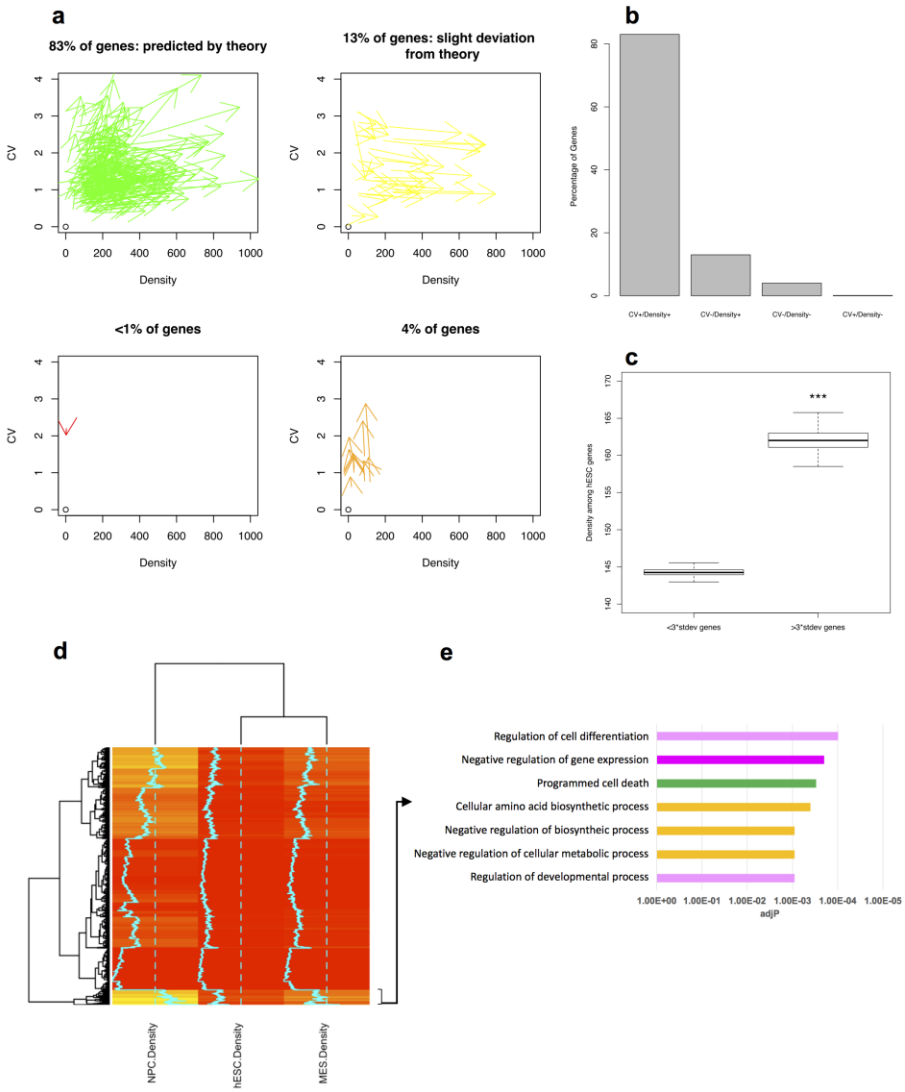


Figure 3.3. Direction of differentiation. (**a**) representation of differentiation arrows in CV-Density coordinates. (**b**) the majority of genes experience increases in both density and CV during NPC differentiation. (**c**) genes with the highest density exhibit the highest CV in hESC (quartiles are estimated using bootstrapping,

*** p-value < 0.001). (**d**) hierarchical clustering shows a functional role for chromatin condensation. (**e**) gene ontology explains the role of chromatin condensation and the gene expression heterogeneity in development, metabolism, and cell death.

the bimodal genes during chromatin condensation which corresponds to the second regime in Fig. 3.1i (a negative correlation between CV and the density is expected). However, since the analysis of scRNA-seq data shows that only a small fraction of genes illustrates a bimodal behavior, we will focus on the 83% of the genes where the CVs and densities are positively correlated. Also it is critical to note that our model doesn't exclude several other factors that can also affect the CVs but instead emphasizes the critical role of the local DNA density on gene expression variation. This observation implies that chromatin condensation is a driving force to render a heterogeneous cell population and thus exciting other silenced cell states. One might argue that this general principle could be employed to study the genes within even one cell type. We tested this hypothesis in hESC by comparing the densities of the genes whose CVs are greater than 3 stdev of all CVs within the corresponding interval (as defined previously) with the rest of the genes. This approach enables us to minimize the impact of the technical noise and gene networks on our conclusions. As predicted by the MMC model (Fig. 3.3c) the set of all genes with high CV in each interval manifest significantly higher densities compared to low CV genes (p-value < 0.001). These results are also confirmed in NPC (Supplementary Fig. 3.1c). In addition, we analyzed the recently published Assay for Transposase-Accessible Chromatin with high throughput sequencing (ATAC-seq) data by Liu et al.[64] to compare the chromatin accessibility of low versus high CV genes in hESCs and did not see a major difference. These results suggest that chromatin accessibility is not related to gene expression variation. We can conclude that one function

of chromatin remodeling proteins is to locally condense chromatin in the vicinity of the genes destined for heterogeneous expression.

Moreover, we perform hierarchical clustering with genes whose DNA densities change during stem cell differentiation. DE genes were excluded across hESC, NPC, and MES using publically available population RNA-seq data by Jang et al. [42]. The first gene cluster (444 genes) (Fig. 3.3d) depicts the highest modification in the densities during stem cell differentiation. Using gene ontology analysis, we found that this gene cluster is significantly associated with cell differentiation, metabolism and cell death (Fig. 3.3e). Namely, several key transcription factors (PRDM16, HES3, HES5, ID3) and epigenetic modifiers (KDM1A) are related to cell differentiation. The mammalian target of rapamycin (mTOR), a master regulator of cellular metabolism is in the term "Negative regulation of cellular metabolic process" together with other important metabolic genes (ENO1, CDA, PEX14). Finally, programmed cell-death related genes include PINK1, CAS9, and MUL1. These results suggest that density changes during stem cell differentiation potentially affect expression variation of genes that are involved in key biological processes.

OCT4, SOX2, and NANOG are core pluripotency genes that play key roles in early stem cell differentiation [53, 54]. OCT4 and NANOG strongly repress NPC differentiation, while promoting differentiation into MES cells. In stark contrast, SOX2 is necessary for a NPC fate, but suppresses MES differentiation. This is consistent with the expression of these genes during stem cell differentiation, with their roles with high OCT4 and NANOG expression in MES and high SOX2 expression in NPC [42]. Given the importance of these genes for stem cell differentiation, we studied how density changes affect single cell expression patterns during NPC or MES differentiation. We limited our analysis to OCT4

and NANOG in MES differentiation and SOX2 in NPC differentiation, to exclude the potential impact of dramatic changes in mean expression levels on gene expression variation. There were no significant changes in the densities of OCT4 and NANOG during MES differentiation (Fig. 3.4a). Consistent with our model, when single cell expression patterns were analyzed by immunostaining, there were no substantial changes in CV and

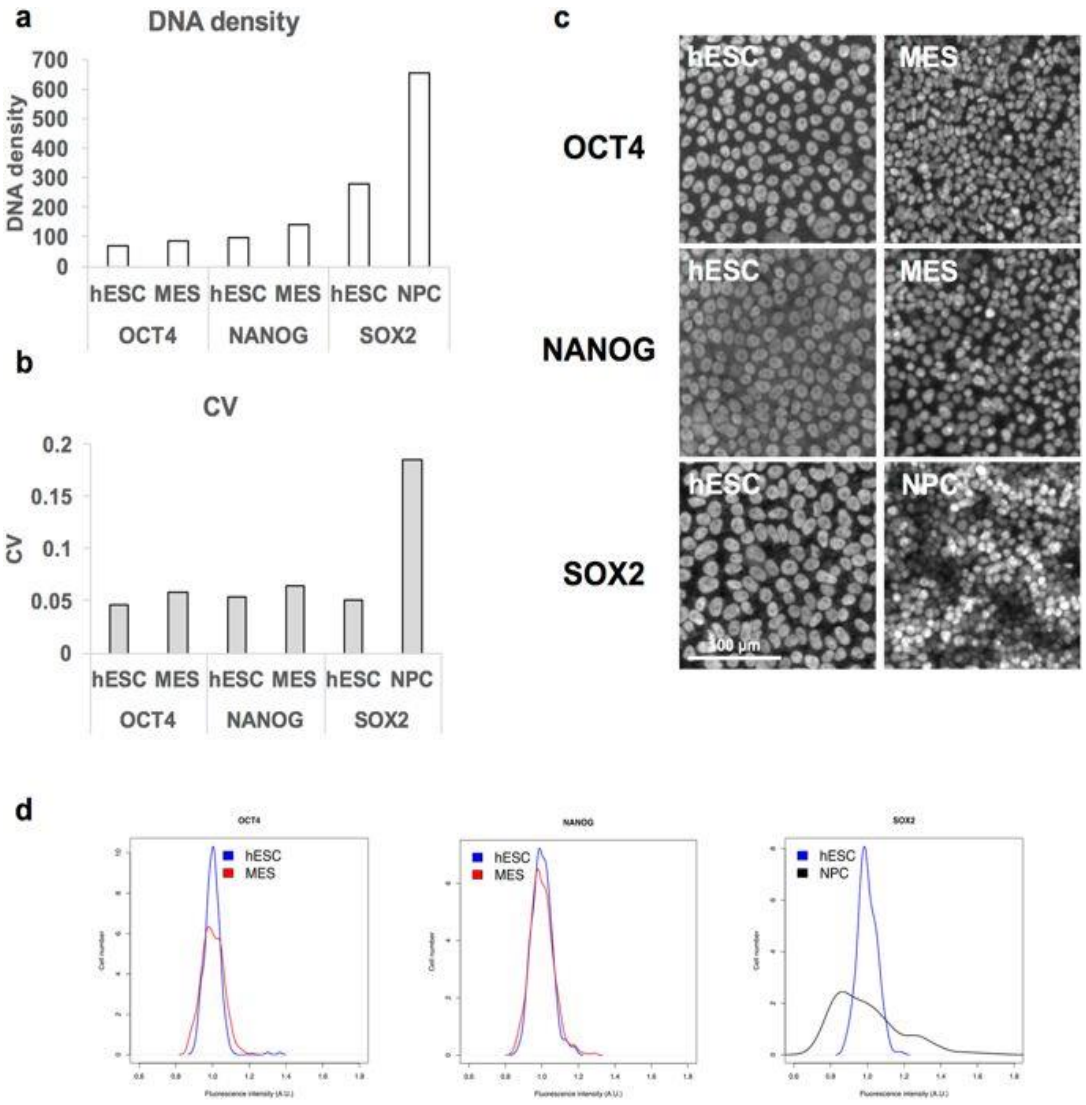


Figure 3.4. Relationship between density changes and gene expression variation in core pluripotency genes. (**a**) The densities of OCT4, NANOG, and SOX2 change during stem cell differentiation. (**b**-**d**) hESC was differentiated to NPC and MES, followed by immunostaining of OCT4, NANOG, and SOX2. CV (**b**) and gene expression distribution (**d**) were calculated from fluorescence intensity measured from individual cells (n=315 for OCT4 in hESC, n=351 for OCT4 in MES, n=294 for NANOG in hESC, n=365 for NANOG in MES, n=308 for SOX2 in hESC, and n=343 for SOX2 in NPC). Representative images are shown in (**c**).

gene expression distributions (Fig. 3.4b-d, Supplementary Fig. 3.2). In contrast, the density of SOX2 underwent a dramatic increase in NPC compared to hESC (Fig. 3.4a). This change was consistent with increased CV and broadened distribution of gene expression during NPC differentiation (Fig. 3.4b-d). NPC derivation is also confirmed by PAX6 expression, suggesting that heterogeneous expression of SOX2 is not due to inefficient NPC differentiation (Supplementary Fig. 3.2b). These results suggest that differentiation-induced local chromatin condensation controls heterogeneous expression of core pluripotency genes.

Discussion:

We have provided substantial evidence that population heterogeneity can be modulated by altering the local chromatin density in single cells. The current model of epigenetics suggests that gene activation and repression occur via modifications of histone H3 methylation and acetylation. Our study elucidates the functional role of higher order chromatin looping, which is regulated by CCCTC-binding factor (CTCF). Our MMC model suggests that aside from the orthodox view of loop formation in down regulation (or up regulation) of the developmental genes such as OCT4 and NANOG in NPC differentiation [55], chromatin looping can modulate expression heterogeneity of other developmental genes (e.g. SOX2 in NPC differentiation) by adjusting the genome-wide long range interactions as well. One particular function of chromatin looping is its impact on the neighboring genes. For example, the expression variation of certain genes can be influenced by inactivation of a neighboring gene, resulting in a new, more crowded environment. It is important to note that our results show that the density not necessarily alters gene expression heterogeneity by

changing the mean expression level, but that, increased densities can increase the gene expression heterogeneity even at similar expression levels.

It is also noteworthy that the current approach to study the chromatin organization by probing TADs is incapable of revealing the vast amount of information hidden in the unoccupied space of the chromatin. As illustrated by our study, chromatin reorganization can alter rebinding kinetics of TFs, through which the gene expression pattern can be modified. Recent studies [56-58] suggest that TADs are highly conserved, not only among different cells of the same clonal population, but also across different cell types. Our study, on the other hand, expresses the local DNA density deviations among distinct cell types, specifically across progenitor and embryonic stem cells, suggesting that the local DNA density changes might play more dynamic roles than TAD reorganization during development.

Methods:

hESC culture and differentiation

H9 hESCs (WiCell) were maintained in mTeSR1 medium (Stem Cell Technologies) on Matrigel (BD Biosciences)-coated tissue culture plates. H9 cells were passaged every 5 days by ReLeSR (Stem Cell Technologies). For NPC differentiation, hESCs were passaged as single cells using Accutase (Life Technologies) on Matrigel with ROCK inhibitor (Y-27632, Millipore) and induced to differentiate with SB431542 (10 µM, Tocris Bioscience) and NOGGIN (200 ng/ml, PeproTech) [52]. To derive MES, hESCs were plated on Matrigel-coated plates as single cells and induced to differentiate with mTeSR1 containing 5 ng/ml hBMP4 (R&D) [59].

Immunofluorescence

Samples were fixed with 4% paraformaldehyde for 15 min, followed by permeabilization with 0.25% Triton X-100. After blocking with 10%FBS in PBS, cells were immunostained with primary antibodies for OCT4 (Santa Cruz), NANOG (R&D), and SOX2 (Millipore) overnight at 4ºC. Staining with secondary antibodies, Alexa Fluor 555-donkey anti-mouse IgG, Alexa Fluor 555-donkey anti-rabbit IgG, Alexa Fluor 488-donkey anti-goat IgG, was done for 1h at room temperature. Images were taken using Olympus IX71 fluorescence microscope.

RNA isolation and quantitative PCR

Total RNA was extracted using TRIzol (Life Technologies), followed by cDNA conversion with the SuperScript III First-Strand Synthesis System (Life Technologies). Quantitative real-time PCR was performed using a QuantStudio 12K Real-Time PCR System (Life Technologies) with Power SYBR Green PCR Master Mix (Applied Biosystems). GAPDH was used as a normalization control.

Data Acquisition and Analysis

Publically available scRNA-seq data for H1 hESC and NPC cell lines were obtained from GEO using accession number GSE75748. In this study expected counts of 212 hESC cells and 173 cells NPC have been used to investigate gene expression variation [33].

Processed FPKM-normalized RNA-seq data from H9 hESC, NPC, and MES (three replicates per cell line) [42] were acquired from GEO using accession number GSE69982.

Hi-C intra-chromosomal interaction frequencies for 5 cell lines (H1 hESC, NPC, MES, MSC, and TRO) have been obtained used using accession number GSE52457, mapped to human genome (hg19) and normalized as described by Dixon et al. [8].

All of the computations were performed using IPython 5.0, R version 3.2.2 and the BioConductor (3.3), qvalue (3.2.3), limma (3.2.4), diptest (3.2.0), GenomicRanges (3.2.3), edgeR (3.2.0), and genefilter (3.2.3). All of the figures were made using basic R graphics and packages gplots (3.2.4), and vioplot (3.2.0). All scripts used here are available from the authors upon request.

Gene ontology analysis has been performed using the web-based gene set analysis toolkit WebGestalt [60].

Details of simulations and computational modeling.

In our study, gene expression is described using the following set of biochemical reactions:

$$\boldsymbol{TF + Promoter_{Repressed} \underset{}{\overset{k_{on},\, k_{off}}{\rightleftharpoons}} Promoter_{Active} \overset{s_A}{\rightarrow} M + Promoter_{Active}} \quad (\mathbf{R1})$$

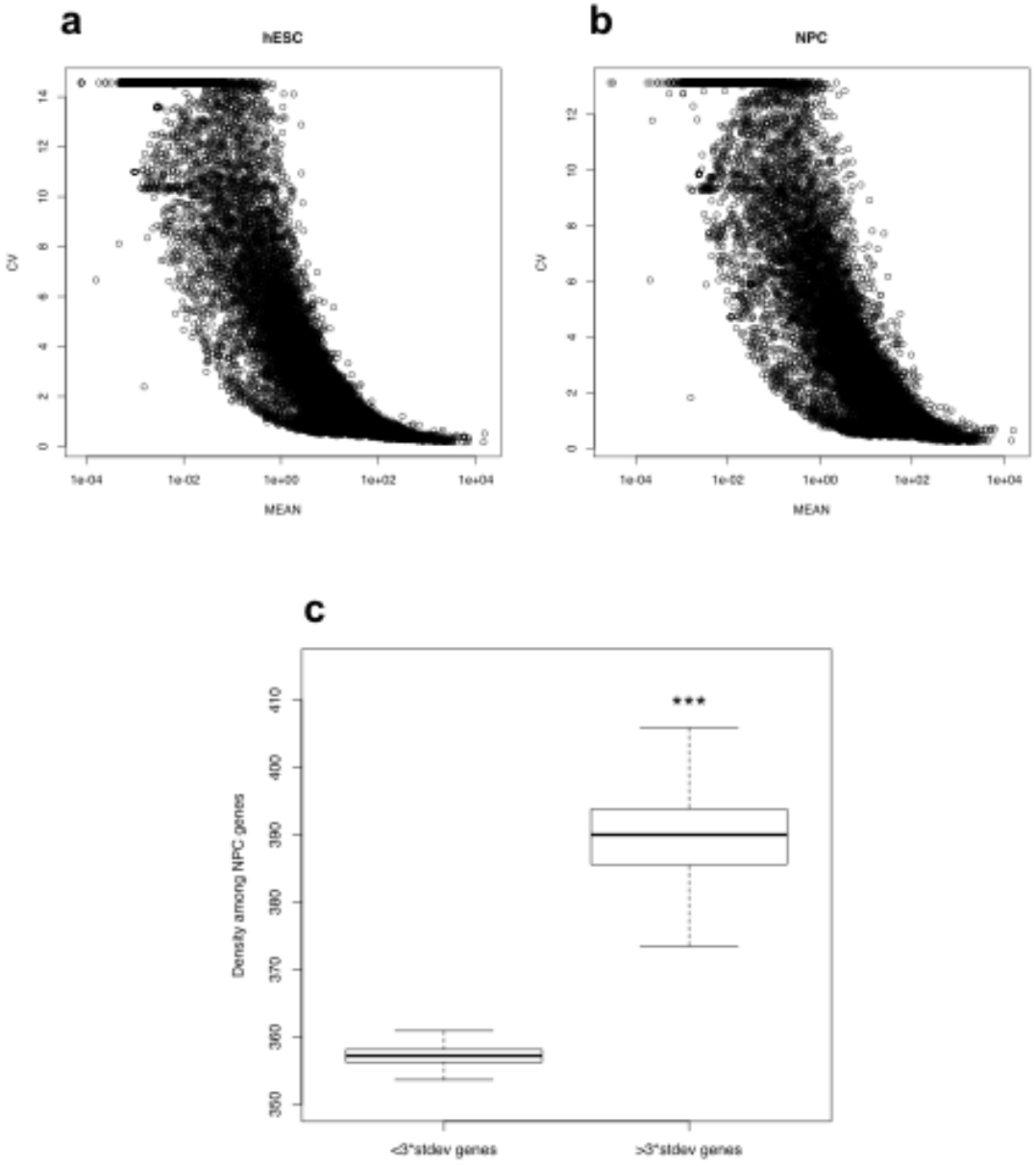
$$\boldsymbol{Promoter_{Repressed} \overset{s_R}{\rightarrow} M + Promoter_{Repressed}} \quad (\mathbf{R2})$$

$$\boldsymbol{M \overset{\delta_M}{\rightarrow} \emptyset} \quad (\mathbf{R3})$$

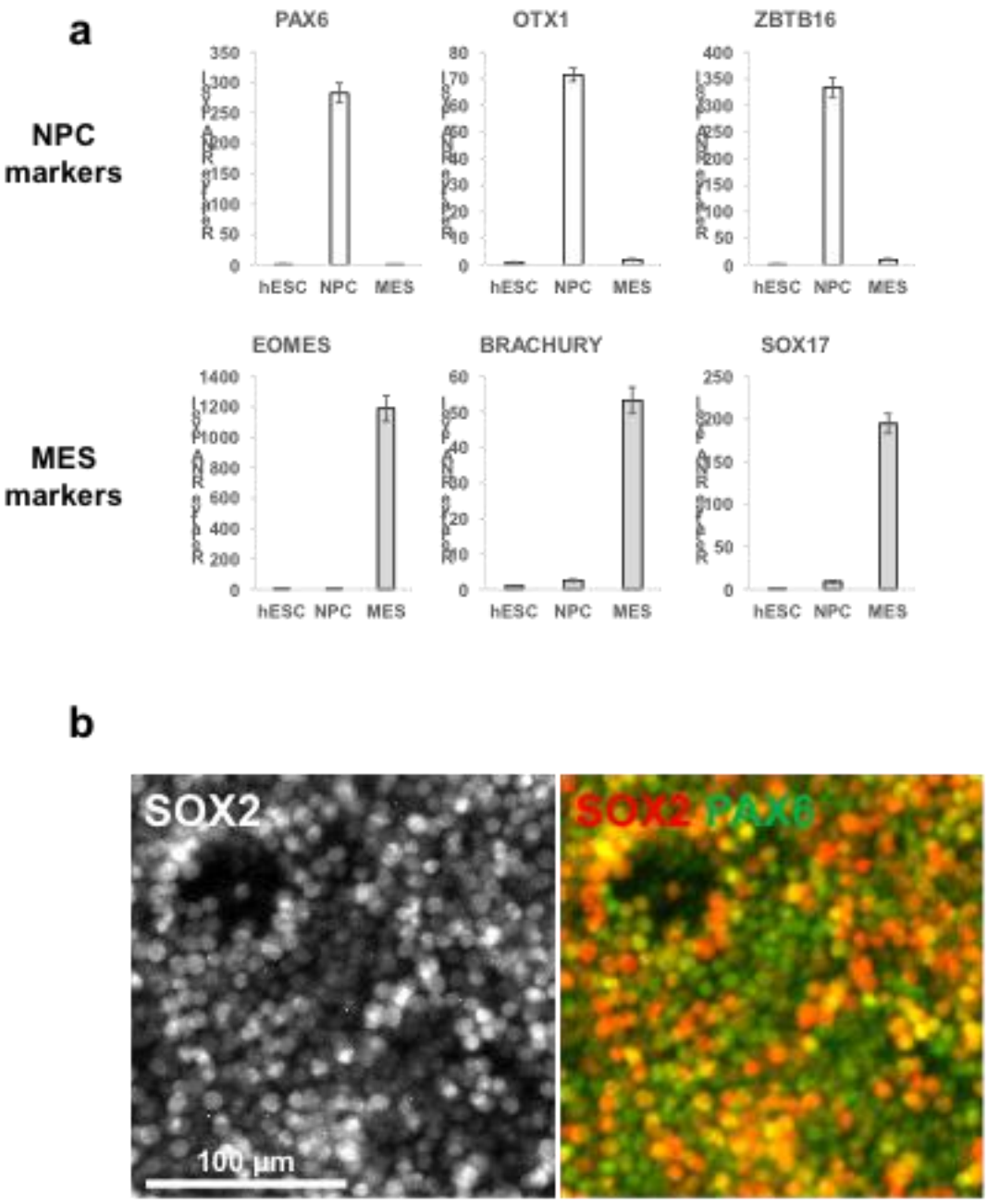
where k_{on} and k_{off} can be found as

$$\frac{k_{on}}{k_{on}^{0}} = \Gamma^{-\gamma} \quad \textbf{and} \quad \frac{k_{off}}{k_{off}^{0}} = \Gamma^{-\gamma-1}$$

The simulation parameters and the initial conditions are provided in the Supplementary Data Table 3.2. The stochastic simulation algorithm (SSA) was used to simulate the above biochemical system using the implementation in GillesPy [61] on the MOLNS software platform [62]. For each Γ ($0 < \Gamma < 100$), 1,000 trajectories are simulated each for 1,000 minutes.



Supplementary Figure 3.1. Analysis of scRNA-seq variations due to mean expression and DNA density variations for hESC and NPC. (a, b) scRNA-seq data show that CV strongly depends on mean expression due to technical noise in both hESC and NPC. (c) genes with the highest density exhibit the highest CV in NPC obtained (quartiles are estimated using bootstrapping, *** p-value < 0.001).



Supplementary Figure 3.2. Validation of lineage-specific differentiation. (a) the expression of NPC (PAX6, OTX1, ZBTB16) and MES (EOMES, BRACHURY, SOX17) markers confirms highly specific differentiation of hESC to each lineage (n=4). Error bars represent SD. (b) NPC was immunostained for SOX2 and PAX6.

References:

1. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *science* **295**, 1306-1311 (2002).
2. Simonis, M. et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nature genetics* **38**, 1348-1354 (2006).
3. Zhao, Z. et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra-and interchromosomal interactions. *Nature genetics* **38**, 1341-1347 (2006).
4. Dostie, J. et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research* **16**, 1299-1309 (2006).
5. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science* **326**, 289-293 (2009).
6. Fullwood, M.J. et al. An oestrogen-receptor-&agr;-bound human chromatin interactome. *Nature* **462**, 58-64 (2009).
7. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64 (2013).
8. Dixon, J.R. et al. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336 (2015).
9. Rao, S.S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).

10. Feng, S. et al. Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in Arabidopsis. *Molecular cell* **55**, 694-707 (2014).

11. Grob, S., Schmid, M.W. & Grossniklaus, U. Hi-C analysis in Arabidopsis identifies the KNOT, a structure with similarities to the flamenco locus of Drosophila. *Molecular cell* **55**, 678-693 (2014).

12. Nakayama, J.-i., Rice, J.C., Strahl, B.D., Allis, C.D. & Grewal, S.I. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* **292**, 110-113 (2001).

13. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics* **33**, 245-254 (2003).

14. Berger, S.L. The complex language of chromatin regulation during transcription. *Nature* **447**, 407-412 (2007).

15. Portela, A. & Esteller, M. Epigenetic modifications and human disease. *Nature biotechnology* **28**, 1057-1068 (2010).

16. Meshorer, E. & Misteli, T. Chromatin in pluripotent embryonic stem cells and differentiation. *Nature reviews Molecular cell biology* **7**, 540-546 (2006).

17. Ho, L. & Crabtree, G.R. Chromatin remodelling during development. *Nature* **463**, 474-484 (2010).

18. Li, E. Chromatin modification and epigenetic reprogramming in mammalian development. *Nature Reviews Genetics* **3**, 662-673 (2002).

19. Elowitz, M.B., Levine, A.J., Siggia, E.D. & Swain, P.S. Stochastic gene expression in a single cell. *Science* **297**, 1183-1186 (2002).

20. Dunlop, M.J., Cox, R.S., Levine, J.H., Murray, R.M. & Elowitz, M.B. Regulatory activity revealed by dynamic correlations in gene expression noise. *Nature genetics* **40**, 1493-1498 (2008).

21. Eldar, A. & Elowitz, M.B. Functional roles for noise in genetic circuits. *Nature* **467**, 167-173 (2010).

22. Blake, W.J., Kærn, M., Cantor, C.R. & Collins, J.J. Noise in eukaryotic gene expression. *Nature* **422**, 633-637 (2003).

23. Suter, D.M. et al. Mammalian genes are transcribed with widely different bursting kinetics. *Science* **332**, 472-474 (2011).

24. Blake, W.J. et al. Phenotypic consequences of promoter-mediated transcriptional noise. *Molecular cell* **24**, 853-865 (2006).

25. Zenklusen, D., Larson, D.R. & Singer, R.H. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature structural & molecular biology* **15**, 1263-1271 (2008).

26. Li, B., Carey, M. & Workman, J.L. The role of chromatin during transcription. *Cell* **128**, 707-719 (2007).

27. Maamar, H., Raj, A. & Dubnau, D. Noise in gene expression determines cell fate in Bacillus subtilis. *Science* **317**, 526-529 (2007).

28. Raser, J.M. & O'Shea, E.K. Noise in gene expression: origins, consequences, and control. *Science* **309**, 2010-2013 (2005).

29. Süel, G.M., Kulkarni, R.P., Dworkin, J., Garcia-Ojalvo, J. & Elowitz, M.B. Tunability and noise dependence in differentiation dynamics. *Science* **315**, 1716-1719 (2007).

30. Chang, H.H., Hemberg, M., Barahona, M., Ingber, D.E. & Huang, S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature* **453**, 544-547 (2008).

31. Graf, T. & Stadtfeld, M. Heterogeneity of embryonic and adult stem cells. *Cell stem cell* **3**, 480-483 (2008).

32. Golkaram, M., Hellander, S., Drawert, B., & Petzold, L.R. Macromolecular Crowding Regulates the Gene Expression Profile by Limiting Diffusion. *PLoS computational biology*, **12**(11), e1005122 (2016).

33. Chu, L.-F. et al. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biology* **17**, 173 (2016).

34. Dekker, J., Marti-Renom, M.A. & Mirny, L.A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics* **14**, 390-403 (2013).

35. Lesne, A., Riposo, J., Roger, P., Cournac, A. & Mozziconacci, J. 3D genome reconstruction from chromosomal contacts. *Nature methods* **11**, 1141-1143 (2014).

36. Trieu, T. & Cheng, J. Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. *Nucleic acids research* **42**, e52-e52 (2014).

37. Trieu, T. & Cheng, J. MOGEN: a tool for reconstructing 3D models of genomes from chromosomal conformation capturing data. *Bioinformatics* **32**, 1286-1292 (2016).

38. Rousseau, M., Fraser, J., Ferraiuolo, M.A., Dostie, J. & Blanchette, M. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC bioinformatics* **12**, 414 (2011).

39. Varoquaux, N., Ay, F., Noble, W.S. & Vert, J.-P. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics* **30**, i26-i33 (2014).

40. Fiszbein, A. & Kornblihtt, A.R. Histone methylation, alternative splicing and neuronal differentiation. *Neurogenesis* **3**, e1204844 (2016).

41. Du, J., Johnson, L.M., Jacobsen, S.E. & Patel, D.J. DNA methylation pathways and their crosstalk with histone methylation. *Nature Reviews Molecular Cell Biology* **16**, 519-532 (2015).

42. Jang, J. et al. Primary Cilium-Autophagy-Nrf2 (PAN) Axis Activation Commits Human Embryonic Stem Cells to a Neuroectoderm Fate. *Cell* **165**, 410-420 (2016).

43. Bancaud, A. et al. Molecular crowding affects diffusion and binding of nuclear proteins in heterochromatin and reveals the fractal organization of chromatin. *The EMBO journal* **28**, 3785-3798 (2009).

44. Hansen, M.M. et al. Macromolecular crowding creates heterogeneous environments of gene expression in picolitre droplets. *Nature nanotechnology* **11**, 191-197 (2016).

45. Tan, C., Saurabh, S., Bruchez, M.P., Schwartz, R. & LeDuc, P. Molecular crowding shapes gene expression in synthetic cellular nanosystems. *Nature nanotechnology* **8**, 602-608 (2013).

46. Muramatsu, N. & Minton, A.P. Tracer diffusion of globular proteins in concentrated protein solutions. *Proceedings of the National Academy of Sciences* **85**, 2984-2988 (1988).

47. Morelli, M.J., Allen, R.J. & Ten Wolde, P.R. Effects of macromolecular crowding on genetic networks. *Biophysical journal* **101**, 2882-2891 (2011).

48. Wunderlich, Z. & Mirny, L.A. Spatial effects on the speed and reliability of protein–DNA search. *Nucleic acids research* **36**, 3570-3578 (2008).

49. Gillespie, D.T. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry* **81**, 2340-2361 (1977).

50. Hartigan, J.A. & Hartigan, P. The dip test of unimodality. *The Annals of Statistics*, 70-84 (1985).

51. Enver, T., Pera, M., Peterson, C. & Andrews, P.W. Stem cell states, fates, and the rules of attraction. *Cell stem cell* **4**, 387-397 (2009).

52. Chambers, S.M. et al. Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nature biotechnology* **27**, 275-280 (2009).

53. Wang, Z., Oron, E., Nelson, B., Razis, S. & Ivanova, N. Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells. *Cell stem cell* **10**, 440-454 (2012).

54. Thomson, M. et al. Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. *Cell* **145**, 875-889 (2011).

55. Li, M., Liu, G.-H. & Belmonte, J.C.I. Navigating the epigenetic landscape of pluripotent stem cells. *Nature Reviews Molecular Cell Biology* **13**, 524-535 (2012).

56. Dixon, J.R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012).

57. Nora, E.P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381-385 (2012).

58. Pope, B.D. et al. Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**, 402-405 (2014).

59. Yu, P., Pan, G., Yu, J. & Thomson, J.A. FGF2 sustains NANOG and switches the outcome of BMP4-induced human embryonic stem cell differentiation. *Cell stem cell* **8**, 326-334 (2011).

60 Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based gene set analysis toolkit (WebGestalt): update 2013. *Nucleic acids research* **41**, W77-W83 (2013).

61. https://github.com/JohnAbel/gillespy

62. Drawert, B., Trogdon, M., Toor, S., Petzold, L.R. & Hellander, A. MOLNs: A Cloud Platform for Interactive, Reproducible, and Scalable Spatial Stochastic Computational Experiments in Systems Biology Using PyURDME. *SIAM Journal on Scientific Computing* **38**, C179-C202 (2016).

63. Freire-Pritchett, P. et al. Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *eLife* **23**;6:e21926 (2017).

64. Liu, Q. et al. Genome-Wide Temporal Profiling of Transcriptome and Open Chromatin of Early Cardiomyocyte Differentiation Derived From hiPSCs and hESCs. *Circulation Research* **121**, 376-391 (2017).

CHAPTER 4

Regulation of Cell-Type-Specific Transcriptomes by miRNA Networks During Human Brain Development

Introduction:

Recent studies utilizing single cell mRNA sequencing (scRNA-seq) to characterize cell-type diversity in tissues have highlighted the need for multi-modal analyses of cellular phenotypes by unbiased classification schemas, particularly in developing systems where complex gene regulatory networks control orthogonal sources of transcriptional variation, including morphology, physiology, maturation, differentiation, and spatial position[1-4]. While mRNA expression levels can be used directly to define putative cell types using unbiased clustering, inferring cell identities and determining cell identity boundaries requires either prior knowledge or additional modalities. MicroRNAs (miRNAs) are an inherently complex network of interactions that can serve as an additional feature of cellular identity with important implications for protein expression. Changes in miRNA expression patterns are characteristic decision nodes during cell differentiation[5], suggesting that their cell type-specific abundance may represent an important parameter in cell type classification and provide insights that extend beyond cell-type classification to the dynamic regulation of differentiation. Previous studies ablating miRNA-processing enzyme Dicer1 emphasized the

pleiotropic roles for this pathway related to tissue specificity, anatomical and cellular compartments, evolutionary relationships, developmental time points, and even specific cell types[6-11], but the underlying framework for these differences are poorly understood. Profiling of miRNA abundance in developing human brain tissue samples suggested developmental regulation of miRNA expression[12], but these studies could neither distinguish cell-type specific patterns of miRNA abundance, nor dynamic cell fate transitions during development at the single cell level. To characterize the *in vivo* miRNA-mRNA interactions during human brain development, and to contextualize these networks in the framework of developmental transitions and cell identity, we leveraged three complementary datasets: high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP)[13] with an AGO2 antibody, simultaneous single cell profiling of mRNAs and miRNAs, and single-cell mRNA sequencing (scRNA-seq) data. Our study revealed a dynamic network involving cell-type specific enrichment of miRNA expression patterns across diverse cell types, and dynamic miRNA target acquisition and loss in which the population of targeted mRNAs keeps pace with the dynamics of tissue development, cell diversity, and lineage progression during human brain development.

Results:

AGO2-HITS-CLIP identifies miRNA-mRNA interactions during prenatal human brain development

To identify the landscape of miRNA-mRNA interactions occurring in developing human brain *in vivo*, we performed high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) [13] with AGO2. AGO2 bound profiles were generated for primary tissue samples of the developing human brain from stages corresponding to peak

neurogenesis (GW15 and 16.5 – early stage) and early gliogenesis (GW19-20.5 – late stage), nine samples in total were harvested from prefrontal cortex (PFC), motor cortex area one (M1), visual cortex area one (V1) and other regions (Supplementary Table 1). AGO2-bound miRNAs and mRNAs were identified after sequencing (Fig. 4.1a, Supplementary Table 4.1-2, see Methods for details). In total, 921 human miRNAs were detected and 10505 Ago2 binding sites were identified from both protein-coding genes and non-coding genes, including lncRNAs (Supplementary Table 4.2). Approximately 43% of sites were in three prime untranslated regions (3'UTR) and 27% of sites were in coding DNA sequence (CDS). For further analysis, we considered only sites identified in the CDS and the 3'UTR, reflecting canonical miRNA-mRNA interactions. We identified 3693 and 2705 genes at early and late stages of development, respectively, actively targeted by miRNAs through CDS or 3'UTR parts of the transcript (Supplementary Fig. 4.1). We validated a subset of the canonical AGO site interactions using luciferase reporter assays in human cells *in vitro* (Supplementary Figure 4.1, Supplementary Table 4.3). Among the detected interactions were previously validated ones, such as miR-9 with FOXG1 and HES1 and miR-210 with CDK7, thereby confirming the strength of the method.

Unbiased enrichment analysis using the total expressed gene set in human developmental brain as the background, revealed that transcription factors, chromatin modifiers, and signaling pathway components were enriched among miRNA targets (Supplementary Table 4.4). Surprisingly, hundreds of *in vivo* miRNA targets we identified were well-established markers of distinct cell types[14-16], regulators of neurogenesis, migration, axonogenesis, synaptogenesis, and neuronal subtype specification. Broadly, single miRNAs target many mRNAs and single mRNAs are targeted by far fewer miRNAs (Supplementary Fig. 4.2),

representing a bipartite network of interactions between miRNAs and their direct target mRNAs (Fig. 4.1b). Using a bipartite community detection

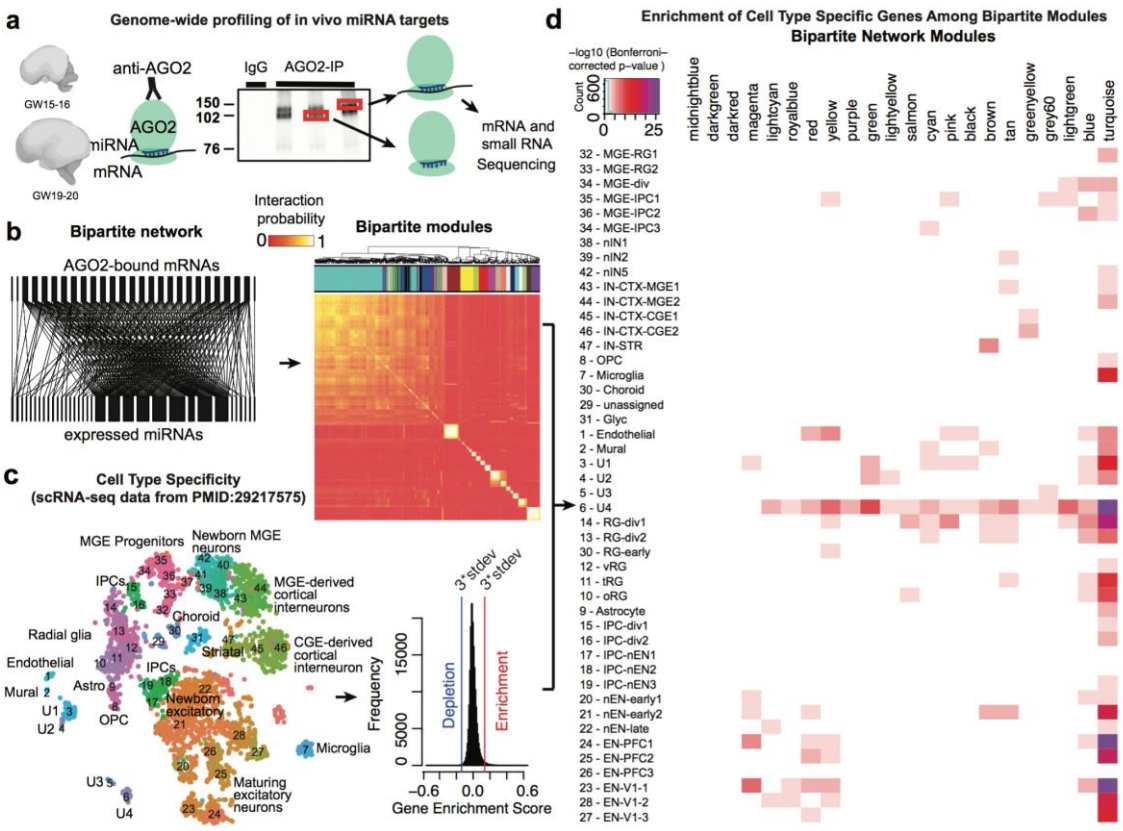


Figure. 4.1: High Throughput Profiling of miRNA-mRNA Interactions. (a) Experimental design. Autoradiogram of ^{32}P-labelled RNA tags crosslinked to AGO2 protein obtained from human prenatal brain homogenates. 110 kDa and 130 kDa bands are visible in samples with AGO2-immunoprecipitation as compared to IgG control. **(b)** The complete bipartite network analysis of miRNA-mRNA interactions shown as a correlation matrix in the lower panel and a segment of the bipartite network shown in the upper panel that illustrates the inhomogeneity of the targeting miRNAs and the relative homogeneity of the targeted mRNAs. **(c-d)** Enrichment of bipartite modules according to cell-type identities. **(c)** Cellular specificity of genes expressed in the developing human brain according to published single-cell mRNA-sequencing dataset, with row names representing cell clusters described in the source study[26]. **(d)** Enrichment of cell-type-specific genes among bipartite network modules. Enrichment scores represent Bonferroni-corrected $-\log_{10}$(p-value).

algorithm[17], we revealed modules of miRNA-mRNA interactions (Fig. 4.1b, Supplementary Fig. 4.2-3, Supplementary Table 4.5, see Methods for details). To contextualize these interactions in the framework of cell diversity we projected bipartite graph modules onto

cell-type specificity information, calculated from published scRNA-seq datasets (Fig. 4.1c-d, Supplementary Table 4.6). Our analysis revealed striking enrichment for cell-type specific transcripts among bipartite graph modules, suggesting that miRNAs acquire targets according to the cognate transcriptional landscape of individual cell-types. Interestingly, the average abundance of a bound miRNA correlated negatively with the total number of different miRNAs bound in each module (Supplementary Fig. 4.4), suggesting that miRNA targeting utilizes two strategies: (a) target with one or a very few abundant miRNAs; (b) target with multiple low abundance miRNAs.

Cell-type enrichment of miRNAs revealed by single-cell miRNA profiling

To investigate further how miRNA-mRNA interactions relate to the emerging diversity of cell types in the primary developing tissues, we developed an innovative protocol for combined detection miRNAs and mRNAs in the same single cells using an automated microfluidic platform to perform automated cell capture, reverse transcription and targeted preamplification of mRNA and miRNA (Fig. 4.2a-c, Supplementary Table 4.7-8). In addition to long-established markers of distinct cell types in the developing cortex, we selected mRNA targets according to the specificity of their expression for distinct cell types (Fig. 4.2b) as determined in Pollen AA et al., [14]. We profiled single cells isolated from primary human cortical tissue samples at gestational week (GW) 14 (deep layer neurogenesis) and GW17 (upper layer neurogenesis). To enrich for progenitor cells and newborn neurons, we microdissected samples from the cortical germinal zone (GZ), and to capture maturing neuron populations and interneurons, we microdissected cortical plate regions (CP).

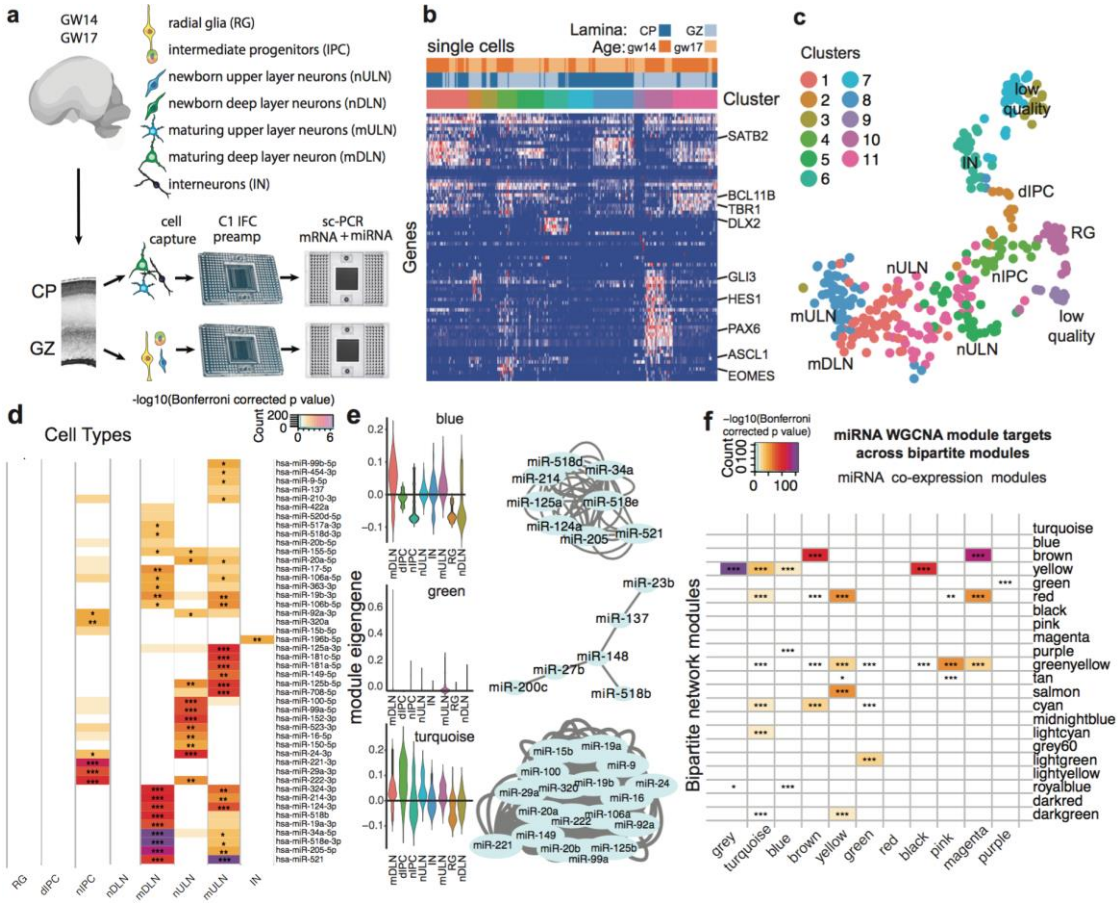


Figure. 4.2. Single-cell miRNA Expression Profiling Reveals Patterns of Cell-type Enrichment. (a-c) sc-qPCR profiling of mRNA and miRNA abundance in the same cell. **(a)** Schematic outlining experimental approach and cell types expected to be enriched in microdissected brain regions. **(b)** tSNE plot of single-cell data generated calculated using Seurat[40] based on mRNA marker gene abundance. Colors represent unbiased clustering (see Methods for details). **(c)** Heatmap of mRNA target genes used to interpret cell identities. **(d)** Heatmap representing cell-type enriched miRNA expressions profiled (one-tailed U-test) **(e)** Weighted gene co-expression network analysis reveals modules of co-expressed miRNAs across single-cells profiled in **a-b**. Network plot shows miRNAs assigned to this network based on correlation of abundance across single cells. Module blue and turquoise are enriched in mDLN and IPC, respectively. **(f)** Enrichment of targets of co-expressed miRNAs across bipartite network modules. * $p<0.05$, ** $p<0.01$, *** $p<0.001$ (Bonferroni corrected for multiple comparison). RG: radial glia, nIPC: IPC-neuron, dIPC: dividing intermediate progenitor cells, nDLN: newborn deep layer neurons, mDLN: maturing deep layer neurons, nULN: newborn upper layer neurons, mULN: maturing upper layer neurons, IN: interneurons.

In total, we retained data from 312 cells with more than 10 genes detected. Clustering analysis performed based on marker gene abundance revealed 11 clusters (Supplementary Table 4.7). We inferred the identities of individual cell clusters as radial glia, intermediate progenitors, upper and deep cortical layer neurons, and interneurons (Fig. 4.2b-c). Furthermore, spatial microdissections supported further refinement of our interpretations with respect to neuronal maturation state (newborn neurons captured from the GZ and maturing neurons captured from the CP) (Fig. 4.2c). Next, for every miRNA profiled, we quantified the abundance in every cell and calculated an expression enrichment score for every cell type (Fig. 4.2d). Surprisingly, the vast majority of miRNAs we profiled showed significant enrichment in at least one cell type, suggesting robust variation in miRNA abundance across closely related cells of the developing brain. For example, miR-221/222 and miR-92a were found enriched in cortical IPCs, in line with recent reports[18] and consistent with their proposed roles in controlling proliferation[19], while miR-124 was enriched in postmitotic neurons, consistent with its proneural role in development[20]. Furthermore, we grouped miRNAs according to the shared pattern of abundance across single cells using weighted gene co-expression network analysis (Supplementary Table 4.8). Some of these specific miRNA coexpression modules operate within specific cell types, whereas others are broadly distributed across multiple cell-types (Fig. 4.2e). Our analysis revealed dynamic changes in miRNA abundance in concordance with neuronal differentiation and maturation, a critical axis of transcriptional variation in the developing brain.

Dual nature of miRNA-mRNA interactions

During mouse brain development, miRNAs are involved in regulating cell-type transitions[7], but analogous regulatory mechanisms during human brain development have

largely not been investigated. To address this limitation, we projected targets of miRNAs found to be co-expressed using sc-qPCR (Fig. 4.2e) onto bipartite co-regulatory modules inferred from HITS-CLIP (Fig. 4.1b). This analysis revealed a striking enrichment of targets of co-expressed miRNAs among bipartite network modules (Fig. 4.2f). Interestingly, we found examples of interactions where miRNAs were enriched in neurons (WGCNA module 'blue', Fig. 4.2e), and their targets fell into a regulatory module enriched for neuronal markers (bipartite module 'yellow', Fig. 4.1d, 4.2f). This interaction includes *DNM1*, a GTPase involved in synaptic vesicle recycling[21], and *NOVA1*, a neuron specific RNA binding protein[22]. Presence of miRNAs and target mRNAs in the same cell type was also observed when we correlated the abundance of miRNAs and their targets across single cells (Supplementary Fig. 4.4b). Similarly, enrichment of miR-92 and its direct target EOMES in the same cell type as reported in the developing cerebral cortex [18,23], suggests a subset of interactions in which control over the target counters the expected effects of miRNA, which usually leads to transcript degradation.

In addition, we also found interactions, in which a miRNA co-expression module was enriched in neurons (WGCNA module 'green', Fig. 4.2e), but their targets show moderate enrichment for proliferating progenitors (bipartite module 'lightgreen', Fig. 4.1d, 2f). This interaction includes genes such as *H2AFZ*, involved in cell cycle regulation[24], and PHGDH, a radial glia specific gene involved in L-serine biosynthesis[25]. Another example, miRNA co-expression module 'turquoise' is enriched in intermediate progenitor cells (Fig. 4.2e), whereas their targets are enriched in bipartite modules that include neuronal and radial glia genes, and genes expressed in both (Fig. 4.1d, 2f). For example, CC2D1B is expressed highly in radial glia and neurons[26], and has been previously implicated in serotonergic

signaling in neurons[27], but also regulates EGFR expression and could regulate cell proliferation[28]. In these cases, miRNAs appear to be suppressing cell identity.

Together, these examples highlight dynamic rewiring of miRNA-mRNA interactions during neuronal differentiation and maturation in the developing human cerebral cortex. In particular, co-modularity miRNA-mRNA interactions as contextualized in the framework of gene and miRNA coexpression seems to follow at least two broad patterns: miRNAs are recruited in a cell type to repress genes not normally expressed in that cell type in some cases or miRNAs are expressed in a cell type to regulate the expression of genes expressed in that same cell type. Further refining these dual roles may emerge from higher resolution temporal data that synchronizes single cell developmental transitions with miRNA target degradation kinetics.

Dynamic changes in miRNA-mRNA network during development

Next, we explored the temporal axis of miRNA-mRNA interactions. We compared the abundance of miRNAs at two stages of development - GW15-16.5 and GW19-20.5 and found 69 differentially expressed miRNAs between these two stages including recently evolved miRNAs (Fig. 4.3a-b, Supplementary Fig. 4.6-11, and Supplementary Table 4.9). Two miRNAs, miR-449a and miR-449b-5p, which control mitotic spindle orientation during mammalian brain development [29,30] showed the highest overall fold change in expression level between GW15-16.5 and GW19-20.5. We confirmed the expression of miR-2115, miR-449c, miR-455 and miR-362 by *in-situ* hybridization (Fig. 4.3b, Supplementary Fig. 4.5-7). In addition, we also identified miRNAs, miR-1286, miR-142 and miR-548aa, as enriched in the occipital lobe compared to the frontal lobe (Supplementary Fig. 4.8-10,

Supplementary Table 4.2) suggesting, in line with recent studies[31], that miRNAs may regulate regionally divergent transcriptional states in the developing human cortex.

By independently performing bipartite network analyses for samples at each of the two stages studied, we found a striking preservation of most co-regulatory modules, as well as a set of distinct interactions present predominantly at one stage (Fig. 4.3c, Supplementary Fig. 4.11a, Supplementary Table 4.10). Interestingly, many of these modules were also highly preserved when compared to adult human brain interactions previously surveyed using the same experimental strategy[31,32] (Supplementary Fig. 4.11b). Together, our findings suggest that miRNA-mediated regulation forms a developmentally dynamic network of interactions related to cell type, developmental stage, and cortical area specificity.

Recent studies suggest that perturbations in miRNA expression may underlie human developmental neuropsychiatric disorders[33,34], but the specific molecular consequences remain poorly understood. Interestingly, genes implicated in autism spectrum disorders (ASD), are enriched in the magenta module (Supplementary Fig. 4.12). In addition, we found that the expression of several miRNAs recently implicated in ASD[33] was biased towards expression in excitatory neurons in the developing mid-gestational human samples (Supplementary Fig. 4.12), although their expression patterns may change over the course of brain development[35]. Genes targeted by miR-137 in developing brain differ greatly from targets identified in adult human brain tissue[32], suggesting that *in vivo* target interactions of these miRNAs may also change substantially during development (Supplementary Table 4.11). To explore this observation more broadly, we considered whether individual miRNAs dynamically change their target landscape during development. We intersected cell type specificity of miRNA targets at either stage of development and found miRNAs whose

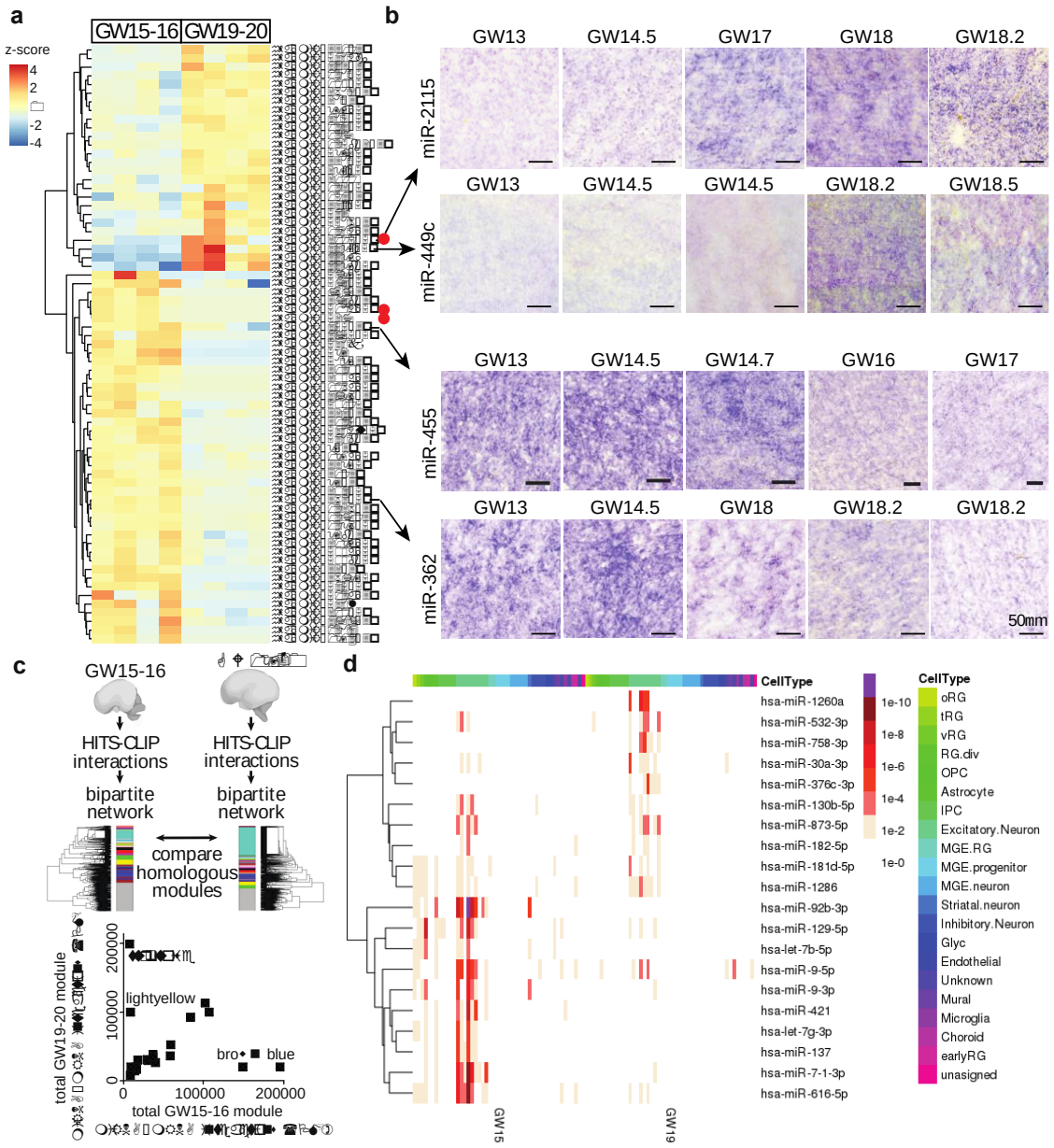


Figure. 4.3. Dynamic changes in miRNA regulatory networks during development. (a) Differential expression analysis identifies miRNAs differentially expressed between GW15-16 and GW19-20 developing human cortex. Red dots indicate primate-specific miRNAs. **(b)** Validation of differentially expressed miRNAs by *in-situ* hybridization in the developing human neocortex sections. Images show staining

in the outer sub-ventricular zone. **(c)** Module preservation analysis demonstrates significant similarity between most modules obtained for networks generated across GW15-16 and across GW19-20 samples. GW15-16 module names are used to compare GW15-16 modules with their homologues in GW19-20. **(d)** Stage-specific changes in miRNA targets according to their specificity to distinct cell types of the developing brain identified using single-cell RNA-seq[26].

targets are enriched in one cell type during early development, and in a different cell type later in development (Fig.4. 3d). We found many miRNAs, such as miR-181d-5p, miR-129-5p, miR-9 and miR-616-5p, whose targets were enriched in different cell types between early developmental stages and late developmental stages thereby indicating changing roles for these miRNAs during brain development. miR-1260a, miR-758-3p and miR-376c-3p targets were enriched in excitatory neurons only in late stages. miR-376c-3p is of interest because it can significantly enhance neural differentiation of *in vitro* pluripotent stem cell models[36]. Targets of miR-92b-3p, let-7-5p, miR-421, and miR-137 were enriched for radial glial markers or excitatory neurons only at early stages. Both miR-92b-3p and miR-130b-5p have been reported to be specifically associated with neural progenitors[12]. These examples further underscore the dynamic remodeling of miRNA interaction networks during development and suggest that further analysis of these interactions may reveal previously unappreciated cellular vulnerabilities of miRNA-mRNA interactions to disease mutations. Understanding cell-type-specific miRNA expression profiles, and their respective targets may highlight cellular patterns of selective vulnerability to disorders affecting miRNA expression by highlighting gene regulatory networks that might be perturbed in disease states.

Recently evolved miR-2115 regulates cell cycle in human radial glia

The developmental transition between GW15-16.5 and GW19-20.5 coincides with changes in proliferation rates of radial glia, and depletion of proliferative capacity in the human ventricular zone[37]. Among the top five miRNAs differentially expressed between these stages, a great ape specific miRNA, miR-2115, was prominently upregulated at GW19-20 in the germinal zones (Fig. 4.3a-b, Supplementary Fig. 4.5, 7). Among miR-2115 targets, ORC4, a known regulator of DNA replication[38], was enriched in radial glia at early stages of development [14,16], and is a member of the turquoise module which is enriched for GW19-20.5 HITS-CLIP interactions within a segment corresponding to a putative miR-2115 response element (Fig. 4.3c, Supplementary Table 4.11). Mutations in ORC4 are linked to Meier-Gorlin syndrome, which is frequently associated with abnormal head circumference, suggesting that this gene may play an important role in normal brain development_ENREF_17[39]. We hypothesized that miR-2115 acts through a radial glia enriched gene regulatory network involving ORC4 to regulate cell cycle dynamics and thereby influences cortical progenitor cell function. To test this hypothesis, we first confirmed the binding of the ORC4 miRNA response element and miR-2115 using reporter assay and inhibitor studies (Fig. 4.4a-d, Supplementary Fig. 4.1). Next, we overexpressed a synthetic mmu-miR-2115 (see Methods) in developing mouse cortex and found an increased proportion of radial glia, but fewer radial glia in mitosis, among the electroporated cells (Supplementary Fig. 4.13). Similarly, manipulation of miR-2115 expression influenced the development of human primary radial glia cells *in vitro*. Both overexpression and inhibition of miR-2115 changed the proportion of cells expressing PAX6, indicating a possible role for this miRNA in proliferation or differentiation (Fig. 4.4c-d). To more specifically test for a possible cell cycle phenotype, we performed a cumulative BrdU incorporation assay (Fig.

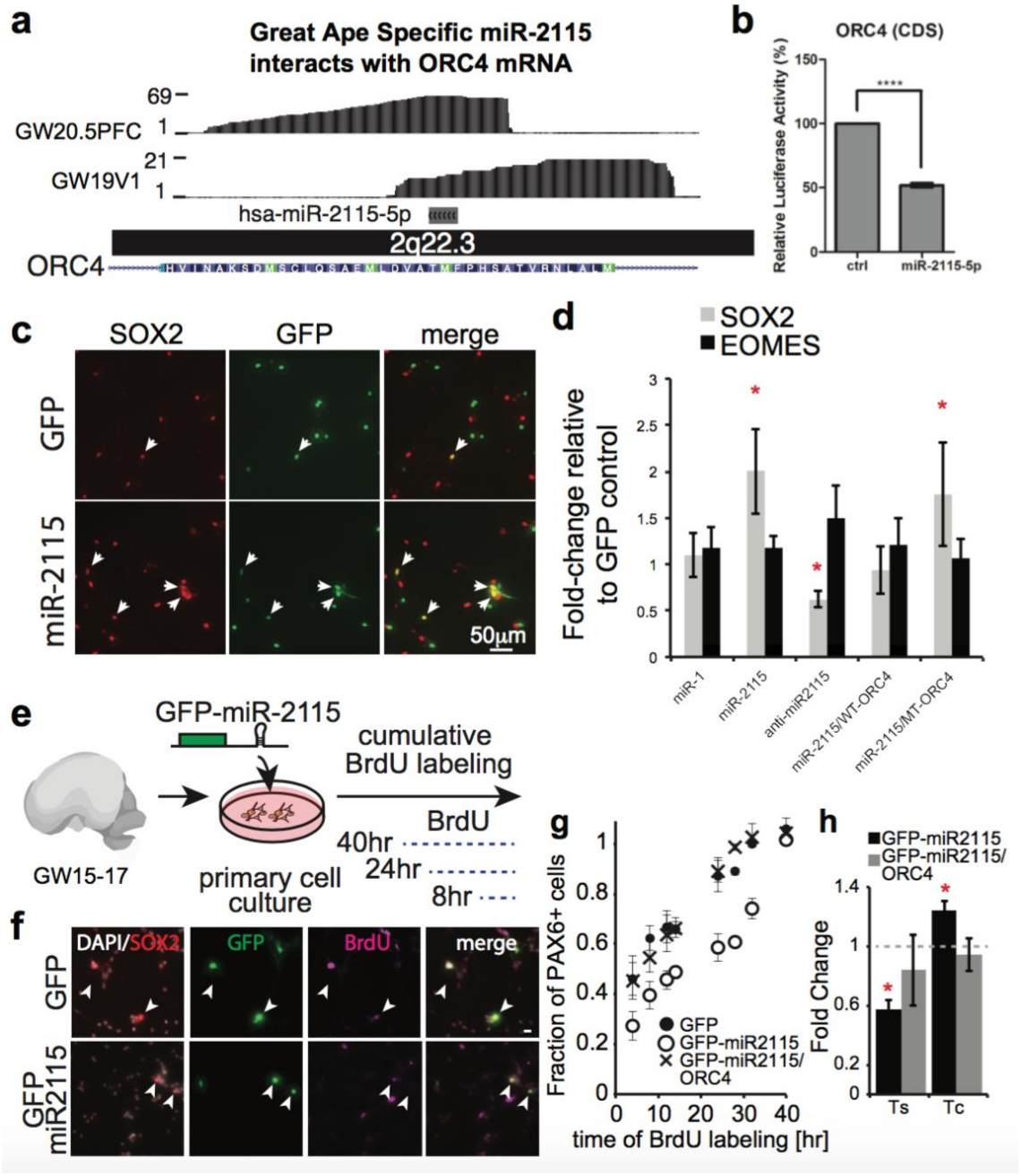


Figure. 4.4. miRNAs contribute to cell-type-specific function. (a) Predicted miR-2115 interaction with *ORC4* mRNA in the CDS. **(b)** Luciferase reporter assay (**** p<0.0001, unpaired t-test). **(c)** miRNA-2115 influences radial glia development in human primary cells. **(d)** Quantification of immunopositive cells (N = 3 biological replicates). "miR-1" – GFP-miR1 small RNA overexpression control construct, "miR-2115" – GFP-miR2115 overexpression construct, "WT-ORC4" – reporter construct with wild type miR-2115 response element containing sequence, "MT-ORC4" – reporter construct lacking miR-2115 response element, "anti-

miR-2115" – miR-2115 inhibitor co-transfected with GFP expression construct. All constructs and reagents are described in the Methods. **(e)** Experimental design of cumulative BrdU labeling in primary human cells *in vitro*. **(f)** Immunostaining of human cultured primary cells. Arrowheads indicate GFP and SOX2 double positive cells. Scale bar 10μm **(g-h)** Quantification of BrdU labeling of PAX6 positive cells **(g)** and estimates of S-phase length (Ts) and cell cycle length (Tc) **(h)** relative to control conditions (N = 3 biological replicates). *-$p<0.05$, student's t-test.

4.4e-g), and showed that miR-2115 expression regulates normal cell cycle duration in primary human radial glia (Fig. 4.4h). Together, these findings suggest that miR-2115 emerged recently in evolution and integrated into post-transcriptional regulatory networks controlling cell cycle dynamics during human cortical development.

Discussion:

Our study revealed three distinct mechanisms by which miRNA regulatory pathways contribute to human brain development. Our novel single cell profiling approach revealed dynamic changes in miRNA expression during neuronal differentiation, suggesting that miRNA abundance profiles may further contribute to unbiased cell type classification in complex tissues. Moreover, by intersecting high throughput profiling of miRNA-mRNA interactions with cell type specific gene expression profiles, we demonstrate that many miRNAs, including those expressed in multiple cell types, can regulate important aspects of cell identity by directly regulating cell-type-specific gene expression. Furthermore, by projecting cell-type-specific miRNA and mRNA expression patterns against the modular framework of the bipartite network of miRNA-mRNA interactions, our study revealed dynamic developmental remodeling of miRNA-mRNA interaction networks involving conserved and recently evolved miRNAs, as well as cell-type-specific miRNA regulatory

networks operating in the developing human brain (Supplementary Table 4.12). Comprehensive understanding of cell-type-specific miRNA-mRNA interactions may reveal previously unappreciated patterns of selective vulnerability of cell-types in neurodevelopmental disorders, including Autism Spectrum Disorders.

References:

1 Tasic, B. Single cell transcriptomics in neuroscience: cell classification and beyond. *Current opinion in neurobiology* **50**, 242-249, doi:10.1016/j.conb.2018.04.021 (2018).

2 Griffiths, J. A., Scialdone, A. & Marioni, J. C. Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular systems biology* **14**, e8046, doi:10.15252/msb.20178046 (2018).

3 Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331-338, doi:10.1038/nature21350 (2017).

4 Zeng, H. & Sanes, J. R. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nature reviews. Neuroscience* **18**, 530-546, doi:10.1038/nrn.2017.85 (2017).

5 Monticelli, S. *et al.* MicroRNA profiling of the murine hematopoietic system. *Genome biology* **6**, R71, doi:10.1186/gb-2005-6-8-r71 (2005).

6 Fineberg, S. K., Kosik, K. S. & Davidson, B. L. MicroRNAs potentiate neural development. *Neuron* **64**, 303-309, doi:10.1016/j.neuron.2009.10.020 (2009).

7 Volvert, M. L., Rogister, F., Moonen, G., Malgrange, B. & Nguyen, L. MicroRNAs tune cerebral cortical neurogenesis. *Cell death and differentiation* **19**, 1573-1581, doi:10.1038/cdd.2012.96 (2012).

8 Berezikov, E. *et al.* Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120**, 21-24, doi:10.1016/j.cell.2004.12.031 (2005).

9 Kapsimali, M. *et al.* MicroRNAs show a wide diversity of expression profiles in the developing and mature central nervous system. *Genome biology* **8**, R173, doi:10.1186/gb-2007-8-8-r173 (2007).

10 Baudet, M. L. *et al.* miR-124 acts through CoREST to control onset of Sema3A sensitivity in navigating retinal growth cones. *Nature neuroscience* **15**, 29-38, doi:10.1038/nn.2979 (2011).

11 Bernstein, E. *et al.* Dicer is essential for mouse development. *Nature genetics* **35**, 215-217, doi:10.1038/ng1253 (2003).

12 Jonsson, M. E. *et al.* Comprehensive analysis of microRNA expression in regionalized human neural progenitor cells reveals microRNA-10 as a caudalizing factor. *Development* **142**, 3166-3177, doi:10.1242/dev.122747 (2015).

13 Moore, M. J. *et al.* Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. *Nature protocols* **9**, 263-293, doi:10.1038/nprot.2014.012 (2014).

14 Pollen, A. A. *et al.* Molecular identity of human outer radial glia during cortical development. *Cell* **163**, 55-67, doi:10.1016/j.cell.2015.09.004 (2015).

15 Camp, J. G. *et al.* Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 15672-15677, doi:10.1073/pnas.1520760112 (2015).

16 Miller, J. A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199-206, doi:10.1038/nature13185 (2014).

17 Liu, X. & Murata, T. Community detection in large-scale bipartite networks. *Information and Media Technologies* **5**, 184-192 (2010).

18 Florio, M. *et al.* Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* **347**, 1465-1470, doi:10.1126/science.aaa1975 (2015).

19 Yu, B. *et al.* miR-221 and miR-222 promote Schwann cell proliferation and migration by targeting LASS2 after sciatic nerve injury. *Journal of cell science* **125**, 2675-2683, doi:10.1242/jcs.098996 (2012).

20 Maiorano, N. A. & Mallamaci, A. Promotion of embryonic cortico-cerebral neuronogenesis by miR-124. *Neural development* **4**, 40, doi:10.1186/1749-8104-4-40 (2009).

21 Boumil, R. M. *et al.* A missense mutation in a highly conserved alternate exon of dynamin-1 causes epilepsy in fitful mice. *PLoS genetics* **6**, doi:10.1371/journal.pgen.1001046 (2010).

22 Buckanovich, R. J., Yang, Y. Y. & Darnell, R. B. The onconeural antigen Nova-1 is a neuron-specific RNA-binding protein, the activity of which is inhibited by paraneoplastic antibodies. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **16**, 1114-1122 (1996).

23 Nowakowski, T. J. *et al.* MicroRNA-92b regulates the development of intermediate cortical progenitors in embryonic mouse brain. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 7056-7061, doi:10.1073/pnas.1219385110 (2013).

24 Magri, L. *et al.* c-Myc-dependent transcriptional regulation of cell cycle and nucleosomal histones during oligodendrocyte differentiation. *Neuroscience* **276**, 72-86, doi:10.1016/j.neuroscience.2014.01.051 (2014).

25 Kawakami, Y. *et al.* Impaired neurogenesis in embryonic spinal cord of Phgdh knockout mice, a serine deficiency disorder model. *Neuroscience research* **63**, 184-193, doi:10.1016/j.neures.2008.12.002 (2009).

26 Nowakowski, T. J. *et al.* Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318-1323, doi:10.1126/science.aap8809 (2017).

27 Hadjighassem, M. R. *et al.* Human Freud-2/CC2D1B: a novel repressor of postsynaptic serotonin-1A receptor expression. *Biological psychiatry* **66**, 214-222, doi:10.1016/j.biopsych.2009.02.033 (2009).

28 Deshar, R., Cho, E. B., Yoon, S. K. & Yoon, J. B. CC2D1A and CC2D1B regulate degradation and signaling of EGFR and TLR4. *Biochemical and biophysical research communications* **480**, 280-287, doi:10.1016/j.bbrc.2016.10.053 (2016).

29 Fededa, J. P. *et al.* MicroRNA-34/449 controls mitotic spindle orientation during mammalian cortex development. *The EMBO journal* **35**, 2386-2398, doi:10.15252/embj.201694056 (2016).

30 Wu, J. *et al.* Two miRNA clusters, miR-34b/c and miR-449, are essential for normal brain development, motile ciliogenesis, and spermatogenesis. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E2851-2857, doi:10.1073/pnas.1407777111 (2014).

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

32 Boudreau, R. L. *et al.* Transcriptome-wide discovery of microRNA binding sites in human brain. *Neuron* **81**, 294-305, doi:10.1016/j.neuron.2013.10.062 (2014).

33 Wu, Y. E., Parikshak, N. N., Belgard, T. G. & Geschwind, D. H. Genome-wide, integrative analysis implicates microRNA dysregulation in autism spectrum disorder. *Nature neuroscience* **19**, 1463-1476, doi:10.1038/nn.4373 (2016).

34 Abu-Elneel, K. *et al.* Heterogeneous dysregulation of microRNAs across the autism spectrum. *Neurogenetics* **9**, 153-161, doi:10.1007/s10048-008-0133-5 (2008).

35 He, M. *et al.* Cell-type-based analysis of microRNA profiles in the mouse brain. *Neuron* **73**, 35-48, doi:10.1016/j.neuron.2011.11.010 (2012).

36 Liu, J. *et al.* A reciprocal antagonism between miR-376c and TGF-beta signaling regulates neural differentiation of human pluripotent stem cells. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **28**, 4642-4656, doi:10.1096/fj.13-249342 (2014).

37 Nowakowski, T. J., Pollen, A. A., Sandoval-Espinosa, C. & Kriegstein, A. R. Transformation of the Radial Glia Scaffold Demarcates Two Stages of Human Cerebral Cortex Development. *Neuron* **91**, 1219-1227, doi:10.1016/j.neuron.2016.09.005 (2016).

38 Guernsey, D. L. *et al.* Mutations in origin recognition complex gene ORC4 cause Meier-Gorlin syndrome. *Nature genetics* **43**, 360-364, doi:10.1038/ng.777 (2011).

39 de Munnik, S. A. *et al.* Meier-Gorlin syndrome: growth and secondary sexual development of a microcephalic primordial dwarfism disorder. *American journal of medical genetics. Part A* **158A**, 2733-2742, doi:10.1002/ajmg.a.35681 (2012).

40 Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology* **33**, 495-502, doi:10.1038/nbt.3192 (2015).

Methods:

Tissue Samples

De-identified human tissue samples were collected with patient consent in strict observance of the legal and institutional ethical regulations from elective pregnancy termination. Protocols were approved by the Human Gamete, Embryo, and Stem Cell Research Committee (UCSF institutional review board) at the University of California, San Francisco. The prenatal brain tissue processing for dissociation as well as fixation, cryosectioning, and long-term storage was performed as described before [1].

Mice

All mice in this study were obtained from Simonsen Laboratories and maintained according to protocols approved by the Institutional Animal Care and Use Committee at UCSF.

AGO2-HITS-CLIP

The experiments were performed as described earlier [2] except for a few modifications. Monoclonal anti-Ago2 antibody (Sigma, 11A9 clone, SAB4200085) was used to perform immunoprecipitation of protein on protein G dynabeads (Invitrogen, 100-03D). For the negative control, goat anti-rat IgG antibody (Sigma, A9037) was used. RNase dilution of 1:50,000 was used after optimization. Primary tissue samples at stages corresponding to peak neurogenesis (gestational weeks GW 15-16.5), and at stages corresponding to upper

layer neurogenesis and early gliogenesis were analyzed. Sample quality and data processing metrics were comparable to previously published results for adult human brain HITS-CLIP study [3]. 3258 genes were actively targeted by miRNAs through protein coding (CDS) or 3' untranslated regions (3'UTR) of the transcript. Strikingly, close to 80% of gene targets were detected in at least two samples, suggesting that we recovered the majority of *in vivo* mRNA targets.

Library Preparation for Sequencing RNA tags on Ion-Torrent

For library preparation, all steps were performed as described before [2], except that the primers with adapter sequences were modified according to the Ion Torrent sequencing platform.

Primers with Adapter sequences -

DSFP5 - 5'-CCATCTCATCCCTGCGTGTCTCCGACTCAGAGGGAGGACGATGCGG -3'

DSFP3 - 5'- CCTCTCTATGGGCAGTCGGTGATCCGCTGGAAGTGACTGACAC -3'

The bands corresponding to AGO2:miRNA:target complexes (130 kDa) were cut. miRNA libraries and target mRNA libraries were made separately.

Plasmid Constructs

gBlocks gene fragments for the respective target sites or miRNAs (table S4.3) with the restriction site(s) at their ends were purchased from IDT. Restriction site cloning was performed using the standard method. ORC4 expression plasmid was generated using gene

synthesis and cloned into CAG-IRES-GFP vector using the GeneArt service (Thermo Fisher).

Luciferase Activity Assay

Luciferase activity assay was performed as described earlier [4] in HEK293 cells. The target sites were cloned in psiCHECK2 (Promega) plasmid (table S4.3) and miR-2115 and miR-9 were cloned in pCAG-GFP (Addgene 11150) plasmid (table S4.2). miRNA mimics (Life Technologies) were used for other miRNA assays.

Preprocessing and Mapping of AGO2-HITS-CLIP Tags

Barcodes were identified and reads were separated into each sample. Adapter sequences at both ends of reads were removed using Cutadapt [5]. Trimmed reads were mapped to the human genome (hg19) with novoalign (http://www.novocraft.com/). Identical alignments were collapsed in each sample to remove PCR replicates. Strand specific read coverage was then calculated using the alignments from each sample.

miRNA profiling and Differential Expression Analysis

Adapter-trimmed AGO2 reads from miRNA libraries were mapped to human miRNA precursors from miRBase version 21 by using miRdeep2 [6]. DESeq2 was used to identify differentially expressed miRNAs between 2 developmental stages [7].

Peak Calling and Identification of Clusters/AGO Footprint

Piranha and zero-truncated negative binomial model (ZTNB) were used to calculate the significance of read coverage at each mapped genomic position in each samples [3,8]. In

summary, the read heights for each mapped genomic position were assumed to be sampled from an underlying ZTNB distribution, and parameters for ZTNB probability density functions were estimated using read height measured at all positions of the genome. P-values were calculated as the probability of observing a read height as large as the height in question and assigned to each position. We used Fisher's method to calculate the joined p-values at each genomic position across all samples. Positions with a joined FDR <5% were deemed significant. Significant positions within 50 nts of one another were merged into a single contiguous interval as single AGO binding site, and resulting regions of less than 50 nts were symmetrically extended to 50 nts as accounting for the AGO binding footprint (*9*). AGO binding sites were then annotated according to their overlapping gene structures from GENCODE annotation Version 19.

Identification of miRNAs for each AGO binding sites

In order to identify miRNAs that bind each AGO site on mRNAs and lncRNAs, all types of canonical binding sites including 7mer-1A,7mer-m8 and 8mer [10] for all prenatal brain expressed miRNAs were searched within the full length AGO binding sites defined as above. miRNAs with miRNA recognition elements (MREs) within each AGO binding site were counted.

Primary human cell dissociation and culture

Primary human cortical cells were dissociated using Papain (Worthington labs) and cultured on matrigel (BD Biosciences) coated tissue culture treated plates. Cells were plated at approximately 100,000-200,000 cells per well of a 12 well plate. Culture media used in this experiment consisted of DMEM (Invitrogen, 11965) supplemented with N2 (Invitrogen,

12587-010) and B27 (17502-048), as well as Penicillin and Streptomycin, but no serum. At the time of plating, culture media was spiked with recombinant human FGF-basic (10ng/ml, Peprotech, AF-100-18B). Approximately 24-48 hours after plating, cells were transfected with plasmids using Lipofectamine 2000 (Life Technologies) following manufacturer protocol. BrdU (Sigma) was diluted in the culture media for dissociated cells (DMEM, supplemented with B27 and N2, with Penicillin-Streptomycin) at 50 μg/ml.

Single-cell qPCR Analysis

The capture of single-cells was done using the C1 Single-Cell Auto Prep Integrated Fluidic Circuit (IFC), which uses a microfluidic chip to capture the cells, perform lysis, reverse transcription and cDNA amplification in nano-liter reaction volumes of miRNA and mRNA species at the same time. The details of the cell capture protocol used are described in protocol 100-6667 at http://www.fluidigm.com/. During the reverse transcription step, miRNAs are reverse-transcribed to cDNA using stem loop RT primers from the Megaplex RT primer pool (Life Technologies) that are specific for mature miRNA species and reagents from the Single-cell-to-CT kit (Life Technologies). mRNA species are reverse-transcribed at the same time during this process using mRNA primers which are present in the Single-Cell VILO RT. Megaplex primers and mRNA primers are added at the recommended concentrations.

During the PCR step, products are uniformly amplified from cDNA templates using the Megaplex PreAmp Primers (Life Technologies), a pool of DELTAgene primers and Single-Cell Preamp mix from the Ambion Single-Cell-to-CT kit (Life Technologies).

RT			
Stage	Temp [C]	Time(min)	Cycles
Anneal	16	2	40
Extend	42	1	
Extend	50	1	
Hold	85	5	1
Hold	4	Hold	Hold

Preamplification			
Stage	Temp [C]	Time(min)	Cycles
Enzyme activ	95	10	1
Hold	55	2	1
Hold	72	2	1
Denature	95	15	18
Anneal/Exter	60	4	
Hold	99.9	10	1
Hold	4	Hold	Hold

After pre-amplification PCR, the amplicons were diluted 1:4 with C1 DNA Dilution Reagent (Fluidigm 100-5317) and stored in -20°C until needed. qPCR was carried out using the 96.96 dynamic array (Fluidigm Corporation) following the manufacturer's protocol (100-3909 and 100-9792). Gene expression analysis was done using the Fluidigm Real-Time PCR Analysis Software (v.3.0.2). Ct values were obtained and then square root normalized to stabilize the variance.

Clustering of single-cells was performed using a recently developed method combining Louvain clustering of single-cell sample coordinates with Jaccard distance metric [11]. T-stochastic neighbor embedding (tSNE) was used to visualize cells in two dimensions.

WGCNA Analysis

To detect groups of co-expressed miRNAs, we used the WGCNA R package [12]. Cells analyzed using sc-qPCR were included in this analysis.

Cell-type-specificity and Enrichment Analysis for miRNAs in single-cell data

We performed one-tailed Wilcox/Mann-Whitney-U Test to detect differential enrichment of each miRNA in each cell-type against the other cell-types. A miRNA is defined to be enriched in a cell-type, if it is expressed significantly higher in that cell-type (FDR<0.05).

Correlation with target mRNA levels

To correlate the relative abundance of co-expressed miRNAs and their targets across the major cell-types of the developing brain, we calculated the average module eigen-gene for modules detected in fig. S4.2 across radial glia, intermediate progenitors, neurons and interneurons. In parallel, we used published scRNA-seq data [13] and the cell-type assignments therein for radial glia, interneurons, intermediate progenitors, and excitatory cortical neurons, to calculate cell-type-wise average expression level of the genes identified as AGO2 bound miRNA targets as well as non-targets. We then correlated the average module eigen-gene with all of the gene targets predicted by HITS-CLIP and non-targets, and we calculated the average correlated for each in Fig. 4.2G. To compare the average correlations, we first converted correlation to z-scores using Fisher transformation. The differences between average z-scores (for targets and non-targets) were divided by the joint standard errors and significance was calculated based on normal distribution.

In-situ hybridization

In situ hybridization in primary tissue sections was performed as described before [14], with the exception that we did not perform a probe linearization step. Locked nucleic acid probes for miRNA were purchased from Exiqon.

.

In-utero electroporation

Survival in utero surgery was performed in strict observance of protocols and recommendations approved by the Institutional Animal Care and Use Committee at UCSF. Plasmids were injected at approximately 1.5 μg/μl as described before [15,16]. Although the ORC4 protein sequence is highly conserved, the miR-2115-5p MRE sequence is not fully conserved in mouse. We generated a mutant miR-2115-5p hairpin sequence (mmu-miR-2115) whose seed would be complementary to the mouse ORC4 mRNA coding sequence at the site homologous to the human miR-2115-5p MRE.

Immunofluorescence

Thin 20 μm cryosections were collected on superfrost slides (VWR) using Leica CM3050S cryostat. Immunohistochemistry based detection of specific antigens was performed according to standard protocols. In short, heat-mediated antigen retrieval was performed in 10mM sodium citrate for 15 min. Cells were permeabilized in phosphate buffered saline (pH = 7.4, PBS) supplemented with 2% Triton X-100. Blocking buffer consisted of PBS supplemented with 10% donkey serum, 0.2% gelatin and 2% Triton X-100. The antibodies used in this study included chicken anti-GFP (1:1000, Aves Labs GFP-1020), rabbit anti-PAX6 (1:300, Covance prb-278p), and mouse anti-pHH3 (1:100, Abcam

ab1791). Secondary antibodies used were from Life Technologies. Nuclei were counterstained with DAPI (Sigma).

After cell fixation, BrdU epitope was unmasked using 2N hydrochloric acid, neutralized using 0.1M boric acid, and stained using a rat anti-BrdU [BU1/75 (ICR1)] antibody (1:50, Abcam ab6326). Coverslips were mounted with Aqua-mount (Lerner Laboratories)

Images were collected with a Leica DMI 4000B microscope using a Leica DFC295 camera Leica TCS SP5 X Confocal microscope. Quantification of immune-positive cells was performed in Adobe Photoshop. Mouse embryonic electroporation cortical staining was quantified as described before [15]. Quantification results for every biological replicate (embryo) represent an average of quantification across three non-adjacent sections. Replicates were drawn from at least two independent litters. Quantification of BrdU incorporation into PAX6 positive primary cells in culture was performed by imaging randomly selected fields in the well using tile-scanning feature. GFP positive cells were first evaluated for expression of PAX6, and only after that, BrdU immunoreactivity was assessed. Between 100 and 200 PAX6/GFP double-positive cells were evaluated per well. Quantification of cell cycle phenotypes was performed by fitting linear regression, and cell cycle parameters were calculated as described before [17].

Gene Enrichment Analysis

Transcriptome data from similar prenatal brain tissues [18] was downloaded and the expressed genes were used as a better background set for gene enrichment analysis. The p-value was calculated with hypergeometric test, and adjusted using the Benjamini–Hochberg (BH) method. To formally demonstrate that cell-type-specific

genes are regulated by miRNAs, we used published scRNA-seq datasets to calculate cell-type specificity scores using ideal vector correlation for every miRNA target identified by HITS-CLIP (Supplementary Table 4.7). Genes with a Pearson correlation above 0.3 were considered cell-type-specific [1]. Strikingly, a high fraction of all genes specific to radial glia or neurons are targeted by miRNAs, with ~75% of radial glia specific genes, and ~32% neuron-specific genes bound by AGO2.

Bipartite Community Detection Analysis

An unweighted (binary) bipartite network was constructed such that there exists an edge between each miRNA-mRNA pair if and only if such interaction is detected by AGO2-HITS-CLIP. First, this network was shown to be scale-free ($P < 0.001$), and then by generating random networks while constraining the number of edges and nodes to the original miRNA-mRNA network and calculating Barber`s modularity score [19] (Null distribution), the miRNA-mRNA network shows to be significantly modular (P for permutation test < 2e-16). Label Propagation followed by the Bipartite Recursively Induced Modularity (LP-BRIM) algorithm [20] was used for community detection in this bipartite network. Due to stochasticity of this method, we obtained robust communities by repeating LP-BRIM 2500 times and determining the overlap among all the iterations.

Bipartite network construction

The bipartite networks were generated using all detected target genes (3463 nodes in mode I) and all miRNA (514 nodes in mode II). An edge exists between only one node in mode I and one node in mode II if such interaction is present in HITS-CLIP data set (Table

S4.2) (no edge is allowed between nodes in mode I (or mode II)). Hence in this study, three distinct networks were built using GW15-16, GW19-20 and the combined GW15-16/19-20 HITS-CLIP data set with 31859, 20734, and 36176 total edges, respectively.

Bipartite network modularity statistics

First, the R "bipartite" package [21] was used to show that the constructed network is scale-free. For consistency with the literature definition of scale-free networks, we showed that this network significantly obeys power law, truncated power law, and exponential distribution (Supplementary Fig. 4.11 A and B, $P < 0.001$). Next to demonstrate that these networks are significantly modular, we first generated a null distribution by randomly shuffling the edges between the nodes and calculating Barber`s modularity score. Finally, we showed that the network is significantly modular using a permutation test ($P < 2e\text{-}16$).

Bipartite community detection

We used R "lpbrim" package [22] to detect communities in the described networks. Due to stochasticity of the method, we ran the community detection algorithm 25 times and obtained the best solution among all 25 runs (with maximum Barber`s modularity score). We repeated this procedure 100 times, and found the consensus clustering as follows:

1) We first defined the Overlap Rate Matrix (ORM) as a symmetric N by N matrix where N is the total number of miRNA and target genes combined. Each entry ORM_{ij} shows the probability (or rate) of which target gene (or miRNA) *i* lies within the same cluster as target gene (or miRNA) *j* upon 100 runs.

2) The obtained ORM was further clustered using hierarchical clustering (Supplementary Fig. 4.11 C and D). Each obtained diagonal block after hierarchical clustering represents one community (contains both target genes and miRNAs that have significant intra-modular interactions compared to their inter-modular interactions with target genes (or miRNAs) in other communities.).
3) Finally, dynamic branch cutting implemented in "dynamicTreeCut" R package was used for tree cutting and assigning cluster ids to each node (Supplementary Fig. 4.11 E and F).
4) We repeated this procedure to show that the obtained clusters are robust and reproducible (Supplementary Fig. 4.11 C and D).

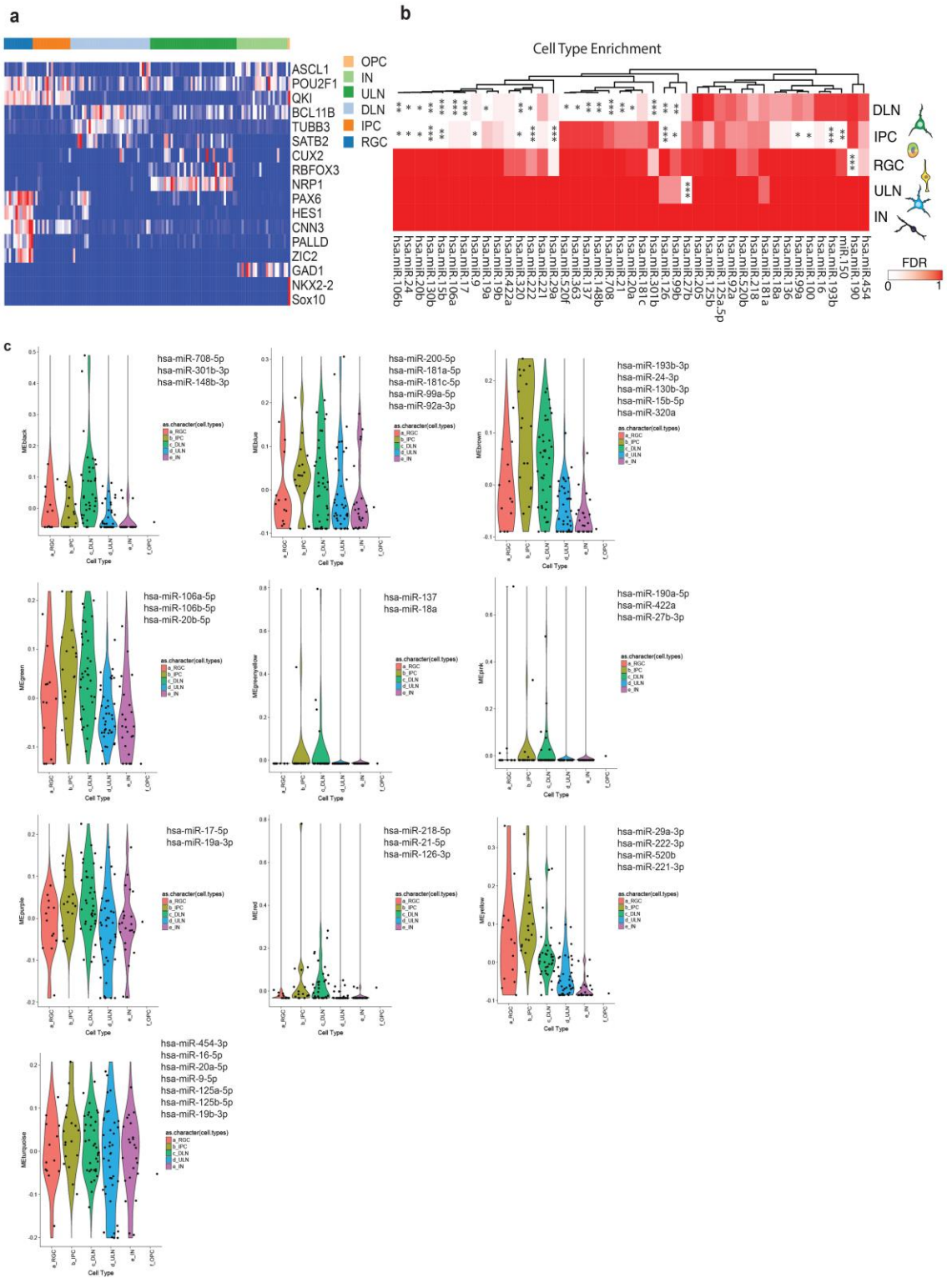
Bipartite module preservation analysis

In order to determine whether each identified module in GW15-16 network is statistically preserved in GW19-20, we first find the closest module in GW19-20 to GW15-16 in terms of the number of shared nodes in that module. Then, using the hypergeometric test, module preservation statistics were obtained and corrected using the BH method for multiple comparison. As shown in Fig. 4.3F, two homolog modules in GW15-16 and GW19-20 have similar interaction levels suggesting preserved topology of modules as well. Note that names of GW15-16 module colors are used to label Fig. 4.3F.

Inference of evolutionary history of miR-2115

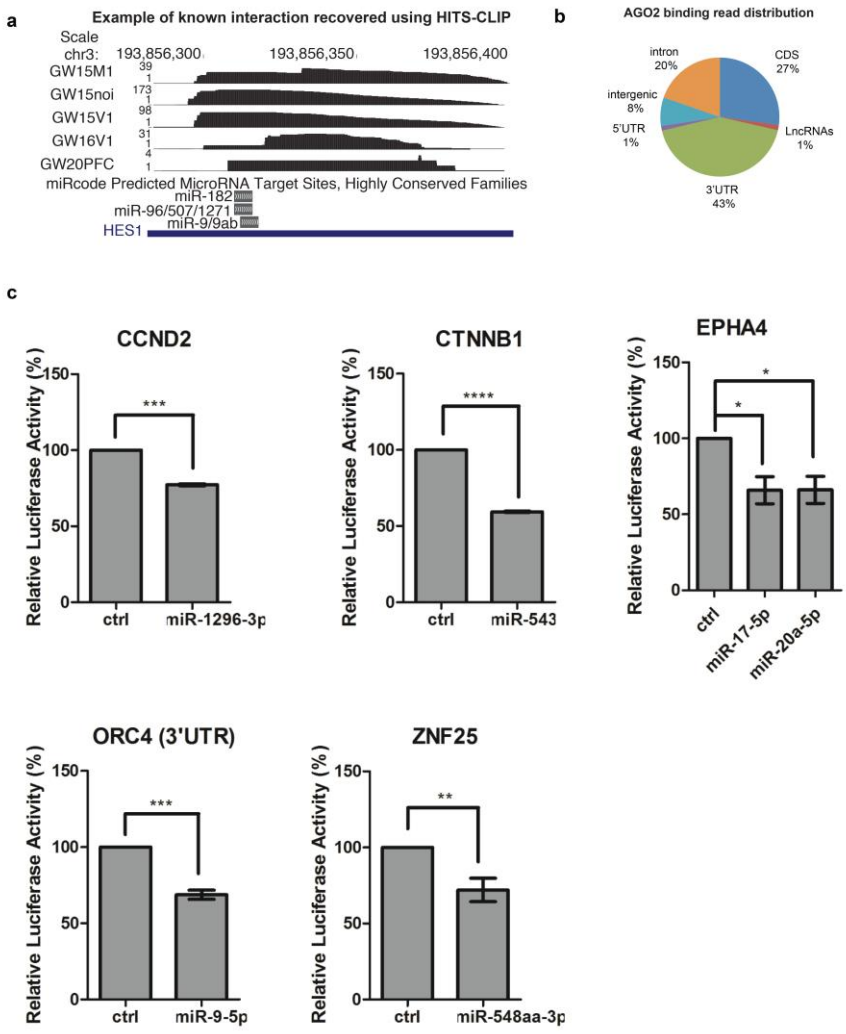
To infer the evolutionary history of miR2115, we used genome sequences obtained from the UCSC Genome Browser (the most recent genome assemblies used). Alignments were

done using the Geneious® bioinformatics platform (version 9.1.8). Human SPINK8 gene sequence is annotated with introns, exons, CDS and UTR regions and miR-2115 location according to "NCBI RefSeq" track on the UCSC Genome Browser. SPINK8 gene orthologs were located in chimpanzee, gorilla, orangutan, and gibbon genomes using "Other RefSeq" track on UCSC Genome Browser. We excluded Bonobo because of the poor quality of genome assembly at the area of interest. Primate SPINK8 sequences were aligned with human SPINK8 sequences individually using the MUSCLE Alignment algorithm. Primate SPINK8 sequences were annotated with introns, exons, CDS and UTR regions and miR-2115 location according to alignment with the annotated human SPINK8 sequence. Originally primate SPINK8 genes were annotated according to the "Other RefSeq" track on UCSC Genome Browser but this yielded varied and unreliable results. Presence or absence of miR-2115 was determined based on alignment of human miR-2115 to the orthologous primate sequence (with mature transcript and seed region taken into consideration). Annotated human, chimpanzee, gorilla, orangutan, and gibbon SPINK8 intron 3/4 (intron miR-2115 is located in) sequences were aligned together to visualize changes in intron sequences between species. Boundaries of insertions and deletions in SPINK8 intron 3/4 occurring between species defined based off evolutionarily chronological alignments of SPINK8 intron 3/4 sequences (e.g. gibbon and orangutan alignment, orangutan and gorilla alignment, etc.). Evolution of the SPINK8 intron 3/4 was predicted using the fewest number of mutations that would give rise to observed insertions and deletions.

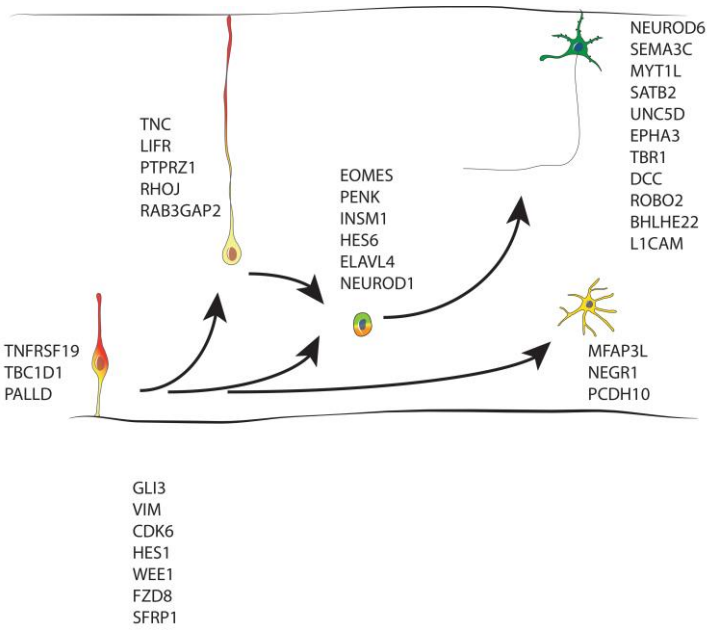


Supplementary Figure. 4.1. Gene co-expression network analysis of miRNAs across single-cells. (a) Heatmap of single-cell qPCR expression data for selected genes informative for cell-type classification. In particular, progenitor cells expressed QKI, POU2F1, CNN3, but radial glia (RGC) further expressed

HES1, PAX6, PALLD[14,23,24], while intermediate progenitors (IPC) expressed higher levels of ASCL1[25,26]. Cells expressing high levels of BCL11B, TUBB3 were interpreted as deep layer neurons (DLN)[27], while cells enriched for the expression of NRP1, CUX2, and RBFOX3 were considered to be upper layer neurons (ULN)[28]. Cells expressing high levels of GAD1 and ACSL1 were considered to represent interneurons (IN)[29], and the single-cell expressing NKX2-2 and SOX10 is considered to be an oligodendrocyte precursor (OPC)[30,31]. **(b)** Heatmap of cell type enriched miRNAs (see also Fig. 4.1d). **(c)** Violin plots of module eigen-gene values of miRNA co-expression networks across cell-types.

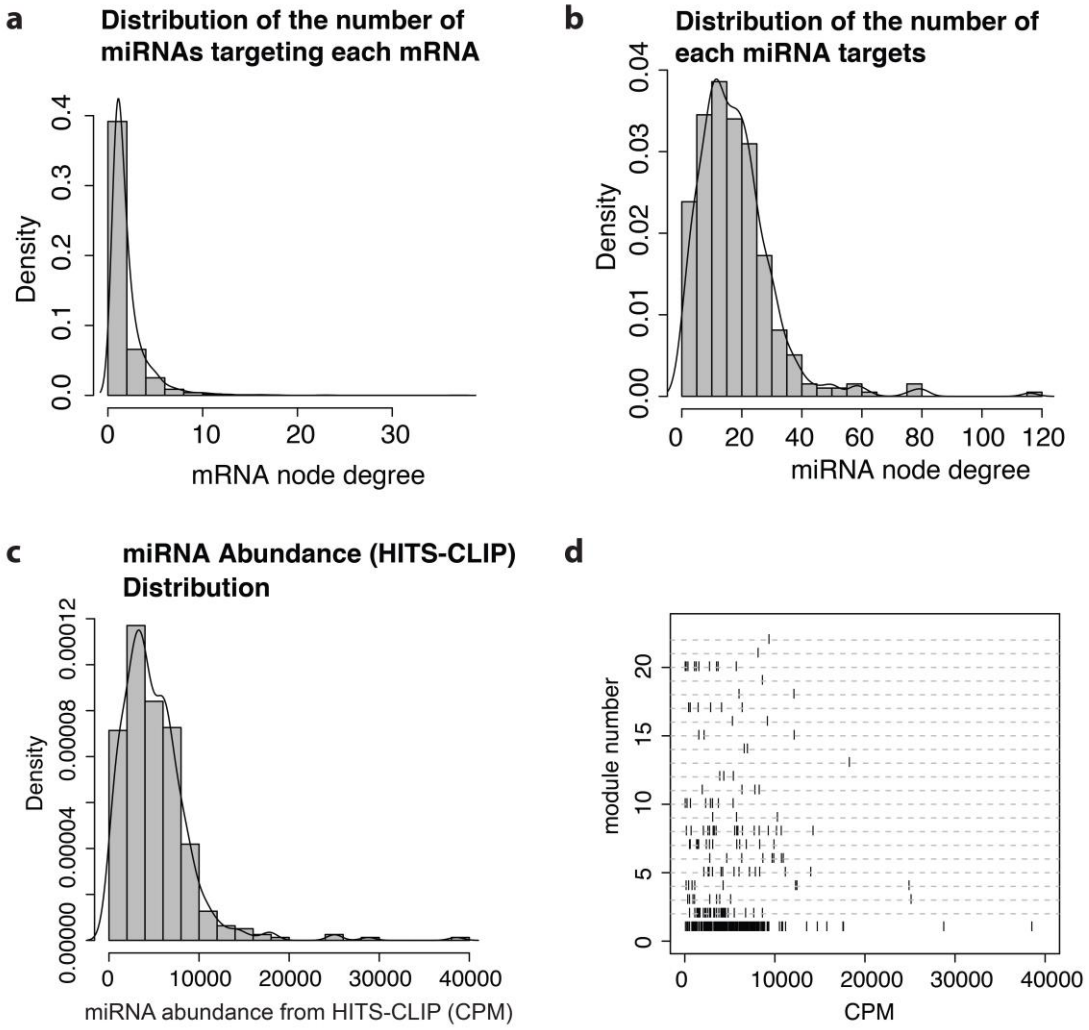


Supplementary Figure. 4.2. Validation of the interaction between miRNA and mRNA identified by HITS-CLIP. (a) Example of AGO2-HITS-CLIP peaks mapping to a previously validated interaction between miR-9 and the 3'UTR of HES1 (*32*). **(b)** Distribution of HITS-CLIP reads. Several long non-coding RNAs (lncRNAs) were also identified among AGO2-bound transcripts (Supplementary Table 2), including 9 lncRNAs previously annotated as cell type specific[18]. **(c)** Luciferase reporter assay showing the down-regulation of luciferase, fused to respective genes, measured 48 h after the overexpression of respective miRNA(s) in HEK293T cells (see Methods for details). *- $p<0.05$, **- $p<0.01$, *** - $p<0.001$, **** - $p<0.0001$.

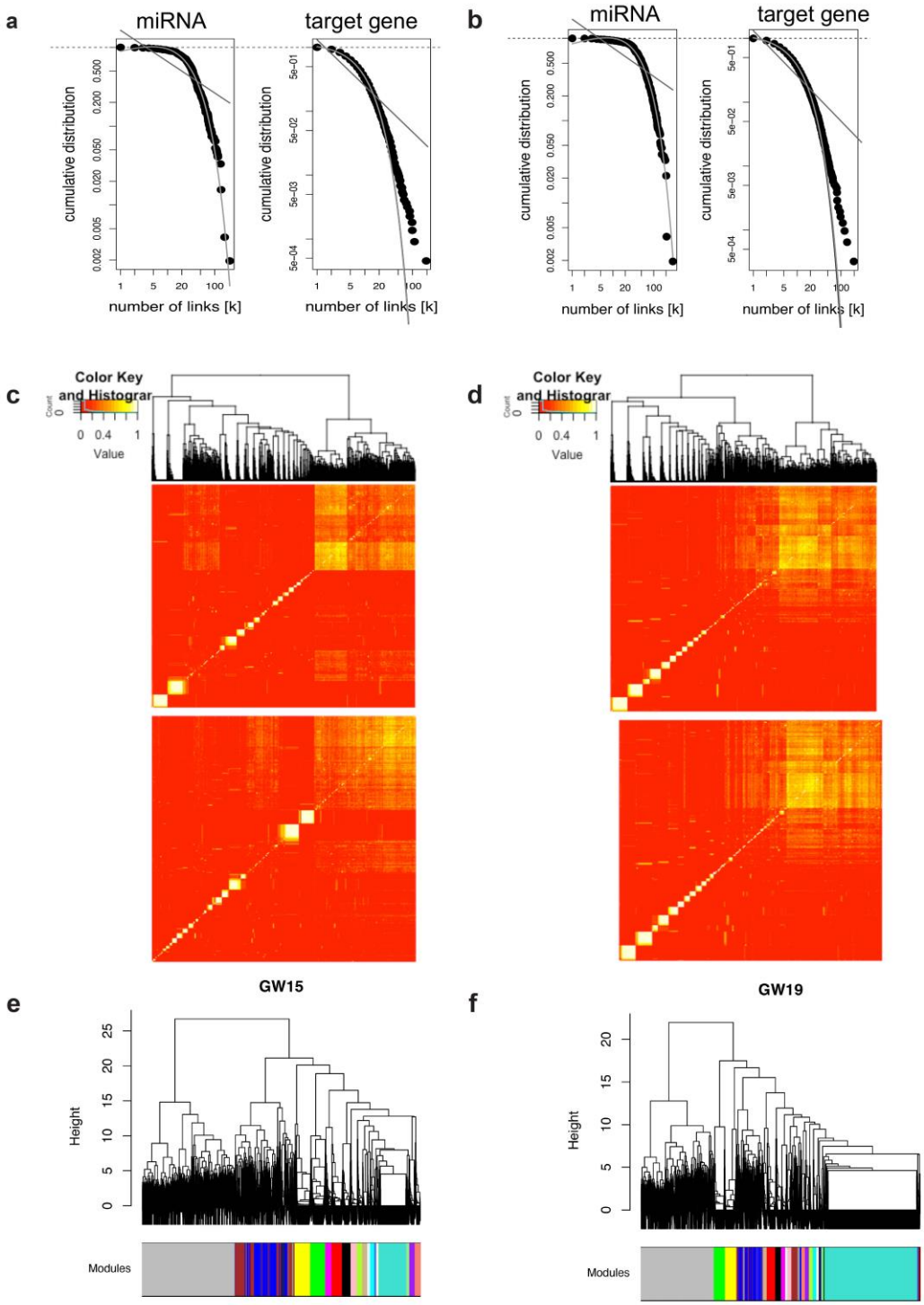


Supplementary Figure. 4.3 Important regulators of cell type identity are targeted by miRNAs. MicroRNAs regulate the expression of genes critical during neuronal differentiation. Examples of cell type enriched genes regulated by microRNAs during peak stages of neurogenesis in human cerebral cortex. Among both ventricular and outer radial glia- enriched genes, microRNAs target many critical components of developmental signaling pathways, including Notch (HES1), Sonic Hedgehog (GLI3, FZD8), Wnt (SFRP1), as well as pathways involved in radial glia subtype specification, such as LIF/LIFR and PTN/PTPRZ1 or TNF/TNFRSF19. Regulation of these pathways by microRNA likely contributes to important aspects of radial glia development. Among intermediate progenitor cell enriched genes, pathways involved in IPC specification and maintenance (EOMEs, HES6, INSM1) are targeted along with proneural factors (NEUROD1, ELAVL4), suggesting that miRNAs may control important aspects of transit amplification and neuronal differentiation. Among excitatory cortical neurons, miRNAs regulate the expression of upper (BHLHE22) and deep layer (TBR1) identity, as well as many genes involved in axon guidance and projection type development (SATB2, SEMA3C, ROBO2, DCC, L1CAM). Interestingly, Both ROBO1 and SLIT1 lay in the **turquoise** module (combined analysis, see Table S8) that which seems to be enriched for cortical interneurons (Fig. 4.3B) in agreement with [33]. Moreover, ROBO1 and ROBO2 both lay in the turquoise module in GW19-20 network. This module is homologous to module **blue,** which is one of the 4 differentially regulated modules during development (Fig. 4.3F) demonstrating their role in developmental processes. In addition, SLIT1 lays in

the grey module in GW19-20, which is not preserved and arises from module **brown** of GW15-16 that is also differentially regulated and has developmental significance (Fig. 4.3f). Overall, all three ROBO1, ROBO2, and SLIT1 are associated with modules that have developmental role.

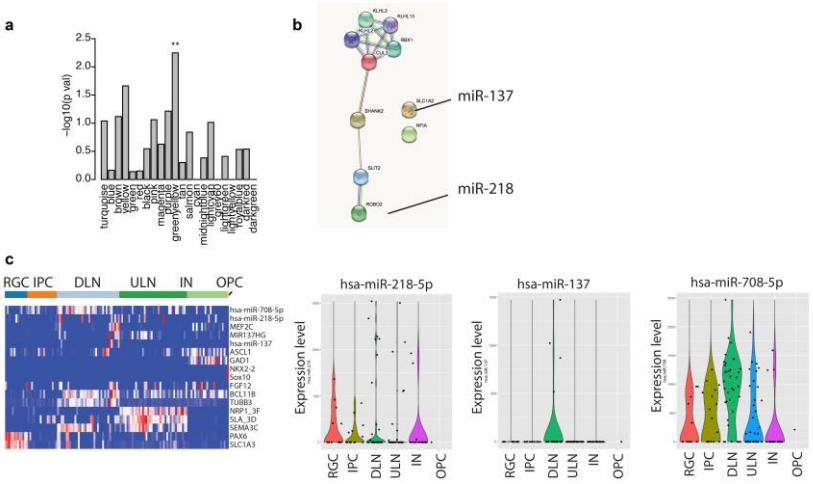


Supplementary Figure. 4.4: miRNA-mRNA node degree distribution. (**a**-**b**) each node in mode I (target genes, **a**) has significantly less connections with the opposite mode than mode II (targeting miRNA, **b**). (**c**) miRNA abundance shows a long-tailed distribution. (**d**) Abundance of miRNA members of each module is illustrated. Modules tend to either recruit several low-abundant miRNAs or few high-abundant miRNAs.

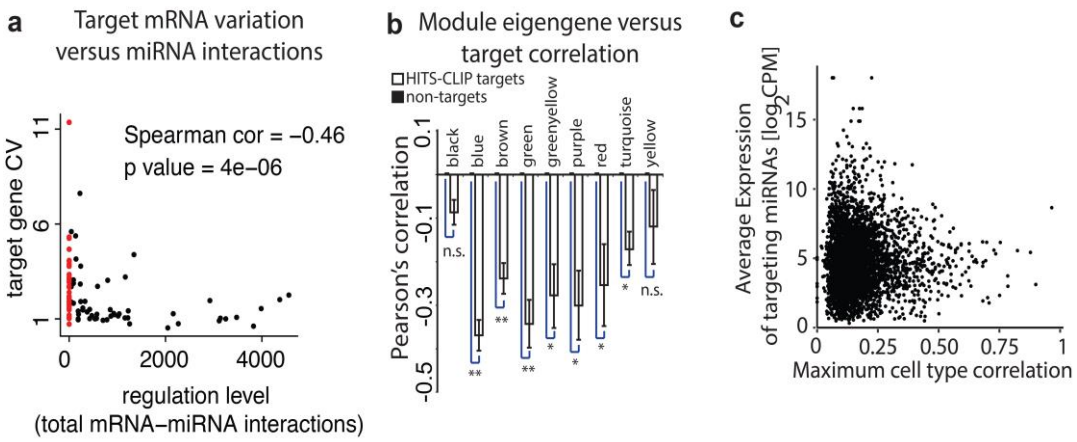


Supplementary Figure. 4.5. Bipartite network analysis. (a-b) Both GW15-16 and GW19-20 networks illustrate a scale-free topology (Dark, median and light grey lines refer to exponential, power and truncated power law, respectively. $P < 0.001$). **(c-d)** Overlap rate matrix after hierarchical clustering

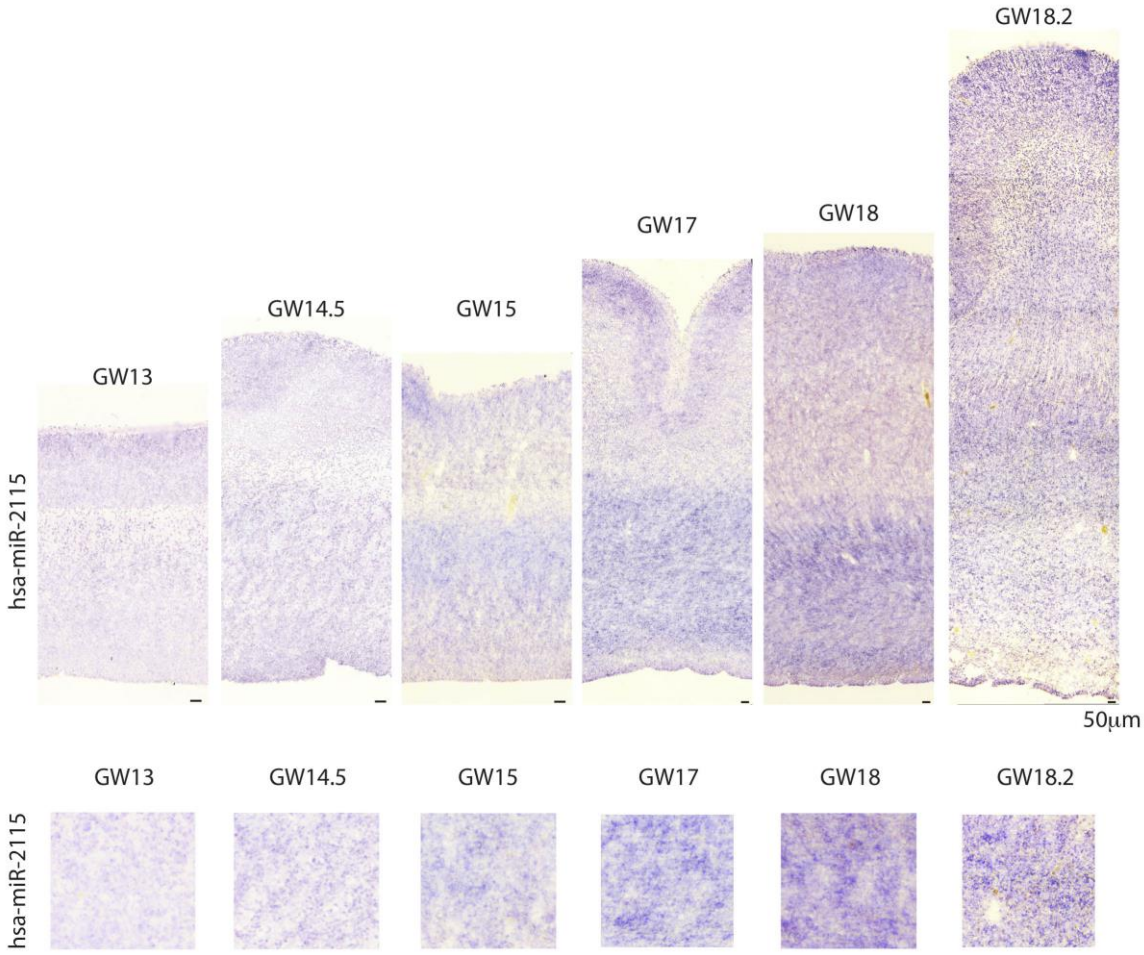
for GW15-16 **(c)** and GW19-20 **(d)**. For each gestational stage two replicates are produced to ensure the robustness of the community detection algorithm. (**e-f)** Dynamic branch cutting of panel **c** and **d**.



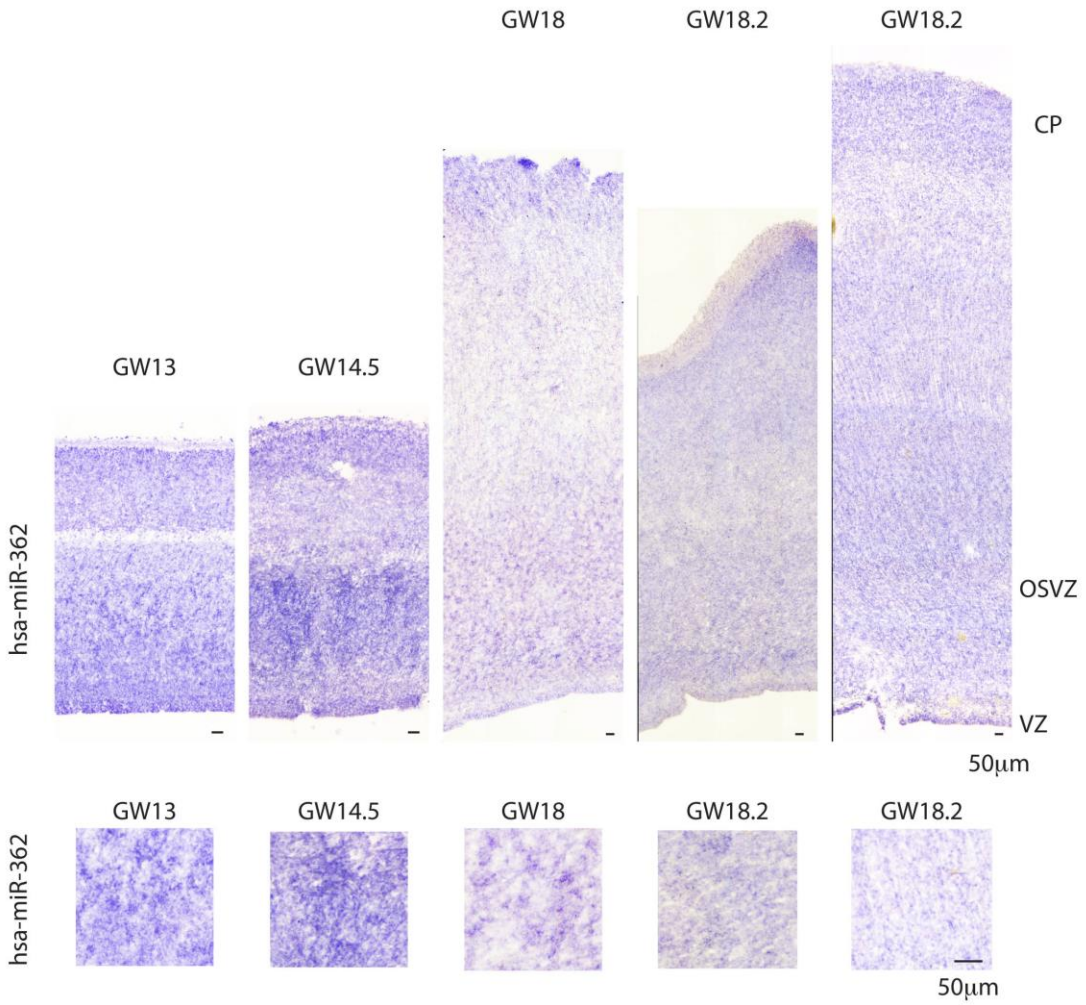
Supplementary Figure. 4.6. miRNA regulation of Autism Spectrum Disorders linked genes. (a) Bipartite network module greenyellow is significantly enriched for genes annotated associated with ASD. **(b)** StringDB representation of ASD relevant genes from bipartite module greenyellow along with highlighted targets of miR-137 and miR-218, recently associated with ASD [34]. (**c**) Heatmap shows sc-qPCR expression values for several of the cell-type markers also shown in fig. S1, and for 3 miRNAs recently associated with ASD [34], miR-708, miR-218, and miR137. Notably, expression of miR-137 and miR-218 is highly correlated and enriched in early maturing (MEF2C- positive) deep layer neurons (TUBB3 and BCL11B double-positive). Expression of miR-708 is enriched in deep layer neurons. Violin plots show distribution of expression levels organized according to inferred cell-types.



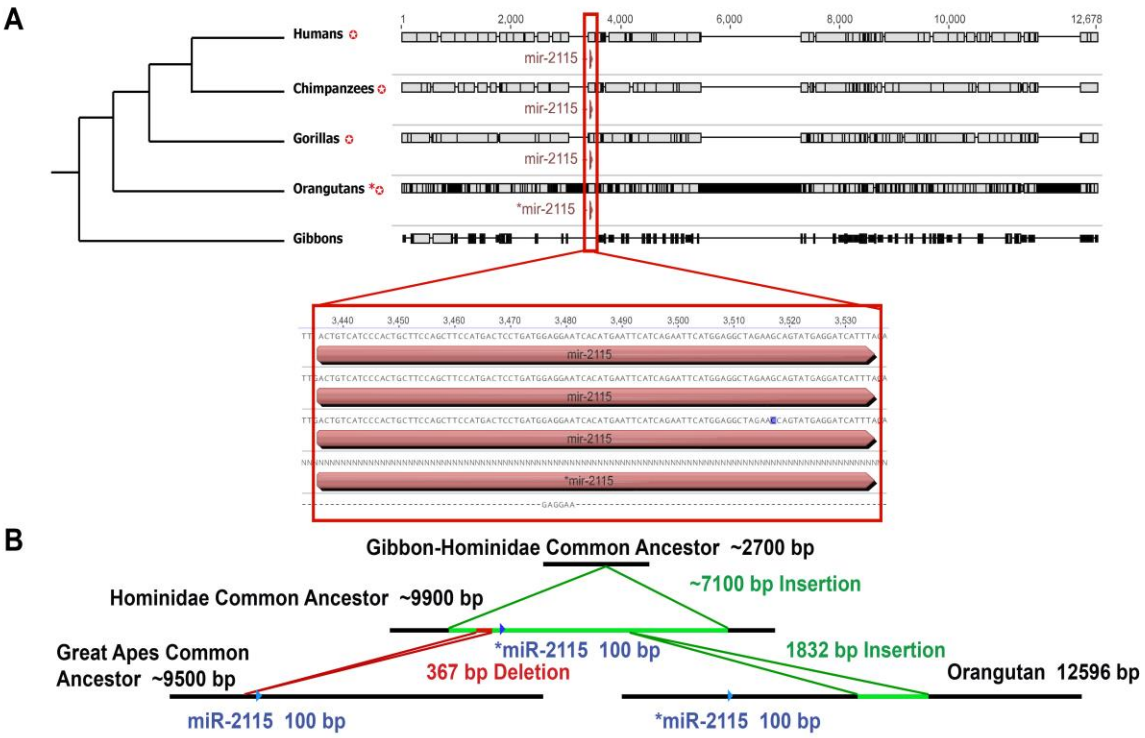
Supplementary Figure. 4.7. Relationship between miRNA and target mRNA abundance. (a) Plot represents coefficient of variation (CV) [35] for every mRNA profiled using sc-qPCR in relation to the number of HITS-CLIP reads reflecting interaction levels. Red points indicate genes not detected in HITS-CLIP. Next, to extend these findings beyond the set of genes we selected for the sc-qPCR, we leveraged published scRNA-seq data [1] to correlate miRNA co-expression module (Supplementary Fig. 4.1) abundance across cell-types with average expression of miRNA targets of that module and non-targets (see Methods). **(b)** Correlation analysis between average module eigengene value across cell-types detected in Fig. S1 and average expression level of HITS-CLIP target and non-target mRNAs across the same cell-types determined using sc-RNAseq in previously published study [1] (see Methods for details). **(c)** Moreover, for each HITS-CLIP associated mRNA we calculated the average abundance of the miRNAs predicted to target that mRNA, and related this average miRNA abundance to the cell type enrichment. Interestingly, cell-type enriched genes are less likely targeted by high abundance miRNAs than genes not enriched in any cell type, suggesting that mRNAs coding for cell-type-specific genes escape suppression by highly abundant miRNAs. Plot representing cellular specificity of HITS-CLIP target mRNA (see Fig. 4.2d) and average abundance of targeting miRNAs.



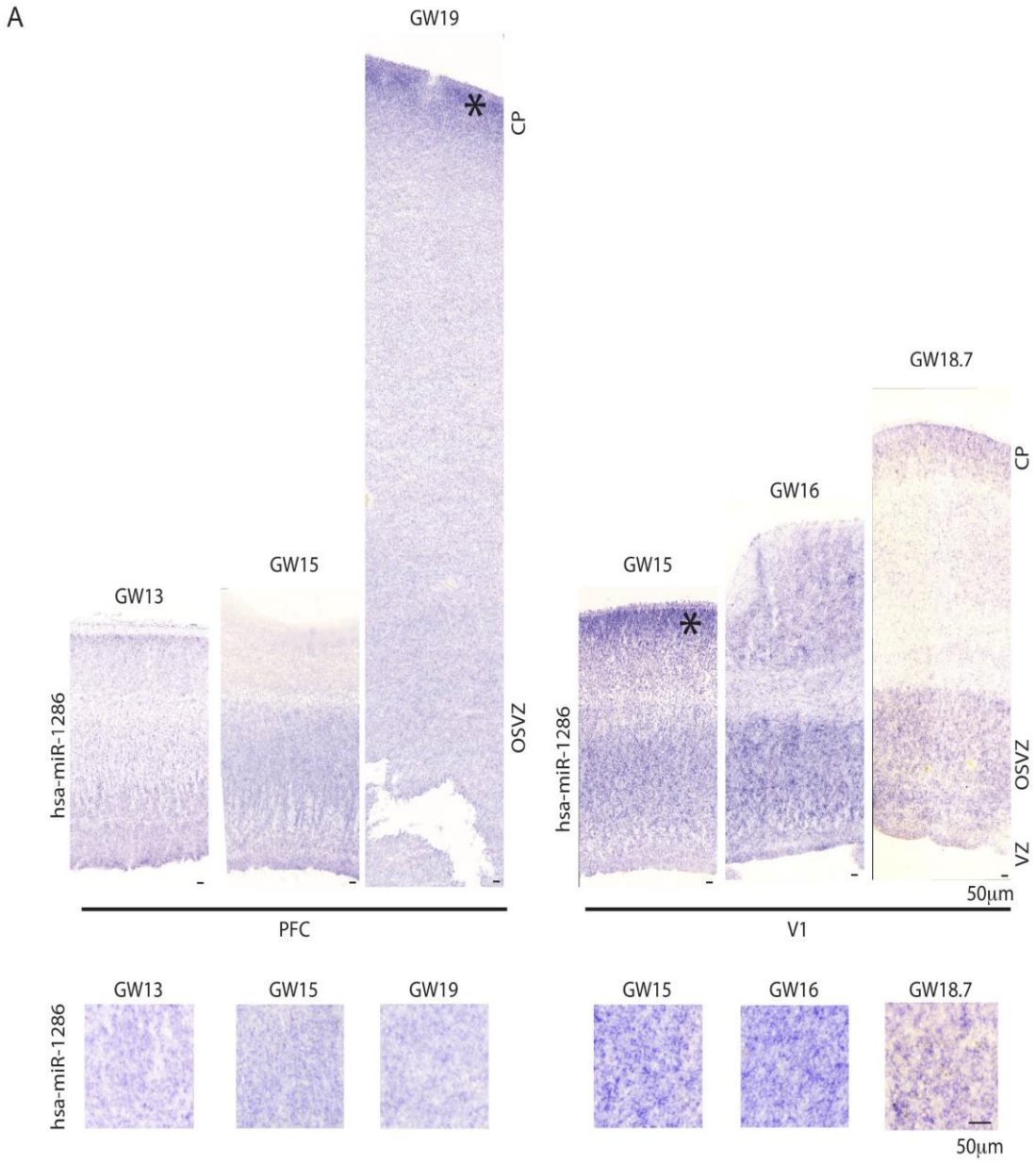
Supplementary Figure. 4.8. Expression of miR-2115 is enriched in late cortical development. In situ hybridization for miR-2115 in human cortex reveals increasing expression levels during development, reflecting sequencing results. Bottom panels, also shown in Fig. 4.3, show magnified view of the cortical OSVZ.



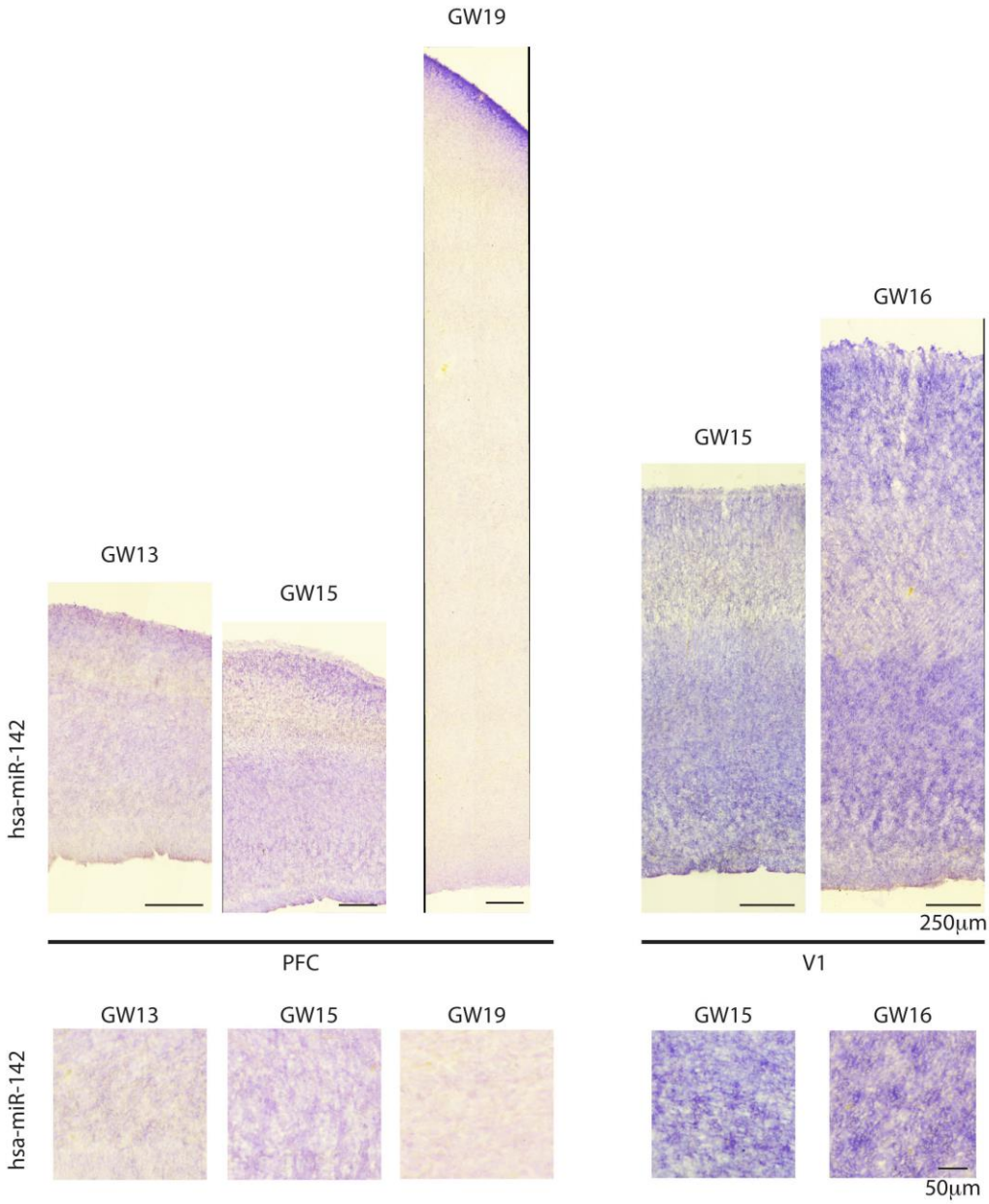
Supplementary Figure. 4.9. Expression of miR-362 is enriched in early cortical development. In situ hybridization for miR-362 in human cortex reveals declining expression levels during development, reflecting sequencing results. Bottom panels, also shown in Fig. 4.3, show magnified view of the cortical OSVZ.


A
Humans
Chimpanzees
Gorillas
Orangutans *
Gibbons
mir-2115
*mir-2115
B
Gibbon-Hominidae Common Ancestor ~2700 bp
~7100 bp Insertion
Hominidae Common Ancestor ~9900 bp
*miR-2115 100 bp
1832 bp Insertion
Great Apes Common Ancestor ~9500 bp
367 bp Deletion
Orangutan 12596 bp
miR-2115 100 bp
*miR-2115 100 bp
C
Sample
GW15M1
GW15V1
GW15NOI
GW16.5V1
GW19M1
GW19.5M1
GW19.5PFC
GW20.5PFC
expression level
row min
row max
miRNA_star
miRNA
hsa-mir-103a-1
hsa-mir-107
hsa-mir-548v
hsa-mir-592
hsa-mir-5682
ID00726_5p
ID00363_3p
ID02736_5p
ID00400_5p
ID01489_3p
ID01678_5p
ID00832_5p
ID03441_3p
ID03111_5p
ID01780_5p
ID02379_3p
ID02791_5p
ID01268_3p
ID03050_3p
ID02594_3p
ID01025_5p
ID01270_3p
ID00055_5p
ID00745_5p
ID02931_5p
ID02376_3p
ID02454_5p
ID01859_5p
ID02367_5p
ID02616_5p
ID02026_5p
ID00638_5p
ID02488_3p
ID02855_5p
ID00726_3p
ID02373_5p
ID00967_5p
ID00194_3p
ID02094_3p
ID00008_3p
ID03264_3p

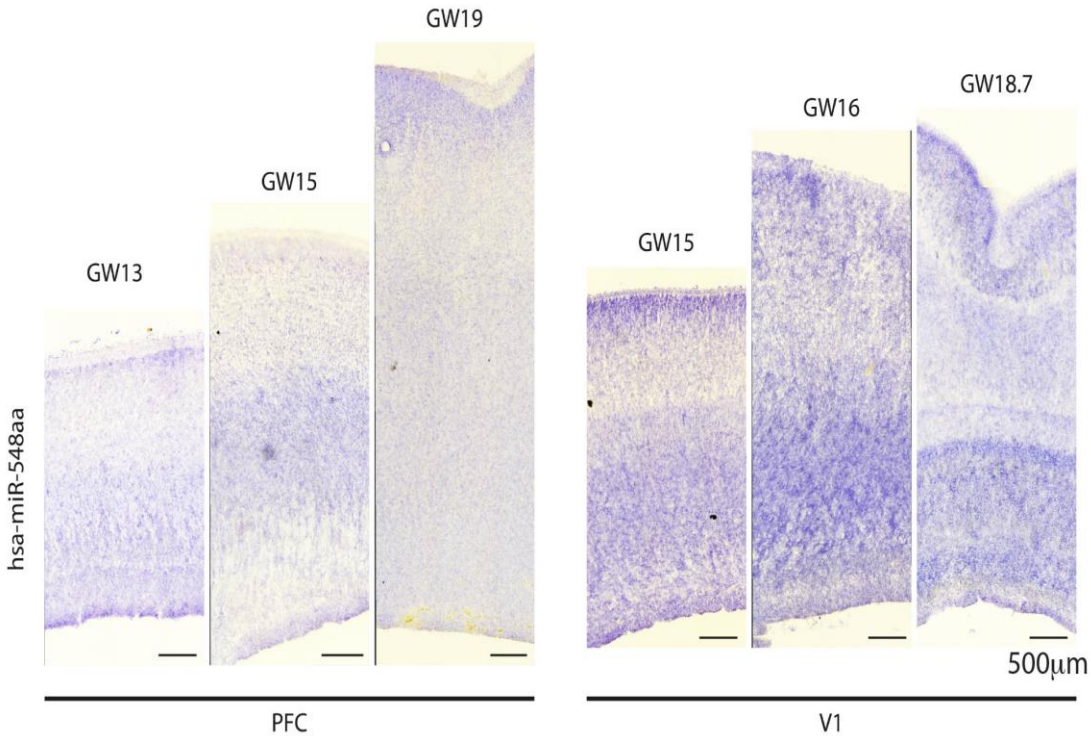
Supplementary Figure. 4.10. Evolution of miRNA-2115 in the primate lineage. (a) Phylogeny tree of upper primates (Bonobo excluded due to poor read sequence quality in region of interest). Star represents presence of miR-2115. I) Alignment of SPINK8 intron 3 - Grey represents agreement to consensus. Black represents disagreement to consensus. Lines indicate gaps compared to consensus. II) Sequence of miR-2115 in higher primates. Highlighting indicates disagreement to consensus. miR-2115 is predicted to be present in Orangutans but is absent in Gibbons. **b)** Intron 3 is significantly smaller in Gibbons than higher primates (~¼ the length) and more closely matches Intron 3 of mice. There is a ~7100 nucleotide section present in Hominidae but absent in Gibbons, likely arising from an insertion occurring in Hominidae after the divergence of Gibbons. This insertion likely carried miR-2115. There is a 367 nucleotide section directly downstream of miR-2115 present in Orangutans but not Great Apes, indicating it was lost in the Great Apes common ancestor after the divergence of Orangutans. There is a 1832 nucleotide section present only in Orangutans occurring after their divergence from the Great Apes. The great mobility in this intron is likely not due to transposable elements judging from the lack of inverted repeats in the area. **(c)** Expression of novel miRNAs that were not annotated in latest miRBase [36] across the samples. In this study, we found 36 of them expressed in prenatal brain samples. Of the expressed miRNAs, 31 of them were either human-specific or primate-specific, and 4 of them were specific expressed at GW15.



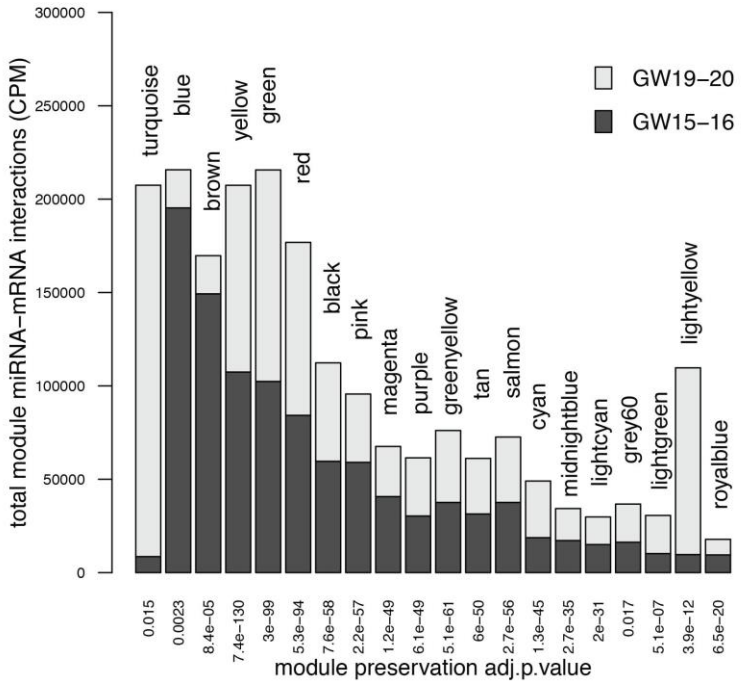
Supplementary Figure. 4.11. Expression of miR-1286 is enriched in human occipital cortex. In situ hybridization for miR-1286 in human prefrontal (PFC) and visual cortex (V1) reveals strong signal in the occipital cortex. Bottom panels show magnified view of the cortical OSVZ.



Supplementary Figure. 4.12. Expression of miR-142 is enriched in human occipital cortex. In situ hybridization for miR-142 in human prefrontal (PFC) and visual cortex (V1) reveals strong signal in the occipital cortex. Bottom panels show magnified view of the cortical OSVZ.



Supplementary Figure. 4.13. Expression of miR-548aa enriched in human occipital cortex. In situ hybridization for miR-548aa in human prefrontal and visual cortex reveals strong signal in the occipital cortex.



Supplementary Figure. 4.14. Bipartite network modules are preserved between GW15-16 samples and GW19-20. Module preservation statistics shown on x-axis suggest a significant preservation of modules (modules with higher interaction levels show higher preservation as well). Modules turquoise and blue are the two least preserved modules, suggesting developmental stage-specific changes. GW15-16 module names are used to compare GW15-16 modules with their homologues in GW19-20 (see Methods for details). Module assignments are listed in Table S4.6.

Supplementary Table S4.1: sc-qPCR primer information

Supplementary Table S4.2: sc-qPCR data of miRNA and mRNA

Supplementary Table S4.3: Sample information, RNA-seq and Mapping information

Supplementary Table S4.4: AGO2 Bound Clusters and miRNAs identified by HITS-CLIP

Supplementary Table S4.5: Target site and miRNA Cloning

Supplementary Table S4.6: Gene Ontology Enrichment Analysis

Supplementary Table S4.7: Cell-type-specificity scores using ideal vector correlation from published scRNA-seq data

Supplementary Table S4.8: Modules obtained from Bipartite Network Analysis of all the miRNA-mRNA interactions identified by HITS-CLIP

Supplementary Table S4.9: miR-137 Targets in Prenatal and Adult Brain Tissue Identified by HITS-CLIP

Supplementary Table S4.10: Differentially expressed miRNAs between two stages of brain development identified by DESeq2.

Supplementary Table S4.11: Analysis of targets of recently evolved miRNAs [36].

Supplementary Table S4.12: Bipartite Network Analysis of miRNA-mRNA interactions identified independently at GW15-16 and GW19-20 stages.

References:

1. A. A. Pollen *et al.*, Molecular identity of human outer radial glia during cortical development. *Cell* **163**, 55 (Sep 24, 2015).
2. M. J. Moore *et al.*, Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. *Nature protocols* **9**, 263 (Feb, 2014).
3. R. L. Boudreau *et al.*, Transcriptome-wide discovery of microRNA binding sites in human brain. *Neuron* **81**, 294 (Jan 22, 2014).
4. N. Rani *et al.*, A Primate lncRNA Mediates Notch Signaling during Neuronal Development by Sequestering miRNA. *Neuron* **90**, 1174 (Jun 15, 2016).
5. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, pp. 10 (2011).

6. M. R. Friedländer, S. D. Mackowiak, N. Li, W. Chen, N. Rajewsky, miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic acids research* **40**, 37 (2011).

7. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550 (2014).

8. P. J. Uren *et al.*, Site identification in high-throughput RNA–protein interaction data. *Bioinformatics* **28**, 3013 (2012).

9. S. W. Chi, J. B. Zang, A. Mele, R. B. Darnell, Ago HITS-CLIP decodes miRNA-mRNA interaction maps. *Nature* **460**, 479 (2009).

10. B. P. Lewis, C. B. Burge, D. P. Bartel, Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15 (2005).

11. K. Shekhar *et al.*, Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell* **166**, 1308 (Aug 25, 2016).

12. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**, 559 (Dec 29, 2008).

13. T. J. Nowakowski, A. A. Pollen, C. Sandoval-Espinosa, A. R. Kriegstein, Transformation of the Radial Glia Scaffold Demarcates Two Stages of Human Cerebral Cortex Development. *Neuron* **91**, 1219 (Sep 21, 2016).

14. A. A. Pollen *et al.*, Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature biotechnology* **32**, 1053 (Oct, 2014).

15. T. J. Nowakowski *et al.*, MicroRNA-92b regulates the development of intermediate cortical progenitors in embryonic mouse brain. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 7056 (Apr 23, 2013).

16. T. Saito, In vivo electroporation in the embryonic mouse central nervous system. *Nature protocols* **1**, 1552 (2006).

17. T. Takahashi, R. S. Nowakowski, V. S. Caviness, Jr., The cell cycle of the pseudostratified ventricular epithelium of the embryonic murine cerebral wall. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **15**, 6046 (Sep, 1995).

18. S. J. Liu *et al.*, Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome biology* **17**, 67 (Apr 14, 2016).

19. M. J. Barber, Modularity and community detection in bipartite networks. *Physical Review E* **76**, 066102 (2007).

20. X. Liu, T. Murata, Community detection in large-scale bipartite networks. *Information and Media Technologies* **5**, 184 (2010).

21. C. F. Dormann, B. Gruber, J. Fründ, Introducing the bipartite package: analysing ecological networks. *interaction* **1**, 0.2413793 (2008).

22. A. Seal, D. J. Wild, Netpredictor: R and Shiny package to perform Drug-Target Bipartite network analysis and prediction of missing links. *bioRxiv*, 080036 (2016).

23. J. M. Gohlke *et al.*, Characterization of the proneural gene regulatory network during mouse telencephalon development. *BMC biology* **6**, 15 (Mar 31, 2008).

24. P. Shu *et al.*, MicroRNA-214 modulates neural progenitor cell differentiation by targeting Quaking during cerebral cortex development. *Scientific reports* **7**, 8014 (Aug 14, 2017).

25. J. Aprea *et al.*, Transcriptome sequencing during mouse brain development identifies long non-coding RNAs functionally involved in neurogenic commitment. *The EMBO journal* **32**, 3145 (Dec 11, 2013).

26. D. V. Hansen, J. H. Lui, P. R. Parker, A. R. Kriegstein, Neurogenic radial glia in the outer subventricular zone of human neocortex. *Nature* **464**, 554 (Mar 25, 2010).

27. P. Arlotta *et al.*, Neuronal subtype-specific genes that control corticospinal motor neuron development in vivo. *Neuron* **45**, 207 (Jan 20, 2005).

28. B. Cubelos *et al.*, Cux1 and Cux2 regulate dendritic branching, spine morphology, and synapses of the upper layer neurons of the cortex. *Neuron* **66**, 523 (May 27, 2010).

29. D. V. Hansen *et al.*, Non-epithelial stem cells and cortical interneuron production in the human ganglionic eminences. *Nature neuroscience* **16**, 1576 (Nov, 2013).

30. N. Takada, S. Kucenas, B. Appel, Sox10 is necessary for oligodendrocyte survival following axon wrapping. *Glia* **58**, 996 (Jun, 2010).

31. H. Fu, M. Qiu, Migration and differentiation of Nkx-2.2+ oligodendrocyte progenitors in embryonic chicken retina. *Brain research. Developmental brain research* **129**, 115 (Jul 23, 2001).

32. B. Bonev, P. Stanley, N. Papalopulu, MicroRNA-9 Modulates Hes1 ultradian oscillations by forming a double-negative feedback loop. *Cell reports* **2**, 10 (Jul 26, 2012).

33. W. Andrews *et al.*, The role of Slit-Robo signaling in the generation, migration and morphological differentiation of cortical interneurons. *Developmental biology* **313**, 648 (Jan 15, 2008).

34. Y. E. Wu, N. N. Parikshak, T. G. Belgard, D. H. Geschwind, Genome-wide, integrative analysis implicates microRNA dysregulation in autism spectrum disorder. *Nature neuroscience* **19**, 1463 (Nov, 2016).

35. S. Garg, P. A. Sharp, GENE EXPRESSION. Single-cell variability guided by microRNAs. *Science* **352**, 1390 (Jun 17, 2016).

36. E. Londin *et al.*, Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. *Proceedings of the National Academy of Sciences of the United States of America* **112**, E1106 (Mar 10, 2015).