Genetic Mechanisms of Regulated Stochastic Gene Expression

Adan Horta

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

ABSTRACT

Genetic Mechanisms of Regulated Stochastic Gene Expression

Adan Horta

The adaptability and robustness of the central nervous system is partially explained by the vast diversity of neuronal identities. Molecular mechanisms generating such heterogeneity have evolved through multiple independent pathways. The olfactory sensory system provides a unique and tractable platform for investigating at least two orthogonal gene expression systems that generate neuronal diversity through stochastic promoter choice: olfactory receptor genes and clustered protocadherins. Olfactory sensory neuron identity is defined by the specific olfactory receptor (OR) gene chosen. Greater than 1300 OR genes are scattered throughout the mouse genome, and expression of an OR defines a unique sensory neuron class that responds to a selective set of odorants. This work further delineated an unprecedented network interchromosomal (*trans*) interactions indispensable for singular OR choice. In a largely orthogonal gene expression system, I sought to understand the molecular mechanisms governing stochastic protocadherin choice. Clustered protocadherins are an evolutionary-conserved system that are involved in cell-cell identification through a series of homo- and heterophilic interactions. This work uncovered a methylation-dependent mechanism for generating stochastic gene expression in the context of *cis* regulatory elements. Overall, this work highlighted divergent *cis* and *trans* transcriptional regulatory mechanisms for generating stochastic gene expression and neuronal diversity.

Table of Contents

Chapter 11:

Appendix I:

Appendix II:

List of Figures

Acknowledgements

Stavros Lomvardas was a spectacular mentor to me. He has a brilliant ability of knowing when and how to motivate people, a key skill amongst great leaders. I first met Stavros while participating in a grant writing course that he led. I was immediately drawn by Stavros' clear-cut logic, far-reaching inferences, and genuine fondness for adventurous scientific endeavors. I saw in Stavros everything I wanted out of graduate school. I was able to convince Stavros to give me a chance in his lab; albeit on "probation" until I could show I was worth my weight in pipette tips. I Immediately felt at home in the Lomvardas Lab. This was certainly fueled by Stavros mentorship style. Stavros would set a bar and *expect* me to meet it. As soon as I got close, he would raise the bar of expectations twice as high as before. This simple yet sophisticated mentorship style motivated me to devour milestones with great confidence and enthusiasm. Once met, I was left with results and most importantly, the confidence to reach bars higher than before. I started 3 different projects that first year. Remarkably, they all yielded positive results. However, the most promising result, and my first ceremonious quarter[1] came in January of 2017, merely 8 months after starting. Using *in situ* HiC, we showed high frequency interchromosomal interactions in mature olfactory sensory neurons, a result that led to the first *Nature* paper for both of us. With a predilection for scientific theatrics, I was sure to save this result for lab meeting. In response, Stavros nearly fell out of his chair with an expletive cry or two. Mission Accomplished.

[1] Officially, Stavros commends scientific breakthroughs with a quarter. I made 50 cents in my PhD.

Enrico Cannavo and **Daniele Canzio** have been my Italian family a few thousand miles from the Tyrrhenian Sea. Note: I am not Italian, but after a few years in the lab, I have a fondness for the *madrepatria*. Late-night science talks with Enrico, kept me excited and up to date on all the exciting discoveries within the scientific community. In addition, he provided guidance for optimizing my experiments that greatly accelerated the pace of my work. I met Daniele later in my PhD. With protocadherins on his mind and nuclear architecture up my sleeve, we quickly made up for lost time. Daniele provided me with historical appreciation for the giants' shoulders we were standing on. His perspective kept me humble. Our work together led to a co-first authorship in *Cell*—a productive friendship indeed! Overall, these two postdocs have contributed greatly to my growth as a scientist and I have become close friends with them. I wish them the very best.

I would also like to thank my parents, **Raquel Uribe** and **Arturo Horta**, and my sisters, **Christina** and **Maryanna**. When I was in elementary school, he bought our family a Packard Bell with Windows 3.11, a testament to my parents' sacrifices in order to further my chances at a successful career. During middle school, we moved to a town with a better school system, despite not having enough bedrooms. My parents lived in the basement for a few years before my dad committed his entire pension to build an entire second story on our house… with his own hands. My father offered me restricted freedom and high expectations as a child that challenged me to think for myself and demand the best. When I was a teenager, he showed me how to drive a stick shift around the neighborhood for less than 30 minutes. To be clear: this wasn't enough time. The following Monday was my first day working in the lab of Bruce Carter, my undergraduate research mentor. During the 30 minutes of bumper-to-bumper traffic on the way home, I learned, at the expense of an annoyed, but then understanding truck driver how to shift out of first gear. I believe that these real-world challenges have made me more resilient and naively courageous in graduate school. My mother on the other hand provided me with care, unconditional love, and a knack for determination. She took great care to teach me Spanish and English, multiplication, reading, and tying my shoes well before I started kindergarten. This gave me a head start in elementary school that has kept me afloat all of these years. My sister, Christina has also been an inspiration to me. Despite not getting the scholarship she needed to go to her dream school for college, she washed down reality with great class. Working hard at the state college, she graduated with honors and got her master's degree in teaching, this time at her dream school. Last, but not least, my youngest sister Maryanna is a champion of kind spirit and determination. I had the privilege of watching her grow since childhood and I can confidently say I have never met a more charitable, tireless, and diligent person. As she moves forward in her life, I am excited to see where she will shine her light. Even when I forget to phone home, I am always appreciative of everything my family has done for me. Thank you everyone.

Adan

As Alexander III of Macedonia:

Τῷ ξίφει τὸν δεσμόν λελύσθαι

“Solve the knot with the sword”

Chapter 1:

Stochastic Gene Regulation

Cellular differentiation is the process by which the fate of a cell is specified by genetic, epigenetic, environmental, and stochastic elements (Goldberg et al., 2007). Much of this program revolves around the genome, and has been conceptualized by Waddington in his Epigenetic Landscape of cellular differentiation (Waddington, 1957) (Fig. 1A). In this model, the fate of a cell is progressively determined through a set of sequential transitions, metaphorically represented as a marble rolling down a hill. Much like rolling a marble down a hill, differentiation is influenced by both deterministic and probabilistic forces. Given the genome's presumably finite instructive role in specifying diverse cell types, a key question in developmental neuroscience emerges: How does the genome generate close to 100 billion unique neurons with merely 20 thousand genes?

Key insights into the emergence of cellular diversity stem from mathematical and experimental revelations that transcription and translation are both inherently and systematically stochastic (Chalancon et al., 2012; Raj and van Oudenaarden, 2008). From this regard, many studies over the years have focused on intrinsic and extrinsic contributions to "noise" at the level of transcripts, proteins, and gene regulatory networks (Elowitz et al., 2002; Jothi et al., 2009; Ozbudak et al., 2002). In essence, transcription and translation can be affected extrinsically by availability of RNA polymerases, transcription factors, tRNA, etc., and by the intrinsic biophysical randomness of enzymatic reactions, such as transcriptional bursting (Fukaya et al., 2016). These variabilities in gene expression can greatly impact the cellular fate of a particular cell, especially given development's strong reliance on protein expression gradients (Jessell, 2000; Spemann and Mangold, 1924).

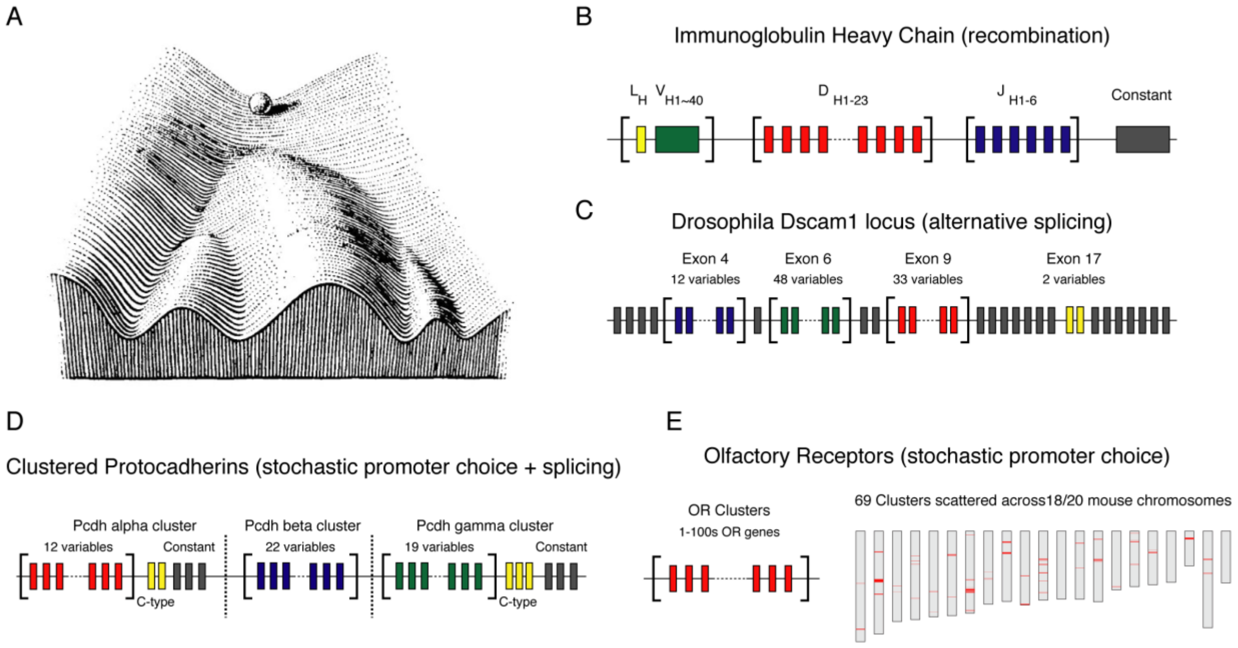

A
B
Immunoglobulin Heavy Chain (recombination)
L H
V H1~40
D H1-23
J H1-6
Constant
C
Drosophila Dscam1 locus (alternative splicing)
Exon 4
12 variables
Exon 6
48 variables
Exon 9
33 variables
Exon 17
2 variables
D
Clustered Protocadherins (stochastic promoter choice + splicing)
Pcdh alpha cluster
12 variables
Constant
C-type
Pcdh beta cluster
22 variables
Pcdh gamma cluster
19 variables
Constant
C-type
E
Olfactory Receptors (stochastic promoter choice)
OR Clusters
1-100s OR genes
69 Clusters scattered across18/20 mouse chromosomes

Figure 1. Stochasticity in gene expression. A. Waddington's Epigenetic Landscape depicts cellular differentiation as a marble rolling down a hill. **B.** The immunoglobulin heavy chain produces thousands of combinations through recombination for the generation of antibodies. **C.** The Drosophila Dscam1 gene is alternatively spliced to generate unique molecular barcodes used in neuronal network assembly. **D.** Clustered protocadherins are cell adhesion molecules that are stochastically chosen for functional self-avoidance in mammalian circuit assembly. **E.** Olfactory receptor genes are stochastically expressed and arranged in arrays scattered throughout the mammalian genome.

In addition to variability in the relative expression levels of genes, evolution has also provided programmed stochastic gene expression (Johnston and Desplan, 2008). In this paradigm, the cell "chooses" a gene or variant amongst an array of "choices", often in a mutually exclusive manner. These stochastic systems appear to be most prevalent in the nervous system and in the immune system, where adaptability and diversity are of the upmost importance. For example, the characteristic arrangement of V(D)J segments in the genome allow for the generation of thousands of unique T cell receptors (TCRs) and immunoglobulins through a process of stochastic recombination (Jain et al., 2018) (Fig. 1B) . In Drosophila, the Dscam locus evolved for the generation of > 38,000 unique axon guidance molecules, indispensable for the formation of neuronal circuits (Schmucker et al., 2000) (Fig. 1C). Interestingly, the Dscam molecules also contain immunoglobulin-like domains and are cell surface recognition molecules, but rely on alternative splicing rather than recombination to generate stochasticity. Somewhat analogous to the Dscam molecules, the clustered protocadherins have evolved to generate single cell diversity in order to facilitate self-avoidance in neurons (Chen and Maniatis, 2013) (Fig. 1D). Unlike TCRs and Dscam, the clustered protocadherins appear to rely on a stochastic promoter choice (Tasic et al., 2002). Still, V(D)J recombination, Dscam alternative splicing, and clustered protocadherin promoter choice promote cellular diversity, albeit through highly divergent mechanisms. Moreover, the genes encoding these receptors are all contained within a single genomic locus, suggesting a heavy reliance on *cis* (or intrachromosomal) gene regulatory mechanisms.

Perhaps the most striking example of stochastic gene expression is olfactory receptor (OR) choice. Like the stochastic gene expression systems aforementioned, the ORs are arranged in arrays that facilitate choice. However, in mouse, the complete OR repertoire is arranged into > 60 arrays of varying size (Fig. 1E). These OR clusters are scattered throughout the genome in a seemingly random fashion. Thus, the possibility of *trans* (or interchromosomal) regulatory mechanisms becomes enticing (Lomvardas et al., 2006).

In this study, I focused on understanding the genomic mechanisms of stochastic gene expression in both olfactory receptor genes and clustered protocadherins. Using cutting-edge chromatin conformation studies and an *in vivo* neuronal differentiation system, we furthered the mechanisms regulating stochastic promoter choice.

Chapter 2:

Nuclear Architecture

The steady flow of scientific discoveries is periodically disrupted in bounds and leaps by technological advances. Although the three-dimensional organization of DNA (chromatin architecture) is a major determinant of cellular fate and function, early studies were limited in scope by available techniques. For example, long-range enhancer-promoter chromatin interactions had been hypothesized for quite some time (Goodbourn et al., 1985; Maniatis et al., 1987), but showing these interactions in their native state remained challenging.

The development of chromosome conformation capture (3C) by Job Dekker was driven by a passion for the structure of mitotic chromosomes (Fig. 2A) (Dekker, 2002; O'Donnell, 2016). In this technique the structure of a chromosome (or nuclear chromatin) can be inferred by the relative frequency of contacts between two genomic loci. Specifically, chromatin is fixed, digested with a restriction enzyme, then the DNA fragments ligated back together. Because fixed protein holds the chromatin in its native conformation, proximity-based ligation occurs between three-dimensional DNA partners. This probabilistic ligation junction can then be assessed by quantitative electrophoresis, RT-qPCR, or next generation sequencing (HiC – high-throughput chromosome conformation capture (Lieberman-Aiden et al., 2009)). With the cost of sequencing plummeting since 2008 and the formation of the NIH 4D Nucleome Consortium aimed at understanding the structure of the nucleus, HiC has become a widely used technology that has rapidly expanded our understanding of nuclear architecture (Fig 3B,C).

Conceptually, *in situ* HiC is similar to 3C except that ligation occurs in intact nuclei and libraries are sequenced deeply to assess the three-dimensional structure of the whole genome (Rao et al., 2014). From this field, two major subnuclear structures have emerged (Fig. 3,4).

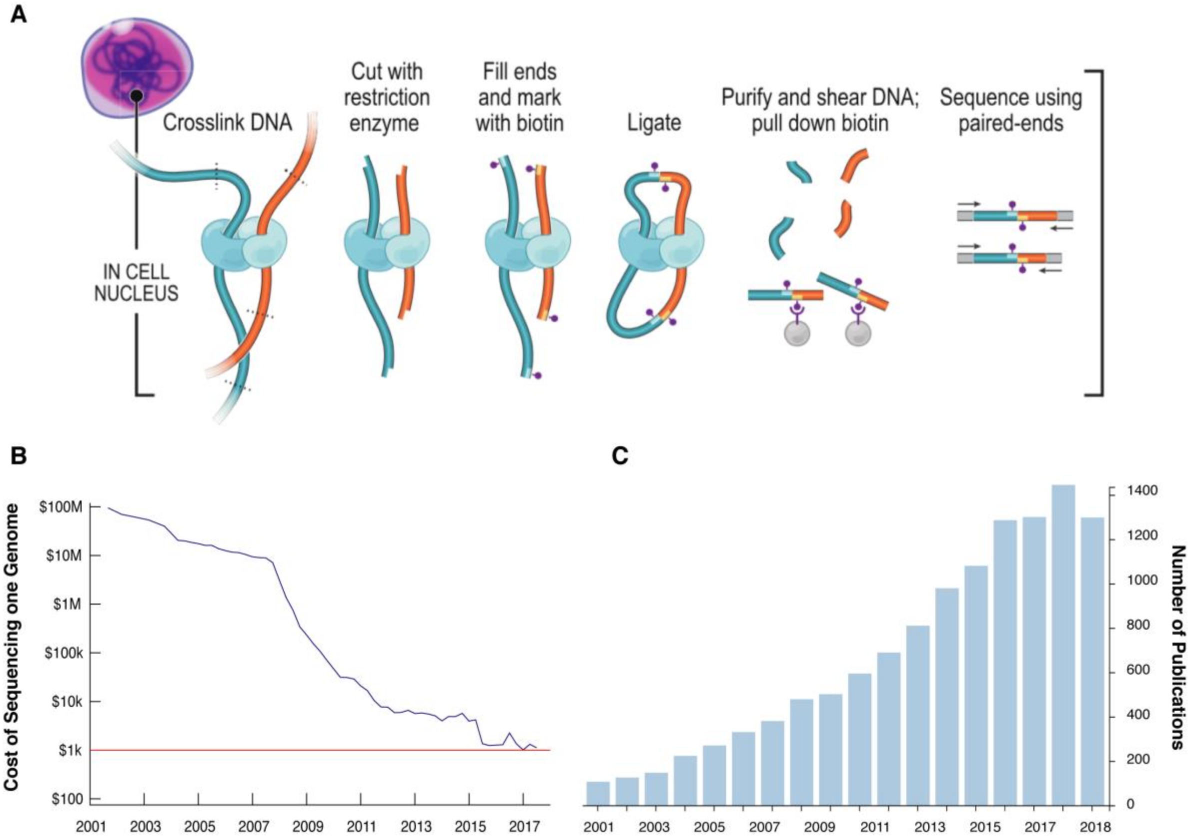


Figure 2. *in situ* HiC and low cost of sequencing have driven genome wide nuclear architecture studies. A. *in situ* HiC protocol. Adapted from Rao et al., 2014. **B.** Cost of sequencing a human genome has fallen much faster than Moore's law would have predicted. Source: NIH. **C.** Exponential increase in publications on the topic of "Genomic Architecture". Source: Web of Science.

The first structure observed, compartments, we know the least about. These compartments were inferred from the increased association frequency between kilobase-sized chromatin domains. Originally, it was believed that the genome was partitioned into two compartments: active (A) and inactive (B) (Lieberman-Aiden et al., 2009). The active compartment contains genes that are actively transcribed, replicated early during S phase of mitosis (Pope et al., 2014), and associated with open chromatin histone marks (Rao et al., 2014). The B compartment, on the other hand, is the foil of the active compartment: heterochromatic, replicated later during S phase, and associated with the repressive nuclear lamina (Dixon et al., 2012). While useful, this simplistic view of compartmentalization has been challenged in recent years (Rao et al., 2014). The general trends hold true (active vs. inactive), but compartments are more complex and nuanced. Our findings suggest that compartments are multiple, dynamically regulated, and cell type-specific (see below for Monahan, Horta, Lomvardas. 2018).

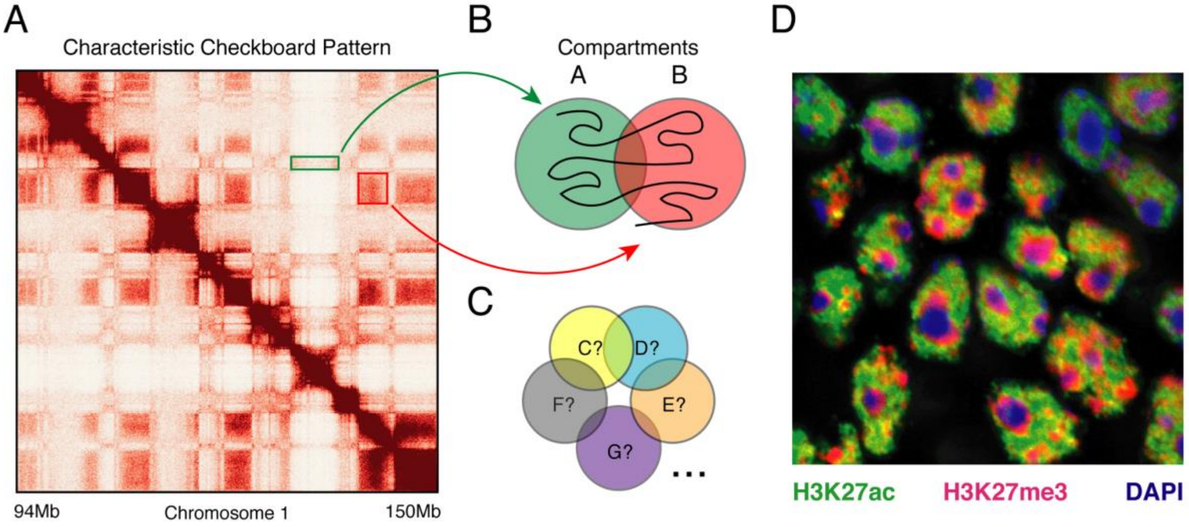


Figure 3. The genome is partitioned into compartments. A. HiC matrix from horizontal basal cells of the olfactory epithelium showing distinct compartments in a locus of chromosome 1. 100kb resolution. In general, heterochromatin always shows increased contact frequency, suggesting it is more tightly packaged at the level of nucleosomes. **B.** Model conceptualizing compartments. **C.** Up to 6 compartments have been described (See Rao et al. 2014), but there are likely more in mature cell states. **D.** Immunofluorescence staining in the main olfactory epithelium showing stark borders between physical “compartments”, suggesting the existence of multiple phases.

Considering the notion that compartmental identities are associated with histone marks, which are specifically bound by proteins, an attractive hypothesis is that the biophysical correlates of HiC compartments are phase separated membrane-less organelles (Boeynaems et al., 2018). Recent work revealed that heterochromatin is functionally and biochemically isolated from the rest of the nucleus through phase separation (Larson et al., 2017; Strom et al., 2017). In these studies, heterochromatin protein 1 was shown to form a gel-like matrix that confers discreet physical properties from the surrounding microenvironment. Even more recently, transcription factors were shown to activate genes through a phase separation mechanism using intrinsically disordered regions (IDR) (Boija et al., 2018). Briefly, Oct4 and other transcription factors with trans activation domains containing IDRs phase separate with mediator *in* *vitro* in a context dependent manner. Moreover, these transactivation domains are indispensable for transcriptional activation. Thus, phase separation may be an important mechanism for the functional sequestration for heterochromatin, euchromatin, and their respective subtypes (Plys et al., 2018). Biophysically, it appears that the formation of protein lattices through the use of unstructured protein domains may be critical for the formation of a phase. Ultimately, the sheer number of transcription factors and nuclear proteins with IDRs, and the requisite nuclear functions (e.g. nucleolus, DNA repair, splicing, replication, etc.) suggest phase separation may be a widespread phenomenon affecting every domain in the genome.

The second major structural feature described by the “HiC field” are loop domains (Dixon et al., 2012). At first conceptualized as topologically associated domains (TADs), deeper sequencing has revealed that loop domains are distinct from ordinary domains that lack a focal “loop” contact (Fig. 4A). In fact, these ordinary domains are likely the building blocks of compartments. The biophysical mechanism of loop domains has been extensively characterized by several groups (Guo et al., 2015; Kagey et al., 2010; Nora et al., 2017; Rao et al., 2017a; Schwarzer et al., 2017; Wendt et al., 2008).

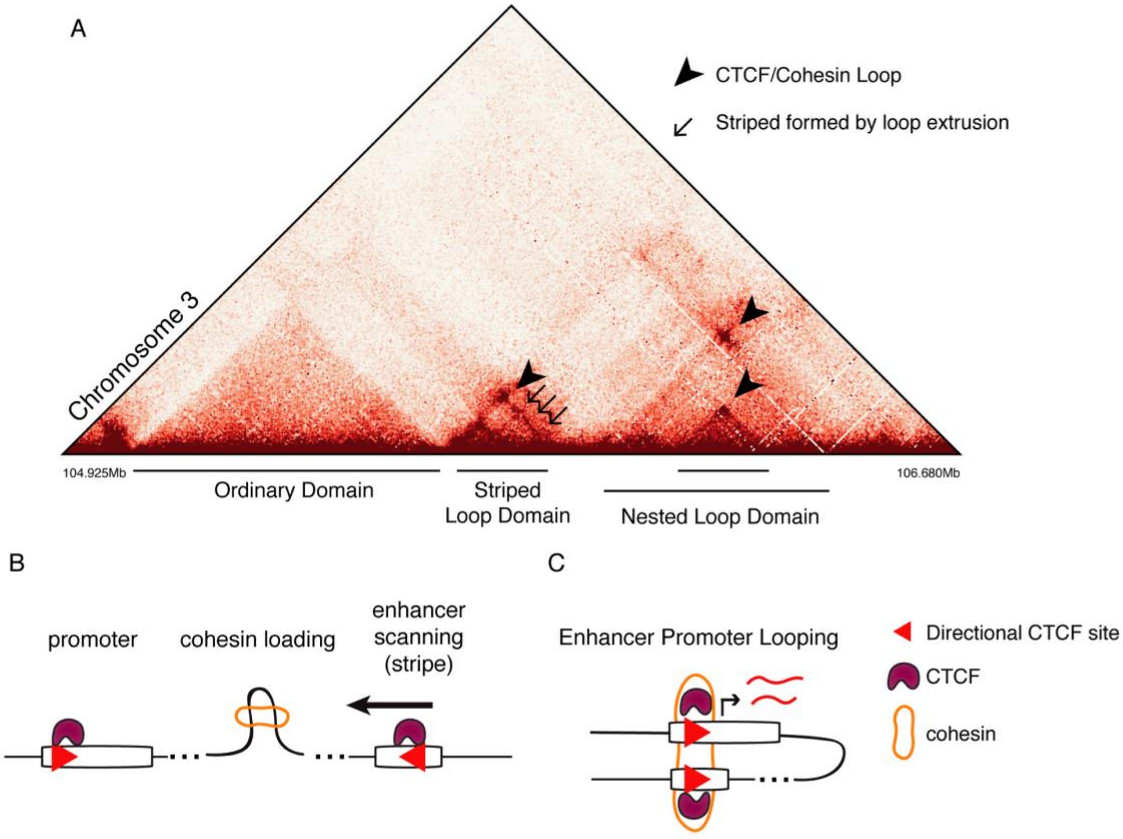


Figure 4. Loop domains are CTCF/Cohesin loops. A. HiC from sorted olfactory sensory neurons showing exemplary cohesin/CTCF loops and stripes. **B.** Mechanistic diagram for cohesin loading between convergent CTCF sites. **C.** Loop extrusion leads to CTCF-CTCF loops by cohesin. In this example, between a promoter and enhancer element.

In summary, loop domains are composed of CTCF and cohesin mediated loops. CTCF is an 11-zinc finger protein that functions as a genomic insulator. That is, it "shields" transcriptional enhancers from activating promoters erroneously. In fact, activating gene expression by perturbing these insulating borders can lead to developmental disorders, including malformation syndromes (Lupiáñez et al., 2015). Moreover, CTCF binding and insulation is directional. This phenomenon is dependent on the orientation of the CTCF binding site on the DNA (Guo et al., 2015).

While CTCF is the border of loop domains, cohesin is the motor protein that functionally realizes these loop domains (Rao et al., 2017b). Cohesin, and also condensin, possess the ability to "extrude" DNA through their loop (Fig. 4B) (Ganji et al., 2018). This ATP-dependent cohesin loop is loaded by Nipbl at super enhancer regions and extrudes through the DNA until it reaches a correctly-oriented CTCF insulator (Fig. 4C) (Schwarzer et al., 2017; Vian et al., 2018). With some frequency, cohesin can pass over a CTCF insulator, but it is normally removed from the DNA by WAPL beforehand (Haarhuis et al., 2017). From these observations, we discover the key functions of four major proteins in formation of loop domains: CTCF is the barrier, cohesin is the ATPase motor, Nipbl is the cohesin loader, and WAPL is the cohesin unloader. Together, these findings of cohesin-mediated loop extrusion greatly motivated key insights into understanding protocadherin alpha promoter choice in this study (see below for Canzio, Nwakeze, Horta, et al. 2018).

Overall, the structure of the three-dimensional genome in time provides a new platform for understanding potential mechanisms for stochastic gene expression, and ultimately cellular diversity. In this study, I set out to understand the nuclear architecture in developing and mature olfactory sensory neurons and how it may contribute to stochastic choice of protocadherins, and olfactory receptors. Lightly speaking, it seems that the most important mechanisms for protocadherin choice involve loop extrusion, while OR choice involves the less charted territory of nuclear compartments.

Chapter 3:

Olfactory Receptor Choice

Although the general framework of the olfactory sensory system at the cellular level was understood since Cajal, it took nearly 100 years and another Nobel Prize in Physiology or Medicine to identify the molecular machines that detect odorants (Buck and Axel, 1991). In 1991 Linda Buck and Richard Axel cloned a multigene family of G-protein coupled receptors that mediate odorant detection (Fig. 2A). Over the ensuing years, the Axel lab and others went on to uncover much more about the biology of these olfactory receptors (ORs):

- ORs are stochastically expressed in a zonal fashion within the olfactory epithelium (Ngai et al., 1993; Vassar et al., 1993);
- only one OR allele is expressed per OSN (Chess et al., 1994);
- OSNs expressing the same OR project to a single glomerulus (Vassar et al., 1994);
- the OR plays an instructive role in the formation of the topographic map within the olfactory bulb (Mombaerts et al., 1996; Wang et al., 1998) (Fig. 2B).

That conditional deletion of the OR dissociates the topographic map for that OR and that lack of OR expression leads to developmental arrest underscores the importance of the OR for mature OSN identity (Lyons et al., 2013). Our lab has largely focused on the mechanisms underlying singular OR choice.

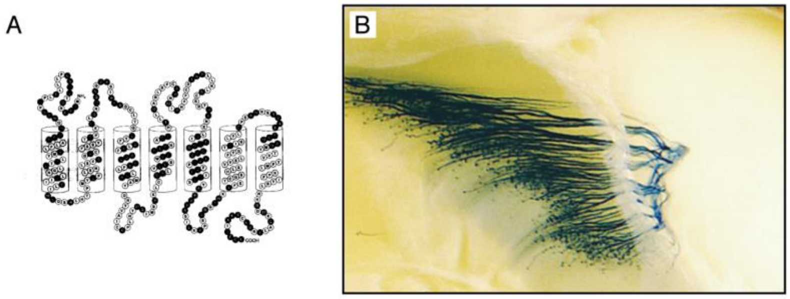


Figure 5. Olfactory Receptors dictate OSN identity. A. The OR superfamily is a seven transmembrane G protein-coupled receptor (GPCR). Frequently mutated residues are highlighted in black. Adapted from Buck and Axel. 1991. **B.** Cells expressing Olfr17 in the olfactory epithelium project to a single glomerulus in the olfactory bulb. Adapted from Wang, et al. 1998.

As discussed in Chapter 1, the complexity of singular OR choice is complicated by the seemingly random arrangement of OR genes across the genome (Fig. 1E). Early studies regarding OR choice focused on the minimum genetic requirements for participating in the competition of choice (Serizawa et al., 2000; Vassalli et al., 2002). By inserting a 2.2kb transgene containing an OR and flanking regions randomly into the genome, it was shown that this small genomic element was able to participate in OR choice and project to the olfactory bulb, in a manner indistinguishable from the homologous endogenous OR gene (Vassalli et al., 2002). This immediately suggests that requirements for participating in choice are contained within a small genomic region. It also suggests that the genome is indifferent to the genomic position of the OR, yet allows for repression of the non-chosen ORs (Serizawa et al., 2000). Perhaps the most striking discovery in OR choice was made by my mentor, Stavros Lomvardas. With a freshly-minted 3C technique (Dekker, 2002) and DNA fluorescent *in situ* hybridization (FISH), it was demonstrated that an OR enhancer, H (Serizawa et al., 2003), contacts various OR genes in *trans* (Lomvardas et al., 2006). This publication was almost immediately met with a legitimate challenge. By deleting the region encompassing the H enhancer, it was convincingly shown that H has mostly *cis* effects on OR choice (Fuss et al., 2007). This directly challenged the hypothesis that H regulates OR genes in *trans*[1].

Still, the finding that H associates with OR genes in *trans* was not refuted, and thus the full story remained at large. The first major solidification of *trans* interactions came from the use of a DNA FISH probe that specifically labeled the majority of the OR genes (Fig. 3B,C) . By labeling OR genes with this DNA FISH probe, it was shown that ORs aggregate in a developmentally regulated manner (Clowney et al., 2012). This aggregation is believed to be due to the heterochromatin patterning that occurs during early during the development of OSNs (Magklara et al., 2011), although the precise temporal dynamics remained unknown. These studies also underscored the importance of the nuclear architecture for the OSN. The unusual "fried-egg" pattern of the mature OSN nucleus, whereby heterochromatin collapses onto the middle of nucleus in a developmentally regulated fashion, is in fact, vital for OR choice and OSN identity (Fig 3A) (Le Gros et al., 2016). The appearance fried-egg nucleus is concomitant with the downregulation of lamin B receptor (Lbr) during differentiation. Overexpression of Lbr in mature OSNs perturbs this macroscopic organization of nuclear DNA and abolishes OR expression (Clowney et al., 2012). Thus, in some cryptic way, OR choice relied on macroscopic nuclear architecture and *trans* interactions.

[1] I am told that Richard Axel called Stavros on his very first day as Principal Investigator at UCSF with the news of the Matters Arising story. Talk about motivation!

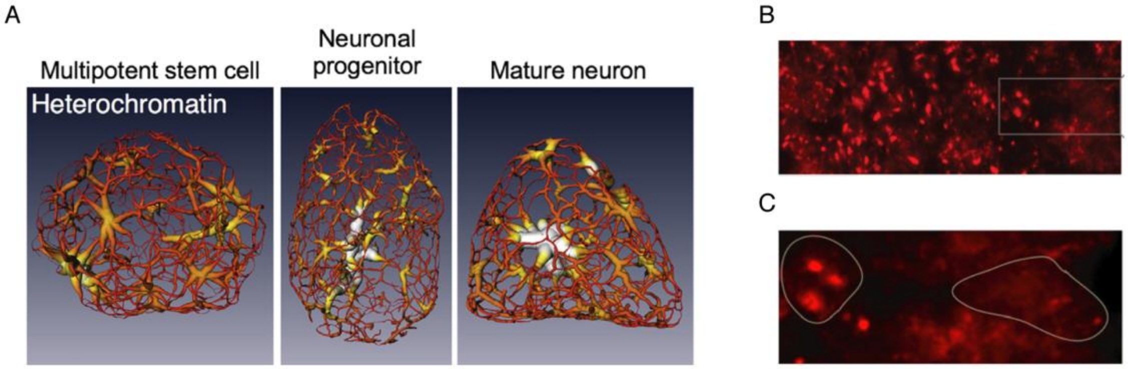


Figure 6. Mature OSNs have an unusual and characteristic "fried-egg" nucleus. A. Soft X-ray tomography reveals that heterochromatin in mature OSNs is concentrated in the center of the nucleus. Adapted from Le Gros et al. 2016. **B.** DNA FISH labeled all OR genes reveals puncta in the mature OSNs, but not other cells of the olfactory epithelium. **C.** Zoomed-in view of B with an outlined OSN (left), and an outlined non-OSN (right). Adapted from Clowney, et al. 2012.

While H was thought to be associating with OR genes in *trans*, the finding that H mostly regulated choice in *cis*, and the discovery of the P element enhancer regulating *cis* ORs, further complicated the models (Bozza et al., 2009; Khan et al., 2011). Clarity into these inconsistencies finally began to emerge with the discovery of more than 60 transcriptional enhancers, termed Greek Islands, scattered throughout the OR clusters. Using 4C and two-color DNA FISH, it was shown that these Greek Islands make frequent contacts with each other and with the active allele (Markenscoff-Papadimitriou et al., 2014). Thus, it was postulated that OR genes are regulated by the Greek Islands in a cooperative fashion. These Greek Islands are characterized by accessibility and binding of Lhx2 and Ebf in most mOSNs (Khan et al., 2011; Monahan et al., 2017). Interestingly, transcription factors regulating OR choice are bound to the Greek Islands at developmental stages before singular OR choice. Thus, this suggests that these enhancers play developmentally dynamic roles.

To answer the questions of specificity of OR cluster and Greek Island contacts with base pair resolution, I set out to optimize *in situ* HiC in the olfactory epithelium. With this genome wide technique, I would be able to understand exactly how wide spread these interchromosomal contacts. By optimizing each step of the protocol, I was able to optimize in situ HiC to work with ~5,000 cells, 100-fold fewer than previously reported. This allowed me to ask deep probing questions about the developmental dynamics of nuclear architecture, and what happens to the chromatin state when an OR gene is chosen.

Chapter 4:

in situ HiC Optimization

The steady flow of scientific discoveries is periodically disrupted in bounds and leaps by technological advances. Although the three-dimensional organization of DNA (chromatin architecture) is a major determinant of cellular fate and function, early studies were limited in scope by available techniques. For example, long-range enhancer-promoter chromatin interactions had been hypothesized for quite some time (Goodbourn et al., 1985; Maniatis et al., 1987), but showing these interactions in their native state remained challenging.

The development of chromosome conformation capture (3C) by Job Dekker was driven by a passion for the structure of mitotic chromosomes (Fig. 4A) (Dekker, 2002; O'Donnell, 2016). In this technique the structure of a chromosome (or nuclear chromatin) can be inferred by the relative frequency of contacts between two genomic loci. Specifically, chromatin is fixed, digested with a restriction enzyme, then the DNA fragments are ligated back together. Because fixed protein holds the chromatin in its native conformation, proximity-based ligation occurs between three-dimensional DNA partners. This probabilistic ligation junction can then be assessed by quantitative electrophoresis, RT-qPCR, or next generation sequencing (HiC – high-throughput chromosome conformation capture (Lieberman-Aiden et al., 2009)). With the cost of sequencing plummeting since 2008 (Fig. 4B,C) and the formation of the NIH 4D Nucleome Consortium aimed at understanding the structure of the nucleus, HiC has become a widely used technology that has rapidly expanded our understanding of nuclear architecture.

Conceptually, *in situ* HiC is similar to 3C except that ligation occurs in intact nuclei and libraries are sequenced deeply to assess the three-dimensional structure of the whole genome (Rao et al., 2014) (Fig. 4A). I have been told that several members of our lab and other labs have failed at optimizing *in situ* HIC for mOSNs. In order to give myself the best chance at using in situ HiC in neurons, I set out to deconstruct the entire protocol and optimize each step. I will delineate this process here. I must also comment that I often changed multiple parameters at once in order to expedite the optimization process. In summary, the optimizations to this protocol yielded robust improvement for application to olfactory sensory neurons.

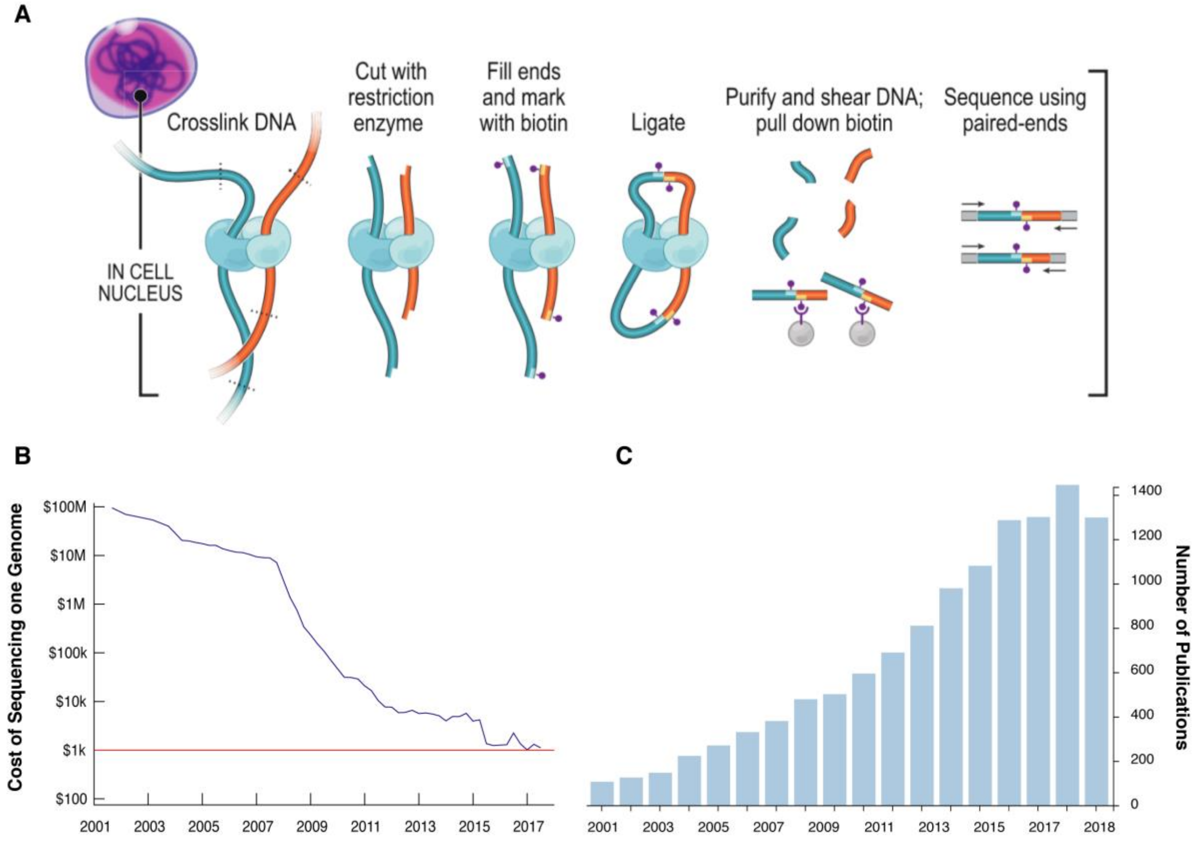
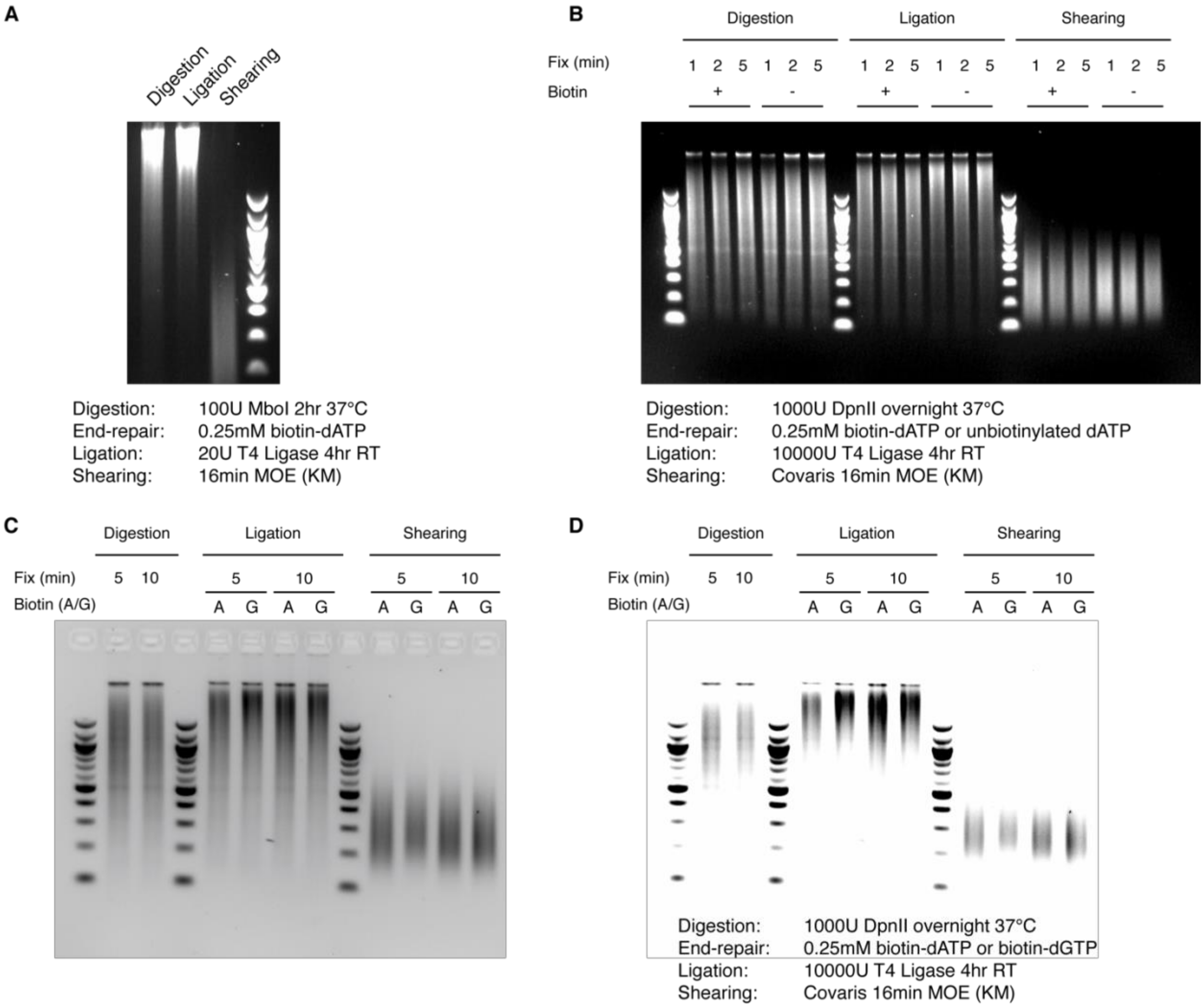


Figure 7. *in situ* HiC and low cost of sequencing have driven genome wide nuclear architecture studies. A. *in situ* HiC protocol. Adapted from Rao et al., 2014. **B.** Cost of sequencing a human genome has fallen much faster than Moore's law would have predicted. Source: NIH. **C.** Exponential increase in publications on the topic of "Genomic Architecture". Source: Web of Science; Nov. 2018.

Fixation and isolation of nuclei

The first step of the *in situ* HiC protocol is to fix cells with 1% formaldehyde for 10 minutes at room temperature, then lyse the cells to isolate pure nuclei. Before this step, I started with FACS-purified neurons as described previously by our lab (Monahan et al., 2017). Fixation is a common practice in molecular biology for techniques like ChIP-seq, immunofluorescence, HiC, etc. Previously, Kevin Monahan had optimized ChIP-seq to work with mOSNs (Monahan et al., 2017). I used the fundamentals of this 2-year optimization as a launching point. Notably our ChIP-seq protocol uses 5 min of fixation, while the *in situ* HiC protocol uses 10 min. In the first few rounds of optimization, I determined that I was having a difficult time digesting my chromatin by gel electrophoresis (Fig 5A). First, I switched the restriction enzyme I was using, but at the same time, I empirically tested fixation times of 1min, 2min, 5min (Fig. 5B). By gel electrophoresis, I found no observable effect of fixation time on digestibility of the chromatin. Therefore, in later experiments, I directly tested 5 min vs. 10 min and again found limited effect of fixation on digestion (Fig. 5C,D). Because everyone in the *in situ* HiC field was using 10 min as their fixation time, and because I found no effect on fixation time on digestibility of the chromatin, I decided to use 10 min in 1% PFA in PBS at room temperature for my experiments.

Notably, I never directly tested the different fixation techniques on the final sequencing data produced. Through this type of analysis and comparison, it remains possible that one can parse out frequency vs. strength of interactions. For example, if a protein brings together two segments of DNA, the frequency of HiC ligations observed would be at least dependent on the frequency on interaction as well as the strength and/or distance of this interaction. By using different fixation conditions, especially weak or selective fixation, one might be able to tease out the weak but frequent interactions from the infrequent but strong interactions. A conceptually similar experimental approach would be to use different types of fixatives. To my knowledge, no one has tried to understand interaction frequencies in HiC using experimental approaches.



Digestion: 100U MboI 2hr 37°C
End-repair: 0.25mM biotin-dATP
Ligation: 20U T4 Ligase 4hr RT
Shearing: 16min MOE (KM)

Digestion: 1000U DpnII overnight 37°C
End-repair: 0.25mM biotin-dATP or unbiotinylated dATP
Ligation: 10000U T4 Ligase 4hr RT
Shearing: Covaris 16min MOE (KM)

Digestion: 1000U DpnII overnight 37°C
End-repair: 0.25mM biotin-dATP or biotin-dGTP
Ligation: 10000U T4 Ligase 4hr RT
Shearing: Covaris 16min MOE (KM)

Figure 8. Optimization process of *in situ* HiC in mOSNs. A. First trial of *in situ* HiC in mOSNs in October of 2016. Notably, there were no positive controls for this experiment and I used the in situ protocol “out-of-the-box”. **B.** First round of optimization included testing fixation times as well as including or excluding biotin to test efficiency of biotin. **C.D.** Same image at different exposures. The second round of optimization focused on again testing fixation times and testing the hypothesis that biotin-dGTP is better for ligation than biotin-dATP, irrespective of fixation time.

For the isolation of fixed nuclei through, I deviated from the HiC field’s detergents and settled for milder detergents as used by our lab for ChIP-seq. As the purpose of this step is to isolate pure nuclei with as minimal disturbance of the chromatin as possible, I did not pursue further optimization of the cell lysis step.

DNA Digestion

DNA digestion is a major step in the *in situ* HiC protocol. After fixing, isolating and permeabilizing nuclei with mild SDS, digestion of chromatin allows for downstream formation of chimeric HiC ligations. As the resolution of data extracted is directly related to how often your restriction enzyme cuts, the standard in the field is to use 4-cutter enzymes like MboI and DpnII[2]. When I started optimizing this protocol, the standard enzyme was MboI. I later learned that MboI is sensitive to CpG methylation if it is overlapping with the GATC site. I switched to DpnII, a methylation-insensitive enzyme. I only tried MboI once (Fig. 5A,B). Later gel electrophoresis and sequencing experiments revealed that DpnII cut DNA more often, however true insights are limited because I also used a more concentrated enzyme stock (10-fold), and digested overnight instead for 2 hours. Moving forward, I always digested with 10000U DpnII at 37°C for at least 18 hours. In the morning of the next day, I would spin down my pellet of nuclei, and replace the solution with fresh buffer and DpnII and digested for an additional 2 hours. By gel electrophoresis, I saw no difference, but I always did this step. It remains possible that this step may be optimized for efficiency and cost-effectiveness. As mentioned above, this digestion strategy was robust to varying fixation times from 1 min to 10 min. Also notable is that other labs in the field have also started to use DpnII as their enzyme of choice.

End-repair with biotin

The next step of the *in situ* HiC protocol is to end repair the digested ends in preparation for blunt-end ligation. In order to enrich for sequencing reads that underwent a chimeric ligation, a biotinylated nucleotide is incorporated at this step. At the time that I began optimizing this protocol, the field used biotinylated dATP, most likely for its widespread availability through Thermo Fisher and other historical/logical reasons. I first started to consider that biotin might be

[2] In recent years, people have tried to use DNAse I in order to fragment the DNA but this is somewhat limited by downstream analysis of the cut sites as well as the propensity of DNA I to cut most frequently in accessible chromatin regions (Ramani et al., 2016).

interfering with ligation efficiency when I came across a paper that avoided the biotin step in order to generate “chromosome walks” for understanding larger genomic structures. Simply speaking, if one generates a 3C/HiC library without the biotin incorporation step and stops just after ligation step, then one is left with a kilo-megabase-long DNA fragment of all the DNA ligated to itself. Tanay and colleagues then sought to sequence each one of these extremely long DNA fragments through single-molecule dilution and whole genome amplification. For me, the key insight was when I noticed that their gel electrophoresis band representing the ligated library was much larger than the ligated libraries I was generating. This immediately prompted me to test the exclusion of a biotinylated dNTP (Fig. 5B). Indeed, biotin was somehow interfering with my ligation efficiency. In the gel, one can appreciate that the unbiotinylated lanes have high molecular weights after the ligation step.

As our goal was to optimize HiC (not chromosome walking), I still needed to include a biotin step. In thinking more deeply about this, I considered the ligation site created by the restriction enzyme: GATC. Rather simply, the A is closer to the interface of the ligation than the G. As sterics and electrostatic forces play a major role in reaction kinetics, I thought to test whether a biotin-dGTP (Perkin Elmer) would be a more effective nucleotide (Fig. 5C,D). Indeed, biotin-dGTP is superior to biotin-dATP for producing a higher molecular weight ligation product.

The experimental logic of this experiment is clever but not perfect: the linkers are different lengths and on different atoms of the nitrogenous base. In biotin-dATP, the linker is 14 bases long and on position 6 of the nitrogenous base. In biotin-dGTP, the linker is 11 bases longer and on position 7 of the nitrogenous base. It is possible that any three of these factors are responsible for the increased efficiency:

1. Distance to ligation site
2. Length of biotin linker
3. Position of biotin linker on nitrogenous base
4. Nucleotide identity

It also remains unknown at which step biotin interferes (fill-in with Klenow fragment vs. ligation with T4 DNA ligase).

Ligation

The next step in the protocol is to ligate the digested and end-repaired DNA to generate chimeric reads. Optimization is limited by the enzymes that are available. In my optimization, I was guided two principles of thermodynamics and enzyme kinetics:

1. More is better
2. Longer is better

Normally, enzyme concentration is negligible by traditional Michaelis-Menten kinetics, however, I have been told on many occasions that T4 ligase is highly sensitive to temperature. Therefore, the assumption that enzyme concentration is constant and negligible in Michaelis-Menten enzyme kinetics may not actually apply to our conditions. For my intervention, I focused on increasing the concentration of T4 ligase. The concentration of T4 DNA ligase is mostly limited by the concentration. T4 DNA ligase is stored in glycerol and glycerol inhibits its activity. New England Biolabs sells T4 ligase at two different concentrations: 400K U/mL and 2M U/mL. Therefore, I was able to increase the concentration of T4 ligase 5-fold without increasing the concentration of rate-inhibiting glycerol. In an effort to keep the protocol under 4 days, I opted to stick with the 4-hour room temperature incubation step. Moving forward, this step could be more thoroughly optimized for time, cost, and effectiveness.

Shearing

Following ligation, the next step is shearing of the DNA fragments for DNA library preparation. Here again I was able to lean on the work of Kevin Monahan. He had previously optimized shearing conditions for ChIP-seq and settled on a robust and reproducible program that uses the Covaris Ultrasonicator to gently shear DNA from mOSNs into ~400bp fragments (examples in Fig. 5). I actually suspect that shearing is less important for HiC than ChIP-seq. The basic principle of ChIP-seq is to pulldown DNA that is bound to a protein of interest through the use of an antibody. If the DNA fragments that are being pulled down are too large, then the ChIP will be "noisy" with DNA fragments that do not represent the DNA binding region of the protein[3]. On the other hand, in HiC, one is not looking for a transcription factor footprint, merely a fragment of DNA that was ligated to another. I have suspected that shearing to different DNA lengths may bias the data one way or another. Similar to the discussion above about using different fixatives, if you shear to short fragments vs. long fragments, you may be enriching for shorter vs. longer range interactions. The consensus is that any fragment of DNA greater than 1kb will not properly cluster on a next-generation sequencer, therefore the range of DNA fragments one could experiment with is between 100-1000 bp. As this represents a 34-340 nm range, it is possible that one may see a short vs. long bias. To my knowledge, no one has investigated this.

Library preparation

In the original *in situ* HiC protocol, the library prep step involves a standard home-brew library prep protocol using NEB reagents. At the time when I was optimizing my protocol, our lab used Nugen Ovation kit for everything from RNA-seq to ChIP-seq because its proprietary

[3] In fact, the Henikoff lab has exploited the idea that smaller fragments make for a better ChIP in their technique: CUT&RUN. In CUT&RUN, an antibody-conjugated micrococcal nuclease "releases" protein-DNA complexes from DNA and allows for better and more efficient footprinting of transcription factors using ~150 bp DNA fragments (Skene and Henikoff, 2017). A similar technique has been employed by the Franklin Pugh lab to improve ChIP-seq footprinting: ChIP-Exo. This technique uses an exonuclease following pulldown to better resolve the footprint of DNA binding proteins (Rhee and Pugh, 2011).

reagents were more efficient with our cells. I used the standard Nugen Ovation V2 kit and protocol with great success.

Results of optimization

In the two months of optimization, I only sequenced two libraries: the first one was my first trial of *in situ* HiC and the second one went into my final paper (Fig. 6). In the first experiment, I followed the protocol as described in the original *in situ* HiC manuscript (Rao et al., 2014). I generated 1 million HiC reads after sequencing > 450 million reads (0.2% HiC reads). At that point I learned that libraries can be sequenced shallowly to validate HiC library quality. Following optimization, I sequenced my second library to 9.6 million reads and got back > 4 million HiC reads, an improvement to over 40% HiC reads. Objectively, that is a greater than 200-fold improvement.

I also tested the protocol in different for required starting material. In initial experiments, I started with 5 million neurons. As experimental constraints demanded, I would decrease the starting material and sequence the result. At the end of my second year doing HiC, I was able to generate complex HiC libraries (> 1 billion unique reads) with only 5,000 cells! That's a 1000-fold improvement in cell number.

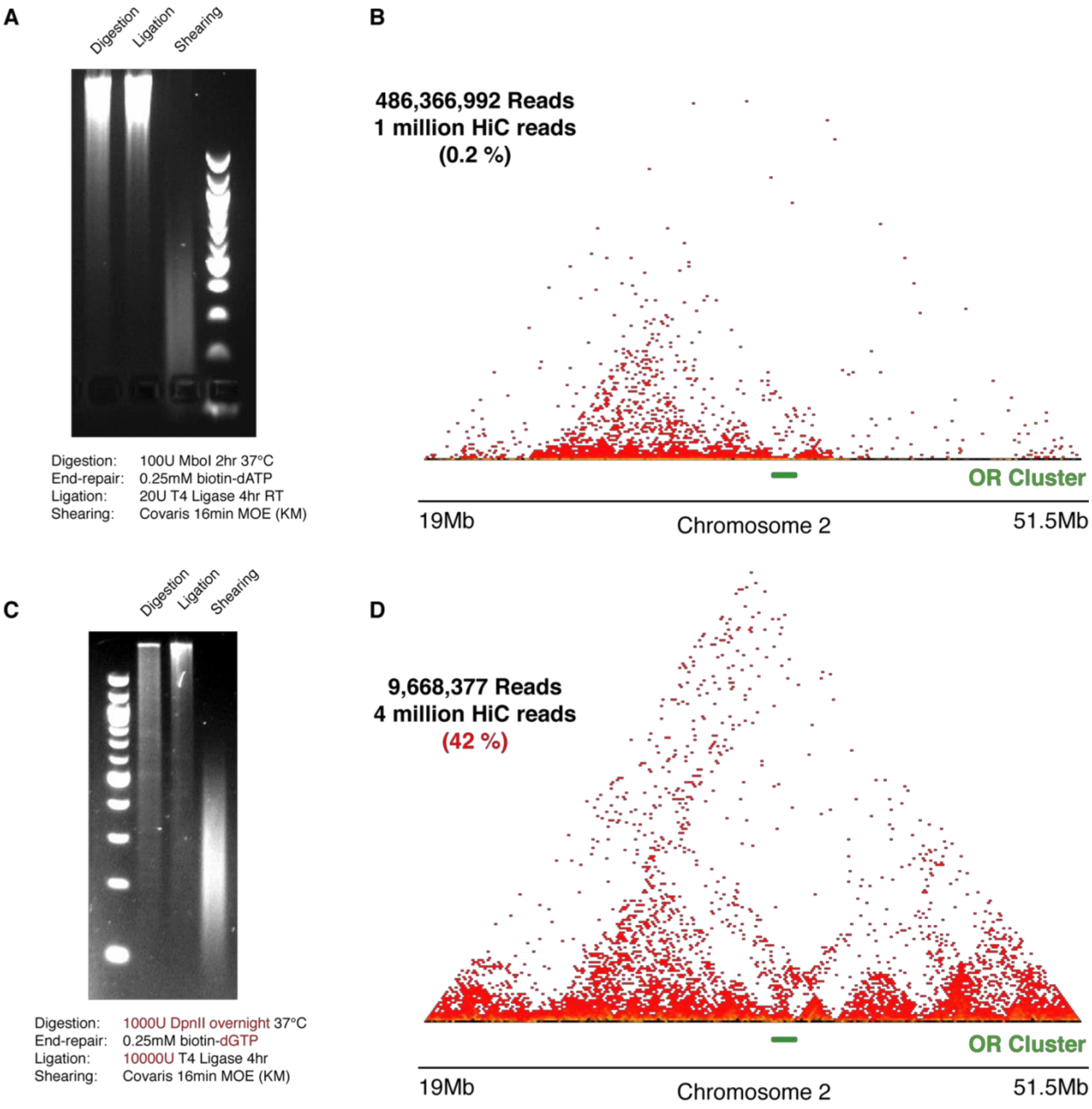


Figure 9. Results from optimizing in situ HIC in mOSNs. A. First trial of *in situ* HiC in mOSNs in October of 2016. There were no positive controls for this experiment. In situ HiC protocol "out-of-the-box" (Rao et al. 2014). **B.** This library was deeply sequenced on a high-output NextSeq kit with over 450 million reads. This library was biased by chromatin accessibility and yielded only 0.2% HIC contacts. The region to the left of the OR cluster with numerous HiC contacts is a highly accessible region by ATAC seq. **C.** Following optimization, this library was generated in January 2017 with the parameters below. Note the improved digestion, high molecular weight ligation and unchanged shearing. **D.** January 2017 library was sequenced shallowly to ~10 million reads and yielded over 4 million HiC yields. The bias to accessibility was quantifiably minimized and computationally corrected with the Knight Ruiz matrix balancing algorithm in downstream analyses. This library was sequenced to 450 million reads and presented in our final manuscript.

Chapter 5:

mOSN specific interchromosomal interactions

Mouse ORs are encoded by a family of ~1400 genes that are organized in 69 heterochromatic genomic clusters distributed across most chromosomes (Fig. 1E). Every mature OSN (mOSN) expresses one OR gene from one allele in a seemingly stochastic fashion (Buck and Axel, 1991; Chess et al., 1994; Monahan and Lomvardas, 2015). Previous work suggested that repressive and activating interchromosomal interactions contribute to the singular OR expression (Clowney et al., 2012; Lomvardas et al., 2006; Markenscoff-Papadimitriou et al., 2014). However, these interactions have only been analyzed with the use of biased and low-throughput approaches (3C, 4C, capture HiC, and DNA FISH), which have either limited genomic resolution or restricted genomic coverage. Thus, it remains unknown how prevalent and specific these interactions are, and how they form in relationship to OSN differentiation and OR expression. Moreover, *in situ* HiC (Rao et al., 2014), which reduces the occurrence of non-specific ligation events observed in dilution HiC, revealed that interchromosomal associations between non-repetitive, genic regions are extremely infrequent (Johanson et al., 2017; Nagano et al., 2015), and only emerge upon depletion of cohesin complexes (Rao et al., 2017a; Schwarzer et al., 2017). Thus, to explore the landscape of interchromosomal interactions in a biological system that likely depends on them, and to provide a conclusive answer into whether interchromosomal contacts actually occur with biologically meaningful frequency and specificity, I performed *in situ* HiC in distinct cell populations of the main olfactory epithelium (MOE).

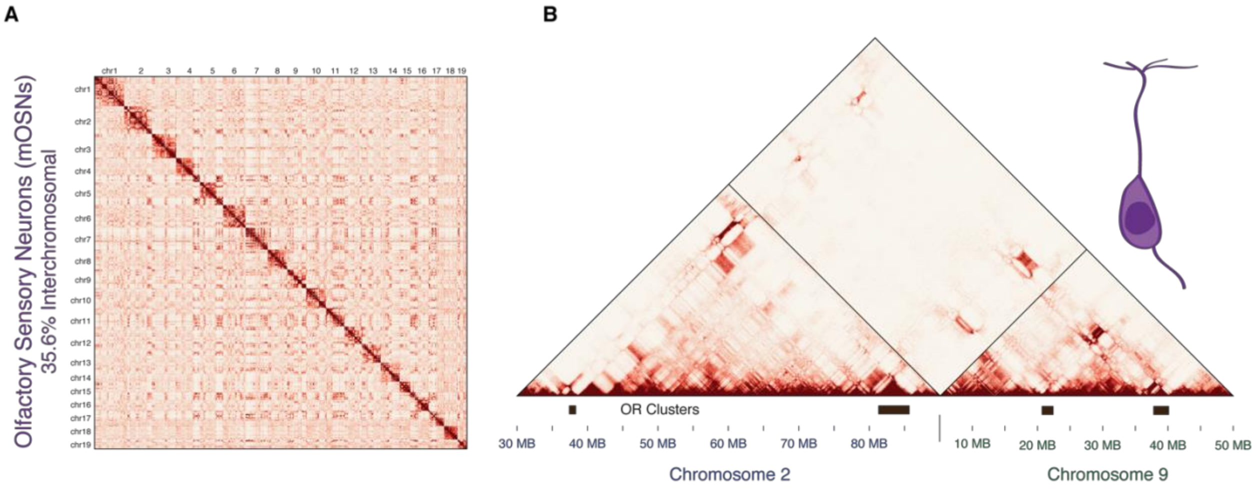


Figure 10. Mature Olfactory Sensory Neurons (mOSNs) make extensive interchromosomal contacts between olfactory receptor (OR) clusters. A. Genome-wide *in situ* HiC contact matrices reveal increased interchromosomal contacts in mOSNs. **B.** Zoomed-in views of chromosome 2 and 9 show highly restricted and frequent contacts between OR gene clusters in *cis* and *trans* in mOSNs.

First, I analyzed FAC-sorted mOSNs, which represent terminally differentiated, post-mitotic neurons that are heterogeneous in regards of the identity of the chosen OR. *In situ* HiC in mOSNs revealed quantitative and qualitative differences from other cell types. Genome-wide, there are extensive and discreet interactions across chromosomes, that correspond to 35.6% of total HiC contacts (Fig. 7A), whereas in B cells (Rao et al., 2014) (20%), ES cells (Yan et al., 2018) (16%) and neocortical neurons (Bonev et al., 2017) (26.2%) these interactions are less frequent and appear more diffuse (Fig. 8). Zoomed in views of chromosomal regions that contain OR gene clusters reveal strong *trans* contacts between these clusters that are undetectable in B cells, and the other cell types analyzed (Fig. 7B, 8).

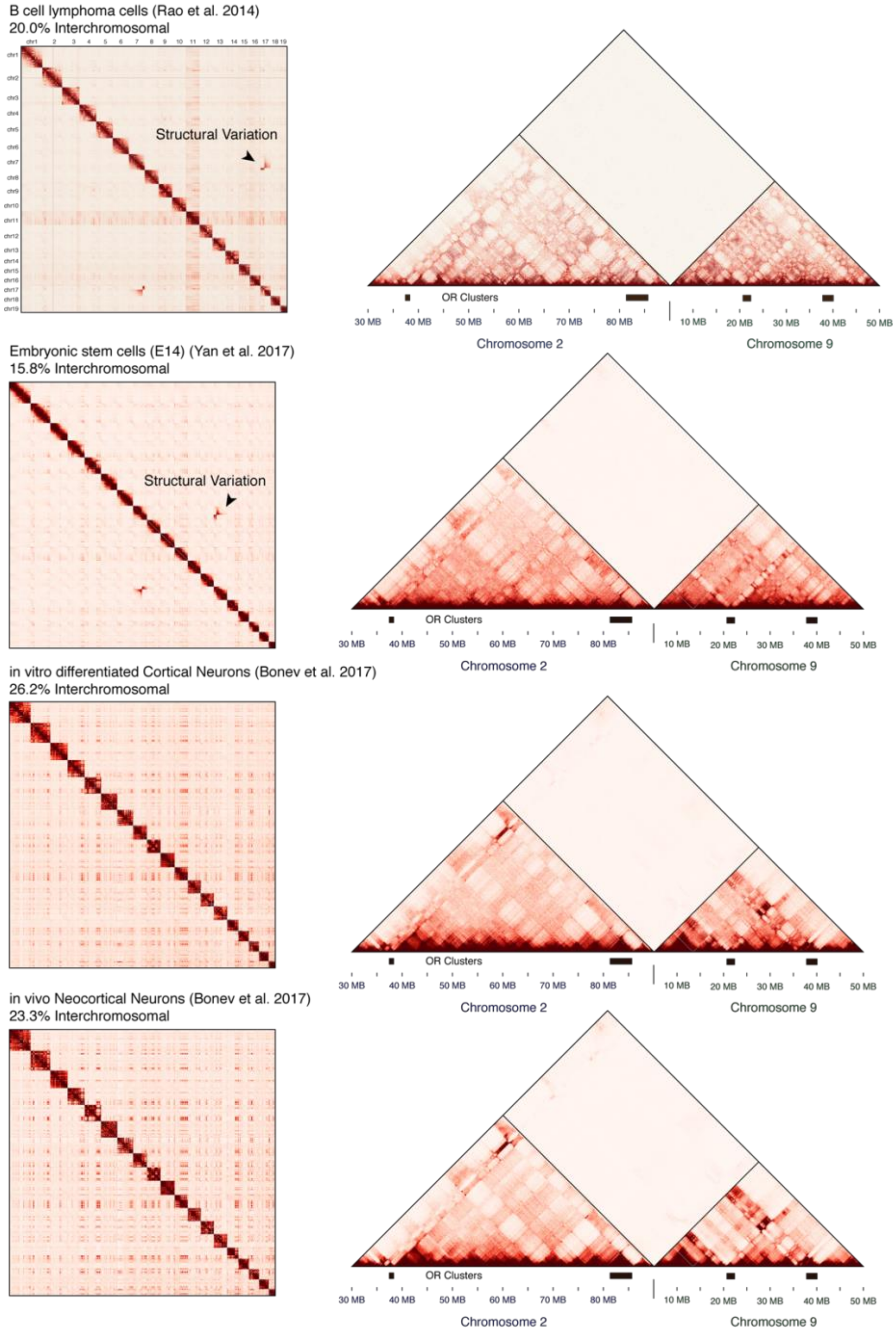


Figure 11. Long-range contacts between OR gene clusters are infrequent in other cell types. A. Genome wide and zoomed-in view of HiC contact matrices reveal decreased genome-wide 5 interchromosomal interactions when compared to mOSNs, as well as lack of specific interchromosomal 6 contacts between OR gene clusters in B cell lymphoma cells, ES-E14 cells, in vitro differentiated neurons, and in vivo cortical neurons. Structural variations are marked by arrow heads.

Genome-wide, OR gene clusters from every chromosome make strong and specific contacts with each other (Fig.9A). Aggregate peak analysis (APA) (Rao et al., 2014) showing highly focused *trans* contacts between OR gene clusters, confirms the specificity of these interactions which is not observed in other cell types (Fig. 9B). Interestingly, in cortical neurons, although OR gene clusters do not interact in *trans*, they form strong *cis* contacts over large genomic distances (Fig 8). However, these interactions are less selective and less prevalent when directly compared with mOSNs (Fig. 10). Finally, unsupervised compartment discovery (Rao et al., 2014) suggests that there are at least 9 distinct compartments, one of which contains OR gene clusters (Fig. 11).

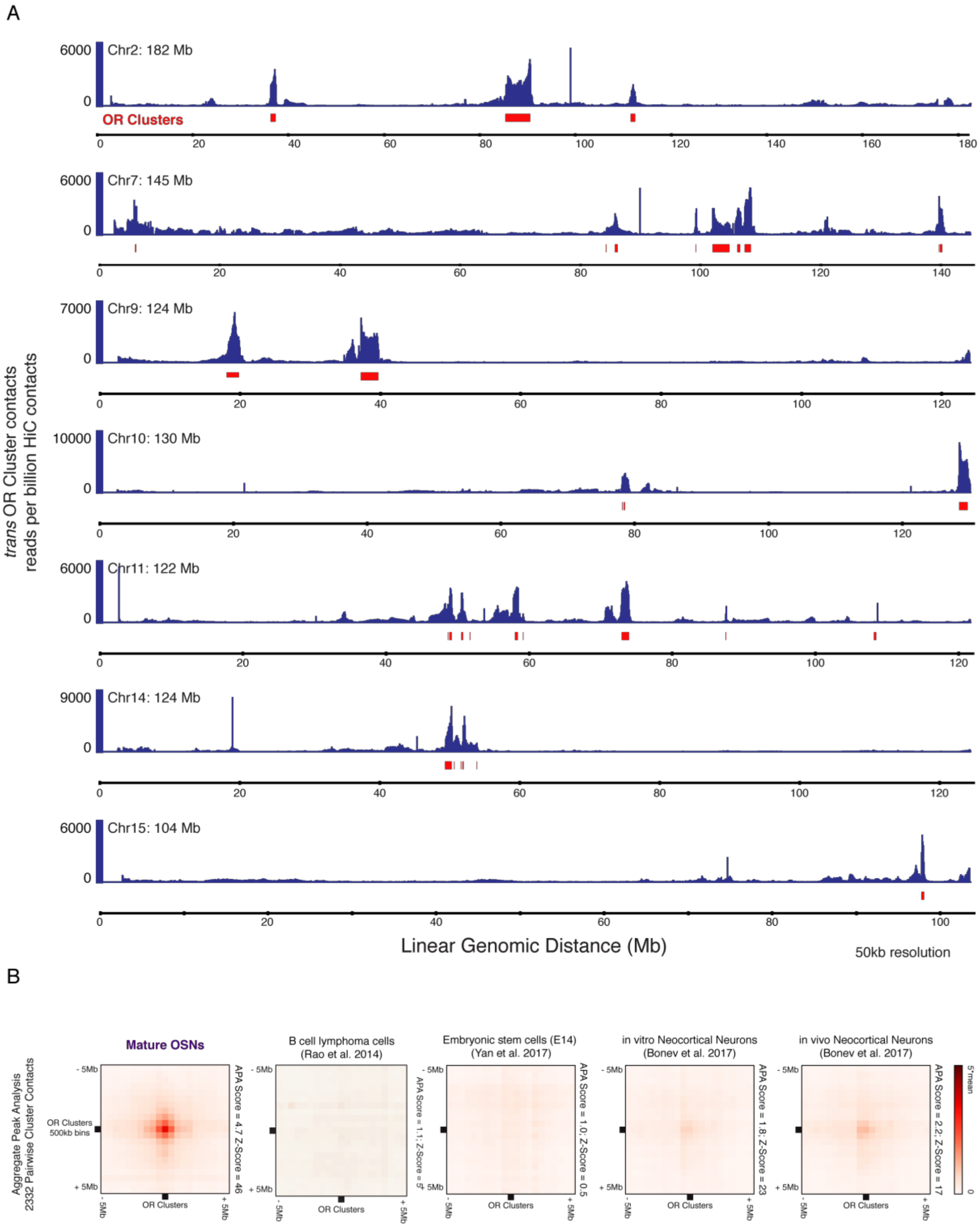


Figure 12. OR gene clusters make specific contacts with other OR gene clusters in *trans*. A. Chromosome-wide views of *trans* OR contacts reveal that contacts are specifically restricted to other OR clusters. **B** Aggregate peak analysis reveals the specificity of OR-OR contacts in mOSNs but not other cell types examined.

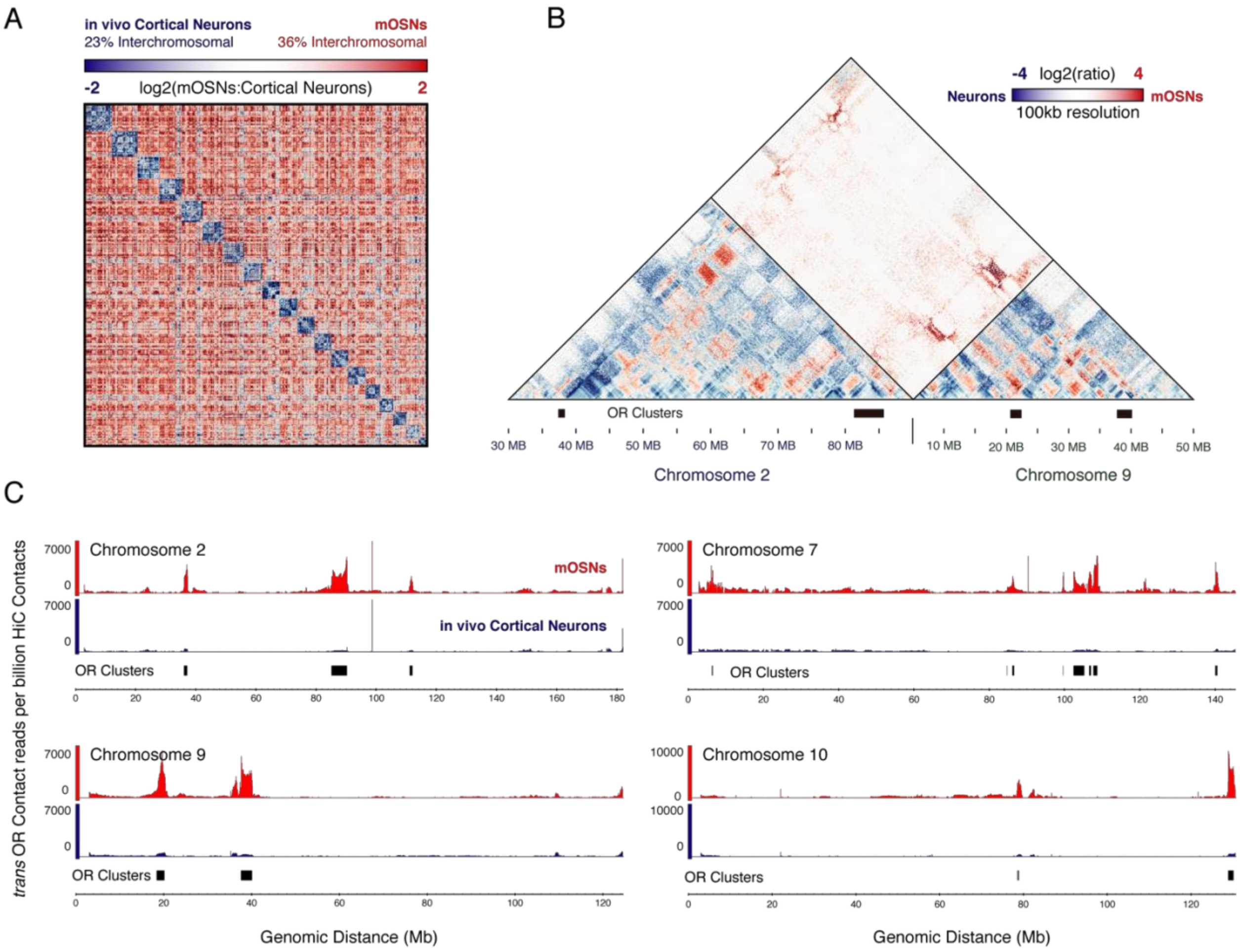


Figure 13. Interchromosomal contacts between OR gene clusters are stronger in mOSNs compared to neocortical neurons. A. Genome wide difference map of HiC contacts between mOSNs and in vivo neocortical neurons. **B.** Zoomed-in view of regions on chromosome 2 and 9 reveal that cis and trans contacts between OR gene clusters are more frequent in mOSNs compared to neocortical neurons. **C.** Cumulative interchromosomal contacts from OR Clusters to 4 different full length chromosomes reveal differences in frequency of contacts between mOSNs (red) and in vivo cortical neurons (blue).

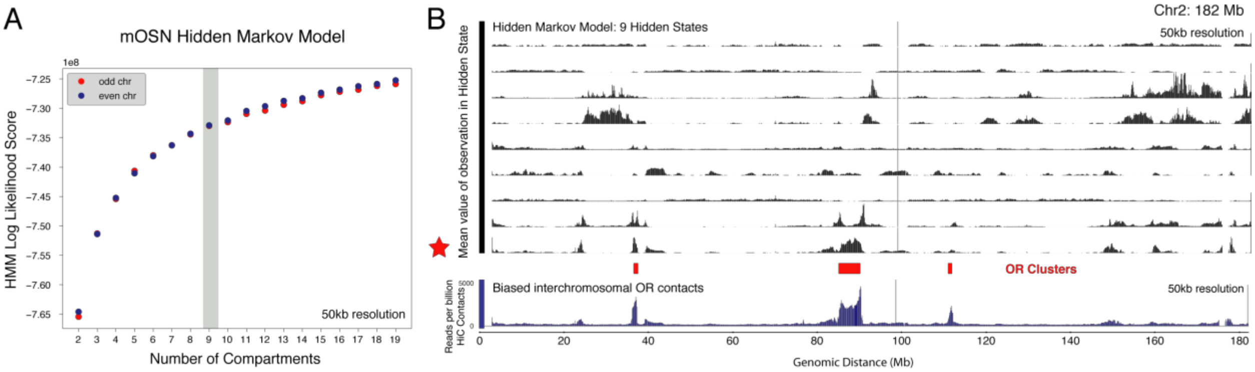


Figure 14. Machine learning recapitulates the biased OR gene compartment. A. Hidden Markov Model (HMM) score for a given number of compartments. 9 compartments were used for further analysis**. B.** 9 HMM-derived compartments reveal the existence of distinct compartments, one of which (black star) corresponds with the biased analysis of contacts from trans OR Clusters. Scale is the average value of a given locus in a given compartment.

Chapter 6:

Gradual compartmentalization during development

Upon establishing the genome-wide, mOSN-specific compartmentalization of OR gene clusters, I sought to identify the differentiation timing of OR compartment formation. I FAC-sorted two progenitor cell populations, Mash1$^+$ and Ngn1$^+$ cells. Mash1$^+$ cells are multipotent, mitotically active OSN progenitors with undetectable levels of OR transcription (Fletcher et al., 2017). Only 17.9% of the total reads in this population correspond to interchromosomal contacts (Fig. 12A). In agreement with this genome-wide pattern, in Mash1$^+$cells interchromosomal contacts between OR clusters are almost undetectable, and *cis* contacts are weak (Fig. 12B). In contrast, in the more differentiated Ngn1$^+$ cells, which are mostly post-mitotic immediate OSN precursors (Fletcher et al., 2017), 32.2% of HiC contacts are interchromosomal (Fig.12C, D). Moreover, I detect both *cis* and *trans* interactions between OR clusters that are weaker than the OR contacts in mOSNs (Fig. 12F), but appear as specific according to an unbiased compartment analysis (Fig. 13). Thus, OR compartments form in a hierarchical fashion during development, with *cis* interactions being detected first, *trans* interactions appearing in more differentiated stages and reaching maximum frequency in mOSNs. Interestingly, the gradual increase of compartmentalization is not restricted to OR clusters, since our HMM-based prediction of genomic compartments shows that the total number of distinct compartments increases with differentiation (Fig. 13) consistent with predictions made by soft X-ray tomography studies on these cells (Le Gros et al., 2016).

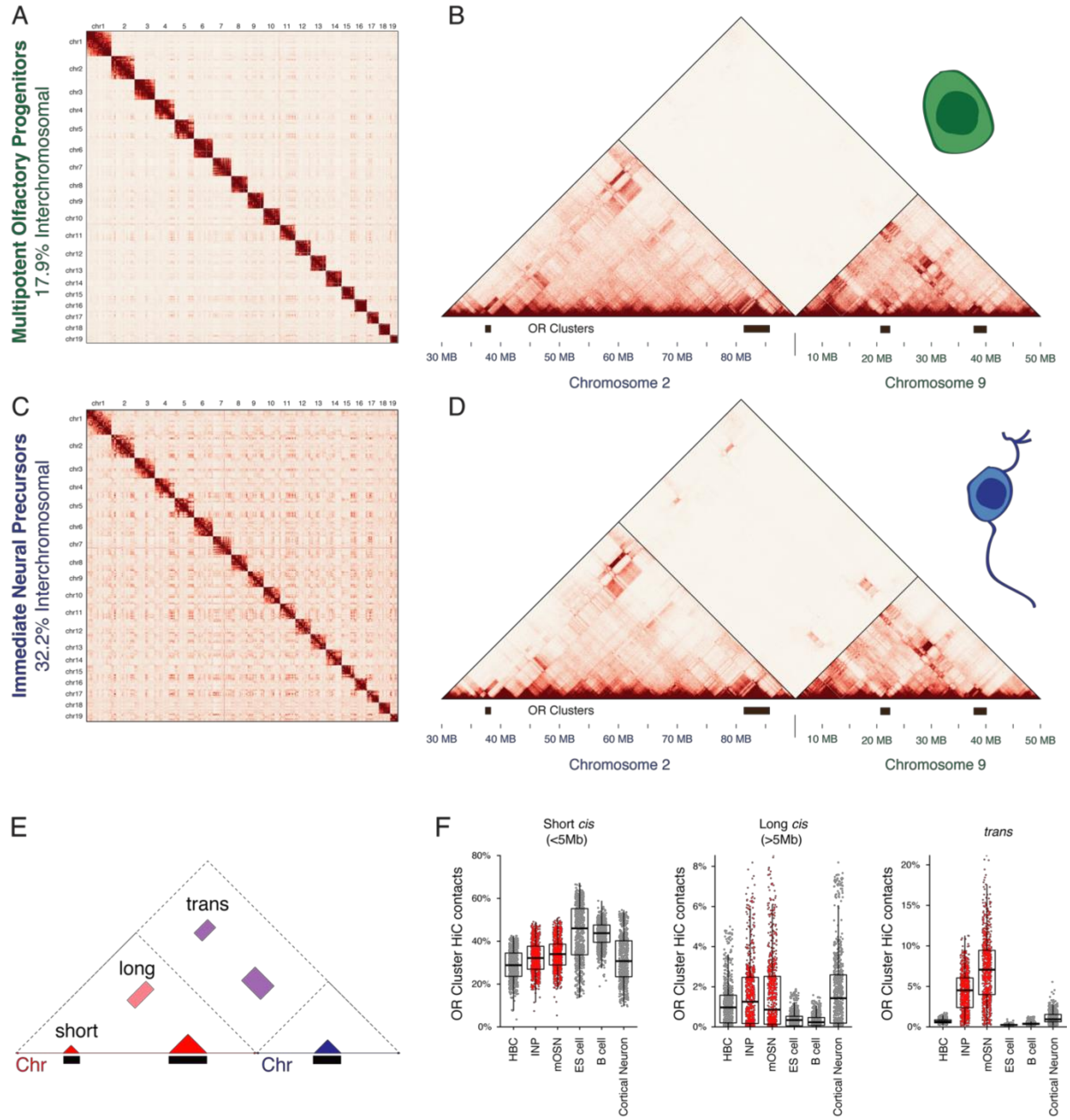


Figure 15. Gradual OR compartmentalization during mOSN differentiation. A. Genome wide *in situ* HiC contact matrices of multipotent olfactory progenitors. **B.** Zoomed-in views of OR gene clusters on chromosome 2 and 9 in multipotent olfactory progenitors. **C**. Genome wide *in situ* HiC contact matrices of immediate neuronal precursors. **D.** Zoomed-in views of OR gene clusters on chromosome 2 and 9 in INPs. **E.** Summary of binning strategy used for quantitative analysis. **F.** OR Cluster *trans* contacts are most pronounced in mOSNs. Short range *cis* contacts are not significantly different for the different cell types. Counts analysis in F done by Kevin Monahan.

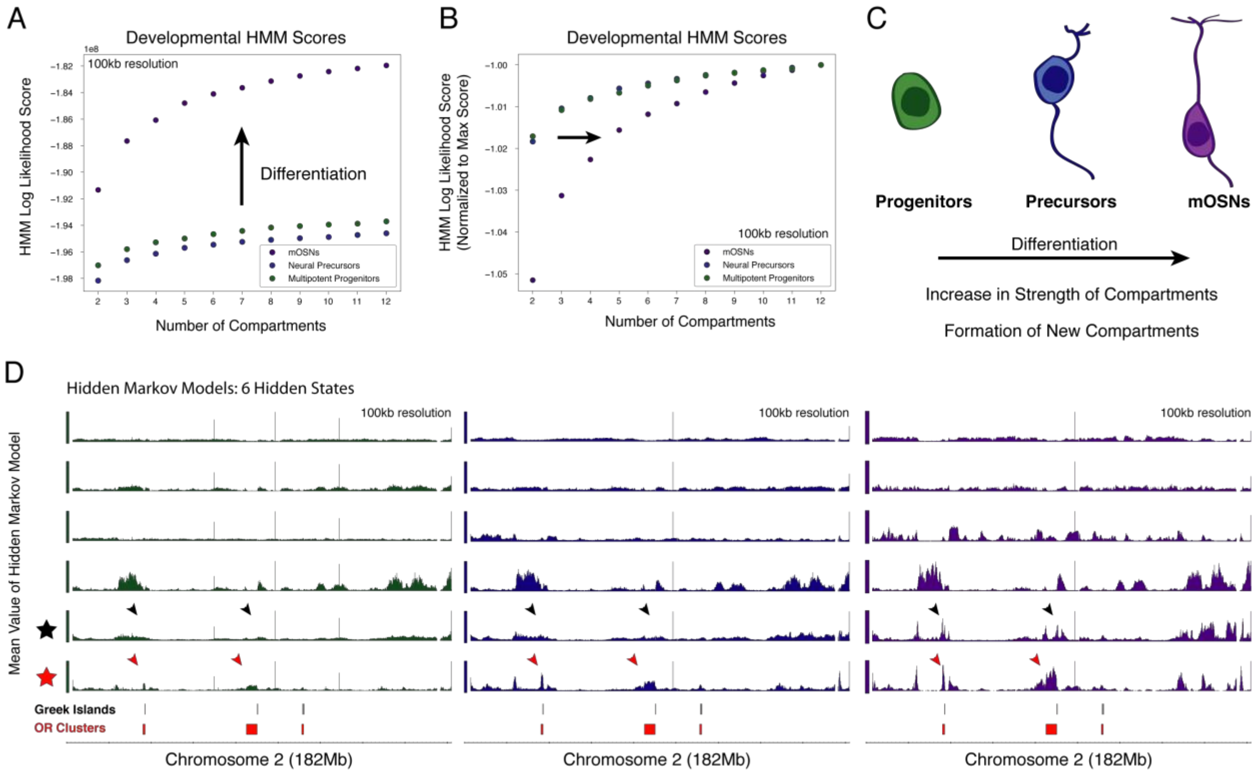


Figure 16: Differentiation of mOSNs leads to new and stronger interchromosomal compartments. A. HMM scores of a compartment analysis of differentiating cells of the olfactory epithelium reveal that interchromosomal compartments become more likely with differentiation. **B.** When normalized to the maximum value, HMM scores reveal a shift in the likelihood curve, suggesting the formation of new compartments with differentiation. **C.** Summary of graphs in A and B. **D.** Close examination of chromosome 2 reveals the strengthening of the OR compartment (red arrowheads) with differentiation, and the formation of a distinct compartment that corresponds with a Greek Island compartment (black arrowheads).

Chapter 7:

Formation of a multi-chromosomal super-enhancer hub

The interactions described thus far involve heterochromatic regions, which may compartmentalize due to phase transition properties of heterochromatin proteins (Larson et al., 2017; Strom et al., 2017). Within the OR clusters, however, reside 63 euchromatic transcriptional enhancers, the Greek Islands, which regulate the transcription of proximal ORs (Markenscoff-Papadimitriou et al., 2014; Monahan et al., 2017). Previous work suggested that these elements interact with high frequency in the MOE (Markenscoff-Papadimitriou et al., 2014), however it is unclear if their associations represent highly specific contacts between these elements or a consequence of surrounding OR interactions. Consistent with the former hypothesis, Greek Island contacts represent HiC “hot spots” suggesting that these elements interact with high specificity with each other (Fig.14A,B). Examination of our HiC data from neuronal OSN precursors (Fig 14C,D) and mitotic progenitors (Fig. 14E,F) shows that Greek Island interactions in trans are undetectable in progenitor cells, first form in OSN precursors and reach maximum frequency and specificity in mOSNs, concomitantly with the peak of OR transcription. This is a general property of Greek Islands and exhibits the same specificity as OR Clusters (Fig. 15D). Quantification of these Greek Island contacts across development underscores the developmental increase in contact frequency.

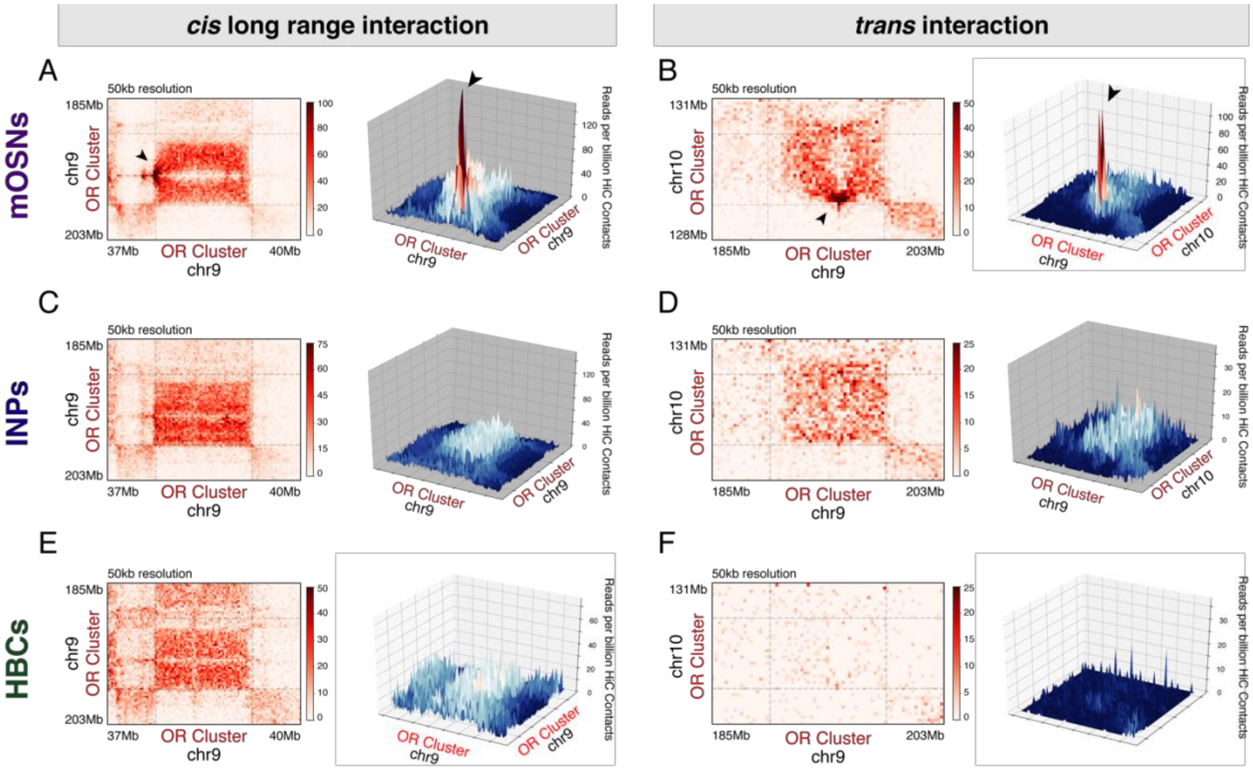


Figure 17. Greek Island-Greek Island contacts form after OR Cluster-OR Cluster contacts. A-F. cis and trans contacts between OR gene clusters reveal contact hotspots in mOSNs (A,B), but not in INPs (C,D) or multipotent progenitors (E,F).

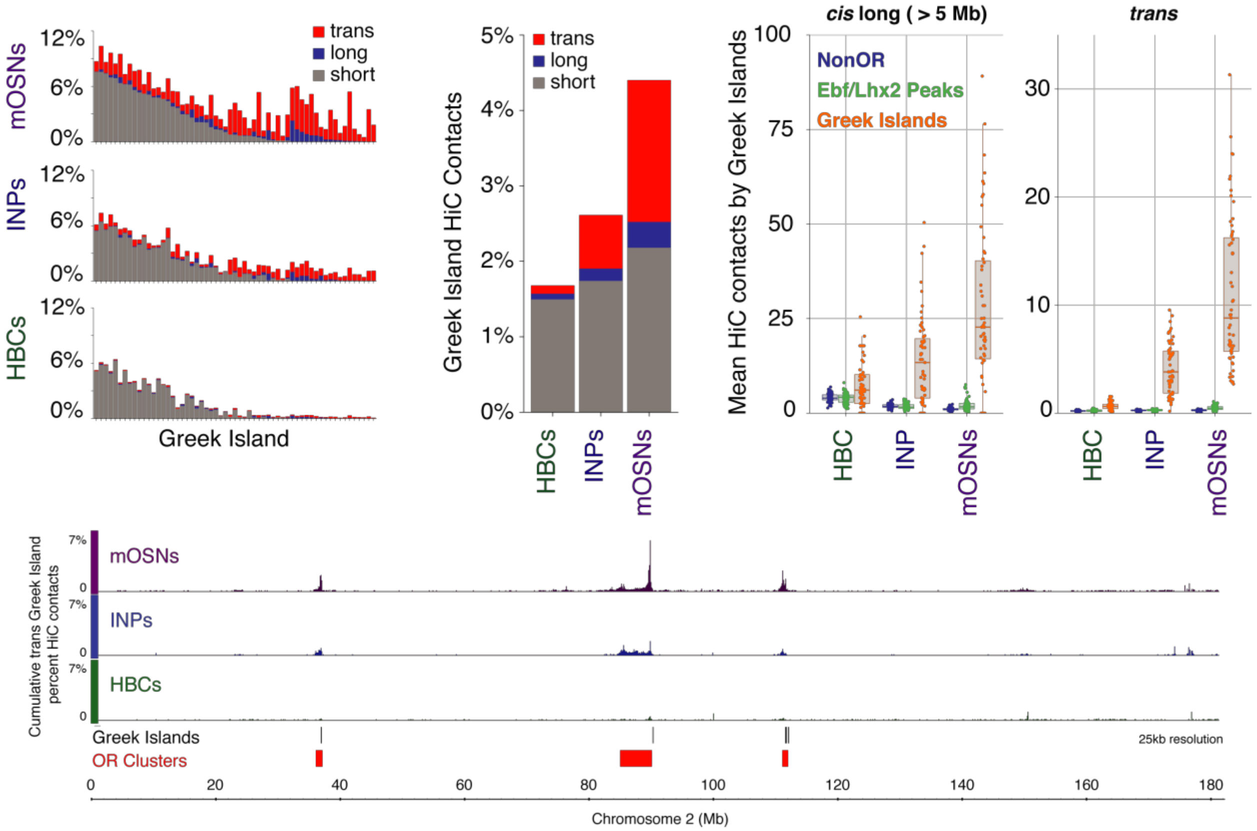


Figure 18. Greek Islands form frequent and specific pairwise contacts with other Greek Islands in mOSNs but not other cell types. A. For each Greek island, the fraction of total Hi-C contacts that are made to other Greek islands located in cis at short range (< 5 Mb apart, grey), long range (> 5 Mb apart, blue) and in trans (red). **B.** mean fraction of Hi-C contacts across all Greek islands (two-sided, paired Wilcoxon signed-rank test, n = 59). **C.** For each Greek island bin (n = 59), the mean number of cis long-range (left) and trans (right) Hi-C contacts per billion made to every non-OR sequence (at 50-kb resolution), intergenic LHX2- and EBF-bound peak (outside OR clusters), or Greek island. Box indicates median, upper, and lower quartiles; whiskers indicate 1.5 × the interquartile range. All panels present pooled data from two independent biological replicates that yielded similar results when analysed separately. **D.** Chromosome-wide views of *trans* Greek Island contacts reveal that contacts are specifically restricted to other Greek Islands in mOSNs. These interactions are less frequent in developmental progenitors. Count analysis in **A**, **B**, and **C** done by Kevin Monahan.

Chapter 8:

The active OR gene contacts the super-enhancer hub

Because Greek Islands are OR transcriptional enhancers that associate at the same developmental time OR genes are transcribed, I sought to investigate their spatial relationship with transcriptionally active OR gene loci. For this I FAC-sorted neurons expressing Olfr16 from chromosome 1, Olfr17 from chromosome 7, and Olfr1507 from chromosome 14 using knock-in iresGFP reporter strains (Bozza et al., 2002; Shykind et al., 2004; Vassalli et al., 2002). First, I compared *cis* interactions made by these OR loci in the OSNs that transcribe them versus OSN subtypes in which they are silent. In each case I find that the transcriptionally active OR locus makes extremely specific contacts with Greek Islands from different OR clusters, residing in separate TADs located more than 1Mb from the transcribed OR (Fig. 16A,E,I).

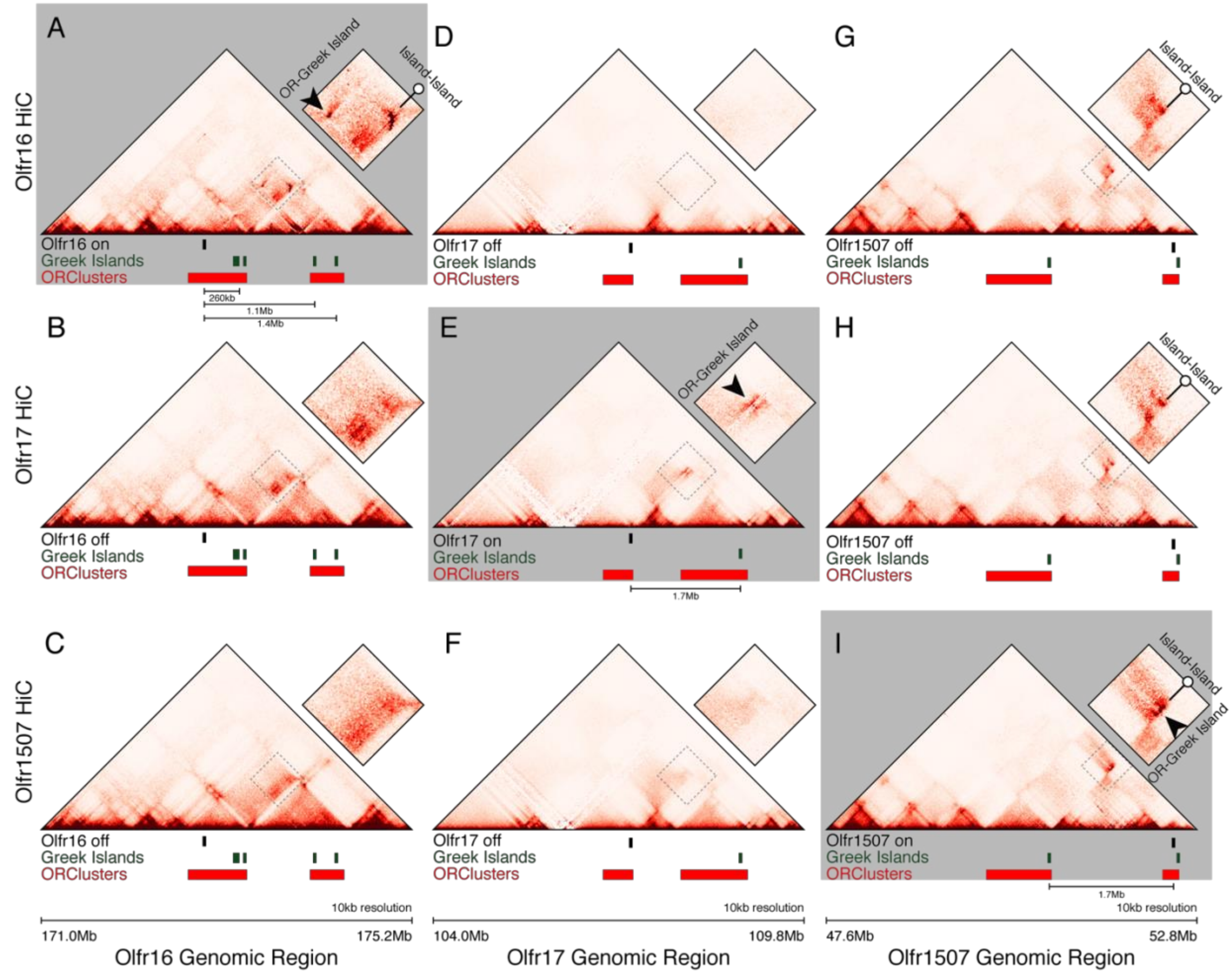


Figure 19. The active OR makes local contacts with Greek Islands in neighboring contact domains. A. In Olfr16+ cells, Olfr16 makes extensive contacts with neighboring Greek Islands up to 1Mb away. **B, C.** Olfr16 does not contact Greek Islands in cells that do not express Olfr16. Similar analyses for Olfr17+ cells and Olfr1507+ cells.

In the case of transcriptionally active Olfr16, I detected a strong and highly specific contact with a Greek Island located ~80Mb apart (Fig. 17A,B), providing the most extreme example of long-range enhancer-promoter *cis* interaction ever described. Interestingly, unlike the three OR loci, Greek Islands make long range that, by and large, are independent of the identity of the transcribed OR (Fig. 16B,C,G,H), consistent with prevalence of Greek Island interactions in mixed mOSN populations. In this vein, in the case of Olfr1507, which is located 50Kb from the Greek Island H (Serizawa et al., 2003), I observe a remarkable example of specificity in genomic contacts. Here, I detect strong interactions between H and the Greek Island Lesvos located 1,7Mb away, which do not extend to the neighboring Olfr1507 unless it is transcriptionally active (Fig. 16. G,H,I).

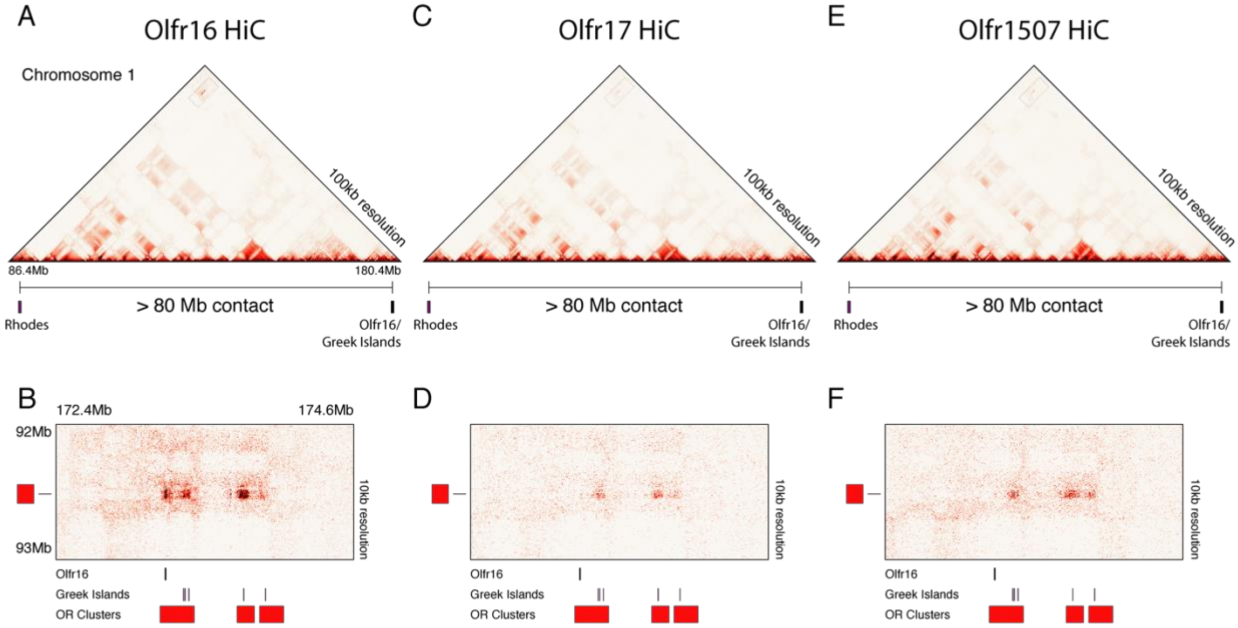


Figure 20. Extremely long-range cis contacts between Greek Islands and the active OR gene. A-F. Contacts that span more than 80 Mb are observed in HiC from Olfr16+ (A), Olfr17+ (C), and Olfr1507+ (E) cells. Close examination of the contacts (dotted boxes) reveals that Greek Islands contact Olfr16+ only in Olfr16+ cells (B). Extremely long-range contacts between Greek Islands in cis are observed also in Olfr17+ and Olfr1507+ cells (D, F).

Finally, I asked if Greek Islands from different chromosomes associate with the active OR gene locus with the same specificity as the *cis* Greek Islands. Indeed, the Olfr16 locus interacts strongly with many Islands in *trans* in Olfr16^{+} OSNs, but has minimal contacts with these elements in Olfr17^{+} or Olfr1507^{+} OSNs (Fig. 18A). Importantly, even in *trans* I detect remarkable specificity in the genomic associations of the transcribed OR that is displayed at multiple genomic scales. First, these interactions are focused on functionally relevant regulatory sequences: Greek Islands preferentially interact with the promoter region of Olfr16, and the promoter of Olfr16 targets the center of the Greek Island bins (Fig. 18A,B,C). Second, at a chromosome-wide scale Olfr16 contacts select Greek Islands but no other sequence in the whole chromosome (Fig. 18D,E). Third, at a genome-wide scale, Olfr16 is the only OR that interacts with many Greek Islands at high frequency. A Manhattan plot depicting normalized aggregate Greek Island-OR interactions shows that the Olfr16-Greek Island contacts are orders of magnitude more significant than the any OR-Greek Island interaction (Fig. 18F). In other words, *in sit*u HiC accurately identifies the transcriptionally active OR from its cumulative interchromosomal interactions with Greek Islands.

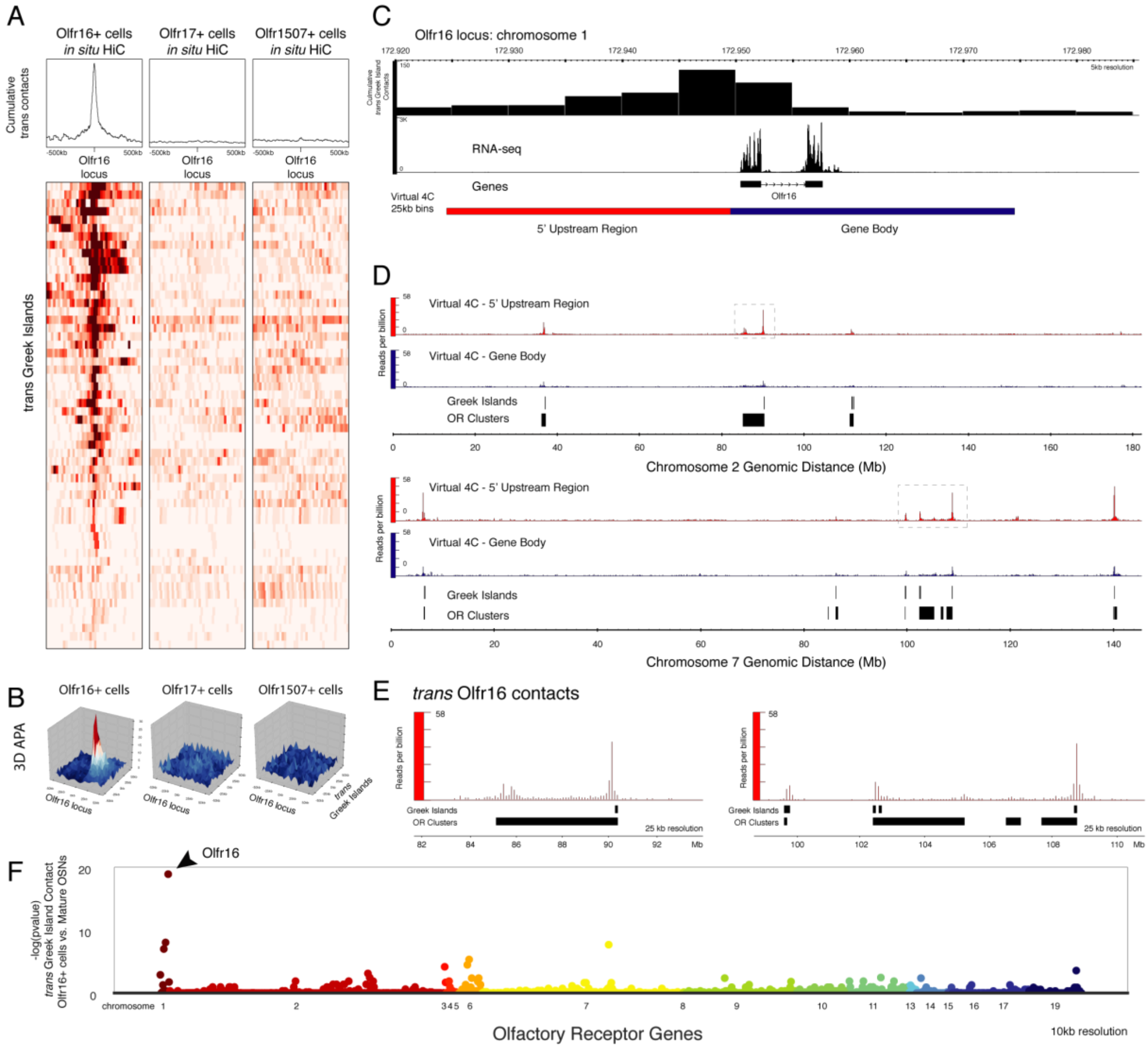


Figure 21. Specific *trans* interactions between the transcriptionally active Olfr16 gene locus and multiple Greek Islands. A. Heatmap depicting interchromosomal contacts between Olfr16 (chromosome 1) and Greek Islands from different chromosomes in *in situ* HiC from Olfr16^{+}, Olfr17^{+} and Olf1507^{+} cells. **B.** APA of the Olfr16 locus and *trans* Greek Islands in the three specific mOSN populations. **C.** *trans* Greek Islands make increased contacts on the 5' end of Olfr16 that contains the promoter of Olfr16. **D.** Virtual 4C from two 25kb bins surrounding the Olfr16 allele (5' end in red, gene body in blue) reveals extremely specific interchromosomal contacts between Olfr16 5' region and Greek Islands in Olfr16+ cells. **E.** Zoomed-in views of dotted boxes in (d). **F.** Manhattan plot of Greek Island contacts onto OR genes reveals that in Olfr16^{+} cells, Greek Islands are most likely to contact Olfr16 when compared to heterogeneous mOSNs.

Similar observations are made for Olfr17 and Olfr1507, which interact with a plethora of Greek Islands in *trans* only in the OSNs that are transcribed (Fig. 19). As described for the *cis* contacts with Lesvos, H makes strong contacts with numerous Greek Islands also in *trans* regardless of the identity of the chosen OR, but the H-proximal Olfr1507 is privy to these interactions only in Olfr1507 OSNs (Fig. 19B,E).

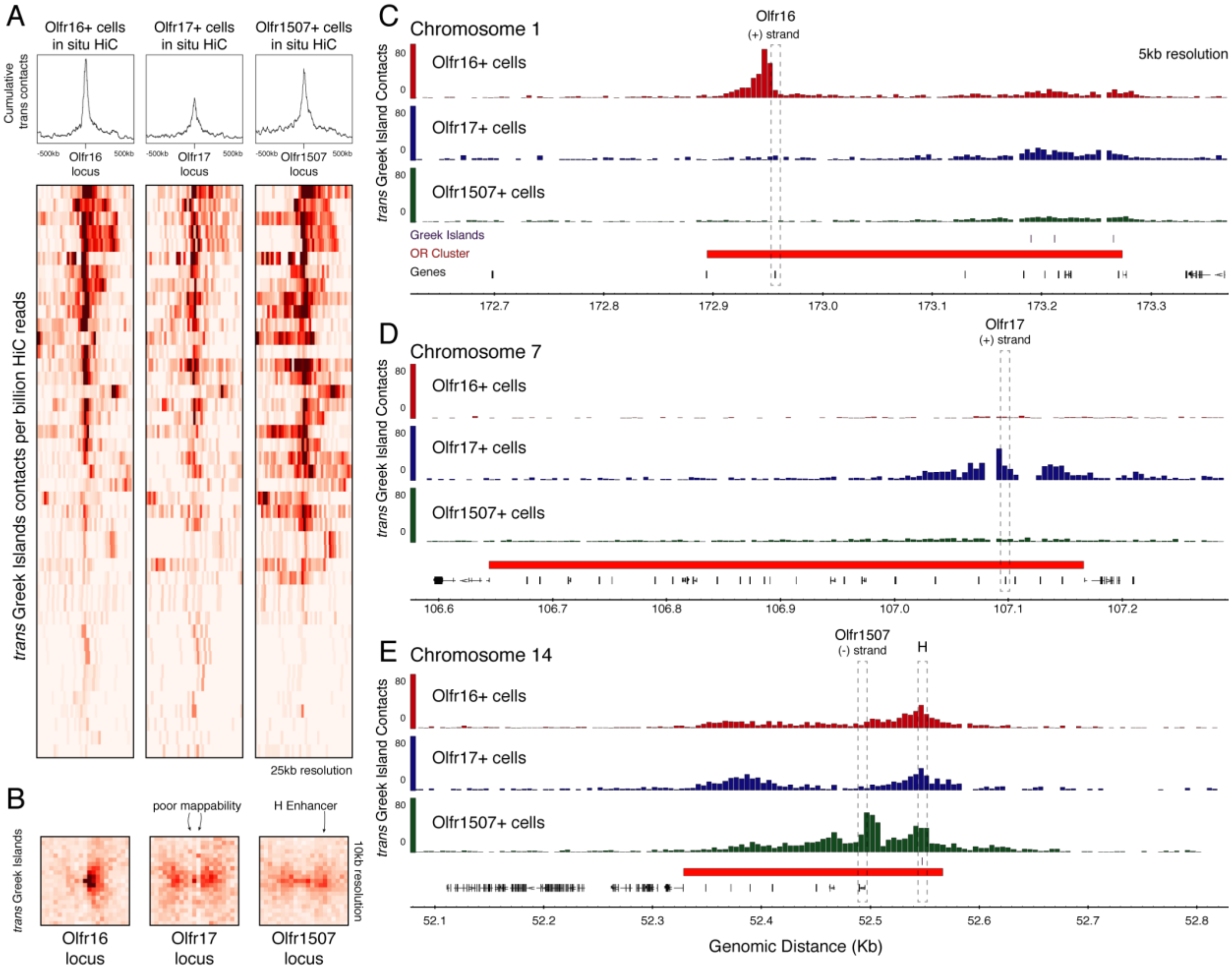


Figure 22. The active OR allele makes contacts with Greek Islands in trans. A. Heatmaps for contacts between Olfr16, Olfr17, or Olfr1507 and trans Greek Islands reveals an accumulation of contacts centered around the active allele. **B.** APA for an OR vs trans Greek Islands shows the accumulation of contacts on the active allele at 10kb resolution. The poor mapability of the Olfr17 locus perturbs the expected focal peak. The presence of the Greek Island, H, 50kb from Olfr1507 also contributes to the perceived "spreading" of Greek Island contacts on the Olfr1507 locus in the OSNs that is not transcribed, however in Olfr1507+ cells there is an increase of trans interactions with the active Olfr1507 gene. **C,D,E**. trans Greek Island contacts accumulate on the 5' end of the active allele at the Olfr16 (c), Olfr17 (d), and Olfr1507 (e).

My experiments show that interchromosomal interactions between genic regions exist, are highly specific, and occur with remarkable stereotypy across OSNs. The exceptionally high frequencies of Greek Island interactions suggest that multiple Islands interact with each other in each mOSN, forming a hub that associates with the active OR locus. Unlike previously proposed transcription factories (Osborne et al., 2004; Schoenfelder et al., 2010), the Greek Island hub is extremely selective in regards to the number of interacting genes, as only a single OR locus makes stereotypic contacts with this hub in a given OSN sub-population. The mechanism that prevents additional OR loci from associating with a Greek Island hub remains unknown and so does the mechanism that instructs the remarkable specificity of Greek Island interactions in *cis* and *trans,* since the factors necessary for these interactions have thousands of peaks in the OSN genome. In any case, specific interactions between Greek Islands in *cis* and *trans* are essential for OR transcription, since genetic manipulations that disrupt this multi-chromosomal Greek Island hub result in significant downregulation of OR transcription (Monahan et al., 2019). Thus, our *in situ HiC* experiments uncover a differentiation dependent transition in nuclear architecture that essentially eliminates topological restrictions imposed by chromosomes, allowing the formation of interchromosomal interactions of unprecedented frequency and specificity. Although these interactions are reproducible enough to be detected in mixed mOSN populations, *in situ* HiC of molecularly identical OSN subtypes reveals subtle differences in the contacts between OR clusters and Greek Islands. OSN subtype-specific nuclear compartmentalization may reduce OR gene choice to a selection of one out of few OR loci that are stochastically placed in the optimal distance from a Greek Island hub, explaining deterministic restrictions in OR gene expression (Ressler et al., 1993; Vassar et al., 1993). Extrapolating our findings to other cell types and gene families, we propose that interchromosomal interactions occurring only within subtypes of, otherwise homogeneous, cell populations, may be responsible for variegated transcription programs that are yet unappreciated (Nagano et al., 2013). Although these interactions, and their presumed transcriptional consequences, are currently viewed as “noise”, there are many examples where increased transcriptional variation is desirable and biologically beneficial (Johnston and Desplan, 2014; Lefebvre et al., 2012; Mountoufaris et al., 2017; Raser and O'Shea, 2005). The nervous system, with astounding numbers of post-mitotic cell types, may offer the ideal setting for this diversity-generating mechanism of gene regulation.

Chapter 9:

Clustered Protocadherins

Clustered protocadherins are a family of cell adhesion molecules used extensively by the nervous system. These proteins are localized to the cell membrane and are critical for neural circuit assembly and maintenance, largely through self-avoidance homophilic interactions (Fig. 20) (Chen et al., 2017; Lefebvre et al., 2012; Mountoufaris et al., 2017; Zipursky and Sanes, 2010). In the olfactory epithelium, deletion of the clustered protocadherins causes axonal arborization defects. Remarkably, forcing the expression of a single protocadherin combination of alpha, beta, and gamma in all OSNs leads to the lack of glomerulus formation (Fig. 20C) (Mountoufaris et al. 2017). Interstingly, this unimolecular pcdh combination does not affect OR choice or OSN maturuation, highlighting its mostly orthogonal function. Collectively, these studies underscore the functional importance of clustered protocadherins in normal circuit assembly.

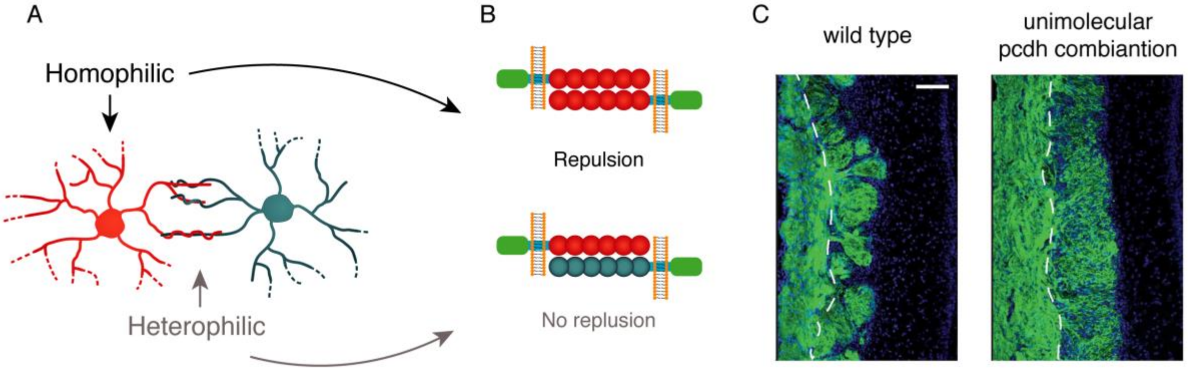


Figure 23. Homophilic repulsion by clustered protocadherins. A. Protocadherins mediate dendritic repulsion. Adapted from Lefebvre et al. 2012. **B.** Homophilic interactions cause repulsion. **C.** Mature OSNs expressing a uniform set of protocadherins fail to form glomeruli in the olfactory bulb. Adapted from Mountoufaris et al. 2017.

As previously discussed (Fig. 1D), protocadherins are stochastically generated from a peculiar genic arrangement (Wu and Maniatis, 1999). The realization that their organization resembles that of the TCRs, suggested that these genes were involved in generating neural diversity. Indeed, protocadherins contribute to diversity at the cell surface, albeit through a different transcriptional mechanism than recombination found in TCRs.

Through a process of stochastic promoter choice and splicing, protocadherins generate remarkable diversity (Chen and Maniatis, 2013; Guo et al., 2012; Monahan et al., 2012; Tasic et al., 2002). To understand the mechanisms generating such diversity, my work focused on protocadherin alpha promoter choice. The heavy conservation from teleost to human permits mechanistic studies in across multiple model systems (Ribich et al., 2006). The alpha protocadherin locus is characterized by 12 alternate exons in mouse, and 13 in human. One of these alternate exons constitutes the extracellular domain and 3 constant exons make up the transmembrane and intracellular domain (Fig. 21). The locus contains several highly conserved hypersensitivity sites throughout, including HS5-1 and HS7 (not shown), which are required for proper expression (Ribich et al., 2006). These transcriptional enhancers are tissue specific, reporting activity only in the nervous system. Further work revealed that HS5-1 binds CTCF together with cohesin at two distinct sites. HS7 was shown to only bind cohesin (Rad21) (Monahan et al., 2012). With the current understanding of *cis* regulatory elements, I speculate that HS7 is the cohesin-loading region since cohesin must load somewhere between two convergent CTCF sites. This would predict that HS7 would also bind Nipbl, the cohesin loading factor.

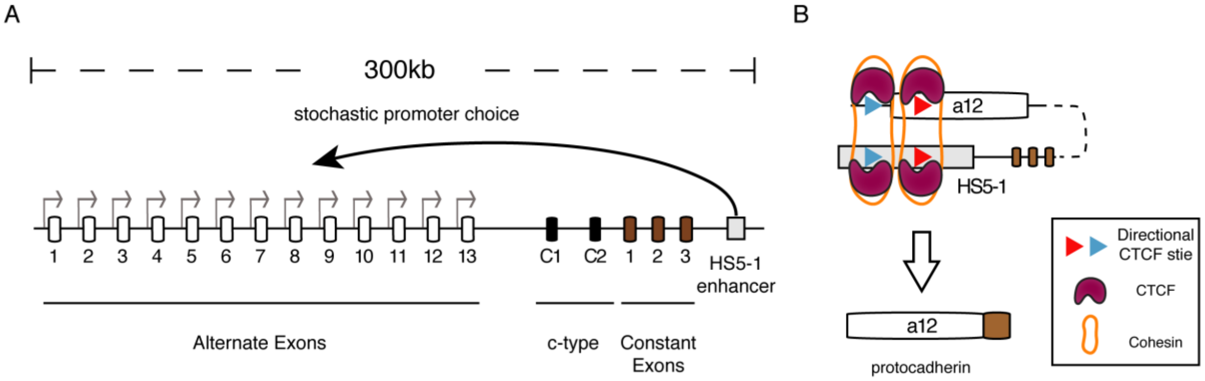


Figure 24. The protocadherin alpha locus is an array of stochastically chosen extracellular domains. A. The 300kb human protocadherin locus contains several notable features, including alternate exons, constant exons, and the HS5-1 enhancer required for expression. **B.** CTCF/cohesin mediated looping regulates protocadherin promoter choice.

Chapter 10:

Stochastic demethylation drives protocadherin choice

Analysis of the Hi-C data from Ngn1^{+} and Omp^{+} cells revealed architectural “stripes” along the Pcdh-alpha gene cluster (Fig. 22,24A), a feature that has been associated with Cohesin activity in the assembly of promoter/enhancer complexes during DNA loop-extrusion (Vian et al., 2018). A prediction of the DNA loop-extrusion model for the assembly of a Pcdh-alpha promoter/enhancer complex is that uncoupling CTCF binding to Pcdh-alpha promoters from DNA looping to the HS5-1 enhancer by the Cohesin complex should result in an overall loss of expression of all Pcdh-alpha exons. To test this possibility, Kevin Monahan conditionally deleted the Cohesin subunit, Rad21, in mouse olfactory sensory neurons (Fig. 23A) using OMPiresCre. With this driver, Rad21 is deleted in post-mitotic, fully differentiated, OSNs in which Pcdh-alpha promoter choice has already occurred (Fig. 23B). However, upon deletion of Rad21, a loss of long-range DNA contacts between the Pcdh-alpha promoters and the HS5-1 enhancer was observed (Fig. 24A,B). More importantly, loss of DNA contacts correlated with a significant loss of expression of all Pcdh-alpha exons as determined by RNA-Seq (Fig. 24C). Thus, continuous Cohesin activity appears to be required for the maintenance of DNA looping in the Pcdh-alpha cluster, even in the absence of cell division.

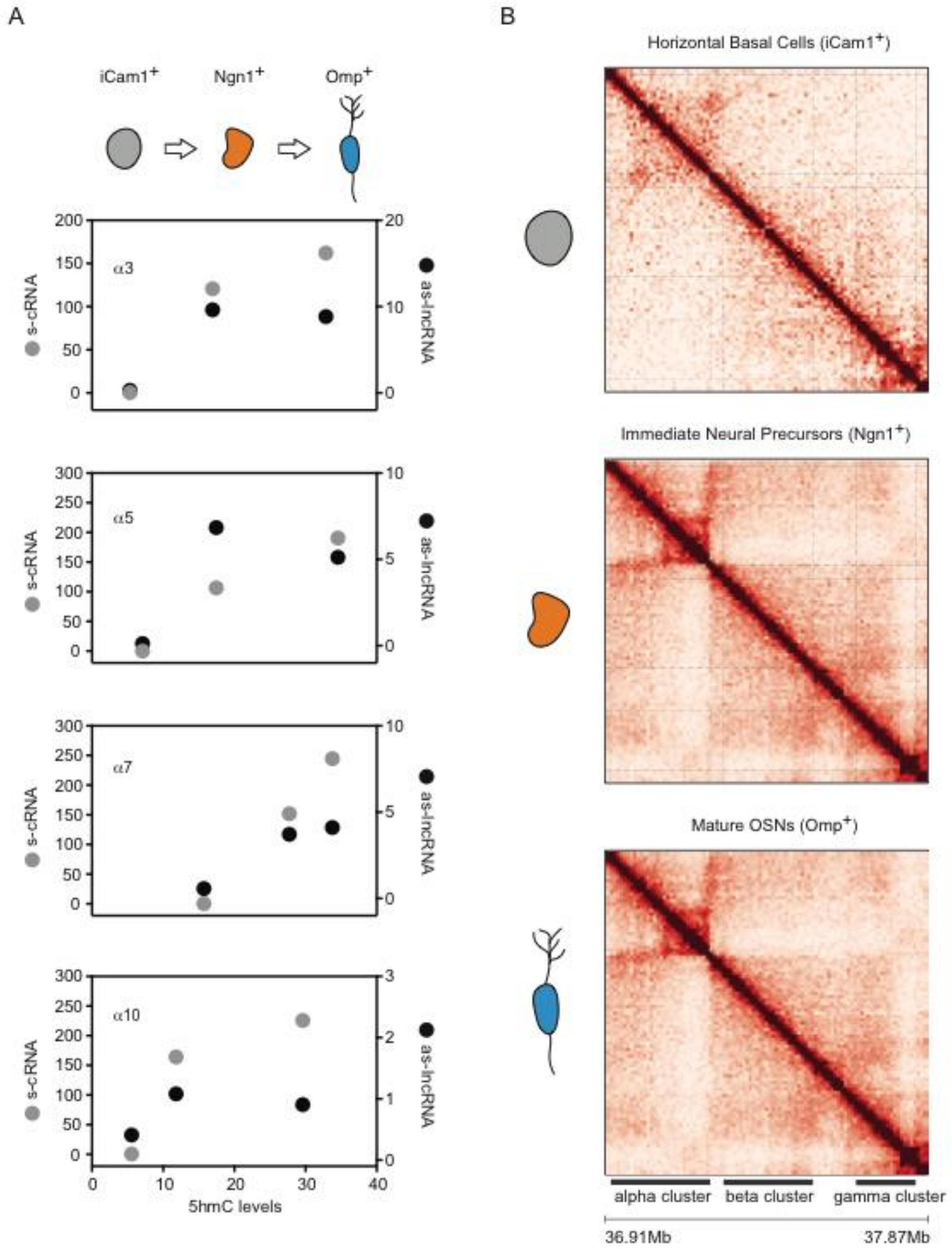


Figure 25. DNA demethylation correlates with Pcdh-alpha expression in vivo. A. Changes in 5hmC (x-axis) relative to the expression of the s-cRNA (left y-axis, grey) and the as-lncRNA (right x-axis, black) during the maturation of olfactory sensory neurons. Data for Pcdh -alpha3, -alpha5, -alpha7 and -alpha10 are shown. **B.** In situ Hi-C contact maps at 10kb resolution for horizontal basal cells (ICAM1+, Top), immediate neural precursors (Ngn1+, Middle) and mature olfactory sensory neurons (Omp+, Bottom). Daniele Canzio generated data for **A.** I generated data in **B**.

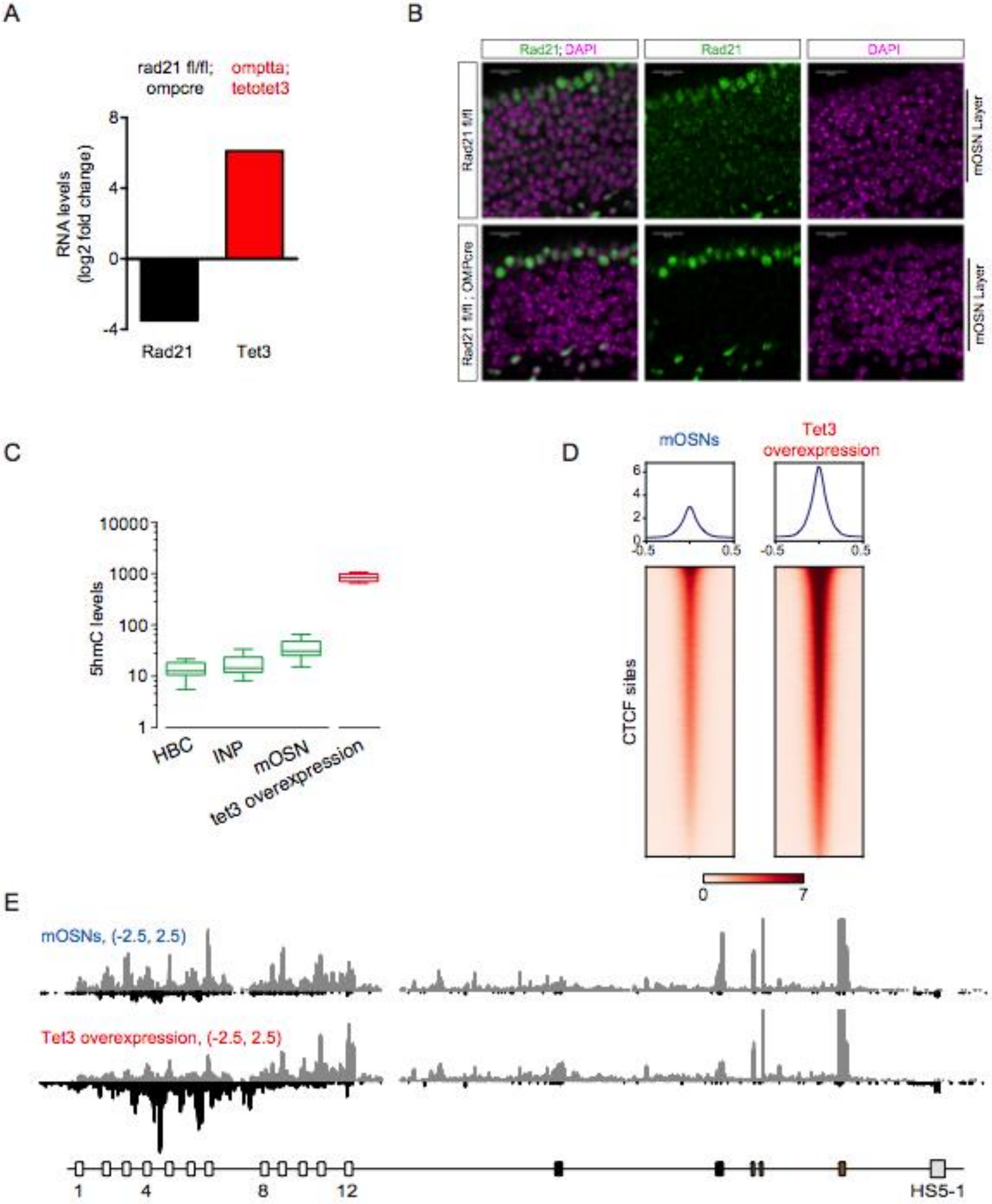

A
rad21 fl/fl;
ompcre
omptta;
tetotet3
RNA levels
(log2 fold change)
8
4
0
-4
Rad21
Tet3
B
Rad21; DAPI
Rad21
DAPI
Rad21 fl/fl
Rad21 fl/fl ; OMPcre
mOSN Layer
mOSN Layer
C
5hmC levels
10000
1000
100
10
1
HBC
INP
mOSN
tet3 overexpression
D
mOSNs
Tet3
overexpression
6
4
2
0
-0.5
0.5
-0.5
0.5
CTCF sites
0
7
E
mOSNs, (-2.5, 2.5)
Tet3 overexpression, (-2.5, 2.5)
1
4
8
12
HS5-1

Figure 26: Rad21 knockout and Tet3 overexpression in mature olfactory sensory neurons. A. Log2 fold change for Rad21 from Rad21fl/fl;OMPcre mice and Tet3 from tetotet3iresGFP;omptta relative to mOSNs from control mice. **B.** Rad21 immunofluorescence (green) in MOE sections from 14-week-old control (Rad21-fl/fl) and Rad21 KO (Rad21 fl/fl;OMPcre) mice. Nuclei are stained with DAPI (magenta). Rad21 is lost from mOSNs but retained in apical sustentactular cells and basal immature cells. Scale bar = 20μm. **C.** Average of cumulative RPM values for the Pcdh-alpha alternate promoters/exons for 5hmC for horizontal basal cells (HBC), immediate neural precursors (INP), and control or Tet3 overexpressing mature olfactory sensory neurons (mOSN). **D.** CTCF profiles in mOSNs (Left) and mOSNs overexpressing Tet3 (Right) as measured by ChIP-Seq. **E**. RNA-Seq profiles for s-cRNA (grey) and as-lncRNA (black) in mOSNs and mOSNs upon Tet3 overexpression. The x-axis represents the linear sequence of the genomic organization of the mouse Pcdh-alpha cluster and the numbers on the left-hand side of each track represent the minimum and maximum densities in read per million. Data generated by Kevin Monahan and Daniele Canzio.

Figure 7

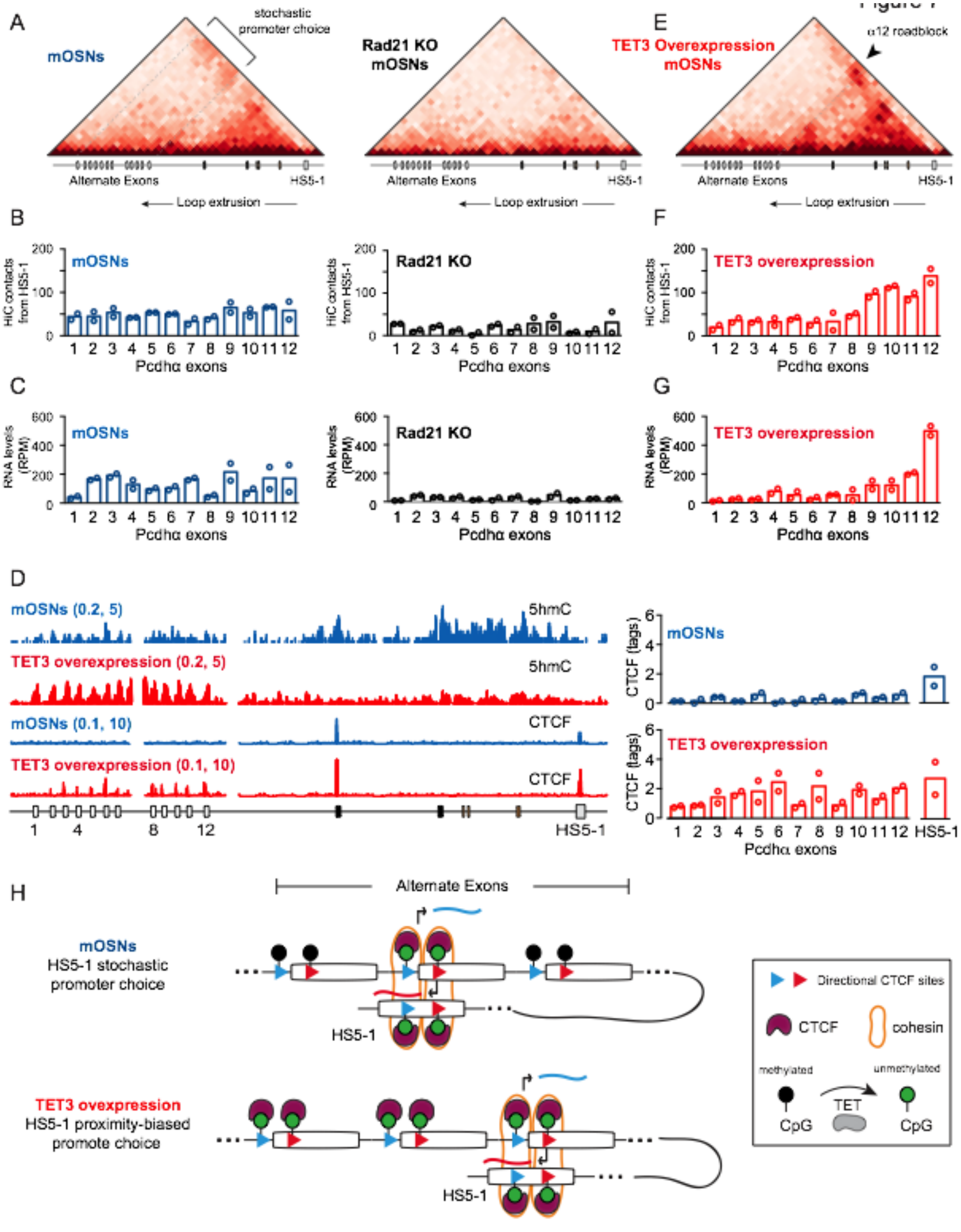

A
mOSNs
stochastic promoter choice
Rad21 KO mOSNs
E
TET3 Overexpression mOSNs
α12 roadblock
Alternate Exons
HS5-1
Loop extrusion
B
HiC contacts from HS5-1
200
100
0
mOSNs
Rad21 KO
F
TET3 overexpression
1 2 3 4 5 6 7 8 9 10 11 12
Pcdhα exons
C
RNA levels (RPM)
600
400
200
0
mOSNs
Rad21 KO
G
TET3 overexpression
D
mOSNs (0.2, 5)
5hmC
TET3 overexpression (0.2, 5)
5hmC
mOSNs (0.1, 10)
CTCF
TET3 overexpression (0.1, 10)
CTCF
1
4
8
12
HS5-1
CTCF (fags)
6
4
2
0
mOSNs
TET3 overexpression
1 2 3 4 5 6 7 8 9 10 11 12 HS5-1
Pcdhα exons
H
Alternate Exons
mOSNs
HS5-1 stochastic promoter choice
HS5-1
TET3 ovexpression
HS5-1 proximity-biased promote choice
HS5-1
Directional CTCF sites
CTCF
cohesin
methylated
unmethylated
TET
CpG
CpG

Figure 27. Stochastic DNA demethylation ensures random Pcdh-alpha promoter choice by the CTCF/Cohesin proteins via DNA loop-extrusion. A. Hi-C contacts maps at 10kb resolution for the Pcdh-alpha cluster in mOSNs (Left) and mOSNs upon Rad21 conditional knockout, Rad21 KO (Right); max: 100 reads per billion Hi-C contacts. **B,C.** Average HiC contacts of the HS5-1 enhancer with the individual Pcdh-alpha promoters (B) and average RPM values of s-cRNA for individual Pcdh-alpha exons (C) in mOSNs (Blue) and mOSNs upon Rad21 conditional knockout (Black). **D.** Left: 5hmC (MeDIP-Seq) and CTCF (ChIP-Seq) profiles in mOSNs (Blue) and mOSNs upon Tet3 overexpression (Red). Right: Quantification of CTCF binding. **E.** Hi-C contact maps at 10kb resolution for the Pcdh-alpha cluster in mOSNs overexpressing Tet3; max: 100 reads per billion Hi-C contacts. **F,G.** Average HiC contacts of the HS5-1 enhancer with the individual Pcdh-alpha promoters (F) and average RPM values of s-cRNA for individual Pcdh-alpha exons (G) mOSNs overexpressing Tet3. **H.** Model for how coupling of as-lncRNA transcription and DNA demethylation ensures a stochastic and HS5-1 distance-independent choice of a Pcdh-alpha promoter. Uncoupling DNA demethylation from as-lncRNA transcription by overexpression of Tet3 results in non-random and HS5-1 distance-biased Pcdh-alpha promoter choice. Kevin Monahan generated Rad21 KO HiC data. Daneiel Canzio generated RNA-seq, CTCF, and 5hmC data. I generated mOSN and Tet3 HiC data.

These data are consistent with a model in which CTCF acts as a boundary element for the Cohesin complex to mediate long-range interactions between Pcdh-alpha promoters and the HS5-1 enhancer. In the context of our methylation data and the mechanistic coupling of demethylation to CTCF binding, this model predicts that formation of long-range DNA contacts between a Pcdh-alpha promoter and the HS5-1 enhancer in individual neurons is stochastic and distance-independent with respect to HS5-1. I propose that this enhancer/promoter engagement is achieved by virtue of random demethylation of Pcdh-alpha promoters. According to this model, random demethylation of one of the Pcdh-alpha exons, as a consequence of as-lncRNA transcription, ensures that only one exon is bound to CTCF, and thus results in the assembly of a specific Pcdh-alpha promoter/HS5-1 enhancer complex. A prediction of this model is that uncoupling DNA demethylation from antisense lncRNA transcription results in a non-random choice of Pcdh-alpha promoters by the HS5-1 enhancer. To uncouple as-lncRNA transcription from DNA demethylation, we overexpressed Tet3 in OSNs (Fig. 23A). Tet3 is the most highly expressed Tet protein in OSNs, and has been shown to associate with the Pcdh-alpha promoters in differentiated neuronal precursor cells (Li et al., 2016). Overexpression of Tet3 resulted in strong demethylation of Pcdh-alpha promoters, as indicated by a large increase in 5hmC levels (Fig. 24D, 23C) and by an increase of CTCF binding to CBS sites genome-wide (Fig. 23D), and to all Pcdh-alpha exons, irrespective of transcription of their cognate as-lncRNAs (Fig. 24D, 23E). To address the function of uncoupling as-lncRNA transcription from stochastic DNA demethylation, I performed Hi-C and RNA-Seq in mOSNs overexpressing Tet3. Remarkably, despite the fact that all Pcdh-alpha exons are bound by CTCF, and that the expression of the as-lncRNAs is maintained (Fig. 24D, 23E), overexpression of Tet3 resulted in a strong bias in Pcdh-alpha promoter/HS5-1 enhancer contacts biased towards the Pcdh-alpha12 promoter (Fig. 24E,F) and a concomitant bias in Pcdh-alpha12 expression relative to all other Pcdh-alpha exons, as determined by RNA-Seq (Fig. 24G). Thus CTCF bound to the CBS sites of Pcdh-alpha12 created a “roadblock” for Cohesin, preventing the HS5-1 enhancer from engaging with any of the upstream Pcdh-alpha promoters. These data are consistent with a model in which coupling antisense lncRNA transcription to DNA demethylation ensures random choice of Pcdh-alpha promoters *in vivo* (Fig. 24H).

Stochastic, combinatorial expression of individual Pcdh protein isoforms in Purkinje (Esumi et al., 2005) and olfactory sensory neurons (Mountoufaris et al., 2017) generates distinct combinations of Protocadherin isoforms that function as a cell-surface identity code for individual neurons. This conclusion has been confirmed more broadly through single cell RNA sequencing studies in a variety of neuronal cell types (Tasic et al., 2018). Here we identify a mechanism by which Pcdh-alpha alternate exon promoters are stochastically activated in individual neurons, and propose a model that may apply more broadly in promoter choice and gene expression in vertebrates.

Chapter 11:

Discussion and Conclusions

The observations made during my thesis work regarding the 3D genome architecture of olfactory neurons raise important questions about the mechanism of singular OR gene choice and the applicability of our findings to other biological systems. In this chapter I will address the most important open questions and provide an overview of experiments that I believe should be performed in the future.

1. Does the assembly of a multi-chromosomal hub facilitate transcriptional singularity?

Bulk *in situ* HiC experiments from FAC-sorted OSNs revealed that we can infer the chosen OR gene from the frequency of contact with the OR enhancers in *trans*. These observations suggest that in every OSN, the expressed OR gene locus interacts with multiple OR enhancers in *cis* and *trans*. However population-wide HiC data cannot exclude the existence of secondary hubs that associate with a different OR, or the possibility that additional OR alleles also associate with a singular hub, but are not detected in population studies because they differ between cells. In other words, our data cannot exclude the possibility that in Olfr16-expressing cells, additional OR alleles associate with the hub that expresses Olfr16 or with additional hubs that may form in each cell. These questions can only be tackled by single-cell experiments that interrogate the 3D distribution of all the OR alleles at high resolution. Unfortunately, this is not currently possible. For example, imaging-based experiments can provide high spatial resolution but cannot visualize all the OR alleles and OR enhancers simultaneously. Ongoing experiments in our lab interrogate the physical association of the transcriptionally active Olfr17 allele with 30 OR enhancers with single molecule DNA FISH, which will reveal the physical distribution of enhancers over the active OR but cannot reveal the presence of additional OR alleles in the hub. Genomic approaches, on the other hand, may provide genome-wide information on OR gene distribution but low spatial resolution due to the scarcity of the data. For example, a recent study performed single cell HiC in olfactory neurons (Tan et al., 2019), confirming the bulk HiC data produced by my thesis, but failing to reveal the exact relationship between OR transcription and enhancer interactions due to limitations imposed by sequence coverage. That said, 3D modeling of chromosomal folding using single cell HiC data suggest that OSNs may contain two dominant multi-enhancer hubs. If future experiments confirm this prediction, then there are two possible models explaining transcriptional singularity in the presence of two multi-enhancer hubs: First, only one of the two hubs may be transcriptionally competent, i.e. a key transcriptional activator may be selectively recruited in only one of the two hubs. Second, both enhancer hubs may be functional, but because OR gene choice operates under kinetic restrictions imposed by the OR-elicited feedback, a “winner takes all” process may prevent the second enhancer hub from removing heterochromatin marks from a second OR allele.

2. Specificity of OR-OR and enhancer-enhancer interactions

The second immediate question emerging from my observations relates to the specificity of the interactions between the inactive OR genes and the active enhancers during OSN differentiation. Even as genomic compartmentalization becomes more elaborate with the realization that repressive compartments are further segregated to compartments containing facultative and constitutive heterochromatin, the generation of gene-specific compartments is unusual. In fact the nucleolus represent the only example of gene-specific compartmentalization, however, in this case the converging gene are under control of a unique set of transcriptional regulators and RNA polymerase subunits.

Although it has been suggested that histone marks are correlated with the specificity of compartments (Rao et al., 2014), our data suggest that this may not be the full story. For example inactive OR genes are coated with H3K9me3, the same histone mark found on constitutive heterochromatin (Magklara et al., 2011), however, we know that the inactive OR gene compartment is distinct from these regions (Clowney et al., 2012). It remains possible that there may be other histone marks and protein readers that increase the combinatorial complexity of these interactions. However, preliminary data from our lab suggest that polygenic low-level expression of OR genes during differentiation may play a role in compartmentalization (Hanchate et al., 2015). Artificially driving high expression of a transgenic Olfr17 with an inducible TetO system during the progenitor stage increases the frequency of interaction with the OR hub and the OR enhancers by *in situ* HiC. This raises the possibility that the transcript or an RNA binding protein may play a crucial role in forming the inactive OR hub.

The second specific compartment uncovered in this work is the active OR enhancer compartment. We showed that these sequences are bound by Ldb1 and require this protein for the stabilization of this hub (Monahan et al., 2019). However, as Ldb1 regulates many genes across the genome, it remains a mystery how Ldb1 facilitates long-range *cis* and *trans* interactions only between OR enhancers. This suggests that the identity of Ldb1 binding to OR enhancers is distinct from the Ldb1 that bind other genomic regions, perhaps through post-transcriptional and/or post-translational modifications. A key insight is that the OR enhancer hub forms after the inactive OR genes aggregate, despite Ldb1 binding OR enhancers earlier during differentiation. This implies that the OR gene hub should form *before* the OR enhancer hub. It also suggests that Ldb1 may be post-translationally modified or pair with co-factors to allow compartmentalization while it is bound to the OR enhancer sequence. This again may involve the OR gene transcript to ensure specificity and temporal regulation of this process. Indeed, driving Olfr17 in the OSN lineage ensures that it is chosen with a higher frequency and stably expressed even after removal of doxycycline in the TetO-inducible system.

3. The contribution of *cis* vs *trans* genomic interactions in OR gene regulation.

The concept of *trans* enhancement is somewhat controversial and has only been genetically demonstrated in flies (transvection). In the olfactory system, deletion of OR enhancers results only in the downregulation of OR alleles that reside in the same chromosome with the deleted Greek Island. However, my HiC data from OSNs expressing a common OR allele, combined with HiC data from triple enhancer KO mice, may explain why OR enhancers are essential in *cis* and redundant in *trans*. Each one of the three Greek Islands for which we have genomic deletions, appears essential for the recruitment of trans and long-range cis enhancers to the OR cluster harboring this Island. Extrapolating these observations to the other 60 Greek Islands, it appears that each OR cluster uses a *cis* enhancer for the recruitment of *trans* enhancers. In contrast, because my HiC data showed that up to 40 enhancers associate with the active OR, deletion of 1 or few OR enhancers should not affect global OR transcription since other enhancers can substitute their function in the hub. In contrast, experiments that led to the physical disruption of the Greek Island hub, through deletion of Ldb1, resulted in significant and widespread downregulation of OR transcription.

4. The generality of our observations.

Finally, an important question posed by my HiC data is whether my observations are applicable to other biological systems. Because OR genes can be found in 18/20 mouse chromosomes, have a clustered genomic arrangement, and constitute the largest gene family (>1100 genes, 68 clusters covering ~36Mb), detection of their interchromosomal contacts by *in situ* HiC is robust and unequivocal, transforming our understanding of genomic compartmentalization (Fig.1). However, the fact that specific interactions between chromosomes are easier to detect when dealing with large genomic clusters does not mean that they are not occurring in other biological systems. For example, recent observations suggest that super-enhancers (SEs) tend to form interchromosomal compartments, a result that was first observed by GAM (Beagrie et al., 2017), and further confirmed by *in situ* HiC (Rao et al., 2017a; Schwarzer et al., 2017). Furthermore, split-pool recognition based methods (SPRITE) revealed robust multi-chromosomal interactions organized by nuclear RNA speckles (Quinodoz et al., 2018), which act as transcriptional amplifiers (Kim et al., 2019). The non-coding RNA Firre has been shown to regulate interchromosomal interactions (Hacisuleyman et al., 2014; Maass et al., 2018), whereas the nascent Ttn RNA coordinates the formation of a mutli-chromosomal gene hub during cardiogenesis, coordinating the alternative splicing of cardiomyocyte-specific genes (Bertero et al., 2019). Moreover, in the case of human antiviral responses, multi-chromosomal transcription factor "repositories" appear essential for stochastic and monoallelic activation of IFN beta (Apostolou and Thanos, 2008; Nikopoulou et al., 2018). As proposed for the OR enhancer hubs, these multi-chromosomal hubs concentrate locally transcription factor NFkappaB, allowing stable binding on the IFN enhanceosome. Similarly, imaging studies in flies suggest that multi-enhancer chromatin hubs formed during development confer transcriptional robustness by concentrating locally ultrabithorax (Ubx) (Tsai et al., 2019). Such multi-enhancer hubs are also forming in ES cells, concentrating transcription factor Sox2 in nuclear sub-compartments (Liu et al., 2014), whereas the Nanog locus itself appears regulated by a multi-chromosomal hub in these cells (Apostolou et al., 2013). Finally, there are multiple classic examples of interchromosomal interactions regulating mutually exclusive choices, such as X-chromosome inactivation (Masui et al., 2011; Xu et al., 2006), photoreceptor gene choice (Johnston and Desplan, 2014), and Th1/Th2 lymphocyte differentiation (Spilianakis et al., 2005). Notably, single cell HiC studies also revealed extensive interchromosomal contacts in mouse photoreceptor neurons (Tan et al., 2019) without a known physiological role. Given that OR compartments and OR enhancer hubs depend on proteins with widespread developmental functions (Caputo et al., 2015; Kiefer et al., 2011; Kim et al., 2016; Landeira et al., 2009; Li et al., 2011; Mangale et al., 2008; Müller et al., 2018; Subramanian et al., 2011; Yun et al., 2009; Zhao et al., 2014, 2007), our findings are likely applicable to other biological systems where robust but not fully deterministic gene choices prescribe cellular identity. On this note, gene regulation by *trans* genomic interactions has immense significance for the health of hundreds of millions of people infected by parasitic protozoa worldwide. For example, multi-chromosomal interactions are used by trypanosome (Müller et al., 2018) and plasmodium (Bunnik et al., 2018, 2019) to regulate the monogenic expression of surface glycoproteins, VSG and VAR genes, respectively. The genomic compartmentalization of these multigene families is essential for a process known as antigenic variation (Landeira et al., 2009; Müller et al., 2018), which constitutes a key mechanism for avoidance immunological detection by infected human hosts. Thus, understanding the genomic and molecular principles that allow convergence of VARs and VSGs, which represent large AT-rich gene families scattered around chromosomes, like the OR genes, and deciphering how one of these genes escapes repressed compartments to become robustly activated, has high clinical significance.

References

1. Apostolou, E., and Thanos, D. (2008). Virus Infection Induces NF-κB-Dependent Interchromosomal Associations Mediating Monoallelic IFN-β Gene Expression. Cell *134*, 85–96.
2. Apostolou, E., Ferrari, F., Walsh, R.M., Bar-Nur, O., Stadtfeld, M., Cheloufi, S., Stuart, H.T., Polo, J.M., Ohsumi, T.K., Borowsky, M.L., et al. (2013). Genome-wide chromatin interactions of the nanog locus in pluripotency, differentiation, and reprogramming. Cell Stem Cell *12*, 699–712.
3. Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C.A., Chotalia, M., Xie, S.Q., Barbieri, M., de Santiago, I., Lavitas, L.-M., Branco, M.R., et al. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. Nature.
4. Bertero, A., Fields, P.A., Ramani, V., Bonora, G., Yardimci, G.G., Reinecke, H., Pabon, L., Noble, W.S., Shendure, J., and Murry, C.E. (2019). Dynamics of genome reorganization during human cardiogenesis reveal an RBM20-dependent splicing factory. Nat. Commun. *10*, 1538.
5. Boeynaems, S., Alberti, S., Fawzi, N.L., Mittag, T., Polymenidou, M., Rousseau, F., Schymkowitz, J., Shorter, J., Wolozin, B., Van Den Bosch, L., et al. (2018). Protein Phase Separation: A New Phase in Cell Biology. Trends Cell Biol. *28*, 420–435.
6. Boija, A., Klein, I.A., Sabari, B.R., Dall'Agnese, A., Coffey, E.L., Zamudio, A. V., Li, C.H., Shrinivas, K., Manteiga, J.C., Hannett, N.M., et al. (2018). Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. Cell 1–14.
7. Bonev, B., Cohen, N.M., Szabo, Q., Hugnot, J., Tanay, A., Cavalli, G., Bonev, B., Cohen, N.M., Szabo, Q., Fritsch, L., et al. (2017). Multiscale 3D Genome Rewiring during Mouse Article Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell *171*,

557.e1-557.e24.

8. Bozza, T., Feinstein, P., Zheng, C., and Mombaerts, P. (2002). Odorant receptor expression defines functional units in the mouse olfactory system. J. Neurosci. *22*, 3033–3043.

9. Bozza, T., Vassalli, A., Fuss, S., Zhang, J.J., Weiland, B., Pacifico, R., Feinstein, P., and Mombaerts, P. (2009). Mapping of Class I and Class II Odorant Receptors to Glomerular Domains by Two Distinct Types of Olfactory Sensory Neurons in the Mouse. Neuron *61*, 220–233.

10. Buck, L., and Axel, R. (1991). A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell *65*, 175–187.

11. Bunnik, E.M., Cook, K.B., Varoquaux, N., Batugedara, G., Prudhomme, J., Cort, A., Shi, L., Andolina, C., Ross, L.S., Brady, D., et al. (2018). Changes in genome organization of parasite-specific gene families during the Plasmodium transmission stages. Nat. Commun. *9*.

12. Bunnik, E.M., Venkat, A., Shao, J., McGovern, K.E., Batugedara, G., Worth, D., Prudhomme, J., Lapp, S.A., Andolina, C., Ross, L.S., et al. (2019). Comparative 3D genome organization in apicomplexan parasites. Proc. Natl. Acad. Sci. *116*, 201810815.

13. Caputo, L., Witzel, H.R., Kolovos, P., Cheedipudi, S., Looso, M., Mylona, A., Van Ijcken, W.F.J., Laugwitz, K.L., Evans, S.M., Braun, T., et al. (2015). The Isl1/Ldb1 Complex Orchestrates Genome-wide Chromatin Organization to Instruct Differentiation of Multipotent Cardiac Progenitors. Cell Stem Cell *17*, 287–299.

14. Chalancon, G., Ravarani, C.N.J., Balaji, S., Martinez-Arias, A., Aravind, L., Jothi, R., and Babu, M.M. (2012). Interplay between gene expression noise and regulatory network architecture. Trends Genet. *28*, 221–232.

15. Chen, W. V., and Maniatis, T. (2013). Clustered protocadherins. Development *140*, 3297–3302.

16. Chen, W. V., Nwakeze, C.L., Denny, C.A., O'Keeffe, S., Rieger, M.A., Mountoufaris, G., Kirner, A., Dougherty, J.D., Hen, R., Wu, Q., et al. (2017). Pcdhαc2 is required for axonal tiling and assembly of serotonergic circuitries in mice. Science (80-.). *356*, 406–411.

17. Chess, A., Simon, I., Cedar, H., and Axel, R. (1994). Allelic inactivation regulates olfactory receptor gene expression. Cell *78*, 823–834.

18. Clowney, E.J., Legros, M.A., Mosley, C.P., Clowney, F.G., Markenskoff-Papadimitriou, E.C., Myllys, M., Barnea, G., Larabell, C.A., and Lomvardas, S. (2012). Nuclear aggregation of olfactory receptor genes governs their monogenic expression. Cell *151*, 724–737.

19. Dekker, J. (2002). Capturing Chromosome Conformation. Science (80-.). *295*, 1306–1311.

20. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

21. Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell: Supporting online material. Science *297*, 1183–1187.

22. Esumi, S., Kakazu, N., Taguchi, Y., Hirayama, T., Sasaki, A., Hirabayashi, T., Koide, T., Kitsukawa, T., Hamada, S., and Yagi, T. (2005). Monoallelic yet combinatorial expression of variable exons of the protocadherin-α gene cluster in single neurons. Nat. Genet. *37*, 171–176.

23. Fletcher, R.B., Das, D., Gadye, L., Street, K.N., Baudhuin, A., Wagner, A., Cole, M.B., Flores, Q., Choi, Y.G., Yosef, N., et al. (2017). Deconstructing Olfactory Stem Cell Trajectories at Single-Cell Resolution. Cell Stem Cell *20*, 817-830.e8.

24. Fukaya, T., Lim, B., and Levine, M. (2016). Enhancer Control of Transcriptional Bursting. Cell *166*, 358–368.

25. Fuss, S.H., Omura, M., and Mombaerts, P. (2007). Local and cis Effects of the H Element on Expression of Odorant Receptor Genes in Mouse. Cell *130*, 373–384.

26. Ganji, M., Shaltiel, I.A., Bisht, S., Kim, E., Kalichava, A., Haering, C.H., and Dekker, C. (2018). Real-time imaging of DNA loop extrusion by condensin. Science (80-.). *360*, 102–105.

27. Goldberg, A.D., Allis, C.D., and Bernstein, E. (2007). Epigenetics: A Landscape Takes Shape. Cell *128*, 635–638.

28. Goodbourn, S., Zinn, K., and Maniatis, T. (1985). Human β-interferon gene expression is regulated by an inducible enhancer element. Cell *41*, 509–520.

29. Guo, Y., Monahan, K., Wu, H., Gertz, J., Varley, K.E., Li, W., Myers, R.M., Maniatis, T., and Wu, Q. (2012). CTCF/cohesin-mediated DNA looping is required for protocadherin promoter choice. Proc. Natl. Acad. Sci. *109*, 21081–21086.

30. Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900–910.

31. Haarhuis, J.H.I., van der Weide, R.H., Blomen, V.A., Yáñez-Cuna, J.O., Amendola, M., van Ruiten, M.S., Krijger, P.H.L., Teunissen, H., Medema, R.H., van Steensel, B., et al. (2017). The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. Cell *169*, 693-707.e14.

32. Hacisuleyman, E., Goff, L.A., Trapnell, C., Williams, A., Henao-Mejia, J., Sun, L., McClanahan, P., Hendrickson, D.G., Sauvageau, M., Kelley, D.R., et al. (2014). Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. Nat. Struct. Mol. Biol. *21*, 198–206.

33. Hanchate, N.K., Kondoh, K., Lu, Z., Kuang, D., Ye, X., Qiu, X., Pachter, L., Trapnell, C., and Buck, L.B. (2015). Single-cell transcriptomics reveals receptor transformations during olfactory neurogenesis. Science (80-.). *350*, 1251–1255.

34. Jain, S., Ba, Z., Zhang, Y., Dai, H.Q., and Alt, F.W. (2018). CTCF-Binding Elements Mediate Accessibility of RAG Substrates During Chromatin Scanning. Cell *174*, 102-

116.e14.

35. Jessell, T.M. (2000). Neuronal specification in the spinal cord: inductive signals and transcriptional codes. Nat. Rev. Genet. *1*, 20–29.

36. Johanson, T.M., Coughlan, H.D., Lun, A.T.L., Bediaga, N.G., Naselli, G., Garnham, A.L., Harrison, L.C., Smyth, G.K., and Allan, R.S. (2017). No kissing in the nucleus: Unbiased analysis reveals no evidence of trans chromosomal regulation of mammalian immune development. BioRxiv.

37. Johnston, R.J., and Desplan, C. (2008). Stochastic neuronal cell fate choices. Curr. Opin. Neurobiol. *18*, 20–27.

38. Johnston, R.J., and Desplan, C. (2014). Interchromosomal Communication Coordinates Intrinsically Stochastic Expression Between Alleles. Science (80-.). *343*, 661–665.

39. Jothi, R., Balaji, S., Wuster, A., Grochow, J.A., Gsponer, J., Przytycka, T.M., Aravind, L., and Babu, M.M. (2009). Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. Mol. Syst. Biol. *5*.

40. Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., Van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

41. Khan, M., Vaes, E., and Mombaerts, P. (2011). Regulation of the probability of mouse odorant receptor gene choice. Cell *147*, 907–921.

42. Kiefer, C.M., Lee, J., Hou, C., Dale, R.K., Lee, Y.T., Meier, E.R., Miller, J.L., and Dean, A. (2011). Distinct Ldb1/NLI complexes orchestrate -globin repression and reactivation through ETO2 in human adult erythroid cells. Blood *118*, 6200–6208.

43. Kim, J., Khanna, N., and Belmont, A.S. (2019). Transcription Enhancement by Nuclear Speckle Association. BioRxiv.

44. Kim, S., Zhao, Y., Lee, J., Kim, W.R., Gorivodsky, M., Westphal, H., and Geum, D. (2016). Ldb1 Is Essential for the Development of Isthmic Organizer and Midbrain Dopaminergic

Neurons. Stem Cells Dev. *25*, 986–994.

45. Landeira, D., Bart, J.M., Van Tyne, D., and Navarro, M. (2009). Cohesin regulates VSG monoallelic expression in trypanosomes. J. Cell Biol. *186*, 243–254.

46. Larson, A.G., Elnatan, D., Keenen, M.M., Trnka, M.J., Johnston, J.B., Burlingame, A.L., Agard, D.A., Redding, S., and Narlikar, G.J. (2017). Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. Nature *547*, 236–240.

47. Le Gros, M.A., Clowney, E.J., Magklara, A., Yen, A., Markenscoff-Papadimitriou, E., Colquitt, B., Myllys, M., Kellis, M., Lomvardas, S., and Larabell, C.A. (2016). Soft X-Ray Tomography Reveals Gradual Chromatin Compaction and Reorganization during Neurogenesis In Vivo. Cell Rep. *17*, 2125–2136.

48. Lefebvre, J.L., Kostadinov, D., Chen, W. V, Maniatis, T., and Sanes, J.R. (2012). Protocadherins mediate dendritic self-avoidance in the mammalian nervous system. Nature *488*, 517–521.

49. Li, L., Jothi, R., Cui, K., Lee, J.Y., Cohen, T., Gorivodsky, M., Tzchori, I., Zhao, Y., Hayes, S.M., Bresnick, E.H., et al. (2011). Nuclear adaptor Ldb1 regulates a transcriptional program essential for the maintenance of hematopoietic stem cells. Nat. Immunol. *12*, 129–136.

50. Li, X., Yue, X., Pastor, W.A., Lin, L., Georges, R., Chavez, L., Evans, S.M., and Rao, A. (2016). Tet proteins influence the balance between neuroectodermal and mesodermal fate choice by inhibiting Wnt signaling. Proc. Natl. Acad. Sci. *113*, E8267–E8276.

51. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science (80-.). *326*, 289–293.

52. Liu, Z., Legant, W.R., Chen, B.-C., Li, L., Grimm, J.B., Lavis, L.D., Betzig, E., and Tjian, R. (2014). 3D imaging of Sox2 enhancer clusters in embryonic stem cells. Elife *3*, 1–29.

53. Lomvardas, S., Barnea, G., Pisapia, D.J., Mendelsohn, M., Kirkland, J., and Axel, R. (2006). Interchromosomal Interactions and Olfactory Receptor Choice. Cell *126*, 403–413.

54. Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.

55. Lyons, D.B., Allen, W.E., Goh, T., Tsai, L., Barnea, G., and Lomvardas, S. (2013). An epigenetic trap stabilizes singular olfactory receptor expression. Cell *154*, 325–336.

56. Maass, P.G., Barutcu, A.R., Weiner, C.L., and Rinn, J.L. (2018). Inter-chromosomal Contact Properties in Live-Cell Imaging and in Hi-C. Mol. Cell *69*, 1039-1045.e3.

57. Magklara, A., Yen, A., Colquitt, B.M., Clowney, E.J., Allen, W., Markenscoff-Papadimitriou, E., Evans, Z.A., Kheradpour, P., Mountoufaris, G., Carey, C., et al. (2011). An epigenetic signature for monoallelic olfactory receptor expression. Cell *145*, 555–570.

58. Mangale, V.S., Hirokawa, K.E., Satyaki, P.R. V., Gokulchandran, N., Chikbire, S., Subramanian, L., Shetty, A.S., Martynoga, B., Paul, J., Mai, M. V., et al. (2008). Lhx2 Selector Activity Specifies Cortical Identity and Suppresses Hippocampal Organizer Fate. Science (80-.). *319*, 304–309.

59. Maniatis, T., Goodbourn, S., and Fischer, J. (1987). Regulation of inducible and tissue-specific gene expression. Science (80-.). *236*, 1237–1245.

60. Markenscoff-Papadimitriou, E., Allen, W.E., Colquitt, B.M., Goh, T., Murphy, K.K., Monahan, K., Mosley, C.P., Ahituv, N., and Lomvardas, S. (2014). Enhancer interaction networks as a means for singular olfactory receptor expression. Cell *159*, 543–557.

61. Masui, O., Bonnet, I., Le Baccon, P., Brito, I., Pollex, T., Murphy, N., Hupé, P., Barillot, E., Belmont, A.S., and Heard, E. (2011). Live-cell chromosome dynamics and outcome of X chromosome pairing events during ES cell differentiation. Cell *145*, 447–458.

62. Mombaerts, P., Wang, F., Dulac, C., Chao, S.K., Nemes, A., Mendelsohn, M., Edmondson, J., and Axel, R. (1996). Visualizing an Olfactory Sensory Map. Cell *87*, 675–686.

63. Monahan, K., and Lomvardas, S. (2015). Monoallelic Expression of Olfactory Receptors. Annu. Rev. Cell Dev. Biol. *31*, annurev-cellbio-100814-125308.

64. Monahan, K., Rudnick, N.D., Kehayova, P.D., Pauli, F., Newberry, K.M., Myers, R.M., and Maniatis, T. (2012). Role of CCCTC binding factor (CTCF) and cohesin in the generation of single-cell diversity of Protocadherin- gene expression. Proc. Natl. Acad. Sci. *109*, 9125–9130.

65. Monahan, K., Schieren, I., Cheung, J., Mumbey-Wafula, A., Monuki, E.S., and Lomvardas, S. (2017). Cooperative interactions enable singular olfactory receptor expression in mouse olfactory neurons. Elife *6*, 1–32.

66. Monahan, K., Horta, A., and Lomvardas, S. (2019). LHX2- and LDB1-mediated trans interactions regulate olfactory receptor choice. Nature.

67. Mountoufaris, G., Chen, W. V, Hirabayashi, Y., O'Keeffe, S., Chevee, M., Nwakeze, C.L., Polleux, F., and Maniatis, T. (2017). Multicluster Pcdh diversity is required for mouse olfactory neural circuit assembly. Science (80-.). *356*, 411–414.

68. Müller, L.S.M., Cosentino, R.O., Förstner, K.U., Guizetti, J., Wedel, C., Kaplan, N., Janzen, C.J., Arampatzi, P., Vogel, J., Steinbiss, S., et al. (2018). Genome organization and DNA accessibility control antigenic variation in trypanosomes. Nature *563*, 121–125.

69. Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature *502*, 59–64.

70. Nagano, T., Várnai, C., Schoenfelder, S., Javierre, B.M., Wingett, S.W., and Fraser, P. (2015). Comparison of Hi-C results using in-solution versus in-nucleus ligation. Genome Biol. *16*, 1–13.

71. Ngai, J., Chess, A., Dowling, M.M., Necles, N., Macagno, E.R., and Axel, R. (1993). Coding of olfactory information: Topography of odorant receptor expression in the catfish olfactory epithelium. Cell *72*, 667–680.

72. Nikopoulou, C., Panagopoulos, G., Sianidis, G., Psarra, E., Ford, E., and Thanos, D. (2018). The Transcription Factor ThPOK Orchestrates Stochastic Interchromosomal Interactions Required for IFNB1 Virus-Inducible Gene Expression. Mol. Cell *71*, 352-361.e5.

73. Nora, E.P., Goloborodko, A., Valton, A.L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., and Bruneau, B.G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell *169*, 930-944.e22.

74. O'Donnell, M.A. (2016). Job Dekker: Hitting the scientific hi-Cs. J. Cell Biol. *215*, 434–435.

75. Osborne, C.S., Chakalova, L., Brown, K.E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J.A., Lopes, S., Reik, W., et al. (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. Nat. Genet. *36*, 1065–1071.

76. Ozbudak, E.M., Thattai, M., Kurtser, I., Grossman, A.D., and van Oudenaarden, A. (2002). Regulation of noise in the expression of a single gene. Nat. Genet. *31*, 69–73.

77. Plys, A.J., Davis, C.P., Kim, J., Rizki, G., Keenen, M.M., Marr, S.K., and Kingston, R.E. (2018). Phase separation and nucleosome compaction are governed by the same domain of Polycomb Repressive Complex 1. BioRxiv *1*, 467316.

78. Pope, B.D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D.L., Wang, Y., Hansen, R.S., Canfield, T.K., et al. (2014). Topologically associating domains are stable units of replication-timing regulation. Nature *515*, 402–405.

79. Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., et al. (2018). Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell *174*, 744-757.e24.

80. Raj, A., and van Oudenaarden, A. (2008). Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. Cell *135*, 216–226.

81. Ramani, V., Cusanovich, D.A., Hause, R.J., Ma, W., Qiu, R., Deng, X., Blau, C.A.,

Disteche, C.M., Noble, W.S., Shendure, J., et al. (2016). Mapping 3D genome architecture through in situ DNase Hi-C. Nat. Protoc. *11*, 2104–2121.

82. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

83. Rao, S.S.P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.-R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., et al. (2017a). Cohesin Loss Eliminates All Loop Domains. Cell *171*, 305-320.e24.

84. Rao, S.S.P., Huang, S.C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., et al. (2017b). Cohesin Loss Eliminates All Loop Domains. Cell *171*, 305-320.e24.

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

86. Ressler, K.J., Sullivan, S.L., and Buck, L.B. (1993). A zonal organization of odorant receptor gene expression in the olfactory epithelium. Cell *73*, 597–609.

87. Rhee, H.S., and Pugh, B.F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell *147*, 1408–1419.

88. Ribich, S., Tasic, B., and Maniatis, T. (2006). Identification of long-range regulatory elements in the protocadherin- gene cluster. Proc. Natl. Acad. Sci. *103*, 19719–19724.

89. Schmucker, D., Clemens, J.C., Shu, H., Worby, C.A., Xiao, J., Muda, M., Dixon, J.E., Zipursky, S.L., Hughes, H., South, C.E.Y., et al. (2000). Drosophila Dscam Is an Axon Guidance Receptor Exhibiting Extraordinary Molecular Diversity. *101*, 671–684.

90. Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N.F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J. a, Umlauf, D., Dimitrova, D.S., et al. (2010). Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nat.

Genet. *42*, 53–61.

91. Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., Haering, C., Mirny, L., et al. (2017). Two independent modes of chromatin organization revealed by cohesin removal. Nature *542*, 377–380.

92. Serizawa, S., Ishii, T., Nakatani, H., Tsuboi, A., Nagawa, F., Asano, M., Sudo, K., Sakagami, J., Sakano, H., Ijiri, T., et al. (2000). Mutually exclusive expression of odorant receptor transgenes. Nat. Neurosci. *3*, 687–693.

93. Serizawa, S., Miyamichi, K., Nakatani, H., Suzuki, M., Saito, M., Yoshihara, Y., and Sakano, H. (2003). Negative feedback regulation ensures the one receptor-one olfactory neuron rule in mouse. Science *302*, 2088–2094.

94. Shykind, B.M., Rohani, S.C., O'Donnell, S., Nemes, A., Mendelsohn, M., Sun, Y., Axel, R., and Barnea, G. (2004). Gene switching and the stability of odorant receptor gene choice. Cell *117*, 801–815.

95. Skene, P.J., and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. Elife *6*, 1–35.

96. Spemann, H., and Mangold, H. (1924). über Induktion von Embryonalanlagen durch Implantation artfremder Organisatoren. Arch. Für Mikroskopische Anat. Und Entwicklungsmechanik *100*, 599–638.

97. Spilianakis, C.G., Lalioti, M.D., Town, T., Lee, G.R., and Flavell, R.A. (2005). Interchromosomal associations between alternatively expressed loci. Nature *435*, 637–645.

98. Strom, A.R., Emelyanov, A. V., Mir, M., Fyodorov, D. V., Darzacq, X., and Karpen, G.H. (2017). Phase separation drives heterochromatin domain formation. Nature *547*, 241–245.

99. Subramanian, L., Sarkar, A., Shetty, A.S., Muralidharan, B., Padmanabhan, H., Piper, M., Monuki, E.S., Bach, I., Gronostajski, R.M., Richards, L.J., et al. (2011). Transcription factor Lhx2 is necessary and sufficient to suppress astrogliogenesis and promote neurogenesis in the developing hippocampus. Proc. Natl. Acad. Sci. *108*, E265–E274.

100. Tan, L., Xing, D., Daley, N., and Xie, X.S. (2019). Three-dimensional genome structures of single sensory neurons in mouse visual and olfactory systems. Nat. Struct. Mol. Biol. *26*.

101. Tasic, B., Nabholz, C.E., Baldwin, K.K., Kim, Y., Rueckert, E.H., Ribich, S.A., Cramer, P., Wu, Q., Axel, R., and Maniatis, T. (2002). Promoter choice determines splice site selection in protocadherin α and γ pre-mRNA splicing. Mol. Cell *10*, 21–33.

102. Tasic, B., Yao, Z., Graybuck, L.T., Smith, K.A., Nguyen, T.N., Bertagnolli, D., Goldy, J., Garren, E., Economo, M.N., Viswanathan, S., et al. (2018). Shared and distinct transcriptomic cell types across neocortical areas. Nature *563*, 72–78.

103. Tsai, A., Alves, M., and Crocker, J. (2019). Multi-enhancer transcriptional hubs confer phenotypic robustness. BioRxiv 575175.

104. Vassalli, A., Rothman, A., Feinstein, P., Zapotocky, M., and Mombaerts, P. (2002). Minigenes Impart Odorant Receptor-Specific Axon Guidance in the Olfactory Bulb. *35*, 681–696.

105. Vassar, R., Ngai, J., and Axel, R. (1993). Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium. Cell *74*, 309–318.

106. Vassar, R., Chao, S.K., Sitcheran, R., Nuiiez, M., Vosshall, L.B., and Axel, R. (1994). Topographic O rganization of Sensory Projection to the O lfactory Bulb. *79*.

107. Vian, L., Pękowska, A., Rao, S.S.P., Kieffer-Kwon, K.R., Jung, S., Baranello, L., Huang, S.C., El Khattabi, L., Dose, M., Pruett, N., et al. (2018). The Energetics and Physiological Impact of Cohesin Extrusion. Cell *173*, 1165-1178.e20.

108. Waddington, C.H. (1957). The Strategy of the Genes (London).

109. Wang, F., Nemes, A., Mendelsohn, M., and Axel, R. (1998). Odorant Receptors Govern the Formation of a Precise Topographic Map. Cell *93*, 47–60.

110. Wendt, K.S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., et al. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature *451*, 796–801.

111. Wu, Q., and Maniatis, T. (1999). A striking organization of a large family of human neural cadherin-like cell adhesion genes. Cell *97*, 779–790.

112. Xu, N., Tsai, C.-L., and Lee, J.T. (2006). Transient Homologous Chromosome Pairing Marks the Onset of X Inactivation. Science (80-.). *311*, 1149–1152.

113. Yan, J., Chen, S.-A.A., Local, A., Liu, T., Qiu, Y., Dorighi, K.M., Preissl, S., Rivera, C.M., Wang, C., Ye, Z., et al. (2018). Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers. Cell Res. 1–17.

114. Yun, S., Saijoh, Y., Hirokawa, K.E., Kopinke, D., Murtaugh, L.C., Monuki, E.S., and Levine, E.M. (2009). Lhx2 links the intrinsic and extrinsic factors that control optic cup formation. Development *136*, 3895–3906.

115. Zhao, Y., Kwan, K., Mailloux, C.M., Lee, W., Grinberg, A., Wurst, W., Behringer, R.R., and Westphal, H. (2007). LIM-homeodomain proteins Lhx1 and Lhx5, and their cofactor Ldb1, control Purkinje cell differentiation in the developing cerebellum. Proc. Natl. Acad. Sci. *104*, 13182–13186.

116. Zhao, Y., Flandin, P., Vogt, D., Blood, A., Hermesz, E., Westphal, H., and L. R. Rubenstein, J. (2014). Ldb1 is essential for development of Nkx2.1 lineage derived GABAergic and cholinergic neurons in the telencephalon. Dev. Biol. *385*, 94–106.

117. Zipursky, S.L., and Sanes, J.R. (2010). Chemoaffinity revisited: Dscams, protocadherins, and neural circuit assembly. Cell *143*, 343–353.

Appendix I:

Lhx2/Ldb1-mediated *trans* interactions regulate olfactory receptor choice

Monahan K[1*], Horta A[2*], and Lomvardas S[&].

[1] Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032

[2] Department of Neuroscience, Columbia University, New York, NY 10032

[3] Zuckerman Mind Brain and Behavior Institute, Columbia University, New York, NY 10027

[&] Corresponding Author sl682@columbia.edu

[*] These authors contributed equally to this work

Summary

The genome is partitioned into topologically associated domains (TADs) and genomic compartments of shared chromatin valance. This architecture is constrained by the DNA polymer, which precludes genic interactions between chromosomes. Here, we report a dramatic divergence from this pattern of nuclear organization that occurs in mouse olfactory sensory neurons (OSNs). *In situ* HiC on FAC-sorted OSNs and their progenitors shows that olfactory receptor (OR) gene clusters from 18 chromosomes make specific and robust interchromosomal contacts that increase with differentiation. These contacts are orchestrated by intergenic OR enhancers, the Greek Islands, which first contribute to the formation of OR compartments and then form a multi-chromosomal super-enhancer that associates with the single active OR. Greek Island-bound transcription factor Lhx2 and adaptor protein Ldb1 regulate the assembly and maintenance of OR compartments, Greek Island hubs, and OR transcription, providing mechanistic insight and functional support for the role of *trans* interactions in gene expression.

Introduction

Mouse ORs are encoded by a family of >1000 genes[1] that are organized in heterochromatic clusters[2] distributed across chromosomes. Every mature OSN (mOSN) expresses only one OR gene in a monoallelic and stochastic fashion[3,4]. OR gene activation requires removal of heterochromatic marks[5] and the concerted action of 63 intergenic enhancers, the “Greek Islands”, which are bound by transcription factors Lhx2 and Ebf[6,7]. Singular OR expression coincides with nuclear convergence of OR gene clusters[8,9], which promotes interchromosomal interactions between Greek Islands and the chosen OR[6]. The specificity by which Greek Islands associate with the active OR allele, as well as the significance of their interchromosomal contacts in OR transcription are uncertain. In this note, although interchromosomal interactions occur in other systems[10-14], unbiased approaches like *in situ* HiC[15] fail to detect robust *trans* contacts between non-repetitive regions[16,17], raising questions about the frequency and biological role of genomic interactions between chromosomes[17,18]. To obtain quantitative and functional insight into the regulation and function of multi-chromosomal interactions we performed *in situ* HiC in 10 distinct wild type and mutant cell types of the main olfactory epithelium (MOE) (Extended Data Fig. 1a-d).

First, we analyzed FAC-sorted mOSNs, which represent terminally differentiated, post-mitotic neurons that are heterogeneous in regards to OR identity. *In situ* HiC in mOSNs revealed extensive interchromosomal interactions corresponding to 35.6% of total HiC contacts (Extended data Fig.1e). Zoomed in genomic views show strong, OSN-specific *trans* contacts between OR clusters (Fig. 1a, b) with the median OR cluster having ~7.5% of all its HiC contacts map to OR clusters from different chromosomes (Extended data Fig. 1f). Aggregate peak analysis (APA) and unbiased compartment prediction[15] (Fig. 1b, Extended data Fig. 1g-h) confirm that most OR clusters participate in the assembly of OR-selective multi-chromosomal compartments. Notably, *trans* OR cluster contacts represent only 0.25% of all the interchromosomal contacts in mOSNs, but account for 50% of the 1000 strongest *trans* HiC contacts (Extended data Fig. 1i). In Horizontal Basal Cells (HBCs), the quiescent stem cells of the MOE, *trans* OR contacts are almost absent, representing only 2% of the strongest 1000 *trans* contacts genomewide, whereas inter-cluster *cis* OR contacts are strong, but less specific than in mOSNs (Extended data Fig. 2a-c,g-j). In the more differentiated immediate neuronal precursors (INPs)[15] *trans* OR contacts are abundant but less frequent than in mOSNs, with *cis* OR cluster interactions occurring at mOSN levels (Extended data

Fig. 2 d-f,j). Thus, OR compartments form in a hierarchical fashion, with *cis* contacts appearing first, and *trans* interactions strengthening with differentiation (Extended Data Fig. 2j-o). *In vitro* BAC HiC and *in silico* HiC assays, show that intra-cluster HiC fragments do not map in other OR clusters, excluding the possibility of homology-derived mapping artifacts (Extended Data Fig. 3).

Within OR compartments the 63 euchromatic Greek Islands represent HiC "hotspots" of specific and frequent *cis* and *trans* contacts (Fig. 1d,e, Extended data Fig. 4a,b). Similar to OR interactions, *trans* Greek Island contacts are not detected in HBCs (Extended data Fig 4e,f), which do not express ORs. In contrast, in INPs, where OR transcription is weak[2] and multigenic[19-21], Greek Islands interact with each other but lack the focal contact distribution detected in mOSNs (Extended Data Fig. 4c,d). The differentiation-dependent enhancement and specification of *trans* interactions is a property of most Greek Islands (Fig. 1f-g, Extended Data Fig. 4g,h). In total, 4.5% of Greek Island HiC contacts in mOSNs are made with the other Greek Islands, with half of these contacts being *trans* (Fig. 1f). Strikingly, this exceeds the mean and cumulative frequency of contacts that Greek Islands make with Lhx2/Ebf-bound intergenic sequences present in *cis* (Fig. 1g, Extended Data Fig. 4i,j), consistent with the differentiation-dependent assembly of a multi-chromosomal enhancer hub composed exclusively of Greek Islands.

For a mechanistic dissection of Greek Island interactions we explored the role of the core sequences of these enhancers. *In situ* HiC in mOSNs carrying homozygous deletions for Islands H[18] (2 Kb), Lipsi[6] (1 Kb), and Sfaktiria (0.6 Kb) shows strong reductions of *trans* interactions between genomic bins containing these deletions and the remaining Greek Islands, an effect that extends over large genomic distance (Fig. 2a-c, Extended Data Fig. 5a-b). Intriguingly, the reduction of cumulative *trans* Greek Island contacts correlates with the transcriptional OR downregulation observed in Greek Island deletions (Fig. 2c). If we exclude Greek Island bins from this analysis, we also observe reduction in *trans* OR contacts (Fig. 2c,d, Extended Data Fig. 5c). Thus, DNA elements as small as 0.6 Kb coordinate genomic contacts extending over hundreds of Kbs, similarly to "ZIP" elements affecting nuclear positioning in yeast[22], or the Igκ enhancer affecting the positioning of immunoglobulin loci in pre-B cells[23]. The partial effects of the triple enhancer deletions on cluster-wide contacts suggest that additional sequences participate in OR cluster interactions.

We then examined the role of Greek Island-bound transcription factors in OR compartmentalization. We deleted Lhx2 in HBCs, which were induced to differentiate with methimazole[24,25]. Using TdTomato intensity as a marker we identified two distinct cell populations, the dimmest of which is comprised of HBC-derived INPs and mOSNs (Extended Data Fig. 5d,e). RNA-seq of the FAC-sorted cells shows that early Lhx2 deletion caused a developmental delay in the OSN lineage and increase of INP-specific markers (Extended Data Fig.5f). With differentiation deficits and possible cell identity changes taken into account, *trans* OR and *trans* Greek Island contacts are strongly reduced in comparison to mOSNs and even INPs (Fig. 3a-d and Extended Data Fig. 5g). The frequency of interchromosomal interactions remains high in the early Lhx2 KO cells, yet OR-OR contacts represent only 16% of the 1000 strongest *trans* contacts (Extended Data Fig. 1e, 5h). Late Lhx2 deletion, in mOSNs[7] (Extended Data Fig. 5i), also reduces *trans* OR contacts, but not as much as the early deletion (Fig. 3a,c). However, late Lhx2 deletion diminishes *trans* and long-range *cis* contacts between Greek Islands (Fig. 3b,d and Extended Data 5j), consistent with widespread OR downregulation[7].

To decipher how Lhx2 stabilizes Greek Island contacts we asked if Lhx2, a LIM domain protein, recruits LIM domain binding proteins[26,27] (Ldb1 and Ldb2), which are known mediators of long-range genomic interactions[28-32]. ChIP-seq for Ldb1[33], which is the only family member expressed in mOSNs (Extended data Fig. 6a,b), reveals close overlap with Lhx2 peaks in mOSNs (Extended data Fig. 6c-e). Consistent with this, every Greek Island is bound by Ldb1, in an Lhx2-dependent fashion (Extended data Fig. 6f). Greek Islands represent some of the strongest Ldb1 peaks in the genome, suggesting synergistic action of Lhx2 and Ebf in Ldb1 recruitment (Extended data Fig. 6g,h). Greek Islands and OR clusters are not bound by CTCF and Rad21 (Extended data Fig. 6i,j), which is not surprising given the inhibitory role of cohesin complexes in formation of genomic compartments[34,35]. Finally, there is very little Ldb1 signal on OR promoters (Extended data Fig. 6k), a result that holds true even for the active Olfr1507 promoter in Olfr1507$^+$ OSNs (Extended Data Fig. 6l). Ldb1 deletion in mOSNs (Extended Data Fig. 7a,b) causes strong reduction in *trans* and long-range *cis* Greek Island interactions (Fig. 4a,b, Extended data 7c-f), a smaller decrease in the *trans* contracts between OR clusters (Extended data Fig. 7g,h), and even weaker genomewide effects in *trans* (Extended data Fig.1e). Importantly, RNA-seq shows that Ldb1 deletion causes widespread OR transcriptional downregulation (Fig. 4c) that appears highly restricted the OR gene family (Fig. 4d, Extended Data Fig. 7i).

To test if Greek Island hubs regulate OR transcription by direct interaction with the chosen OR we performed in situ HiC in OSNs expressing ORs, Olfr16, 17 and 1507. In these OSN populations the overall network of OR cluster and Greek Island interactions is largely the same (Extended Data Fig. 8a-d), but OSN type-specific variability is also observed (Extended Data Fig. 8e-m, 9a,b). However, in each OSN type the transcriptionally active OR consistently forms frequent interactions with Greek Islands. For example, in Olfr16$^+$ OSNs the Olfr16 locus interacts strongly (5% of the total HiC contacts mapped on Olfr16) with long-range *cis* and *trans* Greek Islands (Fig. 5a,b, Extended data Fig. 9c,d), whereas in Olfr17$^+$ and Olfr1507$^+$ OSNs it primarily interacts with nearby Greek Islands (Fig. 5a,b). Importantly, in Olfr16+ cells, Greek Island contacts are enriched specifically over the Olfr16 locus (Fig. 5b) relatively to the full OR repertoire (Fig. 5c). Thus, *in sit*u HiC accurately identifies the transcriptionally active OR from a pool of >1000 genes through its cumulative interactions with Greek Islands (Fig. 5c and Extended data Fig. 9e-h).

Our experiments reveal new types of genomic compartments with multi-chromosomal composition and extraordinary exclusivity. Genomic compartments represent more complex assemblies than segregation products of transcriptionally active and inactive chromatin[15,36]. However the demonstration that >1000 genes from 18 chromosomes form exclusive compartments, implies a precisely regulated process comparable with the assembly of the nucleolus[37]. Unlike the nucleolus, however, OR compartments and Greek Island hubs are regulated by proteins with widespread binding in the OSN genome. Absent of an OR-specific factor that would explain the specificity of OR contacts, we propose that Lhx2/Ebf/Ldb1-bound Greek Islands and OR heterochromatin create a unique molecular “barcode” that assembles OR-specific compartments. These heterochromatic compartments through phase separation properties of Hp1[38,39] may achieve efficient OR silencing, but they also confine in close proximity Greek Islands from different chromosomes[6,8], forcing them to interact. As proposed for super-enhancers[40,41], this confinement may promote an adjacent euchromatic phase consisted of locally concentrated activators. Where the two phases incompatible, the Greek Island hub would insulate the active OR allele from the surrounding repressive environment, resulting in stable OR choice (Extended data Fig. 10). Given that this multi-chromosomal super-enhancer interacts only with the single chosen OR and its disruption perturbs OR transcription, interchromosomal interactions emerge as essential regulators of OR transcription[6,7,42]. This concept of *trans* enhancement was initially challenged by the *cis*-only effects of enhancer deletions[18,43,44]. However, the demonstration that Greek Islands promote OR compartmentalization and recruit *trans* Greek Islands towards proximal ORs, explain why these elements are essential *cis* but redundant *trans* enhancers. With long-range genomic interactions been implicated in transcriptional stochasticity[12,45,46], cell type specific interchromosomal contacts may serve as an additional generator of molecular diversity.

Acknowledgments

We thank Alice Mumbay-Wafula for assistance with mouse colony management, Ira Schieren for assistance with FACS, Dr. Paul Love for sharing the Ldb1fl/fl mice, Dr. Edwin Monuki for sharing the Lhx2fl/fl mice. AH was funded by F31 post-doctoral fellowship DC016785 (NIH) and KM was funded by F32 post-doctoral fellowship GM108474 (NIH). This project was funded by 1U01DA040582, R01DC013560, R01DC015451 (NIH) and the HHMI Faculty Scholar Award. Research reported in this publication was also performed in the CCTI Flow Cytometry Core, supported in part by the Office of the Director, National Institutes of Health under awards S10OD020056. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Author Contributions

K.M, A.H., and S.L. designed the study. K.M. performed in situ HiC in Lhx2 and Ldb1 KO mice, performed ChIP-seq in wild type and Lhx2 KO mOSNs, and performed RNA-seq in Ldb1 KO and control mOSNs and methimazole treated cells form the MOE. A.H. performed in situ HiC in mOSNs, INPs, HBCs, Olfr1507, Olfr16, and Olfr17-expressing cells. Both K.M and A.H. analyzed data with input from S.L. S.L. wrote the manuscript with input from K.M. and A.H.

Competing interests

The authors declare no competing interests.

Corresponding author

Correspondence to Stavros Lomvardas.

References

1 Buck, L. & Axel, R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65**, 175-187 (1991).

2 Magklara, A. *et al.* An epigenetic signature for monoallelic olfactory receptor expression. *Cell* **145**, 555-570, doi:S0092-8674(11)00374-6 [pii]
10.1016/j.cell.2011.03.040 (2011).

3 Chess, A., Simon, I., Cedar, H. & Axel, R. Allelic inactivation regulates olfactory receptor gene expression. *Cell* **78**, 823-834 (1994).

4 Monahan, K. & Lomvardas, S. Monoallelic expression of olfactory receptors. *Annu Rev Cell Dev Biol* **31**, 721-740, doi:10.1146/annurev-cellbio-100814-125308 (2015).

5 Lyons, D. B. *et al.* An epigenetic trap stabilizes singular olfactory receptor expression. *Cell* **154**, 325-336, doi:10.1016/j.cell.2013.06.039 (2013).

6 Markenscoff-Papadimitriou, E. *et al.* Enhancer interaction networks as a means for singular olfactory receptor expression. *Cell* **159**, 543-557, doi:10.1016/j.cell.2014.09.033 (2014).

7 Monahan, K. *et al.* Cooperative interactions enable singular olfactory receptor expression in mouse olfactory neurons. *eLife* **6**, doi:10.7554/eLife.28620 (2017).

8 Clowney, E. J. *et al.* Nuclear aggregation of olfactory receptor genes governs their monogenic expression. *Cell* **151**, 724-737, doi:10.1016/j.cell.2012.09.043 (2012).

9 Armelin-Correa, L. M., Gutiyama, L. M., Brandt, D. Y. & Malnic, B. Nuclear compartmentalization of odorant receptor genes. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 2782-2787, doi:10.1073/pnas.1317036111 (2014).

10 Spilianakis, C. G. & Flavell, R. A. Molecular biology. Managing associations between different chromosomes. *Science* **312**, 207-208, doi:10.1126/science.1126689 (2006).

11 Beagrie, R. A. *et al.* Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519-524, doi:10.1038/nature21411 (2017).

12 Apostolou, E. & Thanos, D. Virus Infection Induces NF-kappaB-dependent interchromosomal associations mediating monoallelic IFN-beta gene expression. *Cell* **134**, 85-96 (2008).

13 Maass, P. G., Barutcu, A. R., Weiner, C. L. & Rinn, J. L. Inter-chromosomal Contact Properties in Live-Cell Imaging and in Hi-C. *Molecular cell* **70**, 188-189, doi:10.1016/j.molcel.2018.03.021 (2018).

14 Maass, P. G., Barutcu, A. R. & Rinn, J. L. Interchromosomal interactions: A genomic love story of kissing chromosomes. *The Journal of cell biology*, doi:10.1083/jcb.201806052 (2018).

15 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).

16 Nagano, T. *et al.* Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biol* **16**, 175, doi:10.1186/s13059-015-0753-7 (2015).

17 Johanson, T. M. *et al.* Genome-wide analysis reveals no evidence of trans chromosomal regulation of mammalian immune development. *PLoS genetics* **14**, e1007431, doi:10.1371/journal.pgen.1007431 (2018).

18 Fuss, S. H., Omura, M. & Mombaerts, P. Local and cis effects of the H element on expression of odorant receptor genes in mouse. *Cell* **130**, 373-384 (2007).

19 Hanchate, N. K. *et al.* Single-cell transcriptomics reveals receptor transformations during olfactory neurogenesis. *Science* **350**, 1251-1255, doi:10.1126/science.aad2456 (2015).

20 Saraiva, L. R. *et al.* Hierarchical deconstruction of mouse olfactory sensory neurons: from whole mucosa to single-cell RNA-seq. *Scientific reports* **5**, 18178, doi:10.1038/srep18178 (2015).

21 Tan, L., Li, Q. & Xie, X. S. Olfactory sensory neurons transiently express multiple olfactory receptors during development. *Molecular systems biology* **11**, 844, doi:10.15252/msb.20156639 (2015).

22 Ahmed, S. *et al.* DNA zip codes control an ancient mechanism for gene targeting to the nuclear periphery. *Nat Cell Biol* **12**, 111-118, doi:10.1038/ncb2011 (2010).

23 Hewitt, S. L. *et al.* Association between the Igk and Igh immunoglobulin loci mediated by the 3' Igk enhancer induces 'decontraction' of the Igh locus in pre-B cells. *Nature immunology* **9**, 396-404, doi:10.1038/ni1567 (2008).

24 Gadye, L. *et al.* Injury Activates Transient Olfactory Stem Cell States with Diverse Lineage Capacities. *Cell Stem Cell* **21**, 775-790 e779, doi:10.1016/j.stem.2017.10.014 (2017).

25 Lin, B. *et al.* Injury Induces Endogenous Reprogramming and Dedifferentiation of Neuronal Progenitors to Multipotency. *Cell Stem Cell* **21**, 761-774 e765, doi:10.1016/j.stem.2017.09.008 (2017).

26 Agulnick, A. D. *et al.* Interactions of the LIM-domain-binding factor Ldb1 with LIM homeodomain proteins. *Nature* **384**, 270-272, doi:10.1038/384270a0 (1996).

27 Bach, I. The LIM domain: regulation by association. *Mechanisms of development* **91**, 5-17 (2000).

28 Krivega, I. & Dean, A. LDB1-mediated enhancer looping can be established independent of mediator and cohesin. *Nucleic acids research* **45**, 8255-8268, doi:10.1093/nar/gkx433 (2017).

29 Lee, J., Krivega, I., Dale, R. K. & Dean, A. The LDB1 Complex Co-opts CTCF for Erythroid Lineage-Specific Long-Range Enhancer Interactions. *Cell reports* **19**, 2490-2502, doi:10.1016/j.celrep.2017.05.072 (2017).

30 Deng, W. *et al.* Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233-1244, doi:10.1016/j.cell.2012.03.051 (2012).

31 Caputo, L. *et al.* The Isl1/Ldb1 Complex Orchestrates Genome-wide Chromatin Organization to Instruct Differentiation of Multipotent Cardiac Progenitors. *Cell Stem Cell* **17**, 287-299, doi:10.1016/j.stem.2015.08.007 (2015).

32 Bronstein, R. *et al.* Transcriptional regulation by CHIP/LDB complexes. *PLoS genetics* **6**, e1001063, doi:10.1371/journal.pgen.1001063 (2010).

33 Matthews, J. M. & Visvader, J. E. LIM-domain-binding protein 1: a multifunctional cofactor that interacts with diverse proteins. *EMBO Rep* **4**, 1132-1137, doi:10.1038/sj.embor.7400030 (2003).

34 Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320 e324, doi:10.1016/j.cell.2017.09.026 (2017).

35 Schwarzer, W. *et al.* Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51-56, doi:10.1038/nature24281 (2017).

36 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).

37 Pederson, T. The nucleolus. *Cold Spring Harb Perspect Biol* **3**, doi:10.1101/cshperspect.a000638 (2011).

38 Larson, A. G. *et al.* Liquid droplet formation by HP1alpha suggests a role for phase separation in heterochromatin. *Nature* **547**, 236-240, doi:10.1038/nature22822 (2017).

39 Strom, A. R. *et al.* Phase separation drives heterochromatin domain formation. *Nature* **547**, 241-245, doi:10.1038/nature22989 (2017).

40 Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A Phase Separation Model for Transcriptional Control. *Cell* **169**, 13-23, doi:10.1016/j.cell.2017.02.007 (2017).

41 Sabari, B. R. *et al.* Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **361**, doi:10.1126/science.aar3958 (2018).

42 Lomvardas, S. *et al.* Interchromosomal interactions and olfactory receptor choice. *Cell* **126**, 403-413, doi:10.1016/j.cell.2006.06.035 (2006).

43 Khan, M., Vaes, E. & Mombaerts, P. Regulation of the probability of mouse odorant receptor gene choice. *Cell* **147**, 907-921, doi:10.1016/j.cell.2011.09.049 (2011).

44 Nishizumi, H., Kumasaka, K., Inoue, N., Nakashima, A. & Sakano, H. Deletion of the core-H region in mice abolishes the expression of three proximal odorant receptor genes in cis. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 20067-20072, doi:10.1073/pnas.0706544105 (2007).

45 Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**, 900-910, doi:10.1016/j.cell.2015.07.038 (2015).

46 Noordermeer, D. *et al.* Variegated gene expression caused by cell-specific long-range DNA interactions. *Nat Cell Biol* **13**, 944-951, doi:10.1038/ncb2278 (2011).

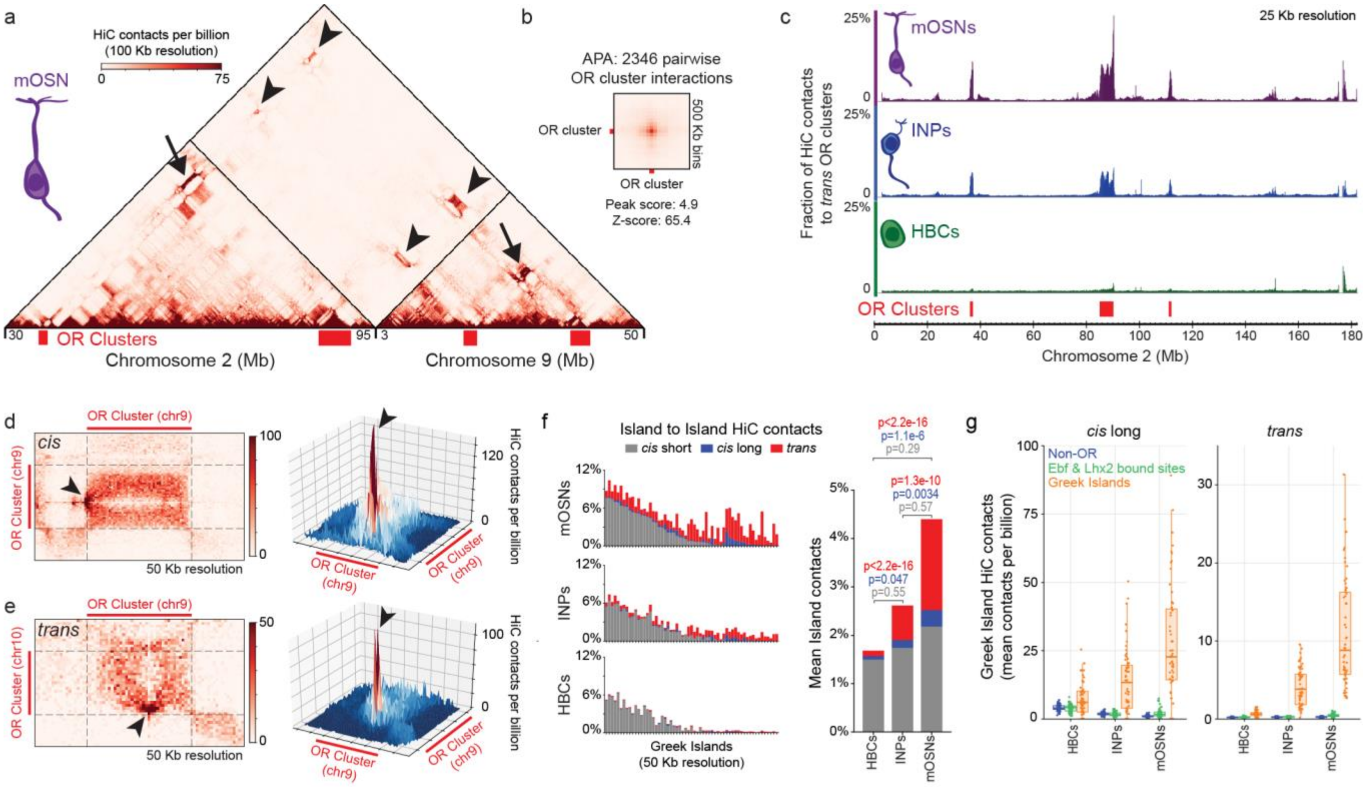


Figure 1: Extensive interchromosomal contacts between OR gene clusters and focal interchromosomal contacts between Greek Islands form over OSN differentiation. **a**, In situ HiC contact matrix of chromosomes 2 and 9 in mOSNs shows highly restricted and frequent contacts between OR gene clusters in *cis* (arrows) and *trans* (arrowheads). **b**. Aggregate Peak Analysis (APA) shows strong focal contacts between OR gene clusters in mOSNs. **c**, The fraction of HiC contacts made to OR clusters located on a different chromosome is shown for every 25 Kb bin along chromosome 2. For OR clusters these contacts increase over differentiation form HBCs (bottom) to INPs (middle) to mOSNs (top). **d-e,** Pairwise views of OR gene clusters reveals a local maximum of *in situ* HiC interactions between Greek Island loci (arrowheads) in *cis* (a) and *trans* (b). **f**, (left) For each Greek Island, the fraction of total HiC contacts that are made to other Greek Islands located in *cis* at short range (<5 Mb apart, grey), long range (>5Mb apart, blue), and in *trans* (red). Top panel represents mOSNs, middle panel INPs, and bottom panel HBCs. (right) Mean fraction of HiC contacts across all Greek Islands (two-sided, paired Wilcoxon signed-rank test, n=59). **g**, For each Greek Island bin (n=59), the mean number of *cis* long range (left) and *trans* (right) HiC contacts per billion made to every non-OR sequence (at 50 Kb resolution), intergenic Lhx2 & Ebf bound peak (outside of OR clusters), or Greek Island. Box indicates median, upper, and lower quartiles while whiskers indicate 1.5 * the interquartile range. All panels present pooled data from 2 independent biological replicates that yielded similar results when analyzed separately.

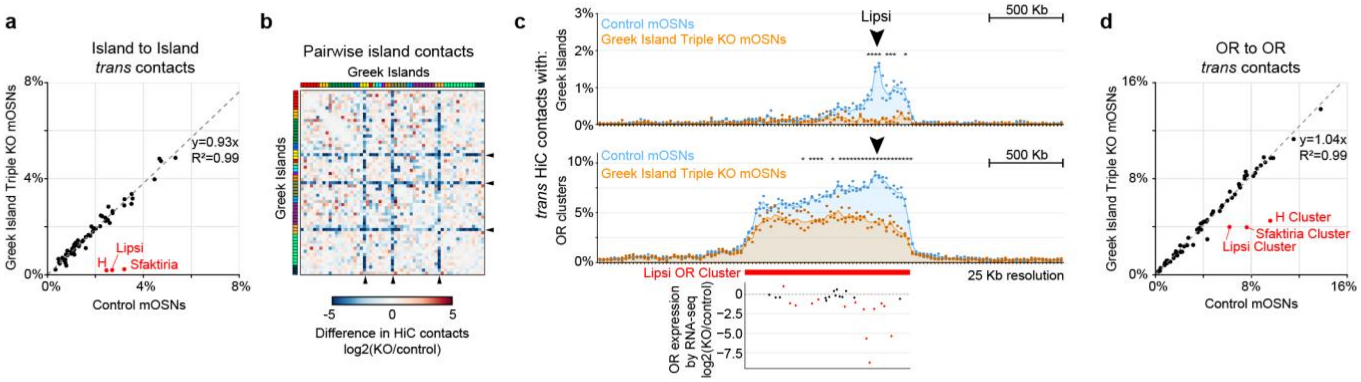


Figure 2: Greek Island deletion disrupts local recruitment of *trans* Greek Islands and impairs OR compartmentalization. a, In mOSNs in which 3 Greek Islands (H, Lipsi, and Sfaktiria) have been homozygously deleted, the 50 Kb regions containing the deleted Islands have reduced *trans* Greek Island contacts, expressed as fraction of total HiC contacts. Interactions among the remaining Islands are not significantly different (p=0.80, two-sided, paired Wilcoxon signed-rank test, n=56). **b,** Pairwise heatmap of Greek Island contacts reveals that the 50 Kb regions containing the deleted Greek Islands (arrowheads) exhibit reduced contacts, plotted as Log2 fold difference, across the full set of Greek Islands. Greek Islands are ordered by genomic position and color bar indicates chromosome. **c,** The OR gene cluster containing Lipsi makes fewer HiC contacts with *trans* Greek Islands and OR gene clusters in KO mOSNs than on control mOSNs. Count data for *trans* Greek Island contacts and *trans* OR cluster contacts from 2 biological replicates were analyzed to identify loci with a significant difference in contacts between conditions (see Extended Materials and Methods). Significantly changed regions, corrected for multiple comparisons, are indicated with an asterisk ($p_{adj} < 0.05$, Wald test). Lower panel shows RNA-seq analysis of the expression of OR genes in KO mOSNs relative to control mOSNs. Significantly changed ORs are red ($p < 0.01$, Wald test, 5 biological replicates for control mOSNs and 4 for KO mOSNs). **d,** OR gene clusters containing the deleted Greek Islands (red) make fewer contacts with *trans* OR gene clusters in KO mOSNs, plotted as fraction of the total HiC contacts. Contacts made by the non-targeted clusters are not significantly different (p=0.79, two-sided, paired Wilcoxon signed-rank test, n=64). Panels a and d present pooled data from 2 independent biological replicates that yielded similar results when analyzed separately.

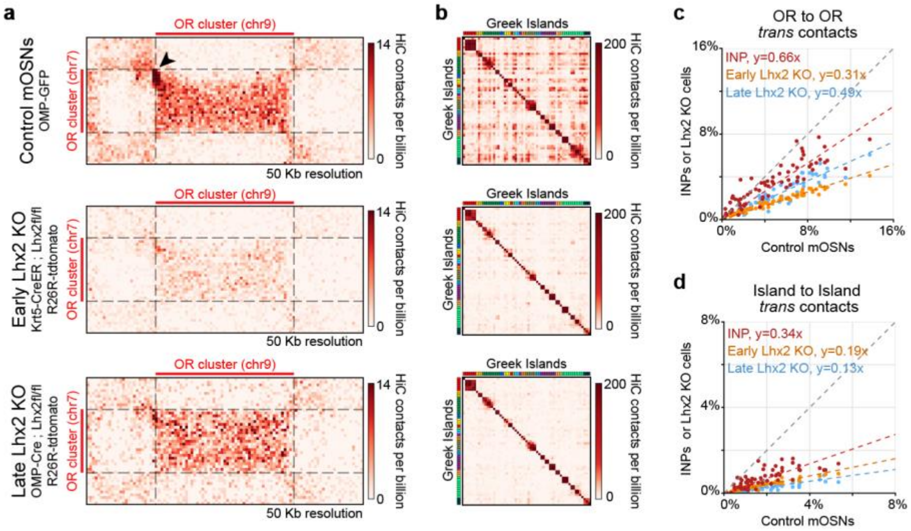


Figure 3: Lhx2 is essential for the formation of OR compartments and the assembly and stability of Greek Island hubs. **a**, Pairwise views of HiC contacts between OR clusters located on different chromosomes in control (top), early Lhx2 KO (middle), and late Lhx2 KO (bottom) OSNs. A HiC hotspot between interacting Greek Islands in control mOSNs (arrowhead) is absent in both early and late Lhx2 KO cells. In addition, a strong reduction in the surrounding OR-OR contacts is observed in the early Lhx2 KO. **b,** Pairwise heatmap of Greek Island contacts reveals reduced HiC contacts across the full set of Greek Islands. **c,** Contacts made by each OR cluster (n=67) to OR clusters located in *trans*, expressed as fraction of the total HiC contacts, in mOSNs versus INPs, early, or late Lhx2 KO cells. Dashed line is a linear fit. **d**, same as c, but for *trans* contacts between Greek Islands (n=59). All panels present pooled data from 2 independent biological replicates that yielded similar results when analyzed separately.

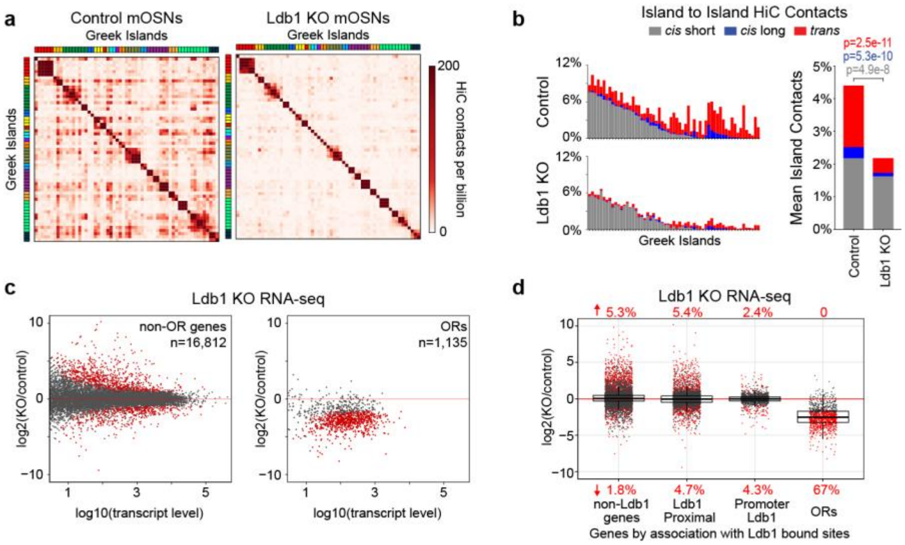


Figure 4: Ldb1 is essential for the stability of Greek Island hubs and for OR transcription. **a,** Pairwise heatmap of Greek Island contacts reveals broad reductions in HiC contacts in Ldb1 KO mOSNs. **b,** (left) For each Greek Island, the fraction of total HiC contacts made to other Greek Islands located in *cis* at short range (<5 Mb apart, grey), long range (>5Mb apart, blue), and in *trans* (red). Top panel represents control mOSNs and bottom panel Ldb1 KO cells. (right) The effect of Ldb1 KO on the mean fraction of HiC contacts across all Greek Islands (two-sided, paired Wilcoxon signed-rank test, n=59). **c**, RNA-seq analysis of gene expression in Ldb1 KO cells relative to control mOSNs. Significantly changed genes are colored red (p_{adj} < 0.05 for greater than 1.5-fold change, Wald test, n=5 for control mOSNs and n=4 Ldb1 KO). **d**, Effect of Ldb1 KO on genes not associated with Ldb1 ChIP peaks (n=9,548), genes located closest to a non-promoter Ldb1 ChIP-seq peak (n=5,624), genes with an Ldb1 ChIP-seq peak within the promoter region (n=1,640), and ORs (n=1,135). The percentage of significantly changed genes in each category is shown (p_{adj} < 0.05 for greater than 1.5-fold change, Wald test, n=5 for control mOSNs and n=4 Ldb1 KO). Box indicates median, upper, and lower quartiles while whiskers indicate 1.5 * the interquartile range. Panels a and b present pooled data from 2 independent biological replicates that yielded similar results when analyzed separately.

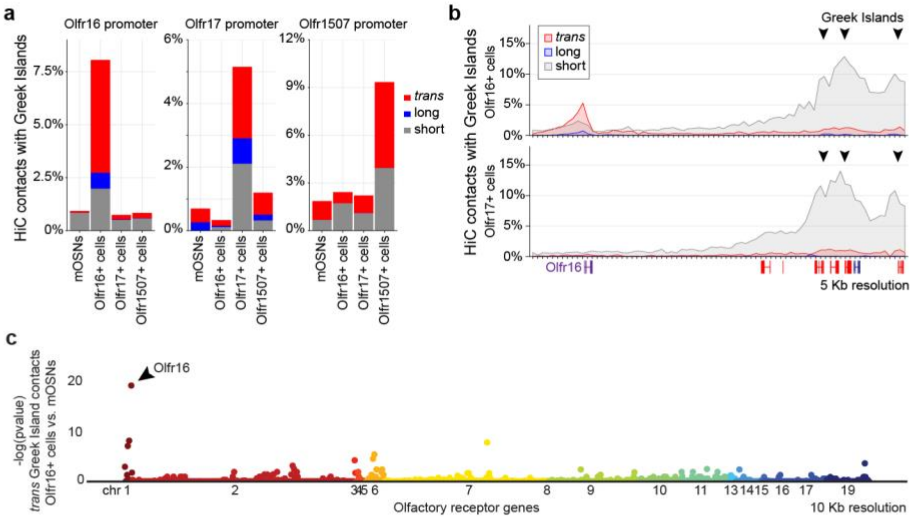


Figure 5: Greek Island hubs interact specifically with the transcriptionally active OR locus. a, Increased contacts between the active OR promoter and Greek Islands located in short range *cis* (<5Mb, grey), long range *cis* (>5Mb, blue) and *trans* (red). Greek Island interactions are expressed as the fraction of the total HiC contacts mapped to each promoter (5 Kb resolution). **b,** Profile of the OR cluster containing Olfr16 reveals increased contacts, expressed as fraction of the total HiC contacts mapped to each position (5 Kb resolution), between the Olfr16 locus and Greek Islands in Olfr16 expressing cells. **c**, Manhattan plot of Greek Island contacts with OR genes reveals that in Olfr16^{+} cells the Olfr16 locus is the OR gene most significantly enriched for Greek Island contacts relative to heterogeneous mOSNs (see Extended Materials and Methods). All panels present pooled data from 2 independent biological replicates that yielded similar results when analyzed separately.

Methods

Mice

Mice were treated in compliance with the rules and regulations of IACUC under protocol number AC-AAAT2450. Mice were sacrificed using CO2 followed by cervical dislocation. Both male and female mice were used for experiments. All experiments were performed on dissected olfactory epithelium tissue or on dissociated cells prepared from whole olfactory epithelium tissue. Dissociated cells were prepared using papain (Worthington Biochemical) and FAC sorted as previously described[7].

This study used several mouse lines to allow isolation of cells at specific stages of olfactory sensory neuron (OSN) development, OSNs that express one of three specific olfactory receptors, and cells with specific targeted mutations. Mature OSNs (mOSNs) were sorted from Omp-IRES-GFP mice[47]. Neural progenitors (INPs) were isolated by sorting the brightest of two GFP populations from Ngn1-GFP mice[2]. The dim population of Ngn1 cells represents a more mature population of OSNs, as determined by RNAseq (data not shown). Multipotent olfactory progenitors (horizontal basal cells) were isolated by injecting perinatal Krt5-CreER[48];B6N.129S6-Gt(ROSA)26Sor$^{\text{tm1(CAG-tdTomato*,-EGFP*)Ees/J}}$ mice[49] with tamoxifen 24 and 48 hours before sorting GFP-positive, tdTomato-negative cells. Olfr17+ cells were sorted from Olfr17-IRES-GFP[47] mice. Olfr1507+ cells were sorted from Olfr1507-IRES-GFP mice[47]. Olfr16+ cells were sorted from Olfr16-IRES-tauGFP (Olfr16$^{\text{tm2Mom}}$)[50]. Triple enhancer knockout mice were generated by crossing mice bearing 3 individual Greek Island deletions (H[38], Lipsi[5], Sfaktiria) and Omp-IRES-GFP and sorting for GFP+ mature OSNs. The Sfaktiria deletion was generated by Biocytogen using TALENs to target the region chr6:42869802-42870400 (mm10).

Conditional deletion of Lhx2 early in mOSN differentiation was achieved by crossing Lhx2 conditional allele mice to mice bearing Krt5-CreER and Cre-inducible tdTomato (ROSA26-tdtomato, Gt(ROSA)26Sor$^{\text{tm14(CAG-tdTomato)Hze/J}}$). At 6-weeks of age, deletion of the conditional allele in horizontal basal cells was induced by two intraperitoneal injections of tamoxifen twenty-four hours apart. One week later, differentiation of horizontal basal cells into olfactory cell types was induced by intraperitoneal injection with methimazole, which triggers ablation of olfactory epithelium and regeneration of the tissue from horizontal basal cells. The olfactory epithelium was allowed to regenerate for 8-weeks, producing bright TdTomato^{+} cells that localized to the basal (HBCs) and apical (Sustentacular cells) layers of the MOE, and dim

TdTomato$^+$ cells that populate the neuronal cell layers of the MOE. FACS of the bright and dim populations separately, followed by RNA-seq confirms that the dim cell population is comprised mostly of mOSNs and INPs (Extended data Fig.6a-c)

Conditional alleles were deleted specifically in mOSNs using OMP-ires-Cre[51] mice. Conditional deletion of Lhx2 in mOSNs was achieved by crossing Lhx2 conditional allele mice[52] (Lhx2-fl: Lhx2^{tm1Monu}) and Cre-inducible tdTomato to OMP-Cre. Similarly, conditional deletion of Ldb1 in mOSNs was achieved by crossing Ldb1 conditional allele mice[53] (Ldb1-fl: Ldb1^{tm2Lmgd}) with Cre-inducible tdTomato and OMP-Cre. Recombined cells were purified by selecting tdTomato positive cells by FACS.

Fluorescence activated cell sorting

Cells were dissociated into a single-cell suspension by incubating freshly dissected main olfactory epithelium with papain for 40min at 37°C according to the Worthington Papain Dissociation System. Following dissociation and filtering three times through a 35µm cell strainer, live cells were sorted by collecting fluorescent, DAPI-negative cells for RNA-seq and ATAC-seq. Alternatively, cells were fixed with 1% PFA in PBS for 5 minutes (ChIP) or 10 minutes (HiC) at room temperature. Fixed fluorescent cells were then sorted on a BD Aria II, BD Influx, or Beckman Coulter MoFlo Astrios EQ cell sorter.

Representative FACS plots for the cells used in this study are available at https://data.4dnucleome.org/search/?lab.display_title=Stavros%20Lomvardas%2C%20COLUMBIA&protocol_type=Cell%20sorting%20protocol&type=Protocol

***in situ* Hi-C**

Depending on the genotype, between 20 thousand and 3 million cells were used for in situ Hi-C. Sorted cells were lysed and intact nuclei were processed through an in situ Hi-C protocol as previously described[15] with a few modifications. Briefly, cells were lysed with 50mM Tris pH 7.5 0.5% Igepal, 0.25% Sodium-deoxychloate 0.1% SDS, 150mM NaCl, and protease inhibitors. Pelleted intact nuclei were then resuspended in 0.5% SDS and incubated 20min 65°C for nuclear permeabilization. After quenching with 1.1% Triton-X for 10min at 37°C, nuclei were digested with 6U/µl DpnII in 1x DpnII buffer overnight at 37°C. Following initial digestion, cells were pelleted (2500g 5min), buffers were replenished to original concentrations and fresh DpnII was added at 37°C for an additional 2 hours of digestion. Following digestion, the restriction enzyme was inactivated at 65°C for 20min. For the 1.5hr fill-in at 37°C, biotinylated dGTP was used instead of dATP to increase ligation efficiency. Ligation was performed at 25°C for 4 hours with rotation. Nuclei were then pelleted and sonicated in 10mM Tris pH 7.5, 1mM EDTA, 0.25% SDS on a Covaris S220 for 16min with 2% duty cycle, 105 intensity, 200 cycles per burst, 1.8-1.85 W, and max temperature of 6°C. DNA was reverse cross-linked overnight at 65°C with proteinase K and RNAse A. Each experiment was performed in biological replicates.

HiC Library preparation and sequencing

Reverse cross-linked DNA was purified with 2x Ampure beads following the standard protocol and eluting in 300µl water. Biotinylated fragments were enriched as previously described using Dynabeads MyOne Strepavidin T1 beads. The biotinylated DNA fragments were prepared for next-generation sequencing directly on the beads by using the Nugen Ovation Ultralow kit protocol with some modifications. Following end repair, magnetic beads were washed twice at 55°C with 0.05% Tween, 1M NaCl in Tris/EDTA pH 7.5, instead of heat-inactivating end-repair enzymes. Residual detergent was removed by washing beads twice in 10mM Tris pH 7.5. End repair buffers were replenished to original concentrations, but the enzyme and enhancer was omitted before adapter ligation. Following adaptor ligation, beads underwent 5 washes with 0.05% Tween, 1M NaCl in Tris/EDTA pH 7.5 at 55°C and two washes with 10mM Tris pH 7.5 to remove ligation enzymes and buffers. DNA was amplified by 10 cycles of PCR. Beads were reclaimed and amplified unbiotinylated DNA fragments were purified with 0.8x Ampure beads. Quality and concentration of libraries were assessed by Agilent Bioanalyzer and KAPA Library Quantification Kit. HiC libraries were sequenced paired-end on NextSeq 500 (2x75bp), or NovaSeq 6000 (2x150bp).

A full protocol and gel electrophoresis of a typical HiC experiment is available at https://data.4dnucleome.org/search/?lab.display_title=Stavros+Lomvardas%2C+COLUMBIA&protocol_type=Experimental+protocol&type=Protocol

Hi-C data processing pipeline

Raw fastq files were processed through use of the Juicer Tools Version 1.76 pipeline[54] with one modification. Reads were aligned to mm10 using BWA 0.7.17 mem[55] algorithm and specifying the -5 option implemented specifically for Hi-C data. The -5 option always takes the leftmost alignment (5') on a read as the primary read. This alignment gets its own alignment score independent of subsequent alignments. Following alignment, independently mapped reads are merged to generate chimeric reads. After reads are aligned, merged, and sorted, chimeras are de-duplicated and finally HiC contact matrices are generated by binning at various resolutions and matrix balancing. Importantly, all reads mapping to multiple locations are discarded as "chimeric ambiguous reads". To remove multi-mappers, we used a stringent cutoff of MAPQ > 30. All data used in this paper, including data generated by other groups, was aligned in this way.

Hi-C data analysis

HiC matrices used in this paper were matrix-balanced using Juicer's built-in Knight-Ruiz (KR) algorithm. Where noted, values were instead normalized to target counts/total HiC contacts for that bin at a specified resolution (e.g. percent OR contacts/total HiC contacts per bin). This accounts for sequencing and alignment depth of a given bin. Matrices were graphed using pandas, seaborn and matplotlib[56-58] packages for python, or R-Studio Server (R version 3.5.1).

Genome wide Hi-C maps were constructed from KR-normalized matrices at 1Mb resolution and normalized to library size. The maximum value of the color scale was set to 1000 reads per billion HiC contacts per 1Mb bin.

Cumulative interchromosomal contacts at the resolutions noted in the text were constructed by calling Juicer Tools dump to extract genome wide un-normalized data from a .hic file. Subsequently, single-ended bins for regions of interest were selected for genome wide interchromosomal counts. Counts pertaining to a particular bin were divided by the total HiC contacts sequenced for the respective bin. These normalized counts were then aggregated per genomic bin to construct a bedGraph and visualized using Integrated Genome Browser[59]. Alternatively, all bins contacted by a bin of interested were categorized by genomic location (e.g. Greek Islands overlapping, OR Cluster overlapping, intergenic Ebf/Lhx2 peak overlapping) and then counts were aggregated by category. For 50 Kb and 25 Kb analyses only the bin directly overlapping a feature (e.g. a Greek Island) was assigned to that category. For 5 Kb resolution analyses the bin containing a feature and the 2 bins directly upstream and downstream were assigned to that feature category. Aggregate counts were converted to fraction of HiC contacts by dividing by the total number of HiC contacts made by the bin of interest. Mean counts per interaction was determined by dividing the aggregate counts for each category (e.g. Greek Island overlapping, OR Cluster overlapping, etc.) by the number of bins matching that category present in *cis* or in *trans*.

Aggregate Peak Analysis (APA) was done through the use of Juicer Tools. Normalized APA matrices were graphed with the maximum scale set to 5 times the mean of the matrix.

OR gene cluster contact matrices were constructed by extracting pairwise contacts between OR gene cluster bins and dividing by the area (size of cluster 1 x size of cluster 2) of the respective pairwise OR gene cluster interaction. The logarithm of these values was then taken to account for the strength of *cis* interactions and plotted.

Specific OR gene cluster contacts were constructed through programmatic access to .hic files using straw for python. These matrix files can also be used to form 3-dimensional contour maps with the same software to better visualize the focal peaks in the contact matrix. KR-normalized matrix values were further normalized by dividing by HiC library size for directly comparing samples.

For box plots quantifying the strength of interchromosomal interactions, the box indicates median and upper and lower quartiles while whiskers indicate 1.5 * the interquartile range. Outliers are not shown.

DESeq2 was used to detect differences between conditions for individual sites[60]. A similar approach has previously been used to analyze count data from 4C-seq[61]. The raw, un-normalized number of HiC contacts mapping to OR clusters located in *trans* or to Greek Islands located in *trans* was determined for every region of the genome at a given resolution (25 Kb bins). For each condition, counts from two biological replicates were analyzed using DESeq2. Regions with zero counts in any condition were excluded. DESeq2 identifies regions where the observed change in counts between conditions is significantly greater than amount of change expected based upon an analysis of variance between replicates. For the analysis of Triple Enhancer Knockout mOSNs compared to control mOSNs (Figure 3, 25 Kb resolution) a total of 22 regions out of 84,592 were found to have significantly changed counts for *trans* Greek Island contacts (padj < 0.05). 21 of these 22 regions map to the OR clusters containing the deleted Greek Islands. Similarly, 117 regions show a significant change in *trans* OR cluster contacts, 62 of them map to OR clusters, and 60 out of those 62 correspond to the OR clusters containing the deleted Greek Islands.

Compartment analysis

A Hidden Markov Model was used to assess the presence of genomic compartments as previously described[15,34] with some minor changes. Briefly, a square matrix of odd vs even chromosome contacts is made (i.e. interchromosomal). Using 2-19 components, HMMs are constructed for odd vs. even chromosomes and a score is calculated using hmmlearn[62]'s built-in score to ascertain the likelihood of the given number of compartments. The same was done for even vs. odd after transposing the matrix. The mean value of a genomic region for a given component (or compartment) was used to construct a bedGraph and visualized with the genome browser. Notably, Rao et al discarded genomic regions with less than 70% of the column filled. We opted to keep all rows because we noticed that many of the specific compartments we are observing (e.g. OR compartment, Greek Island compartment) are inherently sparse in genomic regions not corresponding to their compartment of choice. Throwing out these regions would select for nonspecific (or noisy) compartments.

in vitro BAC HiC

We performed an in vitro HiC on BAC clone RP23-374F2, a 165kb clone containing mostly OR sequences but also non-OR sequences. The HiC protocol is analogous to our experimental HiC. Briefly, we digested the BAC clone with DpnII, filled-in overhangs with DNA Pol I Klenow Fragment, performed a blunt-end ligation with T4 ligase, and sonicated to 300 bp with a Covaris sonicator. In this scenario, we would generate artificial "*cis*" HiC contacts when run through our HiC pipeline without the presence of "*trans*" contacts generated by mismapping.

in silico HIC

To address potential mapping issues by an orthogonal computational approach, we performed in silico HiC. DNA sequences corresponding to 4 of the largest OR Clusters (chr2:36252272-37350072; chr2:85196700-90429754; chr9:18512886-20345134; chr9:37669223-40192314), totaling over 10.5Mb of DNA sequences were retrieved and separately processed through an *in silico* HiC pipeline. In order to emulate digestion by DpnII, DNA sequences were split at GATC stored along with their reverse complements. Each "digested" string was joined with another "digested" string in both the forward and reverse complement orientations with a joining "GATCGATC" in order to emulate the fill-in and ligation. These chimeras, ranging in size from 10s of basepairs to > 4000bp (mode: ~600) were randomly truncated to 300bp to emulate our average library size after shearing and library prep. Following shearing, only fragments with "GATCGATC" were stored, in accordance with experimental biotin pulldown. We then took the first and last 75bp of these strings and wrote them to separate files for each read of the paired end reads. Lastly, to best recapitulate sequencing errors and biases, we used fastq scores from the mOSN HiC experiment used in this manuscript. Following generation of in silico HiC fastqs, we aligned our data using same pipeline we used for all of our datasets.

Chromatin Immunoprecipitation

See ChIP-seq tab of Supplementary Information 1 for a summary of ChIP-seq sequencing data. Chromatin Immunoprecipitation (ChIP) experiments were carried out as previously described[7]. Briefly, 600,000 - 2 million FACS purified cells were used for each experiment. Sheared chromatin was prepared from FACS purified cells using a Covaris S220 Focused-ultrasonicator. ChIP was performed using antibodies for CTCF (Millipore Cat# 07-729, RRID:AB_441965), Rad21 (Abcam Cat# ab992, RRID:AB_2176601), or Ldb1 (Santa Cruz Biotechnology Cat# sc-11198, RRID:AB_2137017). ChIP-seq libraries were prepared using the Nugen Ovation Ultralow Library System v2 (Nugen Cat# 0344-32). All data sets were processed using 50bp of single end; 75bp reads were trimmed to 50bp and only read 1 was used from paired end data. Adapter sequences were removed from raw ChIP-seq data using CutAdapt v1.17 (RRID:SCR_011841) and filtered reads were aligned to the mouse genome (mm10) using Bowtie2[63] v2.3.2 (RRID:SCR_006646) with default settings. Picard (RRID:SCR_006525) was used to identify duplicate reads, which were then removed with Samtools[64] v1.4.1 (RRID:SCR_002105). Samtools was used to select uniquely aligning reads by removing reads with alignment quality alignments below 30 (-q 30). Peaks of ChIP-seq signal were identified using HOMER[65] v4.10.3 (RRID:SCR_010881) in “factor” mode with an input control. Consensus peak sets were generated by selecting peaks that overlapped in at least two biological replicates and extending them to their combined size. Bedtools2[66] v2.26.0 was used to compare peak sets.

For signal tracks, biological replicates were merged and HOMER was used to generate 1bp resolution signal tracks normalized to a library size of 10,000,000 reads. Values in all ChIP-seq signal plots are counts per 10 million reads. Plots of ChIP-seq signal over individual loci were generated using the UCSC Genome Browser. Deeptools2[67] v3.1.1 was used to generate ChIP-seq heatmaps and mean signal plots. For heatmaps, each row of the heatmap is an 8kb region centered on a Greek Island or ChIP-seq peak for the factor shown. For heatmap in Figure 5b, all Greek Islands are shown alongside 500 randomly selected ChIP-seq peaks for each factor. For Figure 5c, each row corresponds to an OR gene with showing

1 Kb upstream of the transcriptional start site, 1 Kb downstream of the transcriptional end site, and the gene body scaled to 2 Kb. Signal plots present average data for all regions each set. Heatmaps are sorted by mean signal

DiffBind[68] v2.8.0 was used to calculate ChIP-seq signal in each peak. For this analysis, Diffbind was used to normalize ChIP-seq scores across biological replicate experiments using the "DBA_SCORE_TMM_READS_EFFECTIVE" scoring system, which normalizes using edgeR and the effective library size. The ChIP-seq signal for each peak was then calculated by averaging the normalized score across biological replicates.

ATACseq

ATAC-seq data were analyzed as previously described[7].

RNA-seq

See RNA-seq tab of Supplementary Information 2 for a summary of RNA-seq sequencing data. RNA-seq experiments were conducted as previously described. Briefly, RNA was extracted from FACS purified cells using Trizol and libraries were prepared using Illumina TruSeq Stranded RNA-seq Gold kits. All data sets were processed using 50bp of single end; 75bp reads were trimmed to 50bp and only read 1 was used from paired end data. CutAdapt was used to remove adapter sequences from raw sequencing data and then filtered reads were aligned to the mouse genome (mm10) using STAR[69] v2.5.3a. Samtools was used to select uniquely aligning reads by removing reads with mapping quality below 30 (-q 30). RSeQC[70] v2.6.4 (RRID:SCR_005275) was used to generate RNA-seq signal tracks with signal normalized to a library size of 10,000,000 reads. RNA-seq data analysis was performed in R with the DESeq2[60] v1.20.0 package. Very low abundance transcripts (genes with fewer than 10 counts combined across all samples) were excluded. DESeq2 was used to calculate normalized counts (regularized log transformed), FPKM values, Log2 fold change values, p-values, and p-values adjusted for multiple comparisons.

Immunofluorescence

MOE was dissected from 6-week Ldb1 KO (Ldb1fl/fl;OMPcre) mice and littermate controls. MOE tissue was embedded in OCT and then coronal cryosections were collected at a thickness 12uM. Tissue sections were prepared and stained as previously described[7]. Tissue sections were stained with primary antibodies for Ldb1 (1:1000 dilution, Santa Cruz Biotechnology Cat# sc-11198, RRID:AB_2137017) and Adcy3 (1:200 dilution, Santa Cruz Biotechnology Cat# sc-588, RRID:AB_630839). DNA was labeled with DAPI (2.5ug/mL, Thermo Fisher Scientific Cat# D3571). Primary antibodies were labeled with the following secondary antibodies: for Ldb1, anti-goat IgG conjugated to Alexa-488 (2ug/mL, Thermo Fisher Scientific Cat# A-11055, RRID:AB_2534102), for Adcy3, anti-rabbit IgG conjugated to Alexa-555 (2ug/mL, Thermo Fisher Scientific Cat# A-31572, RRID:AB_162543). Confocal images were collected with a Zeiss LSM 700 and image processing was carried out with ImageJ (NIH).

Statistics

A sample size of two independent biological replicates was selected for high throughput sequencing experiments. This size was selected because the large number of genes/loci measured in high throughput sequencing data sets allows the analysis and modeling of dispersion and variance within and between replicates, thereby allowing the identification of genes/loci with significant differences between conditions using a limited number of replicates. When possible, additional biological replicates were included.

For ChIP-seq, statistically significant peaks were identified using HOMER on each replicate of each experiment. Candidate peaks were selected by setting a read count threshold based upon an input control false discovery rate of 0.001, and then peaks were filtered based upon the following criteria: Poisson p-value over input < 1.00e-04 and Poisson p-value over local region < 1.00e-04. Consensus peak sets were then generated by selecting peaks that overlapped in at least two biological replicates. A two-tailed Wilcoxon rank-sum test was used to determine whether there was a significant difference in the median ChIP-seq peak strength between sets of peaks. For RNA-seq, five biological replicates of Control mOSNs, four biological replicates of Triple Enhancer KO, and four biological replicates of Ldb1 KO mOSNs were analyzed with DESeq2, which generates two-tailed Wald test p-values, and generates adjusted p-values using the Benjamini-Hochberg method. For HiC data, two independent biological replicates were generated for each condition and analyzed separately. Individual biological replicates yielded similar results and were pooled for the analyses presented here. A paired, two-tailed Wilcoxon rank-sum test was used to determine whether the mean frequency of HiC contacts for the set of Greek Islands was different between conditions.

References

47 Shykind, B. M. *et al.* Gene switching and the stability of odorant receptor gene choice. *Cell* **117**, 801-815 (2004).

48 Rock, J. R. *et al.* Basal cells as stem cells of the mouse trachea and human airway epithelium. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 12771-12775, doi:10.1073/pnas.0906850106 (2009).

49 Madisen, L. *et al.* A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nature neuroscience* **13**, 133-140, doi:10.1038/nn.2467 (2010).

50 Vassalli, A., Rothman, A., Feinstein, P., Zapotocky, M. & Mombaerts, P. Minigenes impart odorant receptor-specific axon guidance in the olfactory bulb. *Neuron* **35**, 681-696, doi:S0896627302007936 [pii] (2002).

51 Eggan, K. *et al.* Mice cloned from olfactory sensory neurons. *Nature* **428**, 44-49, doi:10.1038/nature02375 (2004).

52 Mangale, V. S. *et al.* Lhx2 selector activity specifies cortical identity and suppresses hippocampal organizer fate. *Science* **319**, 304-309, doi:10.1126/science.1151695 (2008).

53 Zhao, Y. *et al.* LIM-homeodomain proteins Lhx1 and Lhx5, and their cofactor Ldb1, control Purkinje cell differentiation in the developing cerebellum. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 13182-13186, doi:10.1073/pnas.0705464104 (2007).

54 Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95-98, doi:10.1016/j.cels.2016.07.002 (2016).

55 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.org>q-bio* **1-3** (2013).

56 Droettboom, M. matplotlib/matplotlib v2.2.2. *ZENODO* **1202077**, doi:doi:10.5281 (2018).

57 Waskom, M. mwaskom/seaborn: v0.8.1. *ZENODO* **883859**, doi:doi:10.5281 (2017).

58 McKinney, W. Data Structures for Statistical Computing in Python. *Proc. 9th Python Sci. Conf.* **1697900**, 51-56 (2010).

59 Freese, N. H., Norris, D. C. & Loraine, A. E. Integrated genome browser: visual analytics platform for genomics. *Bioinformatics* **32**, 2089-2095, doi:10.1093/bioinformatics/btw069 (2016).

60 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

61 Klein, F. A. *et al.* FourCSeq: analysis of 4C sequencing data. *Bioinformatics* **31**, 3085-3091, doi:10.1093/bioinformatics/btv335 (2015).

62 Pedregosa, F. Scikit-learn: Machine Learning in Python. . *J. Mach. Learn. Res.* **12**, 2825–2830 (2012).

63 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

64 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

65 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).

66 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

67 Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids research* **44**, W160-165, doi:10.1093/nar/gkw257 (2016).

68 Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389-393, doi:10.1038/nature10730 (2012).

69 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

70 Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic acids research* **40**, e72, doi:10.1093/nar/gks001 (2012).

71 Yan, J. *et al.* Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers. *Cell Res* **28**, 387, doi:10.1038/cr.2018.18 (2018).

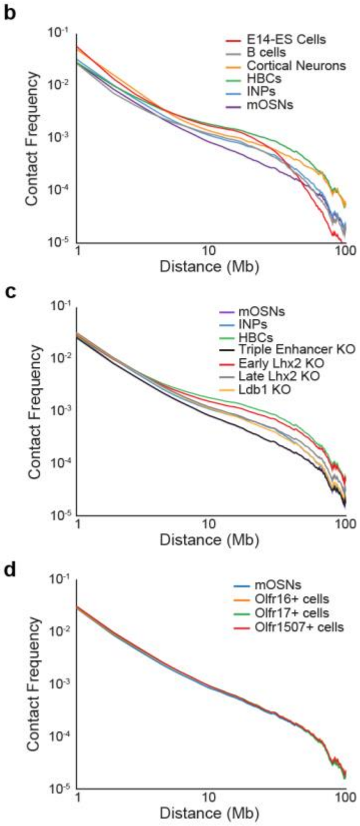
72 Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-572 e524, doi:10.1016/j.cell.2017.09.043 (2017).

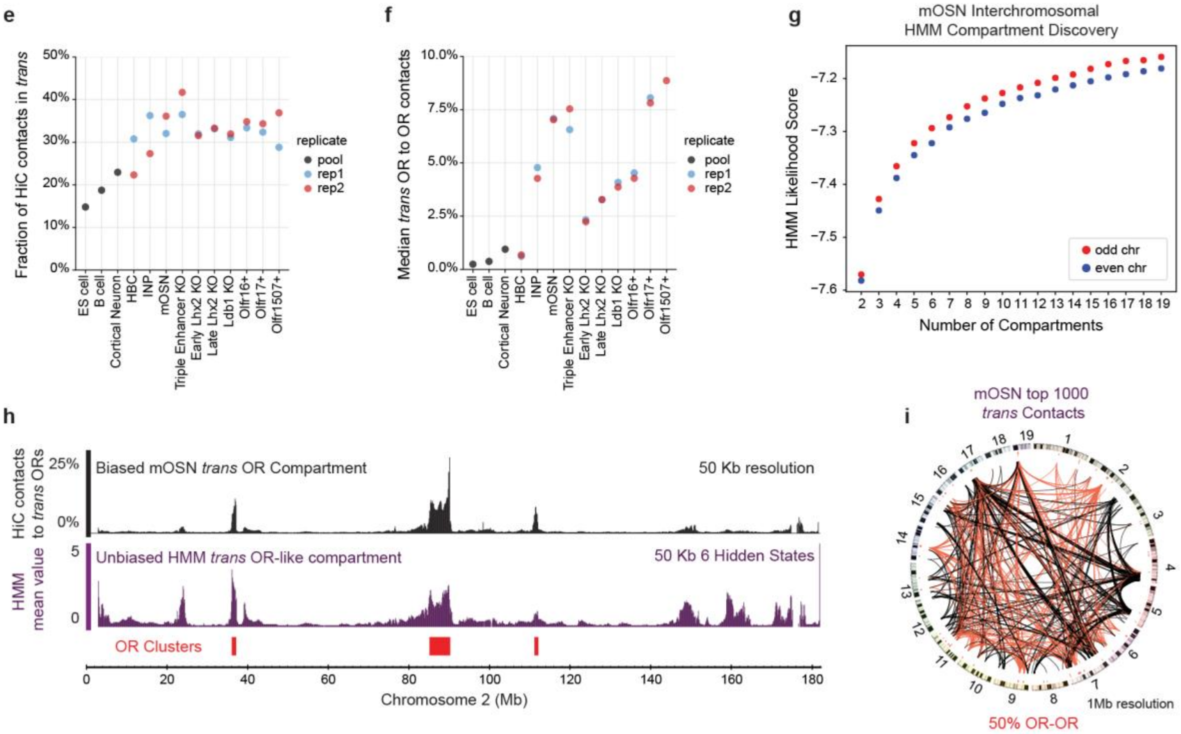
Data Availability Statement

All figures include publicly available data. All ChIP-seq and RNA-seq data reported in this paper (see Supplementary Information 1 and 2) are available from GEO (GSE112153). Additional data (mOSN RNA-seq, mOSN Lhx2 ChIP-seq, mOSN Ebf ChIP-seq, and Olfr1507+ ATAC-seq) were previously described[7] and are available from GEO (GSE93570). All HiC data generated in this study are publicly available at https://data.4dnucleome.org/ under the following accession numbers: 4DNESH4UTRNL, 4DNESNYBDSLY, 4DNES54YB6TQ, 4DNESRE7AK5U, 4DNES425UDGS, 4DNESEPDL6KY.

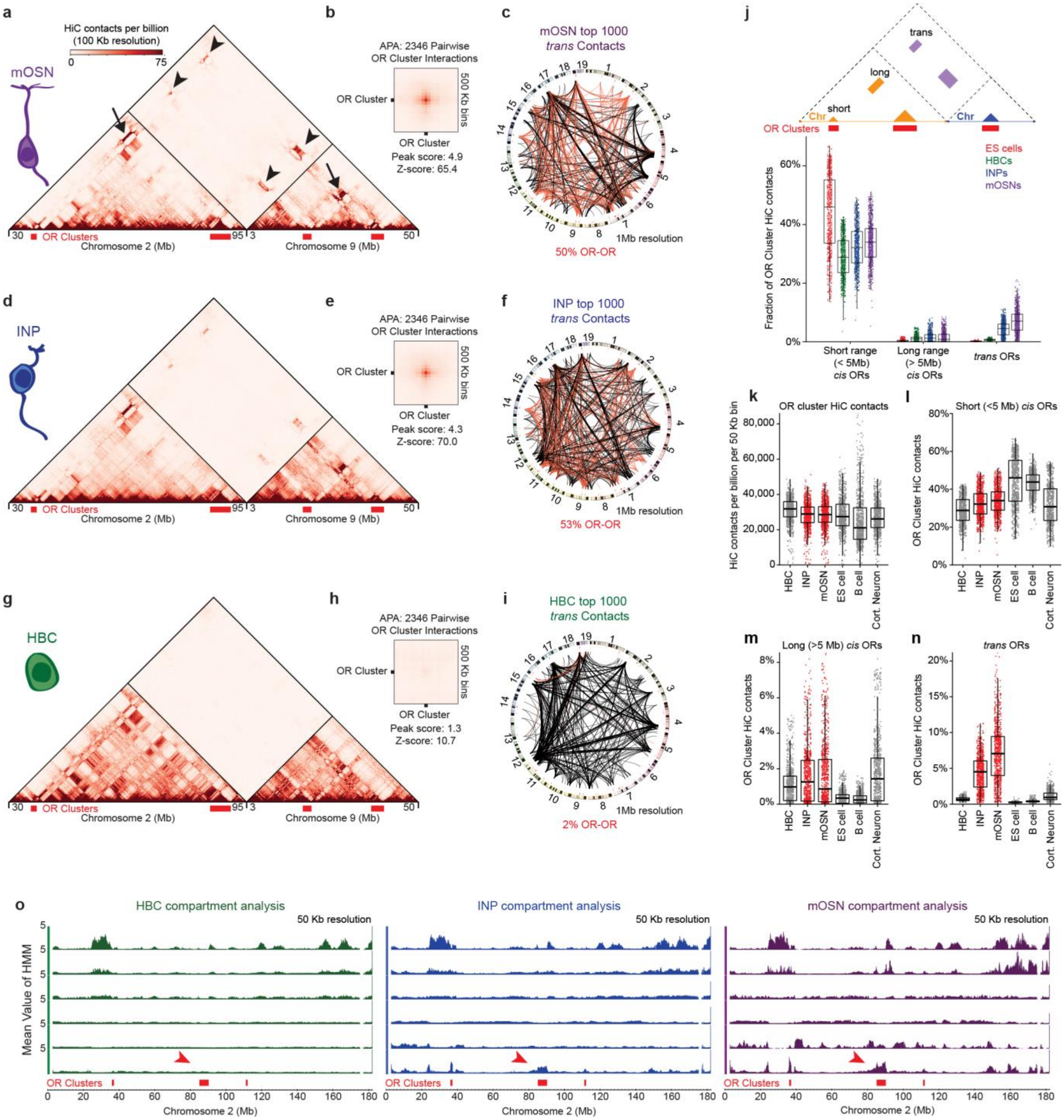
a

Condition	Genotype	Replicate	HiC contacts	
			Total	Interchromosomal
HBC	Krt5Cre;R26R-tdtomato-GFP	1	167,657,359	51,602,222
HBC	Krt5Cre;R26R-tdtomato-GFP	2	225,115,268	50,262,554
Ngn	Ngn1-GFP	1	246,409,005	89,379,271
Ngn	Ngn1-GFP	2	375,978,892	102,730,330
OMP	OMP-ires-GFP	1	252,119,832	80,815,108
OMP	OMP-ires-GFP	2	275,196,670	99,393,678
Triple Enhancer KO	H-/-;Lipsi-/-;Sfaktiria-/-;OMP-ires-GFP	1	229,141,355	83,663,379
Triple Enhancer KO	H-/-;Lipsi-/-;Sfaktiria-/-;OMP-ires-GFP	2	144,819,676	60,397,936
Early Lhx2 KO	Krt5Cre;Lhx2fl/fl;R26R-tdtomato	1	314,430,212	100,500,359
Early Lhx2 KO	Krt5Cre;Lhx2fl/fl;R26R-tdtomato	2	249,476,467	78,746,248
Late Lhx2 KO	OMP-ires-Cre;Lhx2fl/fl;R26R-tdtomato	1	226,965,613	75,233,175
Late Lhx2 KO	OMP-ires-Cre;Lhx2fl/fl;R26R-tdtomato	2	255,920,547	85,214,254
Ldb1 KO	OMP-ires-Cre;Ldb1fl/fl;R26R-tdtomato	1	329,522,878	102,555,125
Ldb1 KO	OMP-ires-Cre;Ldb1fl/fl;R26R-tdtomato	2	218,831,736	69,910,250
Olfr16+	Olfr16-ires-GFP	1	1,143,990,982	381,855,813
Olfr16+	Olfr16-ires-GFP	2	441,663,254	153,807,547
Olfr17+	Olfr17-ires-GFP	1	673,963,559	218,175,148
Olfr17+	Olfr17-ires-GFP	2	545,253,996	187,260,974
Olfr1507+	Olfr1507-ires-GFP	1	542,022,727	156,166,449
Olfr1507+	Olfr1507-ires-GFP	2	1,565,757,634	577,798,102

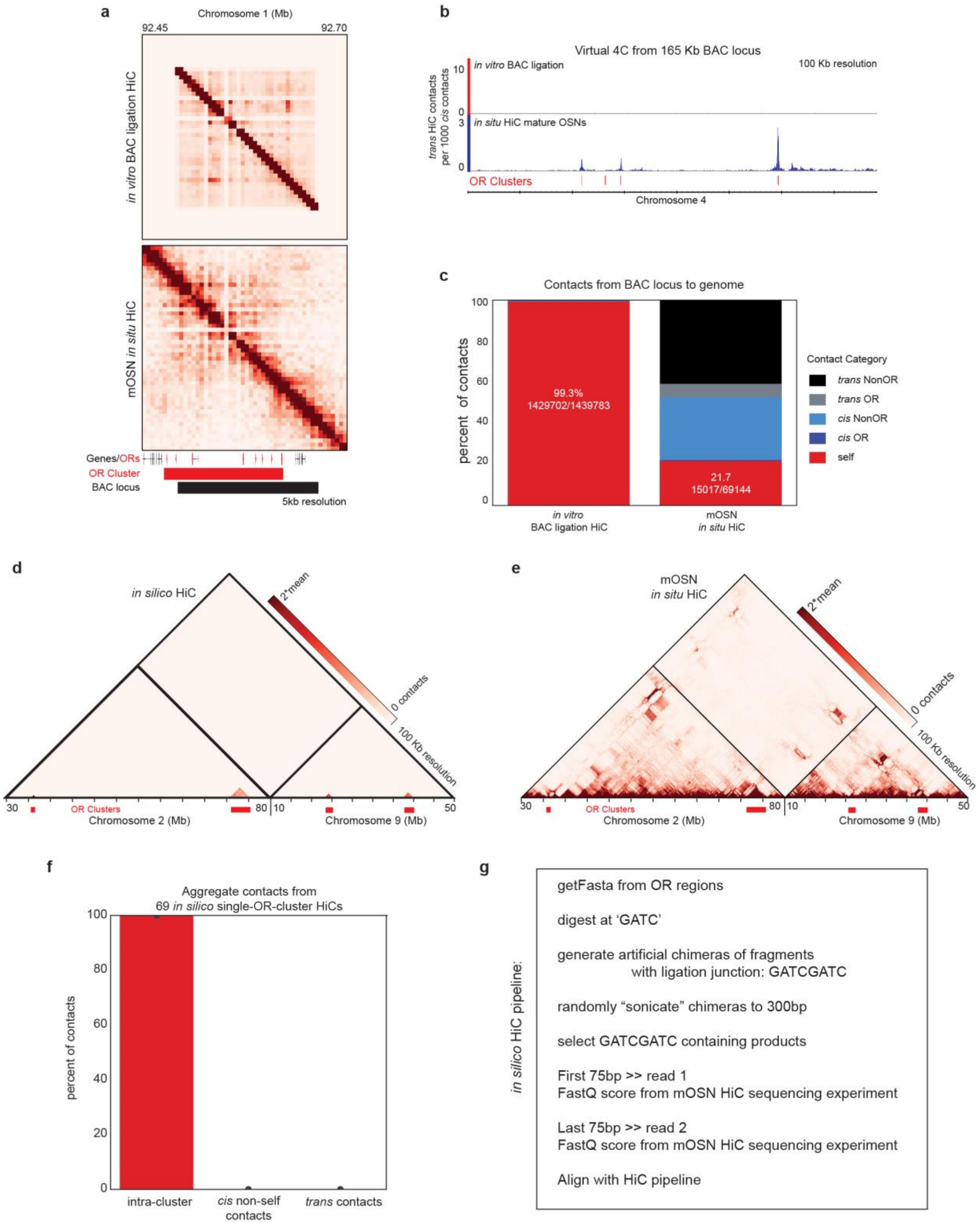




Extended Data Figure 1: HiC on FAC-sorted primary cells from the MOE reveals extensive interchromosomal interactions between OR clusters. a, Table summarizing all HiC experiments in this manuscript separated by biological replicates. The total number of HiC contacts in each replicate and the total number of interchromosomal (*trans*) HiC contacts are shown. **b-d**, HiC contact curves for wild-type conditions (b), for wild-type and mutant MOE populations (c), and for cells sorted based upon the expression of specific OR genes (d). All panels present pooled data from 2 independent biological replicates that yielded similar results when analyzed separately. **e**, graphs showing the proportion of *trans* HiC contacts between replicates of each genotype and cell type. Pooled data from publicly available data sets is shown for ES cells[71], B cells[15], and cortical neurons[72]. **f**, same as e, but showing the median fraction of HiC contacts made to *trans* OR clusters for OR cluster regions divided into 50 Kb bins. **g**, Machine learning Hidden Markov Model (HMM) score for a given number of compartments (see Extended Materials and Methods). 9 compartments were used for further analysis. **h**, From the 9 HMM-derived compartments, one includes predominantly OR clusters (magenta, bottom panel) and overlaps with OR compartments defined by biased analysis of *trans* OR contacts (black top panel). OR gene clusters depicted in red. Scale on the biased analysis represents the percentage of HiC contacts mapped to *trans* OR clusters (pooled data from 2 biological replicates). Scale in the HMM-derived compartments represents the average value of a given locus in a given compartment. **i,** Circos plots depicting the strongest 1000 interchromosomal interactions genomewide at 1 Mb resolution in mOSNs. Red lines represent OR-to-OR contacts and black lines non-OR-to-non-OR contacts. Line thickness increases with contact frequency. Chromosome numbers depicted at the periphery of the circle.


a
HiC contacts per billion
(100 Kb resolution)
0
75
mOSN
30
OR Clusters
Chromosome 2 (Mb)
95
3
Chromosome 9 (Mb)
50
b
APA: 2346 Pairwise
OR Cluster Interactions
OR Cluster
500 Kb bins
OR Cluster
Peak score: 4.9
Z-score: 65.4
c
mOSN top 1000
trans Contacts
1Mb resolution
50% OR-OR
d
INP
e
APA: 2346 Pairwise
OR Cluster Interactions
Peak score: 4.3
Z-score: 70.0
f
INP top 1000
trans Contacts
53% OR-OR
g
HBC
h
APA: 2346 Pairwise
OR Cluster Interactions
Peak score: 1.3
Z-score: 10.7
i
HBC top 1000
trans Contacts
2% OR-OR
j
trans
long
short
Chr
Chr
OR Clusters
ES cells
HBCs
INPs
mOSNs
Fraction of OR Cluster HiC contacts
60%
40%
20%
0%
Short range
(< 5Mb)
cis ORs
Long range
(> 5Mb)
cis ORs
trans ORs
k
OR cluster HiC contacts
HiC contacts per billion per 50 Kb bin
80,000
60,000
40,000
20,000
0
HBC
INP
mOSN
ES cell
B cell
Cort. Neuron
l
Short (<5 Mb) cis ORs
OR Cluster HiC contacts
80%
60%
40%
20%
0%
m
Long (>5 Mb) cis ORs
8%
6%
4%
2%
0%
n
trans ORs
20%
15%
10%
5%
0%
o
HBC compartment analysis
INP compartment analysis
mOSN compartment analysis
50 Kb resolution
Mean Value of HMM
5
OR Clusters
0
20
40
60
80
100
120
140
160
180
Chromosome 2 (Mb)

Extended Data Figure 2: Extensive interchromosomal contacts form between OR gene clusters over OSN differentiation. a-i, In situ HiC contact matrix of chromosomes 2 and 9, Aggregate Peak Analysis (APA), and Circos plot depicting the strongest 1000 interchromosomal interactions genomewide for mOSNs (a-c), INPs (d-f), and HBCs (g-i). All three sets of analyses reveal an increase in *trans* OR cluster interactions over the course of differentiation. **j**, For OR gene clusters (divided into 50 Kb bins, n=768 bins) the frequency of *cis* short (<5Mb distance, including self), *cis* long (>5 Mb), and *trans* contacts with OR clusters is shown, expressed as the fraction of total HiC contacts mapped to each bin. **k,** Number of HiC contacts, normalized to a library size of one billion HiC contacts genomewide, observed for each OR cluster region (divided into 50 Kb bins, n=768 bins) in HBCs, INPs, mOSNs, ES cells, B cells, and cortical neurons. **l-n**, For OR cluster regions (divided into 50 Kb bins, n=768 bins), the fraction of total HiC contacts that are made to ORs clusters located in short range *cis* (l), long range *cis* (m) and *trans* (n). **o,** The 6 most distinct HMM-derived compartments of chromosome 2 in HBCs (green, left), INPs (blue, middle) and mOSNs (magenta, right). OR clusters emerge as distinct compartment in INPs and strengthen in mOSNs. For all boxplots, box indicates median, upper, and lower quartiles while whiskers indicate 1.5 * the interquartile range. All panels present pooled data from 2 independent biological replicates that yielded similar results when analyzed separately.



getFasta from OR regions

digest at 'GATC'

generate artificial chimeras of fragments with ligation junction: GATCGATC

randomly "sonicate" chimeras to 300bp

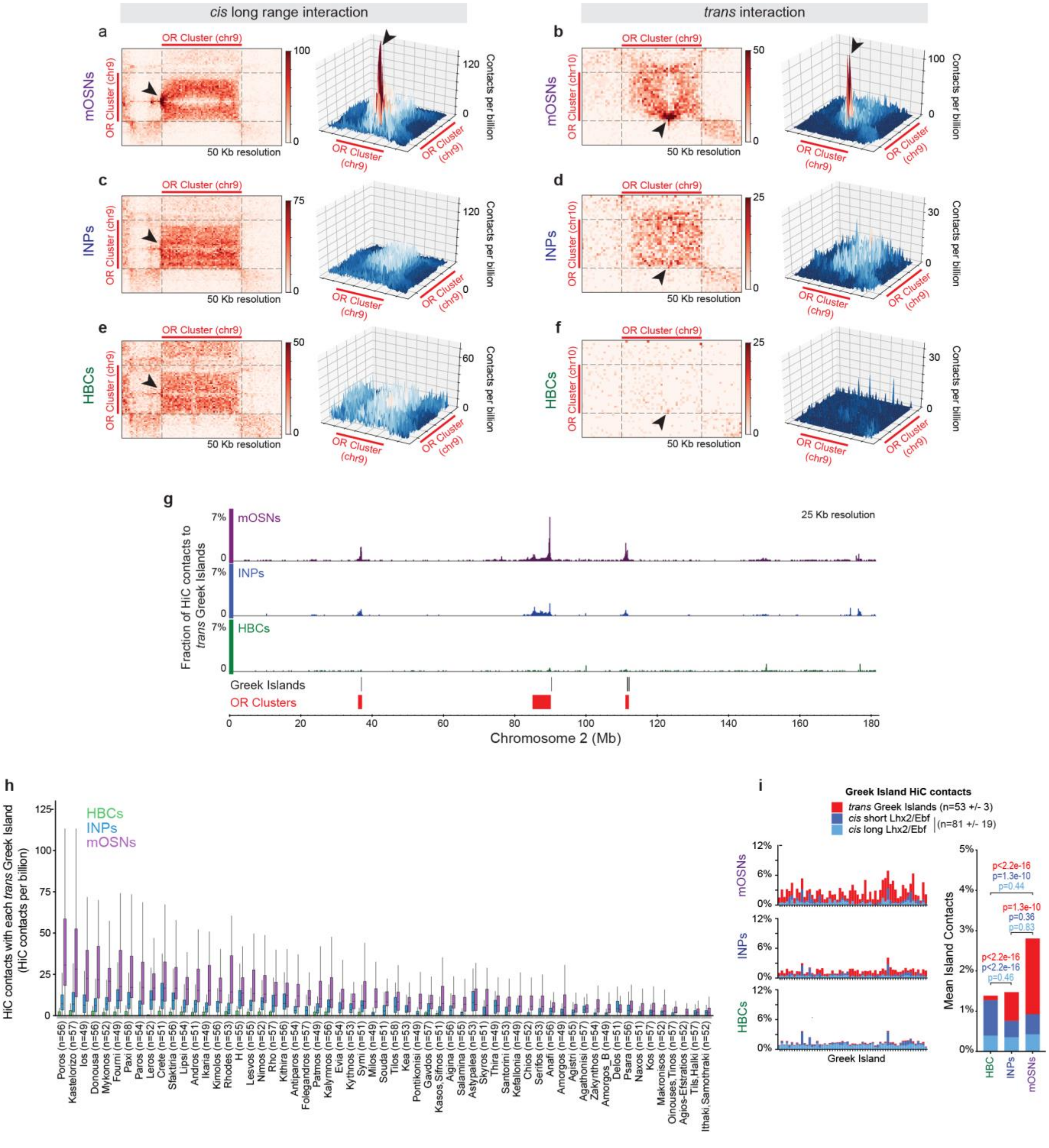
select GATCGATC containing products

First 75bp >> read 1
FastQ score from mOSN HiC sequencing experiment

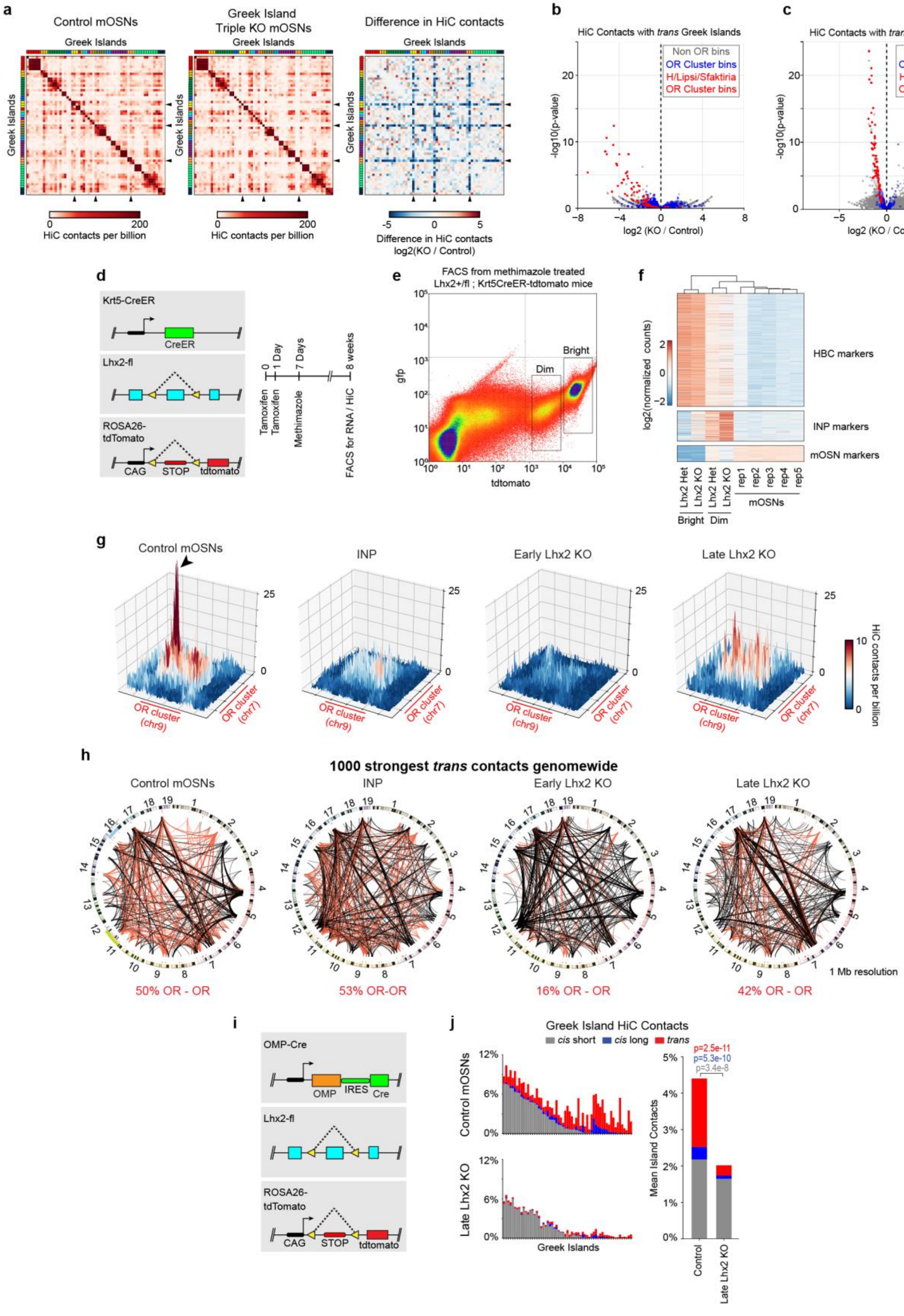
Last 75bp >> read 2
FastQ score from mOSN HiC sequencing experiment

Align with HiC pipeline

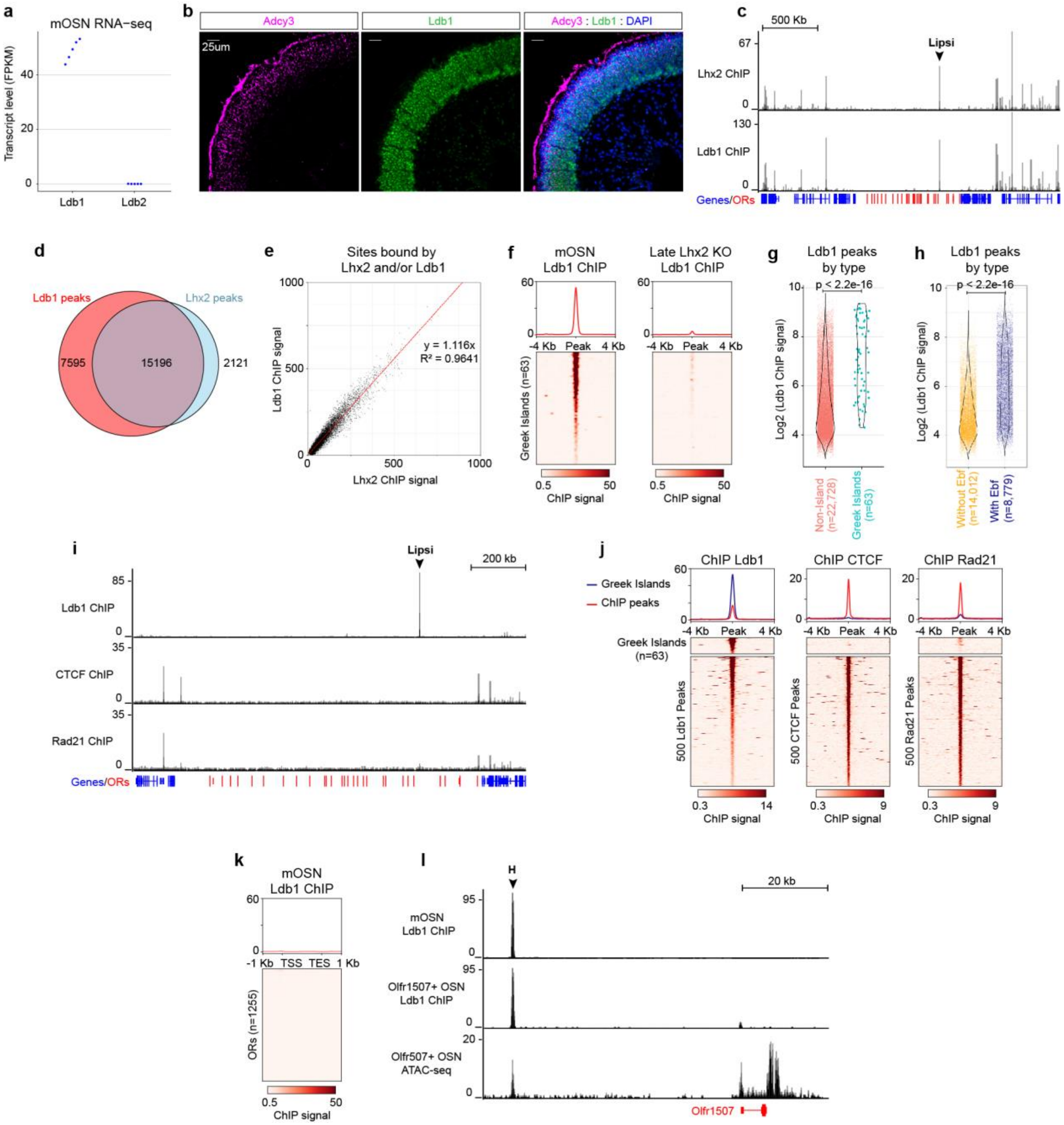
Extended Data Figure 3: *In vitro* and *in silico* HiC experiments show that OR HiC contacts are generated by unique sequences that do not map to other OR clusters. a, Contact matrix from *in vitro* HiC (top) using a 165Kb BAC plasmid containing 7 OR genes from an OR cluster from chromosome 1 and *in situ* HiC from mOSNs (bottom). HiC contacts in the BAC HiC are restricted to the coordinates of the BAC plasmid and do not extend to two OR genes from this cluster that are absent from the BAC. **b**, Virtual 4C from the 165 Kb BAC region to chromosome 2, which contains the highest number of OR genes. On top, virtual 4C from the BAC *in vitro* HiC shows that no reads mapped to ORs from chromosome 2, whereas the same 165 Kb regions makes abundant *trans* contacts with these ORs in mOSNs. **c**, 99.3% of all the BAC HiC contacts map within the BAC, whereas in mOSNs only 21.7% of the BAC region HiC contacts map within the BAC. **d**, *In silico* HiC analysis shows complete absence of mis-mapped reads corresponding to OR clusters under the mapping conditions used throughout the manuscript (removing mapq<30). Each OR cluster was subjected to intra-cluster in silico HiC (g) and then the HiC contacts of the 69 OR clusters were mapped in aggregate to the whole genome. As seen in the contact matrix from chromosomes 2 and 9(d), the in silico reads only map within clusters, with no mis-mapped reads that would erroneously be interpreted as inter-cluster *cis* or *trans* contacts. **e**, For reference, the corresponding *in situ* HiC from mOSNs. **f**, Aggregate analysis for all 69 OR gene clusters shows that our mapping protocol does not mis-map any HiC contacts to the wrong OR cluster. **g**, Brief description of the pipeline used for the *in silico* analysis.


cis long range interaction
trans interaction
a
b
c
d
e
f
mOSNs
INPs
HBCs
OR Cluster (chr9)
OR Cluster (chr10)
50 Kb resolution
100
75
50
25
0
120
60
30
Contacts per billion
OR Cluster (chr9)
g
25 Kb resolution
Fraction of HiC contacts to trans Greek Islands
7%
0
mOSNs
INPs
HBCs
Greek Islands
OR Clusters
0
20
40
60
80
100
120
140
160
180
Chromosome 2 (Mb)
h
HBCs
INPs
mOSNs
HiC contacts with each trans Greek Island (HiC contacts per billion)
125
100
75
50
25
0
Poros (n=56)
Kastelorizo (n=57)
Ios (n=49)
Donousa (n=56)
Mykonos (n=52)
Fourni (n=49)
Paxi (n=58)
Paros (n=54)
Leros (n=52)
Crete (n=51)
Sfaktiria (n=56)
Lipsi (n=54)
Andros (n=51)
Ikaria (n=49)
Kimolos (n=56)
Rhodes (n=53)
H (n=55)
Lesvos (n=55)
Nimos (n=52)
Rho (n=57)
Kithira (n=56)
Antiparos (n=54)
Folegandros (n=57)
Patmos (n=49)
Kalymnos (n=56)
Evia (n=54)
Kythnos (n=53)
Symi (n=51)
Milos (n=49)
Souda (n=51)
Tilos (n=58)
Kea (n=53)
Pontikonisi (n=49)
Gavdos (n=57)
Kasos,Sifnos (n=51)
Aigina (n=56)
Salamina (n=55)
Astypalea (n=53)
Skyros (n=51)
Thira (n=49)
Santorini (n=53)
Kefallonia (n=49)
Chios (n=52)
Serifos (n=53)
Anafi (n=56)
Amorgos (n=49)
Agistri (n=55)
Agathonisi (n=57)
Zakynthos (n=54)
Amorgos_B (n=49)
Delos (n=51)
Psara (n=56)
Naxos (n=51)
Kos (n=57)
Makronisos (n=52)
Oinouses,Tinos (n=57)
Agios-Efstratios (n=52)
Tils,Halki (n=57)
Ithaki,Samothraki (n=52)
i
Greek Island HiC contacts
trans Greek Islands (n=53 +/- 3)
cis short Lhx2/Ebf
cis long Lhx2/Ebf
(n=81 +/- 19)
12%
6%
0%
mOSNs
INPs
HBCs
Greek Island
5%
4%
3%
2%
1%
0%
Mean Island Contacts
p<2.2e-16
p=1.3e-10
p=0.44
p=1.3e-10
p=0.36
p=0.83
p<2.2e-16
p<2.2e-16
p=0.46
HBC
INPs
mOSNs

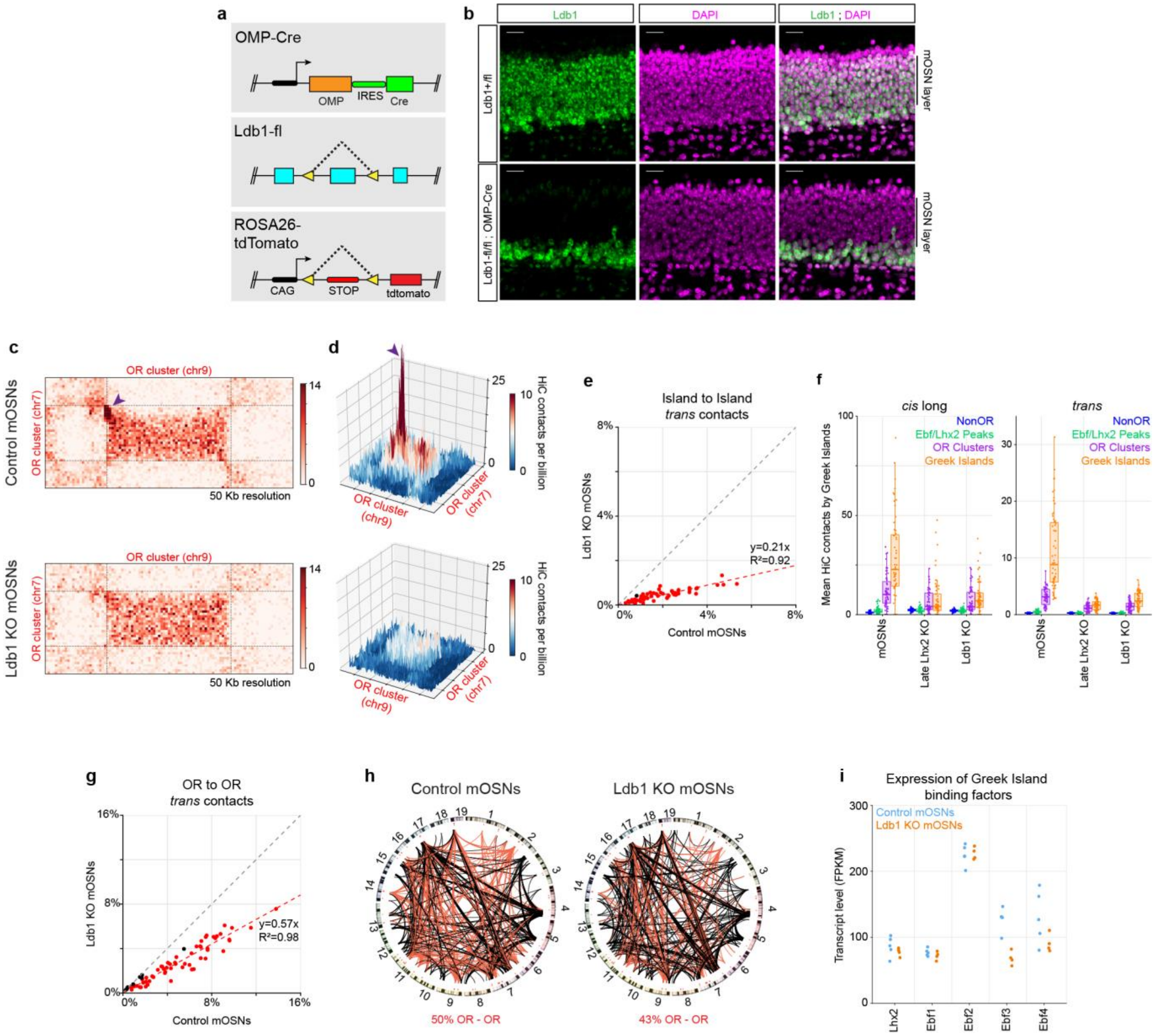
Extended Data Figure 4: Greek Islands make differentiation-dependent contacts with other Greek Islands in *trans* that are stronger than *cis* contacts with Lhx2/Ebf peaks. a-b, Heatmap and 3D projection of HiC contacts between a pair of OR gene clusters in *cis* (a) and *trans* (b) reveals a local maximum of *in situ* HiC interactions between Greek Island loci (arrowheads) in mOSNs. **c-f**, Same as a, b but for immediate neuronal precursors (INPs) and Horizontal Basal Cells (HBCs). **g**, For chromosome 2, fraction of all HiC contacts made to *trans* Greek Islands in mOSNs (top), INPs (middle) and HBCs (bottom). **h,** For each Greek Island, the distribution of HiC contacts, expressed as contacts per billion, made to individual Greek Islands located in *trans* for HBCs, INPs and mOSNs. Box indicates median, upper, and lower quartiles while whiskers indicate 1.5 * the interquartile range. For each Greek Island, the number of *trans* Greek Islands is listed. **i,** (left) Comparison of the total fraction of HiC contacts made by each Greek Islands to intergenic Lhx2/Ebf co-bound peaks present in *cis* versus Greek Islands present in *trans* for HBCs, INPs and mOSNs. For each category we compare roughly equal numbers of peaks (number of *trans* Greek Islands for each Island versus number of *cis* Lhx2/Ebf sites for each Island, mean+/- standard deviation). (right) Mean fraction of HiC contacts across all Greek islands (two-sided, paired Wilcoxon signed-rank test, n=59). Contacts with *trans* Greek Islands (red) constitute a higher fraction of HiC contacts than short-range *cis* (dark blue) or long-range *cis* (light blue) contacts with intergenic Lhx2/Ebf peaks. All panels present pooled data from 2 independent biological replicates that yielded similar results when analyzed separately.


a
Control mOSNs
Greek Islands
Greek Islands
HiC contacts per billion
0
200
Greek Island Triple KO mOSNs
Greek Islands
Greek Islands
HiC contacts per billion
0
200
Difference in HiC contacts
Greek Islands
Greek Islands
-5
0
5
Difference in HiC contacts log2(KO / Control)
b
HiC Contacts with trans Greek Islands
Non OR bins
OR Cluster bins
H/Lipsi/Sfaktiria OR Cluster bins
-log10(p-value)
log2 (KO / Control)
0
10
20
-8
-4
0
4
8
c
HiC Contacts with trans OR Clusters
Non OR bins
OR Cluster bins
H/Lipsi/Sfaktiria OR Cluster bins
-log10(p-value)
log2 (KO / Control)
0
10
20
-5
0
5
d
Krt5-CreER
CreER
Lhx2-fl
ROSA26-tdTomato
CAG
STOP
tdtomato
0
1 Day
7 Days
8 weeks
Tamoxifen
Tamoxifen
Methimazole
FACS for RNA / HiC
e
FACS from methimazole treated Lhx2+/fl ; Krt5CreER-tdtomato mice
gfp
tdtomato
Dim
Bright
f
log2(normalized counts)
2
0
-2
HBC markers
INP markers
mOSN markers
Lhx2 Het
Lhx2 KO
Lhx2 Het
Lhx2 KO
rep1
rep2
rep3
rep4
rep5
Bright
Dim
mOSNs
g
Control mOSNs
INP
Early Lhx2 KO
Late Lhx2 KO
25
0
OR cluster (chr9)
OR cluster (chr7)
HiC contacts per billion
10
0
h
1000 strongest trans contacts genomewide
Control mOSNs
INP
Early Lhx2 KO
Late Lhx2 KO
50% OR - OR
53% OR-OR
16% OR - OR
42% OR - OR
1 Mb resolution
i
OMP-Cre
OMP
IRES
Cre
Lhx2-fl
ROSA26-tdTomato
CAG
STOP
tdtomato
j
Greek Island HiC Contacts
cis short
cis long
trans
Control mOSNs
Late Lhx2 KO
12%
6%
0%
Greek Islands
p=2.5e-11
p=5.3e-10
p=3.4e-8
Mean Island Contacts
5%
4%
3%
2%
1%
0%
Control
Late Lhx2 KO

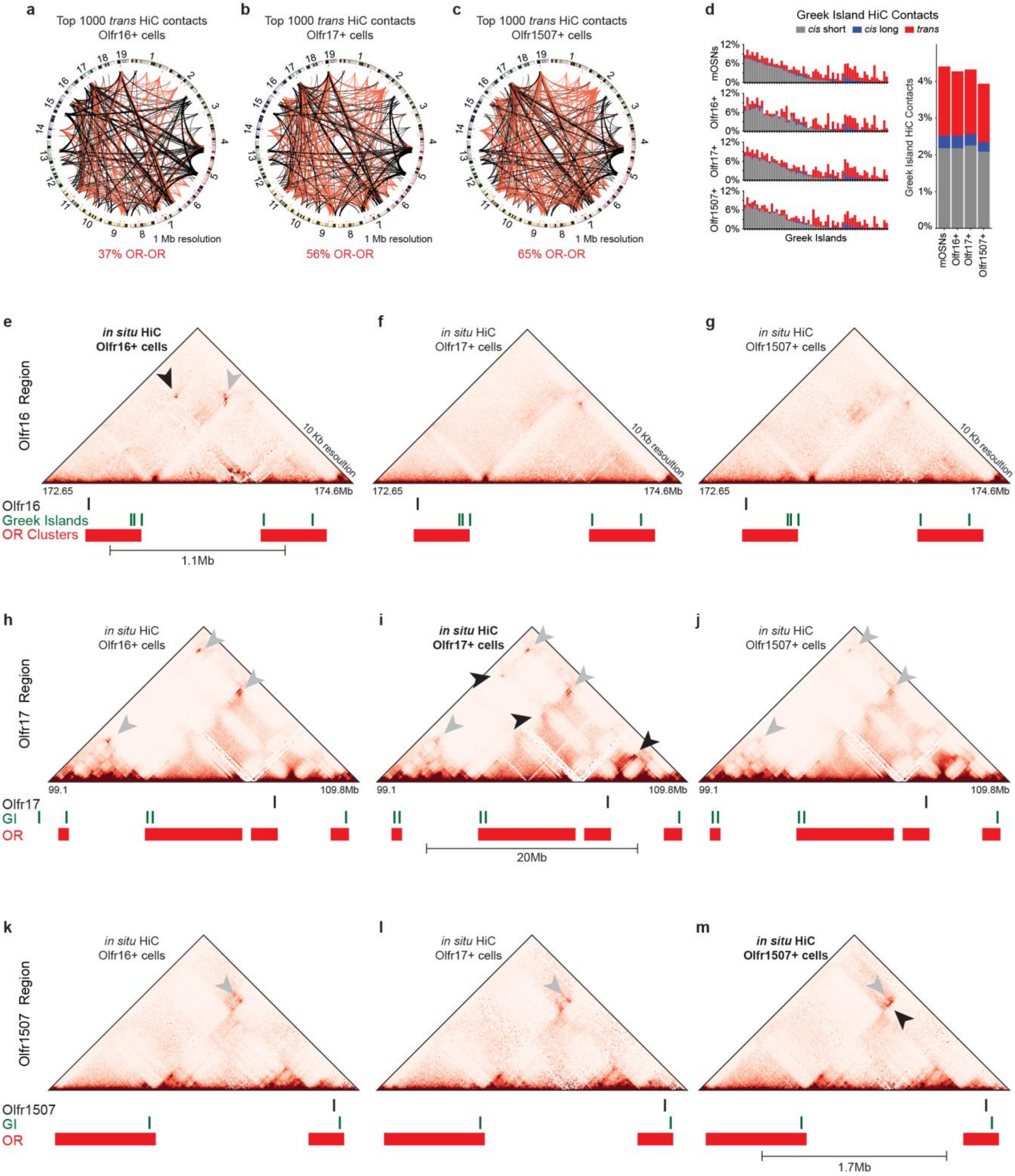
Extended Data Figure 5: Greek Islands and Lhx2 are required for OR compartmentalization in developing OSNs. a, Pairwise HiC contacts between all pairs of Greek Islands ordered by genomic position in Control (left) and Greek Island Triple KO (right) mOSNs. The 50 Kb regions containing the deleted Greek Islands are marked with arrowheads. Plotting the log2 fold difference in HiC contacts (right) reveals that consistent strong reductions are observed for the deleted Islands. Color bar depicts chromosome. **b,c,** The genomic regions exhibiting the most significant reductions in HiC contacts with *trans* OR Greek Islands (b) or *trans* OR clusters (c) in Triple KO mOSN relative to control mOSNs are mostly located within the 3 OR clusters containing the with Greek Island deletions (two biological replicates per condition, see Extended Materials and Methods). **d,** Genetic and experimental strategy for early Lhx2 deletion. Tamoxifen induction with Krt5CreER deletes Lhx2 in HBCs and then methimazole treatment ablates INPs/mOSNs, leading to regeneration from Lhx2-deleted HBCs. **e,** fluorescent labeling of the HBC-derived cells upon methimazole induction reveals two major populations, bright and dim. **f,** By RNA-seq the dim population expresses markers of INPs and mOSNs while the bright population expresses markers of HBCs. Counts are normalized by row. **g,** 3-D projection of HiC contacts between OR clusters located on different chromosomes in control mOSNs (left), INPs, early Lhx2 KO, and late Lhx2 KO (right) cells. A HiC hotspot between interacting Greek Islands is only observed in control mOSNs (arrowhead). In addition, a strong reduction in the surrounding OR-OR contacts relative to mOSNs or INPs is observed in the early Lhx2 KO. **h,** Circos plots depicting the strongest 1000 interchromosomal interactions genomewide at 1 Mb resolution in mOSNs (left), INPs, early Lhx2 KO cells, and late Lhx2 KO cells (right). Red lines represent OR-OR contacts and black lines non-OR-non-OR contacts. Line thickness increases with contact frequency. Chromosome numbers depicted at the periphery of the circle. **i**, Genetic strategy for late Lhx2 deletion and fluorescent marking of Lhx2 KO mOSNs. **j,** (left) For each Greek Island, the fraction of total HiC contacts made to other Greek Islands located in *cis* at short range (<5 Mb apart, grey), long range (>5Mb apart, blue), and in *trans* (red). Top panel represents control mOSNs and bottom panel late Lhx2 KO cells. (right) The effect of late Lhx2 KO on the mean fraction of HiC contacts across all Greek Islands (two-sided, paired Wilcoxon signed-rank test, n=59).


a
mOSN RNA−seq
Transcript level (FPKM)
0
20
40
Ldb1
Ldb2
b
Adcy3
Ldb1
Adcy3 : Ldb1 : DAPI
25um
c
500 Kb
Lipsi
67
Lhx2 ChIP
0
130
Ldb1 ChIP
0
Genes/ORs
d
Ldb1 peaks
Lhx2 peaks
7595
15196
2121
e
Sites bound by Lhx2 and/or Ldb1
Ldb1 ChIP signal
Lhx2 ChIP signal
0
500
1000
y = 1.116x
R² = 0.9641
f
mOSN Ldb1 ChIP
Late Lhx2 KO Ldb1 ChIP
60
0
-4 Kb
Peak
4 Kb
Greek Islands (n=63)
0.5
50
ChIP signal
g
Ldb1 peaks by type
p < 2.2e-16
Log2 (Ldb1 ChIP signal)
4
6
8
10
Non-Island (n=22,728)
Greek Islands (n=63)
h
Ldb1 peaks by type
p < 2.2e-16
Log2 (Ldb1 ChIP signal)
Without Ebf (n=14,012)
With Ebf (n=8,779)
i
Lipsi
200 kb
85
Ldb1 ChIP
0
35
CTCF ChIP
0
35
Rad21 ChIP
0
Genes/ORs
j
ChIP Ldb1
ChIP CTCF
ChIP Rad21
Greek Islands
ChIP peaks
60
20
0
-4 Kb
Peak
4 Kb
Greek Islands (n=63)
500 Ldb1 Peaks
500 CTCF Peaks
500 Rad21 Peaks
0.3
14
9
ChIP signal
k
mOSN Ldb1 ChIP
60
0
-1 Kb
TSS
TES
1 Kb
ORs (n=1255)
0.5
50
ChIP signal
l
H
20 kb
95
mOSN Ldb1 ChIP
0
95
Olfr1507+ OSN Ldb1 ChIP
0
20
Olfr507+ OSN ATAC-seq
0
Olfr1507

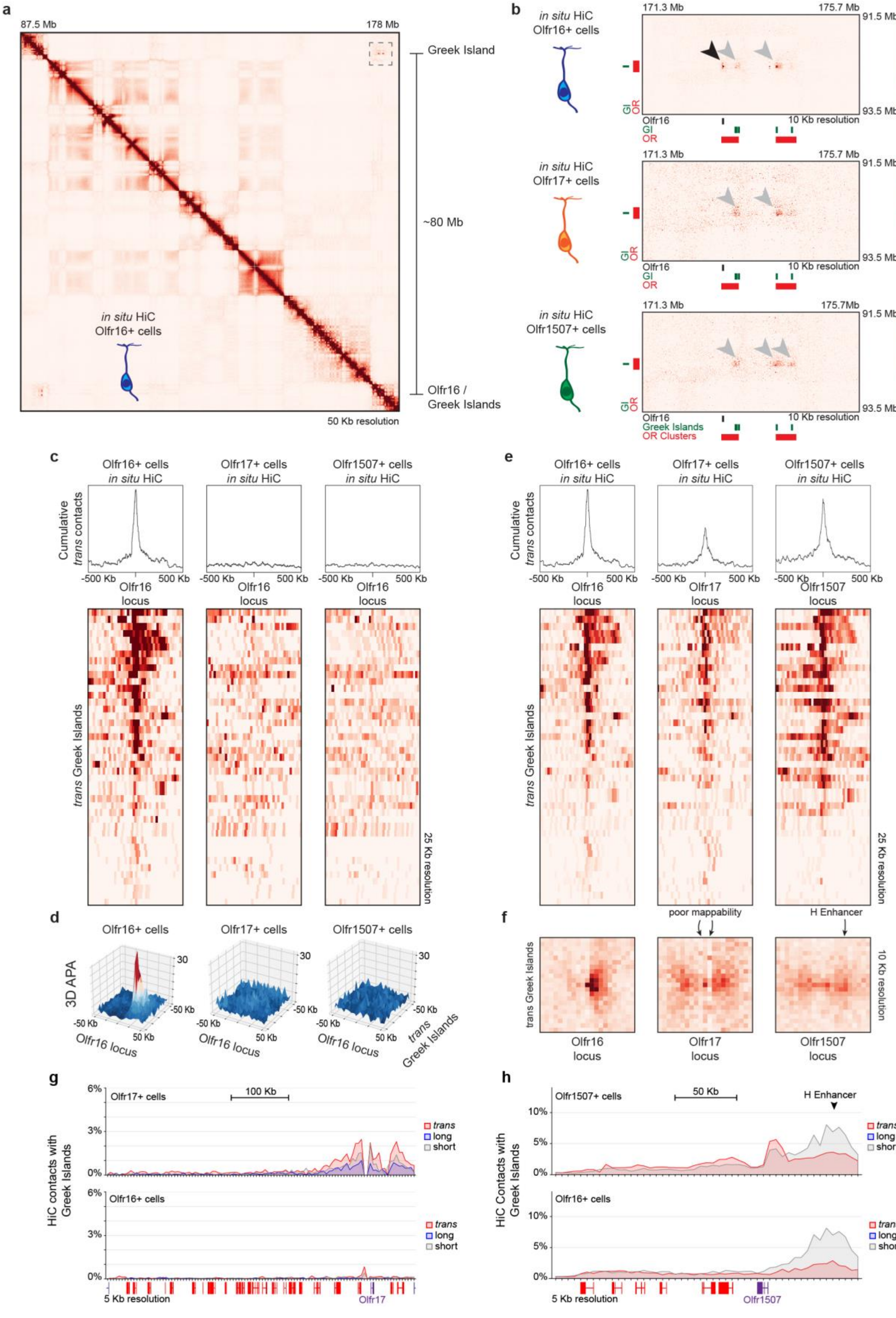
Extended Data Figure 6: Ldb1 expression and genomic distribution in mOSNs. a, Transcript level, expressed as fragments per kilobase per million mapped reads (fpkm), of the two Ldb family members in mOSN RNA-seq data sets (n=5 biological replicates). **b,** Sections of olfactory epithelium stained for Ldb1 (green) and Adcy3 (magenta), a marker for mOSNs. Nuclei are labeled with DAPI (blue). Scale bar = 25um**.** Similar results were obtained from four independent experiments. **c**, Ldb1 and Lhx2 ChIP-seq signal in mOSNs across the OR gene cluster containing the Greek Island Lipsi. OR genes are red and all other genes are blue. Plot shows pooled data from 2 biological replicates for Lhx2 and 3 biological replicates for Ldb1, each of which yielded similar results when analyzed separately. Values are counts per 10 million reads. **d,** Extensive overlap between consensus Lhx2 and Ldb1 ChIP-seq peak sets. **e,** linear relationship between normalized Lhx2 ChIP signal and Ldb1 ChIP signal. Any peak observed in at least two of the 5 experiments (2 for Lhx2 and 3 for Ldb1) was included (n=26,667) and plotted together with a best fit line obtained by linear regression with y-intercept set to 0. **f**, Ldb1 ChIP signal over Greek Islands in mOSNs and Late Lhx2 KO mOSNs. Heatmap shows pooled data from 3 biological replicates for mOSNs and 2 biological replicates for Late Lhx2 KO cells, each of which yielded similar results when analyzed separately. Values are counts per 10 million reads. **g,** Normalized Ldb1 ChIP-seq signal is greater for Ldb1 peaks that overlap Greek Islands than for peaks that do not (p < 2.2e-16, two-sided Wilcoxon rank sum test, n=63 for Greek Islands, n=22,728 for non-Island peaks). Violin plots are scaled to the same area and show density for the full set of points over the full range. **h,** Normalized Ldb1 ChIP-seq signal is greater for Ldb1 peaks that overlap Ebf ChIP peaks than for peaks that do not (p < 2.2e-16, two-sided Wilcoxon rank sum test, n=8,779 for Ldb1 peaks that overlap Ebf peaks, n=14,012 for non-Ebf peaks). Violin plots are scaled to the same area and show density for the full set of points over the full range. **i,** mOSN ChIP-seq for Ldb1, CTCF, and the cohesin-subunit Rad21 across the OR gene cluster containing the Greek Island Lipsi. OR genes are red and all other genes are blue. Plot shows pooled data from 3 biological replicates for Ldb1 and 2 biological replicates CTCF and Rad21. Values are counts per 10 million reads. Analyzing each replicate separately yielded similar results. **j**, mOSN ChIP signal over Greek Islands and non-Geek Island ChIP-seq peaks. For ChIP-seq peaks, the heatmap shows 500 randomly selected peaks and the plot shows data from the full consensus set of peaks (n=22,791 for Ldb1, n=24,883 for CTCF, and n=9,882 for Rad21). Plots show pooled data, similar results were obtained with each replicate (n=3 for Ldb1 ChIP-seq and n=2 for CTCF and Rad21 ChIP-seq). Units are counts per 10 million reads. **k**, As in j, but showing Ldb1 ChIP signal over OR genes (n=1,255) in mOSNs. **l,** Ldb1 ChIP-seq from control mOSNs (top) and Olfr1507-expressing cells (middle). Strong signal is observed on the Greek Island, H, in both populations but only a very weak signal on the Olfr1507 promoter when it is transcriptionally engaged. Pooled data from 3 biological replicates is shown for the mOSNs. One of two biological replicates is shown for Olfr1507+ OSNs; the other replicate yielded similar results but with lower enrichment in peaks. ATAC-seq from the Olfr1507-expressing cells (bottom) shows that the promoter of Olfr1507 has similar accessibility to the H element. ATAC-seq plot shows pooled data from two biological replicates that yielded similar results.


a
OMP-Cre
OMP
IRES
Cre
Ldb1-fl
ROSA26-tdTomato
CAG
STOP
tdtomato
b
Ldb1
DAPI
Ldb1 ; DAPI
Ldb1+/fl
Ldb1-fl/fl ; OMP-Cre
mOSN layer
c
Control mOSNs
Ldb1 KO mOSNs
OR cluster (chr9)
OR cluster (chr7)
50 Kb resolution
14
0
d
HiC contacts per billion
25
10
0
OR cluster (chr9)
OR cluster (chr7)
e
Island to Island
trans contacts
Ldb1 KO mOSNs
Control mOSNs
8%
4%
0%
y=0.21x
R²=0.92
f
cis long
trans
NonOR
Ebf/Lhx2 Peaks
OR Clusters
Greek Islands
Mean HiC contacts by Greek Islands
100
50
0
30
20
10
mOSNs
Late Lhx2 KO
Ldb1 KO
g
OR to OR
trans contacts
Ldb1 KO mOSNs
Control mOSNs
16%
8%
0%
y=0.57x
R²=0.98
h
Control mOSNs
Ldb1 KO mOSNs
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
50% OR - OR
43% OR - OR
i
Expression of Greek Island binding factors
Control mOSNs
Ldb1 KO mOSNs
Transcript level (FPKM)
300
200
100
0
Lhx2
Ebf1
Ebf2
Ebf3
Ebf4

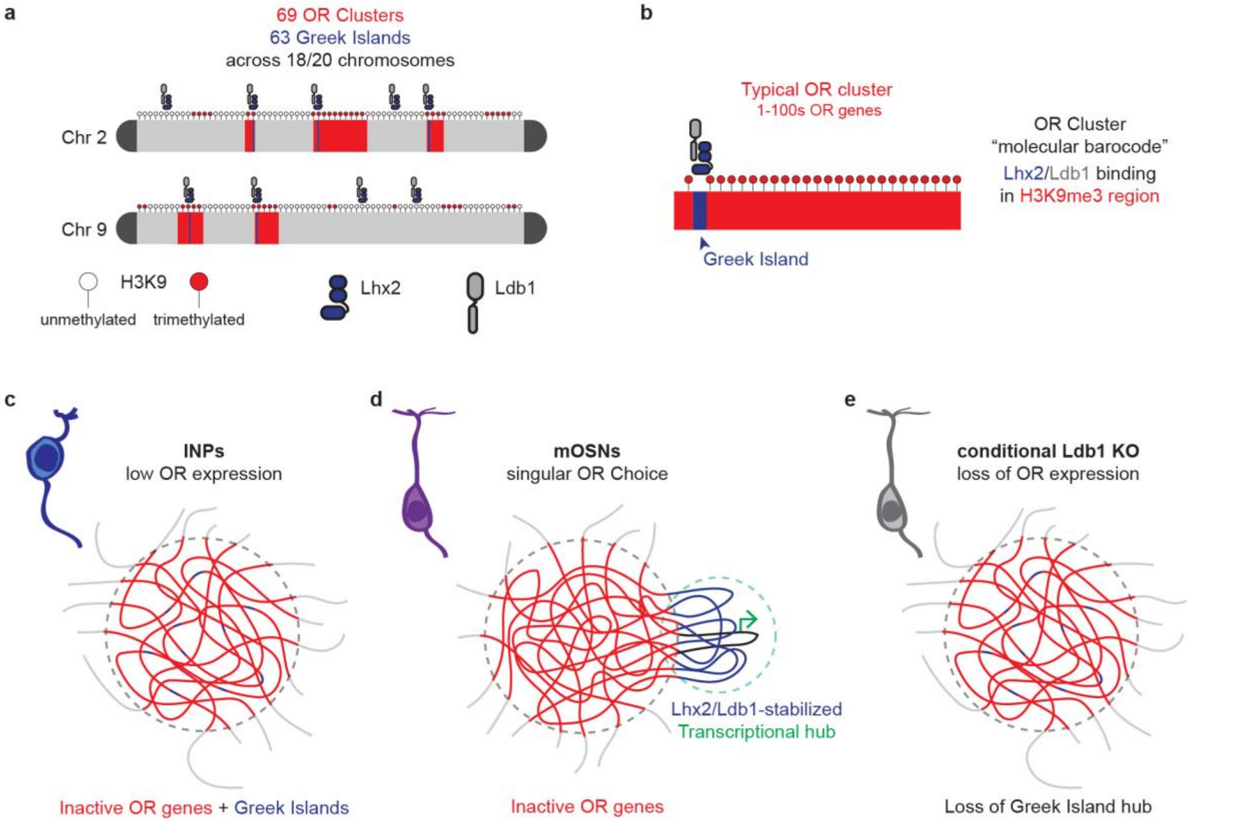
Extended Data Figure 7: Effects of conditional Ldb1 deletion in Greek Island interactions and OR expression. a, Schematic of the genetic strategy used to generate Ldb1 KO mOSNs that are fluorescently labeled **b,** In Ldb1fl/fl;OMP-Cre mice, Ldb1 (green) is lost from mOSNs but retained in basal immature cells. Nuclei are stained with DAPI (magenta). Scale bar = 20um. Similar results were obtained from three independent experiments. **c**, HiC contacts between a pair of OR clusters located on different chromosomes in control (top), and Ldb1 KO (bottom) OSNs. A HiC hotspot between interacting Greek Islands in control mOSNs (arrowheads) is absent in Ldb1 KO OSNs **d,** 3D projection of the same OR cluster pair in control and Ldb1 KO OSNs. **e,** *trans* interactions of each Greek Island (n=59) with the other Greek Islands as fraction of the total HiC contacts in mOSNs versus Ldb1 KO cells. Greek Islands changed more than 2-fold are red. **f**, For each Greek Island, the mean number of *cis* long range (left) and *trans* (right) HiC contacts per billion made to every non-OR sequence (at 50 Kb resolution), intergenic Lhx2 & Ebf bound peak (outside of OR clusters), or Greek Island. Box indicates median, upper, and lower quartiles while whiskers indicate 1.5 * the interquartile range. **g**, same as e but for *trans* contacts between OR gene clusters (n=67). Clusters changed more than 1.5-fold are red. **h,** Circos plots depicting the strongest 1000 interchromosomal interactions genomewide at 1Mb resolution in control mOSNs (left), Ldb1 KO mOSNs (right). Red lines represent OR-OR contacts and black lines non-OR-non-OR contacts. Line thickness increases with contact frequency. Chromosome numbers depicted at the periphery of the circle. **i,** Transcript levels of Greek Island-binding factors in RNA-seq data from control mOSNs and Ldb1 KO mOSNs. Transcript levels of Ebf3 are reduced approximately 2-fold (p = 0.031 for greater than 1.5-fold change, DESeq2 normalized Wald test with n=5 for control mOSNs and n=4 Ldb1 KO). The expression of other factors is not significantly different between conditions.


a
Top 1000 *trans* HiC contacts
Olfr16+ cells
1 Mb resolution
37% OR-OR
b
Top 1000 *trans* HiC contacts
Olfr17+ cells
1 Mb resolution
56% OR-OR
c
Top 1000 *trans* HiC contacts
Olfr1507+ cells
1 Mb resolution
65% OR-OR
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
d
Greek Island HiC Contacts
cis short
cis long
trans
mOSNs
Olfr16+
Olfr17+
Olfr1507+
12%
6%
0%
Greek Islands
Greek Island HiC Contacts
4%
3%
2%
1%
0%
mOSNs
Olfr16+
Olfr17+
Olfr1507+
e
in situ HiC
Olfr16+ cells
Olfr16 Region
10 Kb resoultion
172.65
174.6Mb
Olfr16
Greek Islands
OR Clusters
1.1Mb
f
in situ HiC
Olfr17+ cells
10 Kb resoultion
172.65
174.6Mb
g
in situ HiC
Olfr1507+ cells
10 Kb resoultion
172.65
174.6Mb
h
in situ HiC
Olfr16+ cells
Olfr17 Region
99.1
109.8Mb
Olfr17
GI
OR
i
in situ HiC
Olfr17+ cells
99.1
109.8Mb
20Mb
j
in situ HiC
Olfr1507+ cells
99.1
109.8Mb
k
in situ HiC
Olfr16+ cells
Olfr1507 Region
Olfr1507
GI
OR
l
in situ HiC
Olfr17+ cells
m
in situ HiC
Olfr1507+ cells
1.7Mb

Extended Data Figure 8: Long range interactions in homogeneous OSN subpopulations. a-c, Circos plots representing the 1000 strongest *trans* contacts in Olfr16- (a), Olfr17- (b) and Olfr1507- (c) expressing OSNs. **d,** (left) Comparison of the frequency of local *cis* (grey), long range *cis* (blue) and *trans* (red) Greek Island interactions in mixed mOSNs and OSNs expressing specific OR genes. (right) Mean values for Olfr16+, Olfr17+, and Olfr1507+ cells are not significantly different from those for mixed mOSNs ($p > 0.05$ for all comparisons, two-tailed paired Wilcoxon signed-rank test). **e,** *in situ* HiC contact matrices from Olfr16^{+}, Olfr17^{+} and Olfr1507^{+} cells focused on the Olfr16 gene locus. Arrowhead points to specific long-range contacts between Olfr16 and the Greek Island Astypalea that occur only in Olfr16^{+} cells. Open pin marks Greek Island-Greek Island contacts that also differ between cell types. **f-g,** Similar analysis for the Olfr16 locus in Olfr17+ and Olfr1507+ cells. **h-j,** as in e-g, except for the Olfr17 locus. **k-m**, as in e-g, except for the Olfr1507 locus.


a
87.5 Mb
178 Mb
Greek Island
~80 Mb
in situ HiC
Olfr16+ cells
Olfr16 /
Greek Islands
50 Kb resolution
b
in situ HiC
Olfr16+ cells
171.3 Mb
175.7 Mb
91.5 Mb
93.5 Mb
20
0
HiC contacts per billion
GI
OR
Olfr16
GI
OR
10 Kb resolution
in situ HiC
Olfr17+ cells
171.3 Mb
175.7 Mb
91.5 Mb
93.5 Mb
5
0
HiC contacts per billion
GI
OR
Olfr16
GI
OR
10 Kb resolution
in situ HiC
Olfr1507+ cells
171.3 Mb
175.7Mb
91.5 Mb
93.5 Mb
5
0
HiC contacts per billion
GI
OR
Olfr16
Greek Islands
OR Clusters
10 Kb resolution
c
Olfr16+ cells
in situ HiC
Olfr17+ cells
in situ HiC
Olfr1507+ cells
in situ HiC
Cumulative
trans contacts
-500 Kb
500 Kb
Olfr16
locus
trans Greek Islands
25 Kb resolution
d
Olfr16+ cells
Olfr17+ cells
Olfr1507+ cells
3D APA
30
-50 Kb
50 Kb
Olfr16 locus
trans
Greek Islands
e
Olfr16+ cells
in situ HiC
Olfr17+ cells
in situ HiC
Olfr1507+ cells
in situ HiC
Cumulative
trans contacts
-500 Kb
500 Kb
Olfr16
locus
Olfr17
locus
Olfr1507
locus
trans Greek Islands
25 Kb resolution
f
poor mappability
H Enhancer
trans Greek Islands
10 Kb resolution
Olfr16
locus
Olfr17
locus
Olfr1507
locus
g
Olfr17+ cells
100 Kb
6%
3%
0%
trans
long
short
HiC contacts with
Greek Islands
Olfr16+ cells
6%
3%
0%
trans
long
short
5 Kb resolution
Olfr17
h
Olfr1507+ cells
50 Kb
H Enhancer
10%
5%
0%
trans
long
short
HiC Contacts with
Greek Islands
Olfr16+ cells
10%
5%
0%
trans
long
short
5 Kb resolution
Olfr1507

Extended Data Figure 9: Long-range *cis* and *trans* contacts between Greek Islands and the active OR gene. a, HiC Contacts that span more than 80 Mb are observed between the Olfr16 locus and Greek Islands in Olfr16^{+} cells. **b,** Close examination of the contacts (dashed box from a) reveals that Greek Islands contact Olfr16^{+} only in Olfr16^{+} cells (top, black arrowhead). Extremely long-range contacts between Greek Islands (gray arrowheads), but not involving the Olfr16 locus, are observed also in Olfr17^{+} and Olfr1507^{+} cells (middle, bottom). **c**, Heatmap depicting interchromosomal contacts between Olfr16 (chromosome 1) and Greek Islands from different chromosomes in *in situ* HiC from Olfr16^{+}, Olfr17^{+} and Olf1507^{+} cells. **d**, 3D projection of APA between the Olfr16 locus and *trans* Greek Islands in the three specific mOSN populations. **e,** Heatmaps for contacts between Olfr16, Olfr17, or Olfr1507 and *trans* Greek Islands reveals an accumulation of contacts centered around the active allele. **f**, APA for an OR vs *trans* Greek Islands shows the accumulation of contacts on the active allele at 10 Kb resolution. The poor mappability of the Olfr17 locus and the lower sequencing depth perturbs the expected focal peak. For the Olfr1507 locus, the presence of the Greek Island, H, 50 Kb from Olfr1507 results in HiC contacts spanning a broad area. **g,h,** Short, long, and *trans* contacts with Greek Islands across the OR gene clusters containing Olfr17 (g) and Olfr1507 (h) plotted as fraction of the total HiC contacts mapped to each position (5 Kb resolution). Top panel shows contact in cells in which Olfr17/Olfr1507 is active, and the bottom panel shows data from Olfr16+ cells in which Olfr17/Olfr1507 is silent.



Extended Data Figure 10. A model for specific OR compartmentalization and the generation of mutually exclusive phases regulating OR gene choice. a,b, Coincidence of Lhx2/Ldb1 peaks with H3K9me3 enrichment may generate an OR-enriched molecular barcode that promotes specific interactions between OR gene clusters. **c,** In INPs, where OR compartments first form, Greek Islands do not make specific contacts with each other. **d,** In mOSNs however, Greek Islands through specifically interact with each other through homotypic Ldb1 interactions, forming a multi-enhancer hub that is segregated from the OR compartment. We hypothesize that OR compartments and Greek Island hubs form incompatible liquid phases driven by Hp1 proteins and the unstructured domains of Lhx2 and Ldb1, respectively. **e,** Upon deletion of Ldb1 (or Lhx2) the Greek Island phase falls apart and the Greek Islands become incorporated to the OR compartments, as in the INPs.

Appendix II:

Antisense lncRNA transcription mediates DNA demethylation to drive stochastic Protocadherin α promoter choice

Daniele Canzio[1,2,*], Chiamaka L. Nwakeze[1,2,*], Adan Horta[1,2,*], Sandy M. Rajkumar [1,2], Eliot L. Coffey[3], Erin E. Duffy[4], Rachel Duffié[1,2], Kevin Monahan[1,2], Sean O'Keeffe[1,ω], Matthew D. Simon[4], Stavros Lomvardas[1,2], Tom Maniatis[1,2,5,#]

1. Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, 10032
2. Mortimer B. Zuckerman Mind Brain and Behavior Institute, Columbia University, New York, NY, 10027
3. Whitehead Institute for Biomedical Research, Cambridge, 02142
4. Department of Molecular Biophysics and Biochemistry, Yale University, 06516
5. New York Genome Center, New York, NY 10013

[*] These authors contributed equally to the work

[#] To whom correspondence should be addressed: tm2472@cumc.columbia.edu

[ω] Current address: Celmatix, New York, NY, 10005

The authors declare no competing financial interests.

SUMMARY

Stochastic activation of clustered Protocadherin (Pcdh) α, β, and γ genes generates a cell-surface identity code in individual neurons that functions in neural circuit assembly. Here we show that Pcdhα gene choice involves the activation of an antisense promoter located in the first exon of each Pcdhα alternate gene. Transcription of an antisense long non-coding RNA (lncRNA) from this antisense promoter extends through the sense promoter and mediates DNA demethylation of the CTCF binding sites proximal to each promoter. Demethylation-dependent CTCF binding to both promoters facilitates Cohesin-mediated DNA looping with a distal enhancer (HS5-1), which locks-in the transcriptional state of the chosen Pcdhα gene. Uncoupling DNA demethylation from antisense transcription by Tet3 overexpression in mouse olfactory neurons promotes CTCF binding to all Pcdhα promoters, resulting in proximity-biased DNA looping of the HS5-1 enhancer. Thus, antisense transcription-mediated promoter demethylation functions as a mechanism for distance-independent promoter/enhancer DNA looping to ensure stochastic Pcdhα promoter choice.

Highlights

- Pcdhα alternate exons display convergent sense and antisense promoters proximal to the CBS sites
- Antisense lncRNA transcription leads to DNA demethylation of the CBS sites to promote CTCF binding
- CTCF/Cohesin assemble a Pcdhα promoter/enhancer complex via loop-extrusion
- Coupling lncRNA transcription to DNA demethylation drives stochastic promoter choice

INTRODUCTION

During brain development, individual neurons differentiate into distinct functional cell types, respond to a plethora of guidance molecules, and project into specific regions of the nervous system to form complex neural circuits. A key aspect of this process is the ability of neurites of individual neurons (axons and dendrites) to distinguish between themselves and among neurites from other neurons (self *vs.* non-self) (Grueber and Sagasti, 2010; Lefebvre et al., 2015; Zipursky and Grueber, 2013). This process, which is known as self-avoidance, requires a unique combination of cell-surface homophilic recognition molecules that function as a molecular identity code (Zipursky and Grueber, 2013; Zipursky and Sanes, 2010). In an extraordinary example of convergent evolution, the same cell-surface mechanism involving specific homophilic interactions followed by repulsion is used for self-avoidance in invertebrates and vertebrates. However, in flies, this identity code is generated by the expression of thousands of Dscam1 isoforms by stochastic alternative RNA splicing. In contrast, in mammals, alternate Protocadherin (Pcdh) isoforms are generated by stochastic transcription from alternative Pcdh promoters (Mountoufaris et al., 2018; Zipursky and Grueber, 2013; Zipursky and Sanes, 2010).

Pcdh genes have a unique genomic arrangement of three closely linked clusters (designated as α, β, and γ), and a poorly understood mechanism of stochastic and combinatorial promoter choice (Esumi et al., 2005; Tasic et al., 2002; Wang et al., 2002; Wu and Maniatis, 1999a; Wu et al., 2001). The three Pcdh gene clusters, together, span nearly 1 million base pairs (bp) of genomic DNA, and are organized into variable and constant regions, reminiscent of the organization of immunoglobin and T-cell receptor gene clusters (Wu and Maniatis, 1999b). The variable regions in the Pcdh α and γ cluster are further distinguished as alternate and c-types. The organization of the human Pcdhα gene cluster, which is conserved throughout vertebrate evolution, is illustrated in Figure 1A. Neuron-specific expression of individual Pcdhα genes requires long-range DNA looping between Pcdhα promoters and a transcriptional enhancer, called HS5-1 (hypersensitivity site 5-1) (Guo et al., 2012; 2015; Kehayova et al., 2011; Monahan et al., 2012; Ribich et al., 2006) (Figure 1A). Conserved transcriptional promoter sequences are located immediately proximal to every Pcdhα exon (Tasic et al., 2002) while the HS5-1 enhancer is located downstream of the constant exons, between the Pcdh α and the β clusters (Ribich et al., 2006) (Figure 1A, 1B and S1). These stochastic promoter/enhancer interactions occur independently on each of the two allelic chromosomes in diploid cells and require the binding of the CCCTC-binding protein (CTCF) and the Cohesin protein complex (Guo et al., 2012; Hirayama et al., 2012; Kehayova et al., 2011; Monahan et al., 2012) (Figure 1C). CTCF is an 11 zinc-finger (ZF) domain protein that, together with the Cohesin complex, plays a central role as an insulator of chromatin domains, and mediates genome-wide promoter/enhancer interactions (Carretero et al., 2010; Ghirlando and Felsenfeld, 2016; Ong and Corces, 2014). All Pcdhα alternate exons contain two CTCF binding sites (CBS), one in the promoter (pCBS) and the other in the protein coding sequence in the first exon (eCBS) (Guo et al., 2012; Monahan et al., 2012) (Figure 1B). The two binding sites are separated by approximately 1000 bp, and similarly spaced CBS sites are located in the HS5-1 enhancer (L-CBS and R-

CBS) (Guo et al., 2012; Monahan et al., 2012) (Figure 1B). Interestingly, the CTCF binding sites in Pcdhα promoters and the HS5-1 enhancer are in opposite relative orientations, and inversion of the HS5-1 enhancer results in a significant decrease in Pcdhα gene cluster expression, demonstrating the functional importance of this arrangement (Guo et al., 2015). This opposite relative orientation of promoter and enhancer CBS sites appears to be a general feature of eukaryotic chromosomes genome-wide (Guo et al., 2015; Rao et al., 2014), and has been proposed to play a critical role in promoting the spatial interaction between genes and transcriptional regulatory elements (Merkenschlager and Nora, 2016) by a mechanism known as loop-extrusion (Alipour and Marko, 2012; Fudenberg et al., 2016). In the context of the Pcdhα gene cluster, the loop-extrusion model predicts that the HS5-1 enhancer, bound by CTCF and the Cohesin proteins, scans the Pcdhα exons until it finds the exon bound by CTCF. However, given that the Pcdhα promoters are tandemly arranged in *cis*, and therefore at varying distances from the HS5-1 enhancer, the mechanism by which DNA looping between the HS5-1 enhancer and the promoter by the CTCF and Cohesin proteins is stochastic remains an enigma.

A critical insight into the formation of Pcdhα promoter/enhancer complexes is provided by the observation that there is an inverse relationship between Pcdhα gene expression and DNA methylation of the Pcdhα promoters (Tasic et al., 2002; Toyoda et al., 2014). Specifically, the CTCF/Cohesin complex associates exclusively with transcriptionally active promoters, which are characterized by hypomethylation of the CBS sites, and of the DNA sequences between the two CBS sites (Guo et al., 2012). By contrast, CBS sites and the DNA between them are hypermethylated in inactive promoters, thus preventing CTCF/Cohesin binding (Guo et al., 2012). Although DNA methylation of the CTCF binding sites is likely to play an important role in the mechanism of stochastic Pcdhα promoter choice, the temporal relationship between promoter DNA methylation and promoter choice is not known. That is, it is not known whether promoter methylation is the ground state upon which promoter choice operates, or whether all promoters are initially unmethylated and methylation of the inactive promoters occurs subsequent to stochastic promoter choice (enhancer/promoter engagement).

Here, we use a combination of cell-culture and *in vivo* studies of olfactory sensory neuron differentiation to provide evidence that the ground state of a Pcdhα promoter DNA is methylated and transcriptionally repressed. Moreover, we show that Pcdhα promoter choice requires stochastic transcriptional activation of an antisense promoter located within the first exon, and transcription through the upstream sense strand promoter, which generates a large multiply-spliced, polyadenylated long non-coding RNA (lncRNA). We provide evidence that transcription of this antisense lncRNA leads to the demethylation, de-repression and activation of Pcdhα proximal sense strand promoters, which occurs coordinately with CTCF binding to its CBS sites located proximal to both promoters. This process is driven by the CTCF/Cohesin-dependent long-range DNA looping between the demethylated promoter and the HS5-1 enhancer. These observations are consistent with a promoter scanning mechanism in which the HS5-1 enhancer, bound by CTCF and Cohesin, translocates to the most enhancer-proximal demethylated and CTCF-bound promoter by DNA loop-extrusion. Thus, in the context of chromosome loop-extrusion, stochastic promoter demethylation by antisense transcription "levels the field" by preventing proximity bias in Pcdhα promoter choice. A similar logic was recently demonstrated for V(D)J DNA recombination, whereby Cohesin-mediated DNA loop-extrusion appears to ensure RAG-mediated recombination of the variable Vh exons most proximal to the iEμ enhancer (Jain et al., 2018).

RESULTS

Transcription of sense and antisense RNA from clustered Pcdhα alternate exons

The formation of a promoter/enhancer-CTCF/Cohesin complex plays a critical role in the mechanism of stochastic promoter choice in the Pcdhα gene cluster (Guo et al., 2015; Kehayova et al., 2011; Monahan et al., 2012; Ribich et al., 2006). However, the mechanism by which random Pcdhα promoters are activated is not understood. This mechanism cannot be studied *in vivo*, as each neuron expresses a distinct repertoire of Pcdhα alternate exons. We therefore made use of the well-characterized human neuroblastoma cell line SK-N-SH, which stably expresses a distinct repertoire of Pcdhα isoforms through multiple cell divisions: α4, α8, α12, αc1, and αc2 (Guo et al., 2012) (Figure 1D). This stochastic pattern of expression in cell culture is indistinguishable from that observed in single neurons *in vivo* (Esumi et al., 2005; Mountoufaris et al., 2017). SK-N-SH cells thus provide a multicellular "avatar" for studying single cell expression of Pcdhα genes, and provide internal controls for exons that are transcriptionally silent.

Another challenge to the study of Pcdhα promoter choice is the low level of expression of Pcdh genes. Therefore, to optimize the analysis of Pcdh RNA precursors (pre-mRNA) and mature (mRNA) RNAs in SK-N-SH cells, we employed capture RNA-Sequencing (cRNA-Seq), which affords a two order of magnitude enrichment of Pcdh RNA transcripts (Figure S1). Remarkably, this enrichment revealed a high level of antisense RNA transcription associated with Pcdhα alternate exons containing dual CBSs in SK-N-SH cells (Figure 1D and S1B). By contrast, antisense RNA transcription was not detected within the two c-type exons, αc1 and αc2, which do not contain CBSs within their exons (Figure 1D). Similarly, antisense RNA was not observed in the Pcdh β or γ variable exons in SK-N-SH cells, which also do not contain exonic CBS sites (Figure S1B). We refer to this antisense RNA as as-lncRNA, as this high molecular weight RNA lacks protein-coding sequences, based on analyses of its open reading frames. For clarity, we refer to the sense Pcdh coding RNA as s-cRNA (sense coding RNA).

Convergent promoters in both the Pcdhα alternative exons and HS5-1 enhancer

In order to characterize the nature of the antisense RNAs and to gain mechanistic insights into their function, we first localized their transcription start sites and the location of the promoter-paused RNAPII using Start-Seq (Nechaev et al., 2010). RNA isolated from stalled RNAPII at promoters are approximately 15-45 nucleotides long and contain a 5' 7meG-cap (Figure 2A). Isolation and sequencing of these short RNAs revealed the position of paused RNAPII, thus acting as a proxy for the location of RNAPII-engaged promoters, and the transcriptional start site at a nucleotide-base resolution (Figure S2A). As expected, we observed promoter-proximal RNAPII at the pCBS-proximal promoter of the active Pcdh α4, α8, α12 and αc1 exons, and at the promoter of αc2 in SK-N-SH cells (Figure 2B). To our surprise, however, we also observed promoter-proximal RNAPII just upstream of the eCBS for α4, α8, and α12 in the antisense orientation (Figure 2B). Thus, sequences near the two CBSs in active Pcdhα genes act as convergent promoters, where antisense and sense RNA converge and partially overlap (Figure 2C, Pcdhα4 is shown). This is in contrast to the singular pCBS site in Pcdhαc1, which acts as a more canonical divergent promoter, where transcription of the antisense and sense RNA occurs in opposite directions, and does not overlap (Figure 2C). Remarkably, Start-Seq analysis also identified a similar convergent promoter architecture of the two CBSs in the HS5-1 enhancer (Figure 2B and 2C) associated with the two CBS sites in the enhancer. The position of TSS for Pcdh α4, α8 and α12 are shown in Figure 2D.

Mapping the location of the Pcdhα as-lncRNA promoters with respect to the as-lncRNAs revealed that these nuclear RNA precursors can be as long as 20 kb in length, and are spliced and polyadenylated. As an example, the as-lncRNA that initiates at the eCBS-proximal promoter of Pcdhα4 in SK-N-SH cells is transcribed through the pCBS-proximal promoter of Pcdhα4, and extends in the antisense direction all the way to the intronic sequence between the Pcdh α1 and α2 exons (more than 20 kb) (Figure 2E). By contrast, the antisense RNA that initiates at the eCBS-promoter of Pcdhα12 extends to the Pcdhα11 exon (Figure 2E). In addition, upon close observation of the splicing patterns, we discovered the presence of a highly conserved 5' splice site (5'ss), encoded in the antisense direction about 7 bp upstream of the pCBS core motif (Figure 2F). Usage of that 5'ss results in the most abundant polyadenylated as-lncRNA spliced isoform (Figure 2E). Remarkably, this site is absent from the pCBS of Pcdhαc1, as well as from the pCBS sites of the Pcdh β and γ clusters. These observations suggest that RNA splicing of this promoter-embedded 5' splice site may be coupled to the activation of the pCBS promoter (See Discussion).

Antisense lncRNA and sense coding RNA are transcribed from the same active allele

The cRNA-Seq data obtained from SK-N-SH cells revealed a direct correlation between sense and antisense RNA transcription and transcriptionally active Pcdhα alternate exons. Because transcription of the Pcdhα alternate exons occurs independently on the two allelic chromosomes (Esumi et al., 2005), we sought to determine whether the as-lncRNA and the s-cRNA were transcribed from the same Pcdhα locus allele. To accomplish this, we used CRISPR-Cas9 gene editing to generate SK-N-SH cells heterozygous for the Pcdhα gene cluster, SK-N-SH-αhet (Figure 3A). We isolated two clones (SK-N-SH–αhet 1 and 2) expressing primarily α12, αc1 and αc2 from the remaining copy of the Pcdhα gene cluster (Figure 3B and 3C). Both clones showed expression of the as-lncRNA and s-cRNA from Pcdhα12 (Figure 3B and 3C), confirming that sense and antisense transcription originate from the same allele. For one of the two clones isolated, αhet-1, we also performed chromatin immunoprecipitation sequencing studies (ChIP-Seq) for CTCF and Rad21, a subunit of the Cohesin complex, as well as capture *in situ* high-throughput chromosome conformation capture studies (cHi-C) to examine long-range DNA interactions between the active Pcdhα12 and the HS5-1 enhancer. These studies demonstrated that the Pcdhα alternate exons, from which sense and antisense RNAs are transcribed, are bound by CTCF and Cohesin, and engaged in promoter/HS5-1 enhancer DNA looping (Figure 3C and 3D). We note that the αhet-1 and αhet-2 clones share a 16.7 kb deletion that truncates the Pcdhα8 exon and removes the Pcdh α9 and α10 exons (Figure 3C and 3D). It is interesting to note that this deletion was previously reported as a common feature of individuals from multiple populations of European and East Asian descent with no discernable phenotypic consequence (Noonan et al., 2003).

Taken together, these data clearly demonstrate that transcriptionally active Pcdhα alternative exons express both sense and antisense RNAs, and that these RNAs are transcribed in a convergent orientation. In contrast to SK-N-SH cells, a mixed population of primary neurons, each expressing a distinct repertoire of Pcdhα alternative exons, should collectively express as-lncRNAs from all the Pcdh α1 to α13 exons, but not from Pcdh αc1 and αc2, or from the β or γ exons. As predicted, analysis of RNA from human primary neurons revealed lncRNA expression exclusively from the Pcdhα 1-13 exons, and from the HS5-1 enhancer (Figure S2B). Similarly, analysis of mouse mature olfactory sensory neurons (mOSNs) also revealed lncRNA expression originating from all the Pcdhα alternate exons (Figure S2C). Thus, all Pcdhα alternative exons in human cell lines and human and mouse primary neurons, analyzed in this study, express as-lncRNAs. As in SK-N-SH cells, the as-lncRNA expressed in human and mouse primary neurons are spliced and polyadenylated (Figure 2E, S2B and S2C). However, contrary to SK-N-SH cells, the levels of the as-lncRNAs in both human and mouse primary neurons appeared lower. We speculate that this difference could be a consequence of the mitotic (SK-N-SH) and the post-mitotic (primary neurons) state of the two cell types. We also note that an antisense lncRNA from the Pcdhα12 exon, similar to the one described and characterized above, was reported in human brain samples, but its significance was not understood (Lipovich et al., 2006).

The asymmetric nature of Pcdhα convergent promoters results in asynchronous sense and antisense RNA transcription

Antisense convergent transcription is a widespread phenomenon in the mammalian genome. Yet, its function, as well as the mechanism by which actively transcribing RNA polymerases translocate along the same stretch of DNA, remains unclear (see Discussion). To assess the activity of RNAPII at the pCBS-proximal and eCBS-proximal promoters, we analyzed transcription in SK-N-SH cells using s^{4}UDRB-Seq (Fuchs et al., 2014; Singh and Padgett, 2009). This method combines synchronization of RNAPII at promoters with incorporation of the nucleoside 4-thiouridine (s^{4}U) during RNA synthesis. SK-N-SH cells were treated with 5,6-Dichloro-1-β-D-ribofuranosylbenzimidazole (DRB) to block phosphorylation of the carboxy-terminal domain (CTD) of RNAPII, which is required to release paused RNAPII from promoters in the transition from initiation to productive elongation (Figure 3E). DRB inhibition is reversible, and upon removal from the cell culture media, a wave of newly transcriptionally elongating RNAPII leads to the incorporation of s^{4}U into newly synthesized RNAs (Figure 3E). s^{4}U is rapidly incorporated into living cells without the need of cell lysis or nuclear isolation. Given the thiol-specific reactivity of s^{4}U, s^{4}U-labeled nascent RNA can be covalently and reversibly captured and sequenced (Figure 3E). Consistent with the Start-Seq data, we observed convergent elongating RNAPII from both pCBS-proximal and eCBS-proximal promoters of α4, α8 and α12, and divergent RNAPII from the pCBS-proximal promoter of Pcdhαc1 (Figure 3F). We also observed convergent elongating RNAPII at the HS5-1 enhancer, consistent with the presence of convergent promoters as described above (Figure 3F). These data reveal a remarkable symmetry between the location of CTCF/Cohesin binding sites and sense and antisense transcription from the Pcdhα promoters and the HS5-1 enhancer. However, in contrast to the sense and antisense RNA transcribed from Pcdhα alternate exons, both enhancer RNAs are not polyadenylated in SK-N-SH cells nor in primary neurons, and therefore appear to rapidly turnover over (Figure 1D, 1F and S2B).

Interestingly, quantification of nascent transcription of the antisense and sense RNAs assayed by s^{4}UDRB-Seq revealed that, while RNAPII molecules at the Pcdhα active exons transcribe in a convergent manner, their activity seemed asynchronous. That is, the as-lncRNA is transcribed earlier than the s-cRNA (Figure 3G and 3H). This asynchronous RNAPII activity reveals an intrinsic asymmetry in the activities of the two promoters, a possibility consistent with the observation that the two CBS sites, proximal to the sense and antisense promoters, differ in sequence and in their affinity for CTCF. Specifically, the eCBS appears to be a stronger binding site for CTCF than is the pCBS (Figure S2D). Additional evidence for the asymmetric nature of the convergent Pcdhα promoters was provided by the analysis of published ENCODE DNaseI hypersensitivity and ChIP-Seq data, which revealed the binding of distinct classes of transcription factors (TF) by the pCBS and eCBS sites of transcriptionally active alternate exons in SK-N-SH cells. Specifically, TFs belonging to the ETS family bind to the pCBS-proximal promoter, while TFs belonging to the bHLH family bind to the eCBS-proximal promoter (Figure S2D). It is interesting to note that both of these classes of TFs are implicated in regulating genes involved in neuronal development and differentiation, such as members of the cell-adhesion protein family (Hollenhorst et al., 2011).

Transcription of antisense lncRNAs triggers activation of Pcdhα sense promoters

In order to understand the functional significance of the observed Pcdhα sense and antisense promoter asynchrony, we designed a gain-of-function assay to uncouple transcription of the as-lncRNA from transcription of the sense coding Pcdhα mRNA in the context of the endogenous Pcdhα gene cluster. Specifically, we made use of a catalytic-inactive CRISPR-dCas9 protein fused to a tripartite transcriptional activator (dCas9-VPR) (Chavez et al., 2015) to selectively activate the pCBS-proximal or eCBS-proximal promoters of silent Pcdhα genes (Figure 4A). We chose HEK293T cells, as most Pcdhα genes are transcriptionally silent in this cell line, with the exception of Pcdh α10 and αc2. This property of HEK293T cells, together with the modularity of the CRISPR-dCas9 system, made it possible to selectively design guide RNAs for the transcriptional activation of Pcdh α4, α6, α9 and α12 (Figure S3). As expected, dCas9-VPR activation of the Pcdhα4 sense promoter resulted in robust synthesis of the Pcdhα4 s-cRNA (Figure 4B). Unexpectedly, activation of the Pcdhα4 antisense promoter not only led to high levels of antisense RNA transcription, but high levels of sense RNA transcription were also observed (Figure 4B). This pattern of sense and antisense RNA transcription observed did not depend upon how many dCas9-VPR complexes were used (1 *vs.* 4) nor on their exact position relative to the CBSs (Figure S4A). Most importantly, this pattern of transcription mirrored that of active exons observed in SK-N-SH cells (Figure 1D).

These observations suggested the possibility that transcription of the antisense RNA by the eCBS-proximal promoter activates the cognate pCBS-proximal promoter to generate sense coding RNA. To test this possibility, we measured the levels of histone H3 lysine 4 trimethylation (H3K4me3), a histone post-translational modification that marks transcriptionally active promoters. In the Pcdhα locus, H3K4me3 is present between the two CTCF-bound CBS sites (Figure 1D). Using chromatin immunoprecipitation studies followed by quantitative PCR (ChIP-qPCR), we observed an increase in H3K4me3 upon transcriptional activation of the antisense promoter by dCas9-VPR (Figure 4C). We also observed the same relationship between as-lncRNA transcription and sense transcription of the eCBS-proximal promoters of the Pcdh α6, α9, and α12 exons (Figure 4D and S4B), providing additional support for the conclusion that antisense transcription regulates sense transcription of Pcdhα exons.

Taken together, these data are consistent with a model in which transcription of the antisense RNA by the eCBS-proximal promoter activates its cognate pCBS-proximal promoter, thus generating convergent sense and antisense transcripts. This level of exquisite specificity is remarkable, considering that the as-lncRNA transcribes through multiple upstream sense promoters, yet the only sense promoter activated is the one proximal to the site of initiation of the antisense RNA (see Discussion).

Antisense lncRNA transcription promotes CTCF binding and long-range promoter/enhancer DNA interactions

The expression of Pcdhα sense RNA transcripts requires binding of CTCF and Cohesin to the pCBS and eCBS sites, and long-range DNA looping between active promoters and the HS5-1 enhancer (Guo et al., 2012; 2015). In ChIP-Seq experiments, we observed that both CBSs of Pcdh α4, α6, α9, and α12 in the HEK293T parental cell line used in this study are not bound to CTCF nor to the Cohesin subunit, Rad21 (Figure S3B). We therefore asked whether antisense transcription by the dCas9-VPR gain-of-function assay promotes the binding of CTCF to its binding sites in the activated exon. Consistent with the mechanistic coupling of promoter activation and CTCF/Cohesin binding (Guo et al., 2012; Monahan et al., 2012), we observed a statistically significant enrichment of CTCF occupancy at both the pCBS and eCBS sites upon dCas9-VPR activation of their antisense promoters relative to the activation of their sense promoters (Figure 5A). We note that the levels of CTCF binding at the activated Pcdhα promoters measured by ChIP-qPCR was lower than the one measured for a constitutive promoter such as GAPDH, but significantly higher than an intergenic DNA site (Figure S4C). We reasoned that this lower CTCF enrichment is a consequence of the high degree of cell heterogeneity as a result of transient transfections of the dCas9-VPR constructs.

The binding of CTCF raised the possibility that antisense transcription from the activated exon leads to CTCF/Cohesin-dependent long-range DNA looping between the active promoter and the HS5-1 enhancer. To address this hypothesis, we focused on the Pcdhα12 exon and performed three biologically independent *in situ* cHi-C experiments on HEK293T cells transfected with dCas9-VPR to activate either the Pcdhα12 pCBS-proximal or eCBS-proximal promoter. To enrich for HEK293T cells transfected with the dCas9-VPR activator, we introduced a green fluorescent protein (GFP) reporter into the dCas9-VPR expressing plasmids and sorted cells with the highest GFP signal. Analysis of *in situ* cHi-C data from Pcdhα12 eCBS-activated HEK293T cells showed modest, but statistically significant, increase in DNA contacts between the Pcdhα12 promoter and the HS5-1 enhancer compared to Pcdhα12 pCBS-activated HEK293T cells (Figure 5B). Importantly, dCas9 without the transcriptional activator domain did not result in the formation of Pcdhα12/HS51 contacts (Figure S4D). These data, taken together, support the hypothesis that antisense lncRNA transcription leads to CTCF binding, and that the HS5-1 enhancer scans the Pcdhα locus in *cis* until it reaches the Pcdhα exon bound by CTCF, as predicted by the loop-extrusion model.

Antisense lncRNA transcription promotes DNA demethylation of Pcdhα promoters

The data presented thus far support a model in which stochastic choice of Pcdhα alternate promoters requires coupling between transcription of antisense lncRNAs and the assembly of a promoter/enhancer complex by CTCF and Cohesin proteins. However, the mechanism by which transcription of the antisense lncRNA promotes the recruitment of both CTCF and Cohesin, the assembly of a functional promoter/enhancer complex, and stable transcriptional activation of a Pcdhα sense coding RNA remained to be determined. Given the observation that DNA methylation of the CBS sites blocks CTCF binding (Bell and Felsenfeld, 2000) and that both pCBS and eCBS sequences contain CpG dinucleotides, we reasoned that DNA demethylation could be a mechanism to promote CTCF/Cohesin binding to exons following as-lncRNA transcription.

CBS sites can contain four modules (1, 2, 3, 4) and the 11 zinc fingers (ZFs) domains of CTCF make specific contacts with them (Ong and Corces, 2014). The core motif is embedded in modules 2 and 3 and contacted by ZFs 4-7 (Figure S5A and S5B). DNA methylation of C2, in module 2, and C12, in module 3, are known to result in significant loss of CTCF binding (Wang et al., 2012). In addition to the core motif, ZFs 9-11 and ZFs 1-3 can engage with modules 1 and 4, respectively, to enhance CTCF binding to its CBS. To gain insight into the potential role of DNA methylation in the modularity of CTCF binding to both pCBS and eCBs sites, we obtained nucleotide resolution of the methylation of the CpG dinucleotides within the CBS sites by examining published ENCODE whole genome bisulfite sequencing (WGBS) data from SK-N-SH cells (Figure S5C and S5D, Pcdh α4, α12, α3 and α13 are shown as examples of active and inactive exons). While these data reveal how methylation at position C2 and C12 in the core motif can affect CTCF binding at both CBS sites (Figure S5C and S5D), they also revealed additional methylation sites within module 1 and 4 of the eCBS (Figure S5D), consistent with the observation that the two CBS sites are intrinsically distinct. To better appreciate the overall impact of DNA methylation of pCBS and eCBS on CTCF binding, we quantitated the average CpG methylation level for each Pcdhα exon relative to the occupancy of CTCF at these sites using ChIP-Seq data. Consistent with the inhibitory role of CpG methylation, we observed an anti-correlation between CTCF binding and CpG methylation for both the pCBS and eCBS (Figure S5E).

The two Pcdhα CBS sites are separated on average by about 1000 bp of CpG-rich islands and, in active exons, are enriched for H3K4me3 nucleosomes. We refer to this sequence as “middle”. Using the ENCODE WGBS data, we quantitated the levels of CpG methylation in the middle sequences between Pcdh α4, α12, α3 and α13 and observed how hypermethylation of these sites correlates with inactive exons (Figure S5F). This correlation holds true for all active and inactive exons in the Pcdhα gene cluster (Figure S5G), and is consistent with previous reports on the relationship between methylation and promoter activity (Guo et al., 2012; Kawaguchi et al., 2008; Tasic et al., 2002).

In mammals, 5-methylcytosine (5mC) modified CpG sequences are converted to unmodified cytosine (C) by the activity of TET deoxygenase enzymes, which mediate the oxidation of 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (Wu and Zhang, 2017). Thymine DNA glycosylase (TDG) then converts 5caC to C by a base excision repair mechanism (Wu and Zhang, 2017). 5hmC is a stable oxidation intermediate and its detection is a proxy for a pathway to active demethylation catalyzed by the TET proteins. Therefore, to directly test the possibility that transcription of the as-lncRNA leads to demethylation of CpG elements, we measured the levels of 5mC and 5hmC for the Pcdhα12 in HEK293T cells by Methylated DNA Immunoprecipitation (MeDIP) upon dCas9-VPR-mediated activation of its respective sense and antisense promoters. Consistent with our hypothesis, activation of the Pcdhα12 eCBS-proximal promoter resulted in a decrease of 5mC/5mhC levels at the pCBS, the eCBS and the middle interval between the two CBS sites (Figure 5C). By contrast, activation of the Pcdhα12 pCBS-proximal promoter resulted in a statistically significant decrease of 5mC/5hmC levels only for the pCBS site (Figure 5C). To gain base-pair resolution of the changes occurring at the eCBS site, we performed bisulfite reactions followed by Sanger DNA sequencing. Consistent with the MeDIP experiment, we observed a higher degree of demethylation of all three CpG sites in the eCBS when antisense RNA is transcribed relative to when only sense transcription is initiated (Figure S5H).

Taken together, these data support the hypothesis that transcription of the antisense lncRNA promotes CpG DNA demethylation of both CBS sites, as well as the middle interval between them bearing H3K4me3-marked nucleosomes in active exons. This process promotes stable binding of CTCF and transcriptional activation of Pcdhα promoters.

Demethylation of Pcdhα promoters correlates with activation *in vivo*

The data presented above are consistent with a model in which the ground state of Pcdhα promoter DNA is methylated, and DNA demethylation, targeted by transcription of an antisense lncRNA, controls Pcdhα sense promoter activation. To test this model *in vivo*, we made use of the mouse main olfactory sensory epithelium (mOE), as an *in vivo* developmental system to study the relationship between promoter DNA methylation and Pcdhα gene expression. Previous studies have shown that the Pcdh gene cluster is stochastically and combinatorially expressed in OSNs, and that Pcdhα genes play a fundamental role in OSN wiring (Hasegawa et al., 2008; 2016; Mountoufaris et al., 2017) (Figure 6A). We re-analyzed recently published work carried out to characterize the levels of 5mC and 5hmC in the three cell types that represent discrete neurodevelopmental stages in the mOE: horizontal basal cells (ICAM1$^+$), immediate neural precursors (Ngn1$^+$) and mature olfactory sensory neurons (Omp$^+$) (Figure 6A) (Colquitt et al., 2013). Horizontal basal cells are quiescent multipotent cells that produce all of the cell types present in the mOE. Immediate neural precursors are post-mitotic cell precursors to olfactory sensory neurons. Olfactory sensory neurons are terminally differentiated primary sensory neurons. Consistent with our model, we found that clustered Pcdhα alternate exons and their promoters are enriched in 5mC in iCAM1$^+$ cells, indicating that the pre-neuronal ground state of all Pcdhα alternate promoter DNA is methylated and repressed (Figure 6B and 6C). However, with the development of olfactory sensory neurons (ICAM1$^+$ → Ngn1$^+$ → Omp$^+$), we observed an increase of 5hmC in the Pcdhα alternate promoters and exons (Figure 6B and 6D). To determine whether conversion of 5mC to 5hmC is accompanied by activation of Pcdhα promoters, we performed RNA-Seq experiments in ICAM1$^+$, Ngn1$^+$ and Omp$^+$ cells. Consistent with our hypothesis, conversion of 5mC to 5hmC correlates with the expression of both antisense long noncoding and sense coding Pcdhα RNAs (Figure 6E, 6F and S6A). Finally, we determined whether Pcdhα expression is accompanied by the formation of long-range DNA contacts between the Pcdhα promoters and the HS5-1 enhancer *in vivo,* and performed *in situ* Hi-C experiments in ICAM1$^+$, Ngn1$^+$ and Omp$^+$ cells (Figure S6B). Consistent with our model, we observed a strong increase in alternate promoters/HS5-1 enhancer interactions during neuronal differentiation of the mOE (Figure 6G). These data, collectively, provide *in vivo* confirmation of our observations made in human cell lines.

Stochastic DNA demethylation ensures random Pcdhα promoter choice by the CTCF/Cohesin complex via DNA loop-extrusion

Analysis of the Hi-C data from Ngn1$^+$ and Omp$^+$ cells revealed architectural “stripes” along the Pcdhα gene cluster (Figure S6B and 7A), a feature that has been associated with Cohesin activity in the assembly of promoter/enhancer complexes during DNA loop-extrusion (Vian et al., 2018). A prediction of the DNA loop-extrusion model for the assembly of a Pcdhα promoter/enhancer complex is that uncoupling CTCF binding to Pcdhα promoters from DNA looping to the HS5-1 enhancer by the Cohesin complex should result in an overall loss of expression of all Pcdhα exons. To test this possibility, we conditionally deleted the Cohesin subunit, Rad21, in mouse olfactory sensory neurons (Figure S7A) using OMPiresCre. With this driver, Rad21 is deleted in post-mitotic, fully differentiated, OSNs in which Pcdhα promoter choice has already occurred (Figure 6C-G and S7B). However, upon deletion of Rad21, a loss of long-range DNA contacts between the Pcdhα promoters and the HS5-1 enhancer was observed (Figure 7A and 7B). More importantly, loss of DNA contacts correlated with a significant loss of expression of all Pcdhα exons as determined by RNA-Seq (Figure 7C). Thus, continuous Cohesin activity appears to be required for the maintenance of DNA looping in the Pcdhα cluster, even in the absence of cell division.

These data are consistent with a model in which CTCF acts as a boundary element for the Cohesin complex to mediate long-range interactions between Pcdhα promoters and the HS5-1 enhancer. In the context of our methylation data and the mechanistic coupling of demethylation to CTCF binding, this model predicts that formation of long-range DNA contacts between a Pcdhα promoter and the HS5-1 enhancer in individual neurons is stochastic and distance-independent with respect to HS5-1. We propose that this enhancer/promoter engagement is achieved by virtue of random demethylation of Pcdhα promoters. According to this model, random demethylation of one of the Pcdhα exons, as a consequence of as-lncRNA transcription, ensures that only one exon is bound to CTCF, and thus results in the assembly of a specific Pcdhα promoter/HS5-1 enhancer complex. A prediction of this model is that uncoupling DNA demethylation from antisense lncRNA transcription results in a non-random choice of Pcdhα promoters by the HS5-1 enhancer. To uncouple as-lncRNA transcription from DNA demethylation, we overexpressed Tet3 in OSNs (Figure S7A). Tet3 is the most highly expressed Tet protein in OSNs, and has been shown to associate with the Pcdhα promoters in differentiated neuronal precursor cells (Li et al., 2016). Overexpression of Tet3 resulted in strong demethylation of Pcdhα promoters, as indicated by a large increase in 5hmC levels (Figure 7D and S7C) and by an increase of CTCF binding to CBS sites genome-wide (Figure S7D), and to all Pcdhα exons, irrespective of transcription of their cognate as-lncRNAs (Figure 7D and S7E). To address the function of uncoupling as-lncRNA transcription from stochastic DNA demethylation, we performed Hi-C and RNA-Seq in mOSNs overexpressing Tet3. Remarkably, despite the fact that all Pcdhα exons are bound by CTCF, and that the expression of the as-lncRNAs is maintained (Figure 7D and S7E), overexpression of Tet3 resulted in a strong bias in Pcdhα promoter/HS5-1 enhancer contacts biased towards the Pcdhα12 promoter (Figure 7E and 7F) and a concomitant bias in Pcdhα12 expression relative to all other Pcdhα exons, as determined by RNA-Seq (Figure 7G). Thus CTCF bound to the CBS sites of Pcdhα12 created a “roadblock” for Cohesin, preventing the HS5-1 enhancer from engaging with any of the upstream Pcdhα promoters. These data are consistent with a model in which coupling antisense lncRNA transcription to DNA demethylation ensures random choice of Pcdhα promoters *in vivo* (Figure 7H).

DISCUSSION

Stochastic, combinatorial expression of individual Pcdh protein isoforms in Purkinje (Esumi et al., 2005) and olfactory sensory neurons (Mountoufaris et al., 2017) generates distinct combinations of Protocadherin isoforms that function as a cell-surface identity code for individual neurons (Mountoufaris et al., 2018). This conclusion has been confirmed more broadly through single cell RNA sequencing studies in a variety of neuronal cell types (Tasic et al., 2018). Here we identify a mechanism by which Pcdhα alternate exon promoters are stochastically activated in individual neurons, and propose a model that may apply more broadly in promoter choice and gene expression in vertebrates.

Insights into the mechanism of stochastic Pcdhα promoter choice

We provide evidence that stochastic activation of individual Pcdhα alternate promoters requires mechanistic coupling between transcription of an antisense lncRNA and DNA demethylation of the Pcdhα promoters and CTCF binding sites (Figure 7H). Specifically, each Pcdhα alternate exon bears two convergent promoters located proximal to the CBS sites, pCBS and eCBS. The former is located 5' to the Pcdhα protein coding sequence, and the latter, within the adjacent coding sequence. We have shown that the eCBS-proximal promoter initiates transcription of a long noncoding RNA that extends through the pCBS-proximal promoter, and into upstream intergenic sequences, leading to transcriptional activation of the pCBS-proximal promoter. This process is accompanied by DNA demethylation of the CBS sites and the sequences between them, and to the binding of CTCF to its two CBS sites. CTCF, together with the Cohesin complex, mediates long-range DNA looping between the active promoter and the Pcdhα cluster-specific transcriptional enhancer, HS5-1, via DNA loop-extrusion. We propose that the translocating Cohesin complex stalls at the transcriptionally active promoters bound by CTCF. Formation of this promoter/enhancer complex commits Pcdhα sense strand promoter activation, and thus leads to the stochastic production of a specific Pcdhα mRNA (Figure 7H). We noted above that the as-lncRNA initiated at a Pcdhα eCBS-proximal promoter transcribes through its cognate pCBS-proximal promoter and extends in the antisense direction through upstream sense promoters. However, the only sense promoter that is activated in this process is the proximal promoter. We speculate that this proximal specificity is a consequence of functional coupling between transcription and RNA processing mediated by the carboxy-terminal (CTD) of the RNAPII, the cap-binding complex and the Spliceosome (Maniatis and Reed, 2002). In support of this hypothesis, we identified a highly conserved 5’ss just upstream of each pCBS site in the Pcdhα alternate exons (Figure 2F). This splice site is active and contributes to the processing of the as-lncRNA (Figure 2E). Thus, the Spliceosome may be recruited to the vicinity of the sense promoter by transcriptional read-through. While functional coupling between Tet-mediated DNA demethylation, CTCF and the Spliceosome has been reported elsewhere (Marina and Oberdoerffer, 2016), additional studies will be required to test this hypothesis in the context of Pcdhα promoters.

A fundamental question raised by our model is how antisense promoters are stochastically activated in individual neurons during development. Given the observation that the ground state of the Pcdhα gene cluster is inactive and marked by 5mC in horizontal basal cells, we speculate that activation of eCBS-proximal promoters in the Pcdhα gene cluster is regulated by the presence of transcription factors capable of binding methylated DNA, consistent with our observation that distinct sets of transcription factors have been shown to bind to the pCBS-proximal and eCBS-proximal promoters (Figure S2D). In contrast to the Pcdhα gene cluster, the alternate exons in the Pcdh β and γ clusters bear a single CBS site in their promoters (pCBS), but lack a CTCF/Cohesin binding site, as well as an antisense promoter in the downstream exon. Thus, antisense lncRNAs are not detected in either the Pcdh β or γ gene clusters. Nevertheless, Pcdh β and γ promoter choice is stochastic (Esumi et al., 2005; Mountoufaris et al., 2017) and transcriptional enhancer elements, similar to the HS5-1 enhancer, located distal to the Pcdhγ gene cluster are required for their transcription (Yokota et al., 2011). Thus, the mechanism of random promoter choice in these gene clusters remains unknown, and is likely to be cell type-specific. Indeed, in contrast to the Pcdhα gene cluster, which is expressed exclusively in the nervous system, the Pcdh β and γ gene clusters are expressed more broadly in other cell types (https://www.gtexportal.org/home/).

The molecular logic of convergent promoters

Bi-directional RNA transcription is a common feature of mammalian promoters and enhancers (Core et al., 2014; Wu and Sharp, 2013). The transcripts can be divergent, thus non-overlapping, or convergent, as is the case of the Pcdhα gene cluster, which produces overlapping complementary RNAs. As we have seen in the case of the Pcdhαc1 exon, divergent transcription at promoters usually produces upstream non-coding RNAs, transcribed toward the 5' end of the gene, that are on average 50 to 2000 nucleotides long and relatively unstable (Wu and Sharp, 2013). In contrast, convergent transcription, as the one here described for the Pcdhα alternate exons, can produce long and stable antisense noncoding RNAs that overlap with the sense coding RNA (Brown et al., 2018; Mayer et al., 2015). In general, these antisense RNAs can function to either activate or repress transcription of the coding RNA from the sense promoter, in a process known as transcription interference (Bonasio and Shiekhattar, 2014). Interestingly, genes that are activated by antisense convergent RNA are characterized by an overall low level of expression of sense and antisense RNAs and a unique chromatin signature that facilitate their transcription (Brown et al., 2018; Mayer et al., 2015; Murray and Mellor, 2016; Scruggs et al., 2015). It has been proposed that it is the act of antisense RNA transcription that actively shapes unique chromatin environments as a crucial step in promoting transcription of the cognate sense RNA. We speculate that, at least in the case described here, low levels of RNA expression, together with differences in the chromatin environment in the two convergent promoters, permits the two convergent RNAPII to productively translocate along DNA without significant interference. However, an alternate possibility is that the antisense promoter shuts down upon the activation of the sense promoter. A test of this possibility would require single cell transcriptional analysis of extremely low levels of antisense RNA.

The example of convergent transcription described here also suggests a model in which noncoding antisense RNA transcription couples RNAPII activity to a DNA deoxygenase TET enzyme activity and the insulator CTCF/Cohesin complex. We note that there are precedents for a transcription-dependent mechanism of transcriptional activation coupled to DNA demethylation. Specifically, transcription of the tumor suppressor gene, TCF21, was shown to be activated by an antisense RNA whose transcription is initiated at an intronic promoter sequence located within the TCF21 gene (Arab et al., 2014). Like the mechanism proposed here, transcription through the TCF21 promoter leads to TET-mediated DNA demethylation and activation of the TCF21 sense strand promoter. Here, we propose that this mechanism is used for stochastic choice of Pcdhα promoters, which has profound implications in neuronal circuit assembly during development (Mountoufaris et al., 2018) .

A general mechanism for stochastic promoter activation to generate transcriptional diversity

We used the differentiating mouse olfactory epithelium as an *in vivo* model system for stochastic Pcdhα gene activation. Thus, we could not ignore the striking similarities in the regulatory logic between Pcdhα and olfactory receptor (OR) promoter choice. In both cases, the ground state of the stochastically chosen promoters is repressed and inaccessible to transcriptional activator proteins. In the case of the Pcdhα gene cluster, this repression is mediated predominantly by DNA methylation (Tasic et al., 2002; Toyoda et al., 2014), while OR genes are repressed by the assembly of constitutive heterochromatin (Magklara et al., 2011; Monahan et al., 2017). In both of these cases, however, repressive DNA or histone modifications are replaced by activating marks, concomitantly with selective binding of transcription factors that promote DNA looping between promoters and distant transcriptional start sites. As all the Pcdhα genes are clustered in a single chromosome, stochastic Pcdhα choice is accomplished in *cis* via DNA looping to the enhancer. This mechanism of promoter choice differs from OR promoter choice, which has been shown to require the formation of a multi-chromosomal, multi-enhancer hub that activates only one out of 2800 OR alleles distributed throughout the genome (Horta et al., 2018; Markenscoff-Papadimitriou et al., 2014; Monahan et al., 2018). Most likely, reliance on *cis* versus *trans* interactions also explains why Pcdhα and OR genes require distinct mechanisms to achieve transcriptional stochasticity. In the case of Pcdhα genes, CTCF and Cohesin are critical for stochastic enhancer/promoter interactions. In this case, the loop-extrusion mechanism allows the HS5-1 enhancer to scan the gene cluster locally for the most proximal promoter bound by CTCF. In contrast, OR enhancers cannot deploy loop extrusion mechanisms to activate OR transcription because this process cannot accommodate *trans* chromosomal interactions, which may explain the absence of CTCF and Cohesin binding sites in OR enhancers and promoters (Monahan et al., 2018). Consequently, as Pcdhα choice relies on stable CTCF promoter binding, DNA demethylation provides an effective mechanism for stochastic promoter activation. An important consequence of this mechanism is that, since antisense transcription and DNA demethylation are coupled and appear to occur in a stochastic fashion, DNA loop-extrusion will not create a bias toward the selection of the Pcdhα promoter most proximal to the enhancer (Pcdhα13 and Pcdhα12 in human and mouse, respectively). Rather, DNA loop-extrusion identifies the promoter bound to CTCF, providing an elegant mechanism to overcome selection biases driven by genomic proximity. In fact, we have shown that such a bias occurs if as-lncRNA transcription and DNA demethylation are uncoupled. Finally, our experiments highlight another important property of the loop-extrusion-mediated promoter/enhancer complex mechanism: the dynamic nature of enhancer promoter interactions that requires continuous Cohesin expression even in post-mitotic cells. This observation is reminiscent of the cell-division-independent role of Cohesin in the expression of the T-cell receptor α locus (Seitan et al., 2011). Continual maintenance of promoter enhancer interactions is further highlighted by the striking observation that demethylation of all the Pcdhα promoters, after one is chosen, results in bias towards the HS5-1-proximal alternate promoters. These observations suggest that if Pcdhα promoter choice must be stable for the life of OSNs, then a feedback mechanism must be in place to prevent demethylation of the non-chosen promoters.

It remains to be seen if the proposed mechanism of stochastic Pcdhα choice is applicable to other clustered gene families where stochastic gene expression occurs. As noted in the introduction, an interesting example of promoter stochasticity is the process of V(D)J recombination, whereby Cohesin-mediated loop-extrusion appears to bias RAG-mediated recombination of the variable Vh exons that are most proximal to the iEμ enhancer (Jain et al., 2018). However, even in this system, there is a set of Vh exons that recombine in a distance-independent fashion, which could be accomplished by similar molecular mechanisms as the ones described here, ensuring optimal diversity in the generation of immunoglubulins.

AUTHORS CONTRIBUTIONS

D.C. and T.M. identified, developed and addressed the core questions regarding promoter choice. D.C. performed the bulk of the experiments with help from S.M.R. and E.L.C. C.L.N helped D.C. in the establishment of the dCas9-VPR activation assays. A.H. and S.L. helped to develop chromosome conformation studies, and A.H. performed the experiments and analyzed the data. E.E.D. and M.D.S. helped develop the RNAPII elongation studies and E.E.D. and D.C. performed the experiments. R.D. performed RNA sequencing experiments in the developing mouse olfactory epithelium. K.M. generated the Rad21 conditional knock out mouse line and performed HiC in the Rad21KO mOSNs. S.O. trained D.C. in the bioinformatics analysis of the data. D.C. and T.M. wrote the paper with the help of all the authors.

ACKNOWLEDGEMENTS

We thank Dr. Axel, Dr. Zuker, Dr. Max Gottesman, Dr. David Hirsh, Dr. Germano Cecere, Dr. Mountoufaris, Dr. Enrico Cannavo and members of the Maniatis, Lomvardas and Simon labs for critical discussions and suggestions on the manuscript. D.C. would like to thank Dr. Karen Adelman for advice with the Start-Seq experiments. D.C. and T.M. would like to thank Dr. Ye Zhang and Dr. Ben Barres for the generous gift of human brain neurons and Dr. Victor Lobanenkov and Dr. Elena Pugacheva for the generous gift of the CTCF monoclonal antibody. We also thank Ira Schieren for assistance with flow cytometry. This work was supported by a Helen Hay Whitney Postdoctoral Fellowship (D.C. and R.D.), an NIH F32 Postdoctoral Fellowship GM108474 (K.M.), an NIH F31 Predoctoral Fellowship DC016785 (A.H.), an NIH Path to Independence Award K99/R00 K99GM121815 (D.C.), NIH New Innovator Award DP2 HD083992-01 (M.D.S.), NIH grants 1R01MH108579 and 5R01NS088476 (T.M.), and NIH grant R01DC013560 and HHMI Faculty Scholar (SL).

REFERENCES

Alipour, E., and Marko, J.F. (2012). Self-organization of domain structures by DNA-loop-extruding enzymes. Nucleic Acids Research *40*, 11202–11212.

Arab, K., Park, Y.J., Lindroth, A.M., Schäfer, A., Oakes, C., Weichenhan, D., Lukanova, A., Lundin, E., Risch, A., Meister, M., et al. (2014). Long noncoding RNA TARID directs demethylation and activation of the tumor suppressor TCF21 via GADD45A. Mol. Cell *55*, 604–614.

Bell, A.C., and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature *405*, 482–485.

Bonasio, R., and Shiekhattar, R. (2014). Regulation of Transcription by Long Noncoding RNAs. Annu. Rev. Genet. *48*, 433–455.

Brown, T., Howe, F.S., Murray, S.C., Wouters, M., Lorenz, P., Seward, E., Rata, S., Angel, A., and Mellor, J. (2018). Antisense transcription-dependent chromatin signature modulates sense transcript dynamics. Mol. Syst. Biol. *14*, e8007.

Carretero, M., Remeseiro, S., and Losada, A. (2010). Cohesin ties up the genome. Current Opinion in Cell Biology *22*, 781–787.

Chavez, A., Scheiman, J., Vora, S., Pruitt, B.W., Tuttle, M., P R Iyer, E., Lin, S., Kiani, S., Guzman, C.D., Wiegand, D.J., et al. (2015). Highly efficient Cas9-mediated transcriptional programming. Nat Meth *12*, 326–328.

Colquitt, B.M., Allen, W.E., Barnea, G., and Lomvardas, S. (2013). Alteration of genic 5-hydroxymethylcytosine patterning in olfactory neurons correlates with changes in gene expression and cell identity. Proc. Natl. Acad. Sci. U.S.a. *110*, 14682–14687.

Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat Genet *46*, 1311–1320.

Duffy, E.E., and Simon, M.D. (2009). Enriching s4U-RNA Using Methane Thiosulfonate (MTS) Chemistry (Hoboken, NJ, USA: John Wiley & Sons, Inc.).

Duffy, E.E., Rutenberg-Schoenberg, M., Stark, C.D., Kitchen, R.R., Gerstein, M.B., and Simon, M.D. (2015). Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry. Mol. Cell *59*, 858–866.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Systems *3*, 95–98.

Esumi, S., Kakazu, N., Taguchi, Y., Hirayama, T., Sasaki, A., Hirabayashi, T., Koide, T., Kitsukawa, T., Hamada, S., and Yagi, T. (2005). Monoallelic yet combinatorial expression of variable exons of the protocadherin-alpha gene cluster in single neurons. Nat Genet *37*, 171–176.

Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M., and Joung, J.K. (2014). Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. Nature Biotechnology *32*, 279–284.

Fuchs, G., Voichek, Y., Benjamin, S., Gilad, S., Amit, I., and Oren, M. (2014). 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. Genome Biol *15*, R69.

Fuchs, G., Voichek, Y., Rabani, M., Benjamin, S., Gilad, S., Amit, I., and Oren, M. (2015). Simultaneous measurement of genome-wide transcription elongation speeds and rates of RNA polymerase II transition into active elongation with 4sUDRB-seq. Nat Protoc *10*, 605–618.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L.A. (2016). Formation of Chromosomal Domains by Loop Extrusion. Cell Reports *15*, 2038–2049.

Ghirlando, R., and Felsenfeld, G. (2016). CTCF: making the right connections. Genes & Development *30*, 881–891.

González, F., Zhu, Z., Shi, Z.-D., Lelli, K., Verma, N., Li, Q.V., and Huangfu, D. (2014). An iCRISPR Platform for Rapid, Multiplexable, and Inducible Genome Editing in Human Pluripotent Stem Cells. Stem Cell 1–31.

Grueber, W.B., and Sagasti, A. (2010). Self-avoidance and tiling: Mechanisms of dendrite and axon spacing. Cold Spring Harbor Perspectives in Biology *2*, a001750.

Guo, Y., Maniatis, T., Monahan, K., Myers, R.M., Monahan, K., Wu, H., Gertz, J., Varley, K.E., Li, W., Myers, R.M., et al. (2012). CTCF/cohesin-mediated DNA looping is required for protocadherin   promoter choice. Proceedings of the National Academy of Sciences *109*, 21081–21086.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900–910.

Hasegawa, S., Hamada, S., Kumode, Y., Esumi, S., Katori, S., Fukuda, E., Uchiyama, Y., Hirabayashi, T., Mombaerts, P., and Yagi, T. (2008). The protocadherin-alpha family is involved in axonal coalescence of olfactory sensory neurons into glomeruli of the olfactory bulb in mouse. Mol. Cell. Neurosci. *38*, 66–79.

Hasegawa, S., Kumagai, M., Hagihara, M., Nishimaru, H., Hirano, K., Kaneko, R., Okayama, A., Hirayama, T., Sanbo, M., Hirabayashi, M., et al. (2016). Distinct and Cooperative Functions for the Protocadherin-α, -β and -γ Clusters in Neuronal Survival and Axon Targeting. Front. Mol. Neurosci. *9*, 529.

Hirayama, T., Tarusawa, E., Yoshimura, Y., Galjart, N., and Yagi, T. (2012). CTCF is required for neural development and stochastic expression of clustered Pcdh genes in neurons. Cell Reports *2*, 345–357.

Hollenhorst, P.C., McIntosh, L.P., and Graves, B.J. (2011). Genomic and biochemical insights into the specificity of ETS transcription factors. Annu. Rev. Biochem. *80*, 437–471.

Horta, A., Monahan, K., Bashkirova, L., and Lomvardas, S. (2018). Cell type-specific interchromosomal interactions as a mechanism for transcriptional diversity. bioRxiv 287532.

Jain, S., Ba, Z., Zhang, Y., Dai, H.-Q., and Alt, F.W. (2018). CTCF-Binding Elements Mediate Accessibility of RAG Substrates During Chromatin Scanning. Cell.

Kawaguchi, M., Toyama, T., Kaneko, R., Hirayama, T., Kawamura, Y., and Yagi, T. (2008). Relationship between DNA methylation states and transcription of individual isoforms encoded by the protocadherin-alpha gene cluster. Journal of Biological Chemistry *283*, 12064–12075.

Kehayova, P., Monahan, K., Chen, W., and Maniatis, T. (2011). Regulatory elements required for the activation and repression of the protocadherin-alpha gene cluster. Proceedings of the National Academy of Sciences *108*, 17195–17200.

Knight, P.A., and Ruiz, D. (2013). A fast algorithm for matrix balancing. IMA J Numer Anal *33*, 1029–1047.

Lefebvre, J.L., Sanes, J.R., and Kay, J.N. (2015). Development of Dendritic Form and Function. Annu. Rev. Cell Dev. Biol. *31*, 741–777.

Li, X., Yue, X., Pastor, W.A., Lin, L., Georges, R., Chavez, L., Evans, S.M., and Rao, A. (2016). Tet proteins influence the balance between neuroectodermal and mesodermal fate choice by inhibiting Wnt signaling. Proc. Natl. Acad. Sci. U.S.a. *113*, E8267–E8276.

Lipovich, L., Vanisri, R.R., Kong, S.L., Lin, C.-Y., and Liu, E.T. (2006). Primate-specific endogenous cis-antisense transcription in the human 5q31 protocadherin gene cluster. J. Mol. Evol. *62*, 73–88.

Magklara, A., Yen, A., Colquitt, B.M., Clowney, E.J., Magklara, A., Markenscoff-Papadimitriou, E., Evans, Z.A., Kheradpour, P., Mountoufaris, G., Carey, C., et al. (2011). An epigenetic signature for monoallelic olfactory receptor expression. Cell *145*, 555–570.

Maniatis, T., and Reed, R. (2002). An extensive network of coupling among gene expression machines. Nature *416*, 499–506.

Marina, R.J., and Oberdoerffer, S. (2016). Epigenomics meets splicing through the TETs and CTCF. Cc *15*, 1397–1399.

Markenscoff-Papadimitriou, E., Allen, W.E., Colquitt, B.M., Goh, T., Murphy, K.K., Monahan, K., Mosley, C.P., Ahituv, N., and Lomvardas, S. (2014). Enhancer interaction networks as a means for singular olfactory receptor expression. Cell *159*, 543–557.

Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J.A., and Churchman, L.S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell *161*, 541–554.

Merkenschlager, M., and Nora, E.P. (2016). CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. Annu Rev Genomics Hum Genet *17*, 17–43.

Monahan, K., Horta, A., Mumbay-Wafula, A., Li, L., Zhao, Y., Love, P., and Lomvardas, S. (2018). Ldb1 mediates trans enhancement in mammals. bioRxiv 287524.

Monahan, K., Rudnick, N.D., Kehayova, P.D., Pauli, F., Newberry, K.M., Myers, R.M., and Maniatis, T. (2012). Role of CCCTC binding factor (CTCF) and cohesin in the generation of single-cell diversity of protocadherin-α gene expression. Proc. Natl. Acad. Sci. U.S.a. *109*, 9125–9130.

Monahan, K., Schieren, I., Cheung, J., Mumbey-Wafula, A., Monuki, E.S., and Lomvardas, S. (2017). Cooperative interactions enable singular olfactory receptor expression in mouse olfactory neurons. Elife *6*, 1083.

Mountoufaris, G., Canzio, D., Nwakeze, C.L., Chen, W.V., and Maniatis, T. (2018). Writing, Reading, and Translating the Clustered Protocadherin Cell Surface Recognition Code for Neural Circuit Assembly. Annu. Rev. Cell Dev. Biol. *34*, 471–493.

Mountoufaris, G., Chen, W.V., Hirabayashi, Y., O'Keeffe, S., Chevee, M., Nwakeze, C.L., Polleux, F., and Maniatis, T. (2017). Multicluster Pcdh diversity is required for mouse olfactory neural circuit assembly. Science *356*, 411–414.

Murray, S.C., and Mellor, J. (2016). Using both strands: The fundamental nature of antisense transcription. BioArchitecture *6*, 12–21.

Nechaev, S., Fargo, D.C., Santos, dos, G., Liu, L., Gao, Y., and Adelman, K. (2010). Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science *327*, 335–338.

Noonan, J.P., Li, J., Nguyen, L., Caoile, C., Dickson, M., Grimwood, J., Schmutz, J., Feldman, M.W., and Myers, R.M. (2003). Extensive linkage disequilibrium, a common 16.7-kilobase deletion, and evidence of balancing selection in the human protocadherin alpha cluster. Am. J. Hum. Genet. *72*, 621–635.

Ong, C.-T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nat Rev Genet *15*, 234–246.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell *159*, 1665–1680.

Ribich, S., Tasic, B., and Maniatis, T. (2006). Identification of long-range regulatory elements in the protocadherin-alpha gene cluster. Proceedings of the National Academy of Sciences *103*, 19719–19724.

Scruggs, B.S., Gilchrist, D.A., Nechaev, S., Muse, G.W., Burkholder, A., Fargo, D.C., and Adelman, K. (2015). Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. Mol. Cell *58*, 1101–1112.

Seitan, V.C., Hao, B., Tachibana-Konwalski, K., Lavagnolli, T., Mira-Bontenbal, H., Brown, K.E., Teng, G., Carroll, T., Terry, A., Horan, K., et al. (2011). A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation. Nature *476*, 467–471.

Shykind, B.M., Rohani, S.C., O'Donnell, S., Nemes, A., Mendelsohn, M., Sun, Y., Axel, R., and Barnea, G. (2004). Gene switching and the stability of odorant receptor gene choice. Cell *117*, 801–815.

Singh, J., and Padgett, R.A. (2009). Rates of in situ transcription and splicing in large human genes. Nat

Struct Mol Biol *16*, 1128–1133.

Tasic, B., Nabholz, C.E., Baldwin, K.K., Kim, Y., Rueckert, E.H., Ribich, S.A., Cramer, P., Wu, Q., Axel, R., and Maniatis, T. (2002). Promoter choice determines splice site selection in protocadherin alpha and gamma pre-mRNA splicing. Mol. Cell *10*, 21–33.

Tasic, B., Yao, Z., Smith, K.A., Graybuck, L., Nguyen, T.N., Bertagnolli, D., Goldy, J., Garren, E., Economo, M.N., Viswanathan, S., et al. (2018). Shared and distinct transcriptomic cell types across neocortical areas. Nature *563*, 229542.

Toyoda, S., Okano, M., Tarusawa, E., Kawaguchi, M., Hirabayashi, M., Kobayashi, T., Toyama, T., Oda, M., Nakauchi, H., Yoshimura, Y., et al. (2014). Developmental epigenetic modification regulates stochastic expression of clustered protocadherin genes, generating single neuron diversity. Neuron *82*, 94–108.

Vian, L., Pękowska, A., Rao, S.S.P., Kieffer-Kwon, K.-R., Jung, S., Baranello, L., Huang, S.-C., Khattabi, El, L., Dose, M., Pruett, N., et al. (2018). The Energetics and Physiological Impact of Cohesin Extrusion. Cell *173*, 1165–1178.e20.

Wang, X., Su, H., and Bradley, A. (2002). Molecular mechanisms governing Pcdh-gamma gene expression: evidence for a multiple promoter and cis-alternative splicing model. Genes & Development *16*, 1890–1905.

Wu, Q., and Maniatis, T. (1999a). A striking organization of a large family of human neural cadherin-like cell adhesion genes. Cell *97*, 779–790.

Wu, Q., and Maniatis, T. (1999b). A striking organization of a large family of human neural cadherin-like cell adhesion genes. Cell *97*, 779–790.

Wu, Q., Maniatis, T., Noonan, J.P., Zhang, T., Cheng, J.F., Kim, Y., Grimwood, J., Schmutz, J., Dickson, M., Zhang, M.Q., et al. (2001). Comparative DNA Sequence Analysis of Mouse and Human Protocadherin Gene Clusters. Genome Research *11*, 389–404.

Wu, X., and Zhang, Y. (2017). TET-mediated active DNA demethylation: mechanism, function and beyond. Nat Rev Genet *18*, 517–534.

Wu, X., and Sharp, P.A. (2013). Divergent transcription: a driving force for new gene origination? Cell *155*, 990–996.

Yokota, S., Hirayama, T., Hirano, K., Kaneko, R., Toyoda, S., Kawamura, Y., Hirabayashi, M., Hirabayashi, T., and Yagi, T. (2011). Identification of the cluster control region for the protocadherin-beta genes located beyond the protocadherin-gamma cluster. J. Biol. Chem. *286*, 31885–31895.

Zhu, Z., González, F., and Huangfu, D. (2014). The iCRISPR platform for rapid genome editing in human pluripotent stem cells. Meth. Enzymol. *546*, 215–250.

Zipursky, S.L., and Grueber, W.B. (2013). The molecular basis of self-avoidance. Annu. Rev. Neurosci. *36*, 547–568.

Zipursky, S.L., and Sanes, J.R. (2010). Chemoaffinity revisited: dscams, protocadherins, and neural circuit assembly. Cell *143*, 343–353.

MAIN FIGURE LEGENDS

Figure 1: Transcription of sense and antisense RNA from Pcdhα alternative exons

(A) Genomic organization of the human Pcdhα gene cluster. 13 alternate exons, each with its own promoter, are equally likely to be stochastically activated by the HS5-1 enhancer located downstream of the cluster. c1 and c2 indicate c-type exons and 1-3 are the constant exons encoding the intracellular domain of Pcdh proteins. The promoters of the individual alternate exons is shown with an arrow. The arrow from the HS5-1 to the alternate promoters indicates the stochastic choice by the HS5-1 enhancer. (B) Location and orientation (indicated by the arrows) of the promoter and the exonic CBS sites on alternate exons and the left and right CBS sites in the HS5-1 enhancer. (C) An example of a Pcdhα promoter/HS5-1 enhancer complex mediated by the CTCF and Cohesin proteins. (D) Sense (grey) and antisense (black) RNA (Total RNA, cRNA-Seq) from the Pcdhα cluster in SK-N-SH cells. CTCF, Rad21 (a Cohesin subunit) and H3K4me3 (ChIP-Seq) reveal transcriptionally active exons in SK-N-SH cells. Pcdhαc2 is active but not bound by CTCF or Rad21. Active exons are highlighted in yellow. Virtual 4C (cHiC) is shown on top. HS5-1 is used as a viewpoint. Exons α4, α8, α12 and αc1 are contacted by the HS5-1 enhancer. The x-axis represents the linear sequence of the genomic organization of the Pcdhα cluster. The numbers on the left-hand side of each track represent the minimum and maximum densities in reads per million.

Figure 2: Convergent promoters in the Pcdhα alternative exons and HS5-1 enhancer

(A) Schematic diagram of Start-Seq. (B) Paused RNAPII (Start-Seq) relative to total RNA (cRNA-Seq), and CTCF, H3K4me3 and H3K27ac (ChIP-Seq) in SK-N-SH cells. (C) Promoter architectures for Pcdhα4 (convergent), Pcdhαc1 (divergent) and the HS5-1 enhancer (convergent). (D) Location of the TSS of the as-lncRNA and s-cRNA from Pcdh α4, α8 and α12 in SK-N-SH cells. (E) RNA splicing patterns of the polyadenylated as-lncRNA initiated from the active Pcdhα4 and Pcdhα12 as indicated by the splice junctions in reads mapping to the as-lncRNAs, relative to CTCF and H3K4me3. The PolyA RNA is asseyed by cRNA-Seq. Red triangles denote the antisense 5' splice site described in (F). (F) Position, sequence and conservation of the antisense 5' splice site located upstream of the pCBS (blue). CTCF is in violet. The bar graph indicates the distribution of the distance of the 5'ss from the pCBS.

For B and E, the numbers on the left-hand side of each track represent the minimum and maximum densities in read per million. The x-axis represents the linear sequence of the genomic organization of the Pcdhα cluster and the arrows in (E) indicate the position of transcription start sites as determined by Start-Seq.

Figure 3: Antisense lncRNA transcription precedes sense cRNA transcription from the same active allele

(A) Generation of a SK-N-SH cell line bearing a single copy of the Pcdhα gene cluster by CRISPR-Cas9. Scissors indicate the location of the gRNAs and the PCR confirms the deletion. (B) Expression of Pcdh α4 and α12 relative to RPLPO in SK-N-SH-αh□□ 1 and 2 clonal cells compared to SK-N-SH-WT cells (RT-qPCR). (C) Total RNA (RNA-Seq) relative to the location of Rad21, CTCF and H3K4me3 (ChIP-Seq) in SK-N-SH-αh□□-1. (D) *In situ* cHi-C contact maps at 10kb resolution for SK-N-SH-αhet-1 (Left) and SK-N-SH-WT (Right) cells. Coordinates: 140,780,000-141,050,000, chr5 (hg38). (E) Schematic diagram of s^4U-DRB-cRNA-Seq. (F) Nascent transcription at 20 minutes after the release of RNAPII (s^4U-DRB-cRNA-Seq) (G) Quantification of nascent transcription by RNAPII of the as-lncRNA and s-cRNA from Pcdh α4 (Left) and α12 (Right). The -s^4U is used as a control for specific enrichment of nascent RNA labeled with s^4U. Errors (n=3) represent s.e.m. (H) Schematic diagram describing the asynchronous activity of RNAPII.

For C and F, the numbers on the left-hand side (C) and right side (F) of each track represent the minimum and maximum densities in read per million.

Figure 4: Transcription of the antisense lncRNA triggers activation of sense promoters

(A) Schematics of dCas9-VPR-mediated activation of pCBS-proximal and eCBS-proximal promoters. (B) Activation of the sense and antisense transcription in Pcdhα4 by dCas9-VPR (RNA-Seq). (C) Enrichment of H3K4me3 at the Pcdhα4 promoter (ChIP-qPCR). Errors (n=3) represent s.e.m. and statistical significance was calculated with a Student unpaired *t*-test. (D) Transcription of sense and antisense RNA upon activation of the eCBS-promoters of Pcdh α6, α9, α12 by dCas9-VPR (cRNA-Seq). Side boxes show a zoom-in of the convergent transcription at the activated exons.

For (B and D), the x-axis represents the linear sequence of the genomic organization of the Pcdhα cluster. Arrows indicate the initiation of transcription and the numbers on the left-hand side of each track represent the minimum and maximum densities in read per million.

Figure 5: Antisense lncRNA transcription promotes CTCF binding and promoter/HS5-1 enhancer DNA interactions by DNA demethylation of the CBS sites

(A) Enrichment of CTCF occupancy at the pCBS and the eCBS sites of Pcdh α4, α6, α9, α12 upon activation of either the pCBS-proximal (grey) or the eCBS-proximal (green) promoter by dCas9-VPR (ChIP-qPCR). (B) Left: Virtual 4C with Pcdhα12 promoter as a viewpoint for HEK293T cells activated with dCas9-VPR targeting the pCBS-proximal promoter (grey) or the eCBS-proximal promoter (green) of Pcdhα12. The specific interaction between the Pcdhα12 promoter and the HS5-1 enhancer is highlighted by a black arrow. Right: Quantification of the specific HiC contacts of the Pcdhα12 exon to the HS5-1 enhancer from three biologically independent experiments. The specificity score indicates the signal-to-noise ratio of the interaction in a 6 kb window at 2 kb resolution. (C) Relative levels of 5mC and 5hmC at the pCBS, eCBS and middle sequences of Pcdhα12 in HEK293T cells (white) and HEK293T cells transfected with dCas9-VPR to activate either the sense (grey) or antisense (green) promoters of Pcdhα12.

Errors (n=3) represent s.e.m. and statistical significance was calculated with a Student unpaired *t*-test.

Figure 6: DNA demethylation at Pcdhα promoters correlates with their activation *in vivo*

(A) Top: Schematics showing the maturation of the mouse main olfactory epithelium (OE): horizontal basal cells (HBC), immediate neural precursors (INP), mature olfactory sensory neurons (mOSNs). mOSNs assemble into a functional neural circuit (glomerulus). Bottom: Schematics of stochastic Pcdhα promoter choice in individual mOSNs. (B) 5mC (Black) and 5hmC (Green) profiles of the Pcdhα alternate promoters and exons in HBC (ICAM$^+$), INP (Ngn1$^+$) and mOSN (Omp$^+$) of the mouse main olfactory epithelium. The x-axis represents the linear sequence of the genomic organization of the mouse Pcdhα cluster. The numbers on the left-hand side of each track represent the minimum and maximum read densities in read per million. (C-F) Average of cumulative RPM values for the Pcdhα alternate promoters/exons for 5mC (C), 5hmC (D), as-lncRNAs (E) and s-cRNAs (F) measured during maturation of the OE. (G) Average of cumulative *in situ* Hi-C contacts for the Pcdhα alternate promoters/exons measured during maturation of the OE.

For (C-G), data are represented in Box and whiskers. Error bars represent minimum and maximal values and statistical significance was calculated with one-way ANOVA.

Figure 7: Stochastic DNA demethylation ensures random Pcdhα promoter choice by the CTCF/Cohesin proteins *via* DNA loop-extrusion

(A) Hi-C contacts maps at 10kb resolution for the Pcdhα cluster in mOSNs (Left) and mOSNs upon Rad21 conditional knockout, Rad21 KO (Right); max: 100 reads per billion Hi-C contacts. (B and C) Average HiC contacts of the HS5-1 enhancer with the individual Pcdhα promoters (B) and average RPM values of s-cRNA for individual Pcdhα exons (C) in mOSNs (Blue) and mOSNs upon Rad21 conditional knockout (Black). (D) Left: 5hmC (MeDIP-Seq) and CTCF (ChIP-Seq) profiles in mOSNs (Blue) and mOSNs upon Tet3 overexpression (Red). Right: Quantification of CTCF binding. (E) Hi-C contact maps at 10kb resolution for the Pcdhα cluster in mOSNs overexpressing Tet3; max: 100 reads per billion Hi-C contacts. (F and G) Average HiC contacts of the HS5-1 enhancer with the individual Pcdhα promoters (F) and average RPM values of s-cRNA for individual Pcdhα exons (G) mOSNs overexpressing Tet3. (H) Model for how coupling of as-lncRNA transcription and DNA demethylation ensures a stochastic and HS5-1 distance-independent choice of a Pcdhα promoter. Uncoupling DNA demethylation from as-lncRNA transcription by overexpression of Tet3 results in non-random and HS5-1 distance-biased Pcdhα promoter choice.

SUPPLEMENTAL FIGURE LEGENDS

Figure S1: RNA-Sequencing and Capture RNA-Sequencing

(A) Schematic diagram of Capture RNA-Sequencing (cRNA-Seq). The white, pink and blue bars indicate RNA from the Pcdh α, β and γ gene clusters, respectively. The brown bars indicate RNA from the rest of the genome. (B) RNA-Seq and cRNA-Seq from total RNA isolated from SK-N-SH cells. Red bar: myBaits for the Pcdh α and γ clusters. (C) Enrichment of Pcdh α and γ RNAs by cRNA-Seq. The CBX5 gene was used as a positive control for our capture procedure as we developed myBaits probes to enrich for RNA molecules expressed from the CBX5 locus as well. (D) Sense and antisense RNA reads sequenced by either RNA-Seq or cRNA-Seq. (E) Expression values (RPKM) for as-lncRNA and s-cRNA expressed from Pcdhα4 and Pcdhα12 in SK-N-SH cells.

Figure S2: Expression of convergent Pcdhα antisense and sense RNA in human and mouse primary neurons

(A) Start-Seq signal from two biological replicate experiments in SK-N-SH cells ranked by decreasing read density relative to known transcriptional start sites (TSS) genome-wide. (B) Polyadenylated (PolyA) RNA and Total RNA from human primary neurons (cRNA-Seq). (C) Polyadenylated (PolyA) RNA and Total RNA from mouse olfactory sensory neurons (RNA-Seq). (D) DNaseI hypersensitivity and ChiP-Seq data for distinct transcription factors associated with the active exons by the pCBS-proximal and eCBS-proximal promoters in SK-N-SH cells.

For (B) and (C), the x-axis represents the linear sequence of the genomic organization of the human (B) and mouse (C) Pcdhα gene cluster and the numbers on the left-hand side of each track represent the minimum and maximum densities in read per million.

Figure S3: Recruitment of dCas9-VPR to Pcdhα sense and antisense promoters

(A) Location of the gRNAs used to activate Pcdh α4, α6, α9 and α12, relative to their respective pCBS and the eCBS sites. (B) dCas9-VPR is recruited at the Pcdh α4 and α12 pCBS-proximal and eCBS-proximal promoters. H3K4me3, Rad21 and CTCF (ChIP-Seq) from parental HEK293T cells. The x-axis represents the linear sequence of the genomic organization of the Pcdhα human cluster and the numbers on the left-hand side of each track represent the minimum and maximum densities in read per million. (B) Zoom-in of the dCas9-VPR ChIP-Seq tracks from (B) for Pcdh α4 (Left) and Pcdh α12 (Right).

Figure S4: Functional outcomes of the activation of sense and antisense promoters by dCas9-VPR

(A) Activation of the pCBS-proximal and eCBS-proximal Pcdhα4 promoters by a single dCas9-VPR protein. (B) Percent of uniquely aligned reads from cRNA-Seq for the Pcdh α and γ gene cluster and the CBX5 locus for HEK293T cells (black) and HEK293T cells transfected with gRNA activating the eCBS-proximal promoter of α6 (Red), α9 (green) and α12 (blue). The primary data are shown in Figure 4C. (B) Percent input of CTCF occupancy, as determined by ChIP-qPCR, at the GAPDH promoter (positive control) and at an intergenic DNA region (negative control) for the experiments shown in Figure 5A where the pCBS-proximal (grey) and the eCBS-proximal (green) promoters of Pcdh α4, α6, α9, α12 are activated by dCas9-VPR. Errors (n=3) represent s.e.m. and statistical significance was calculated with a Student unpaired *t*-test. (D) Hi-C contacts between the Pcdhα12 promoter and the HS5-1 enhancer: pCBS-proximal promoter activation (grey); eCBS-proximal promoter activation (green); recruitment of dCas9 (without the VPR activator) to the eCBS-proximal promoter (pink). Y-axis indicated total HiC contacts.

Figure S5: Antisense lncRNA transcription mediates DNA demethylation of Pcdhα promoters

(A and B) Top: Schematics of the pCBS and the eCBS relative to the 11 Zinc fingers of the CTCF protein. Module 2 and 3 represent the core CBS motif. Bottom: DNA Logo for the human Pcdhα pCBS and eCBS sites. CTCF binding to the core CBS motif is significantly affected by DNA methylation at position 2 and 12. (C and D) Nucleotide resolution of the percent CpG DNA methylation of the pCBS and eCBS of Pcdh α4, α12 (ON exons) and α3, α13 (OFF exons) in SK-N-SH cells as determined by whole-genome bisulfite sequencing (WGBS). (E) Average of percentage of CpG methylation at the pCBS (Top) and eCBS (Bottom) of active (ON) and inactive Pcdhα exons (OFF) assayed by WGBS relative to CTCF occupancy of those sites assayed by ChIP-Seq in SK-N-SH cells. (F) Percent CpG DNA methylation of the DNA sequence between the two CBSs (middle) in Pcdh α4, α12 (ON) and α3, α13 (OFF) in SK-N-SH cells. (G) Average of percent CpG methylation of the DNA sequence between the two CBSs of active (ON) and inactive Pcdhα exons (OFF). (H) Nucleotide resolution of the percent CpG DNA methylation of the eCBS of Pcdhα12 in HEK293T cells upon sense promoter activation (Top) and antisense promoter activation (Bottom) assayed by bisulfite sequencing.

Figure S6: DNA demethylation correlates with Pcdhα expression *in vivo*

(A) Changes in 5hmC (x-axis) relative to the expression of the s-cRNA (left y-axis, grey) and the as-lncRNA (right x-axis, black) during the maturation of olfactory sensory neurons. Data for Pcdh α3, α5, α7 and α10 are shown. (B) *In situ* Hi-C contact maps at 10kb resolution for horizontal basal cells (ICAM1$^+$, Top), immediate neural precursors (Ngn1$^+$, Middle) and mature olfactory sensory neurons (Omp$^+$, Bottom).

Figure S7: Rad21 knockout and Tet3 overexpression in mature olfactory sensory neurons

(A) Log2 fold change for Rad21 from Rad21fl/fl;OMPcre mice and Tet3 from tetotet3iresGFP;omptta relative to mOSNs from control mice. (B) Rad21 immunofluorescence (green) in MOE sections from 14-week-old control (Rad21-fl/fl) and Rad21 KO (Rad21 fl/fl;OMPcre) mice. Nuclei are stained with DAPI (magenta). Rad21 is lost from mOSNs but retained in apical sustentactular cells and basal immature cells. Scale bar = 20 μm. (C) Average of cumulative RPM values for the Pcdhα alternate promoters/exons for 5hmC for horizontal basal cells (HBC), immediate neural precursors (INP), and control or Tet3 overexpressing mature olfactory sensory neurons (mOSN). (D) CTCF profiles in mOSNs (Left) and mOSNs overexpressing Tet3 (Right) as measured by ChIP-Seq. (E) RNA-Seq profiles for s-cRNA (grey) and as-lncRNA (black) in mOSNs and mOSNs upon Tet3 overexpression. The x-axis represents the linear sequence of the genomic organization of the mouse Pcdhα cluster and the numbers on the left-hand side of each track represent the minimum and maximum densities in read per million.

STAR METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and request for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Tom Maniatis (tm2472@cumc.columbia.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines and Cell culture

SK-N-SH cells were purchased from ATCC and cultured in RPMI-1640 supplemented with 10% (vol/vol) FBS, 1X GlutaMax, 1mM sodium pyruvate, 1X non-essential amino acids, and 1% penicillin-streptomycin.

HEK293T cells were purchased from ATCC and cultured in DMEM supplemented with 10% (vol/vol) FBS, 1X GlutaMax, 1mM sodium pyruvate, 1X non-essential amino acids, and 1% penicillin-streptomycin. Cells were maintained at 37°C in a 5% (vol/vol) CO_2 incubator.

Generation of a CRISPR-inducible SK-N-SH cell line (SK-N-SH-iCas9)

CRISPR-inducible SK-N-SH cells were generated as previously described for Human pluripotent stem cells (hPSCs) (González et al., 2014; Zhu et al., 2014) with the following differences: (1) the Puro-Cas9 donor plasmid was substituted with a GFP-Cas9 donor plasmid and (2) the Neo-M2rtTA donor plasmid was substituted with a mCherry-M2rtTA donor plasmid. Dual color cells were sorted by flow cytometry and genotyped by PCR and further karyotyped.

Generation of SK-N-SH heterozygous for the Pcdhα cluster (SK-N-SH-αhet)

SK-N-SH-iCas9 cells were plated at 50% density in a 6-well dish, dox-induced (at a concentration of 2 mg/mL) for 48 hours (refresh Media with 1X RPMI with Dox for every day of induction). On days 3 and 5, the cells were transfected with 1 μg (total) of sgRNAs. On day 6, the GFP/mCherry positive and DAPI negative were single cells sorted on plates pre-coated with MEF feeder cells. The cells were allowed to grow for a month until visible colonies were observed, replica plated and genotyped by PCR. We isolated two clones (1 and 2) and named this cell line as SK-N-SH-αhet. Deletion of one copy of the Pcdhα cluster in the SK-N-SH-αhet1 clone was further confirmed by Sanger DNA sequencing and further karyotyped.

Animals

Mice were treated in compliance with the rules and regulations of IACUC under protocol number AC-AAAO3902. All experiments were performed on primary FACS-sorted cells from dissected main olfactory epithelium. HBC cells were sorted from keratin5-creER;rt-gfp mice, INP cells were sorted from the brightest GFP populations of ngn1-GFP mice, OSNs were sorted from omp-IRES-GFP mice (Shykind et al., 2004). Rad21 conditional knockout mOSNs was achieved by crossing Rad21 conditional allele mice (Seitan et al., 2011) to OMP-ires-Cre mice ($\mathrm{Omp}^{\mathrm{tm1(cre)Jae}}$). Recombined cells were purified by including a Cre-inducible tdTomato allele (ROSA26-tdtomato, $\mathrm{Gt(ROSA)26Sor}^{\mathrm{tm14(CAG\text{-}tdTomato)Hze/J}}$) in the cross and selecting tdTomato positive cells by FACS. Overexpression of Tet3 in mOSNs was achieved by crossing tetotet3-IRES-GFP to omptta mice to obtain tetotet3-IRES-GFP;omptta mice. Control mice were achieved by crossing tetoGFP to omptta mice to obtain tetoGFP;omptta mice. GFP positive cells were sorted by FACS for both tetotet3-IRES-GFP;omptta and tetoGFP;omptta mice. In the text and the figures, we refer to the Rad21 conditional knockout in mOSNs as Rad21 KO and the Tet3 overexpression in mOSNs as Tet3 overexpression.

METHODS DETAILS

Fluorescence activated cell sorting of HBCs, INPs and mOSNs

Cells were dissociated into a single-cell suspension by incubating freshly dissected main olfactory epithelium with papain for 40 minutes at 37°C according to the Worthington Papain Dissociation System. Following dissociation and filtering for three times through a 35 µm cell strainer, cells were resuspended in 1X PBS with 5% FBS. For *in situ* Hi-C and ChiP-Seq experiments, upon dissociation, cells were fixed with 1% formaldehyde for 10 minutes at room temperature. Formaldehyde was quenched by adding glycine to a final concentration of 0.125 M for 5 minutes at room temperature. Cells were then washed with 1X cold PBS and resuspended in 1X PBS with 5% FBS. Fluorescent cells were then sorted on a BD Aria II or Influx cell sorter.

Transfections of plasmids into HEK293T cells

One day prior to lipid-mediated transfection, HEK293T cells were seeded in a 6-well plate at a density of about 2 million cells per well. For plasmid DNA transfections, 3 µg of total DNA was added to 125 µL of Opti-MEM containing 5 µL of P300 reagent, followed by an addition 125 µL of Opti-MEM containing 7.5 µL of Lipofectamine 3000 per well. The two solutions were mixed and incubated at room temperature for 5 minutes and the solution was added dropwise to cells. Plates were then incubated at 37°C for 48 or 72 hours in a 5% CO_2 incubator. After incubation, cells were harvested in 1 mL of TRIzol.

RNA isolation and sequencing

RNA was isolated using TRIzol. Cell lysate was extracted with bromo-chloropropane and RNA was precipitated with 100% isopropanol supplemented with 10 μg of glycoblue for 10 min at room temperature and then pelleted at 16,000 x g for 30 min at 4C. The RNA pellet was washed once with 75% ethanol and then resuspended in RNase-free water to a maximal concentration of 200ng/μl. Genomic DNA contaminants were removed by Turbo DNase. Removal of Turbo DNase was performed by phenol:chloroform extraction and RNA was precipitated as described above and resuspended in RNase-free water and stored at -80C.

Sequencing libraries for total RNA and polyadenylated RNA from SK-N-SH cells and human neurons were made using the NEBNext Ultra II Directional RNA Library Prep Kit. Sequencing libraries for total RNA from HEK293T cells and the SK-N-SH-αhet clones were made using the SMARTer Stranded Total RNA-Seq Pico input mammalian RNA kit. The quality of all the libraries was assessed by bioanalyzer and quantified using a combination of bioanalyzer and qubit. Libraries were sequenced on a NEXT-Seq 500/550.

Design of the myBaits Capture Library

To overcome the low level of Pcdh expression in both primary neurons and SK-N-SH cells, we made use of an RNA-based enrichment strategy to capture pre-processed and mature RNA species. We refer to this approach as Capture RNA-Sequencing (cRNA-Seq) (see also Figure S1 for a schematic of the myBaits enrichment procedure).

myBaits targeted capture kits were designed and purchase from MYcroarray (Arbor Biosciences, http://www.arborbiosci.com). A total of 16,357 biotinylated RNA probes covering about 90.42% of the Pcdh α (chr5: 140159476-140429082, hg19) and γ (chr5:140705658-140911381, hg19) clusters were synthesized. We also designed baits for the CBX5 locus (chr12:54624724-54673956, hg19) to serve as a positive control for our enrichment protocol. Baits were design satisfying at least one of the following conditions:

- No blast hit with a T_m above 60°C
- No more than 2 hits at 62.5-65°C or 10 hits in the same interval and at least one neighbor candidate being rejected

- No more than 2 hits at 65-67.5°C and 10 hits at 62.5-65°C and two neighbor candidates on at least one side being rejected
- No more than a single hit at or above 70°C and no more than 1 hit at 65-67.5°C and 2 hits at 62.5-65°C and two neighbor candidates on at least one side being rejected

Sequencing libraries from RNA-Seq or HiC-Seq were multiplexed at the desired ratio and captured using the myBaits Capture Library protocol for 18 hours at 65°C. Captured libraries were eluted in RNase-free water and further amplified. The quality of captured libraries was assessed by bioanalyzer and quantified using a combination of bioanalyzer and qubit. Libraries were sequenced on a NEXT-Seq 500/550.

RNAPII pausing

Start-Seq experiments were previously described (Nechaev et al., 2010) with the following changes: (1) about 10 million SK-N-SH cells were used for each replicate experiment, (2) the 2 μl of RNA 5' Pyrophosphohydrolase, RppH, (NEB M0356S, 5 U/μl) was used in conjunction with ThermoPol Buffer (NEB B9004) to remove the 5'cap to the short-RNAs for 1 hr at 37°C,

(3) RNA-Seq libraries were prepared with the NEXTflex small RNA kit v3. Start-RNA libraries were sequenced using single-end 75-nt cycles on an Illumina NextSeq 500/550 instrument.

The location of promoter-proximal RNAPII and the transcriptional start sites (TSS) were determined by analysis of the full-length reads.

RNAPII elongation

SK-N-SH cells were treated with 100 μM DRB or DMSO for 6 hours. s^{4}UDRB experiments were performed as previously described (Fuchs et al., 2014; 2015) with the following changes: 1 mM s^{4}U was added to media 20 min before cells were harvested. After 6h, DRB and s^{4}U-containing media was removed and replaced with s^{4}U-containing media, and cells were harvested with TRIzol after 0, 8, or 20 min after DRB removal. Cells were flash frozen and stored at -80°C. A no DRB and a no s^{4}U controls were also performed. Total RNA was purified and s^{4}U-RNA was enriched using MTS-biotin chemistry (Duffy and Simon, 2009; Duffy et al., 2015). Briefly, cells were lysed in TRIzol, extracted once with chloroform and the nucleic acids were precipitated with isopropanol. DNA was removed with Turbo DNase. DNase protein was removed by phenol:chloroform:isoamylalcohol extraction, and the RNA was isolated using isopropanol precipitation. RNA was sheared to ~200 bp by adding shearing buffer (150 mM Tris-HCl pH 8.3, 225 mM KCl, 9 mM $MgCl_2$) and heating to 94 °C for 4 min, followed by quenching on ice with EDTA. Sheared RNA was purified using a modified protocol with the RNeasy Mini Kit (Qiagen). To biotinylate the s^4U-RNA, 150 µg sheared RNA was incubated with 60 µg MTS-biotin in biotinylation buffer (150 µL total volume) for 30 min. Excess biotin was removed via chloroform extraction using Phase-Lock Gel Tubes. RNA was precipitated with a 1:10 volume of 3 M NaOAc and an equal volume of isopropanol and centrifuged at 20,000 x g for 20 min. The pellet was washed with an equal volume of 75% ethanol. Purified RNA was dissolved in 200 µl RNase-free water. Biotinylated RNA was separated from non-labeled RNA using glycogen-blocked Dynabeads Streptavidin C1 Beads (Invitrogen). Beads (200 µl) were added to each sample and incubated for 15 min at room temperature, then washed three times with high salt wash buffer (1 ml each, 100 mM Tris-HCl (pH 7.4), 10 mM EDTA, 1 M NaCl, and 0.1% Tween-20). In order to improve the stringency of the washes, an additional three washes with buffer TE (10 mM Tris pH 7.4, 1 mM EDTA) at 55 °C were performed. s^4U-RNA was eluted from Dynabeads with 200 µl freshly prepared elution buffer (10 mM DTT, 100 mM NaCl, 10 mM Tris pH 7.4, 1 mM EDTA) and incubated for 15 min. Enriched RNA was purified by ethanol precipitation and re-biotinylated as above. Excess biotin was removed via chloroform extraction using Phase-Lock Gel Tubes and RNA was purified by RNeasy Mini Kit. s^4U-RNA was enriched on streptavidin beads as above and beads were washed three times with high salt wash buffer. s^4U-RNA was eluted as above and spiked with 200 pg *Schizosaccharomyces pombe* total RNA. 10 ng total RNA from input and enriched RNA samples was used for library preparation with the SMARTer Stranded Total RNA-seq Kit Pico Input Mammalian (Clontech) according to the manufacturer’s instructions. Input and enriched samples were multiplexed with Illumina barcodes and sequenced using paired-end 2 × 75-nt cycles on an Illumina NextSeq 500/550 instrument.

Chromatin Immunoprecipitation (ChIP-Seq and ChIP-qPCR)

The following antibodies were used for chromatin immunoprecipitation studies: CTCF (donated by Victor Lobanenkov), Rad21 (Abcam ab992), Histone H3 Lysine 4 tri-methyl (ThermoFisher PA5-27029), Histone H3 Lysine 27 acetylation (Abcam ab4729), FLAG (Sigma F1804). With the exception of ChIP-Seq experiments for CTCF performed in mOSNs where ~1 million sorted cells were used per IP, about 5 million cells were used. Cells were crosslinked with 1% formaldehyde for 10 minutes at room temperature. Formaldehyde was quenched by adding glycine to a final concentration of 0.125 M for 5 minutes at room temperature. Cells were then washed with 1X cold PBS with protein inhibitors twice and pelleted. Cell pellets were stored at -80C till use. Cells were lysed in lysis buffer (50 mM Tris pH 7.5, 140 mM NaCl, 0.1% SDS, 0.1% sodium deoxycholate, 1% Triton X-100) for 10 minutes. Nuclei were span for 10 minutes at 1000g and resuspended in the sonication buffer (10 mM Tris pH 7.5, 0.5% SDS) as 5^6 nuclei per 300 μl sonication buffer. Chromatin was sheared by Bioruptor for 30 cycles at cycling condition 30/30 (ON/OFF time in seconds). Following a spin at 13,000g for 10 minutes to remove debris, the sheared chromatin was diluted such as the final binding buffer concentration was 15 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS) and incubated for 2 hours with dynabeads G pre-equilibrated in the binding buffer for pre-clearing of the chromatin. Post-cleared chromatin was then incubated with the specific antibody overnight (1 μg of antibody was used per 5^6 nuclei). The next day, dynabeads G were added to the chromatin-antibody mix for 2 hours. A total of four washes with 1X wash buffer (100 mM Tris pH 7.5, 500 mM LiCl, 1% NP-40, 1% sodium deoxycholate) and one wash with TE buffer (10 mM Tris pH 7.5, 1 mM EDTA) were performed. The elution was performed at 65°C for 1 hour in the elution buffer (1% SDS, 250 mM NaCl, 2 mM DTT). All steps, with the exception of the elution, were performed at 4°C. All buffers, with the exception of the TE and elution buffer contained 1X protease inhibitors. The eluted chromatin was reverse-crosslinked overnight at 65°C and the DNA was purified with the Zymo DNA kit.

Libraries for ChIP-Seq were prepared using the NEBNext Ultra II DNA Library Prep Kit. The quality of the libraries was assessed by bioanalyzer and quantified using a combination of bioanalyzer and qubit. Libraries were sequenced on a NEXT-Seq 500/550.

In situ Chromatin Capture Conformation (Hi-C)

HEK293T cells transfected with dCas9-VPR-GFP plasmids were fixed with 1% formaldehyde and GFP-positive cells were FACS-sorted. About 500,000 cells (SK-N-SH or HEK293T) were lysed and intact nuclei were processed through an *in situ* Hi-C protocol as previously described with a few modifications (Rao et al., 2014). Briefly, cells were lysed with 50 mM Tris pH 7.5 0.5% Igepal, 0.25% Sodium-deoxychloate, 0.1%

SDS, 150 mM NaCl, and protease inhibitors. Pelleted intact nuclei were then resuspended in 0.5% SDS and incubated for 20 minutes at 65°C for nuclear permeabilization. After quenching with 1.1% Triton-X for 10 minutes at 37°C, nuclei were digested with 6 U/µl of DpnII in 1x DpnII buffer overnight at 37°C. Following initial digestion, a second DpnII digestion was performed at 37°C for 2 hours. DpnII was heat-inactivated at 65°C for 20 minutes. For the 1.5hr fill-in at 37°C, biotinylated dGTP was used instead of dATP to increase ligation efficiency. Ligation was performed at 25°C for 4 hours. Nuclei were then pelleted and sonicated in 10 mM Tris pH 7.5, 1 mM EDTA, 0.25% SDS on a Covaris S220 for 16 minutes with 2% duty cycle, 105 intensity, 200 cycles per burst, 1.8-1.85 W, and max temperature of 6°C. DNA was reverse cross-linked overnight at 65°C with proteinase K and RNAse A.

Reverse cross-linked DNA was purified with 2x Ampure beads following the standard protocol. Biotinylated fragments were enriched using Dynabeads MyOne Strepavidin T1 beads. The biotinylated DNA fragments were prepared for next-generation sequencing on the beads by using the Nugen Ovation Ultralow kit protocol with some modifications. Following end repair, magnetic beads were washed twice at 55°C with 0.05% Tween, 1 M NaCl in Tris/EDTA pH 7.5. Residual detergent was removed by washing the beads twice in 10 mM Tris pH 7.5. End repair buffers were replenished to original concentrations, but the enzyme and enhancer was omitted before adapter ligation. Following adaptor ligation, beads underwent five washes with 0.05% Tween, 1 M NaCl in Tris/EDTA pH 7.5 at 55°C and two washes with 10mM Tris pH 7.5. DNA was amplified by 10 cycles of PCR, irrespective of starting material. Beads were reclaimed and amplified unbiotinylated DNA fragments were purified with 0.8x Ampure beads. Quality and concentration of libraries were assessed by Agilent Bioanalyzer and Qubit. *In situ* Hi-C libraries from SK-N-SH and HEK293T cells were size-selected and enriched as described above using the myBaits Capture Library protocol described above and sequenced paired-end on NextSeq 500 (2x75bp).

Methylated DNA Immunoprecipitation (MeDIP)

The following antibodies were used: 5-Methylcytosine (5-mC) antibody (Active Motif 39649) and 5-Hydroxymethylcytosine (5-hmC) antibody (Active Motif 39791).

HEK293T cells were transfected with the appropriate set of dCas9 plasmids and incubated at 37°C for 72 hours in a 5% CO_2 incubator. Genomic DNA was extracted using the PureLink Genomic DNA Mini Kit

(Invitrogen). A total of 2 μg of DNA was diluted into 300 μl TE sonication buffer (10 mM Tris pH 7.5, 1 mM EDTA). Genomic DNA was sheared by Bioruptor for 18 cycles at cycling condition 30/90 (ON/OFF time in seconds). The sheared DNA was diluted to a final IP buffer of 15 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100 and incubated overnight with 1 μg of antibody. The next day, a mixture of dynabeads A and G were added to the DNA-antibody mix for 2 hours. A total of three washes with 1X IP buffer were performed. The elution was performed at 55°C for 3 hours with rigorous shaking in the elution buffer (1% SDS, 250 mM NaCl). All steps, with the exception of the elution, were performed at 4°C. The eluted DNA was purified with the Zymo DNA kit.

Bisulfite DNA Reactions

Bisulfite DNA reactions were performed using the TrueMethyl oxBS module, Nugen, following the steps indicated by the protocol. Primers were designed using the MethPrimer. PCR products were cloned and sequenced (at least 15 clones per condition). Data were analyzed using QUMA (http://quma.cdb.riken.jp).

Immunofluorescence

The MOE was dissected from 14-week old Rad21 KO (Rad21-fl/fl;OMP-cre) mice and littermate controls (Rad21-fl/fl). Tissue was embedded in OCT and then coronal cryosections were collected at a thickness 12 μM. Tissue sections were air dried on slides for 10 minutes and then fixed with cold 4 % PFA for 10 minutes. After fixation, slides were washed with PBST (PBS with 0.1 % Triton X-100) and then stained with primary antibody for Rad21 (1:1000 dilution, Abcam Cat# ab42522, RRID: AB_945133) in PBST-DS overnight at 4°C. Slides were then washed, stained with DAPI (2.5 μg/mL) and the secondary antibody (Donkey anti-rabbit IgG conjugated to Alexa-488, diluted 1:1000, Thermo Fisher Scientific Cat# A-21206, RRID:AB_2535792) in PBST-DS for 1 hour, washed, and then mounted with Vectashield. Confocal images were collected with a Zeiss LSM 700 and image processing was carried out with ImageJ (NIH).

Bioinformatic Analysis of Sequencing Data

For RNA-Seq experiments, raw FASTQ files were aligned with either Tophat or STAR using hg19 or mm10 reference genomes. When libraries were made following the SMARTer Stranded Total RNA-Seq, the initial 4 base pairs of both paired reads were trimmed prior to alignment.

For ChiP-Seq experiments, raw FASTQ files were aligned using Bowtie2 using hg19 reference genome upon adapter sequences removal using CutAdapt. Uniquely aligning reads were selected using Samtools and reads with alignment quality below 30 (-q 30) were removed. The HOMER software package was used to generate signal tracks.

For *in situ* Hi-C experiments, raw FASTQ files were processed through use of the Juicer Tools Version 1.76 pipeline (Durand et al., 2016) with one modification. Reads were aligned to hg38 using BWA 0.7.17 mem algorithm and specifying the -5 option implemented specifically for *in situ* Hi-C data. For captured Hi-C libraries, contact matrices were normalized to 2kb resolution by first reporting counts as reads per billion Hi-C contacts, then by normalizing with the Knight Ruiz (KR) matrix balancing algorithm (Knight and Ruiz, 2013) focused on the alpha Pcdh cluster (chr5:140780000-141046000; hg38). For uncaptured libraries (mm10 Hi-C), matrices were KR normalized genome wide.

For generating a contact matrix, scales were set to a minimum of 0 reads and a maximum of 2*(mean normalized reads) in order to report a relative enrichment of contacts.

DNaseI and ChIP data for H3K4me3, CTCF, Rad21, ELF1, GABP, TCF12, MAX, YY1 in SK-N-SH cells were obtained from the ENCODE data matrix.

For Start-Seq experiments, raw FASTQ files were aligned using Bowtie2. TSS peaks were determined using Homer and the most abundant TSS reported in Figure 2.

In situ Hi-C data for INP and OSN cells were obtained from (Horta et al., 2018).

CRISPR gRNA design

All guide RNA (gRNAs) were designed as truncated 18mer long sequences to increase their binding specificity as previously described (Fu et al., 2014) using the CRISPR design web tool (http://crispr.mit.edu). With the exception of the Pcdhα9, where a total of two gRNAs were used to activate either the pCBS-proximal or the eCBS-proximal promoters, we used four gRNAs for the activation of the pCBS-proximal and eCBS-proximal promoters of Pcdh α4, α6, α12.

In vitro transcription of gRNAs

The gRNAs were transcribed using the MEGAshortscript T7 Transcription Kit by Life Technologies (AM1354M), purified by phenol-chloroform and transfected in the SK-N-SH-iCas9 cells by RNAimax lipofectamine reagent.

QUANTIFICATION AND STATISTICS

The statistical tests used in this study are indicated in the respective figure legends. In general, data with single independent experiments were analyzed by Student unpaired *t*-test to determine statistical significant effects ($p < 0.05$). Data with multiple independent experiments were analyzed by one-way ANOVA to determine statistical significant effects ($p < 0.05$).

DATA AND SOFTWARE AVAILABILITY

The data discussed in this work have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE115862.

(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE115862)

The following secure token has been created to allow review of record GSE115862 while it remains in private status: chgbwswuxzyfzkx

SUPPLEMENTAL DATA TABLES

Supplemental Data Table 1: Table of the primers used in this study

Supplemental Data Table 2: Table of the guide RNAs (gRNAs) used in this study

Supplemental Data Table 3: Table of sequence processing steps for the Hi-C experiments

Figure 1

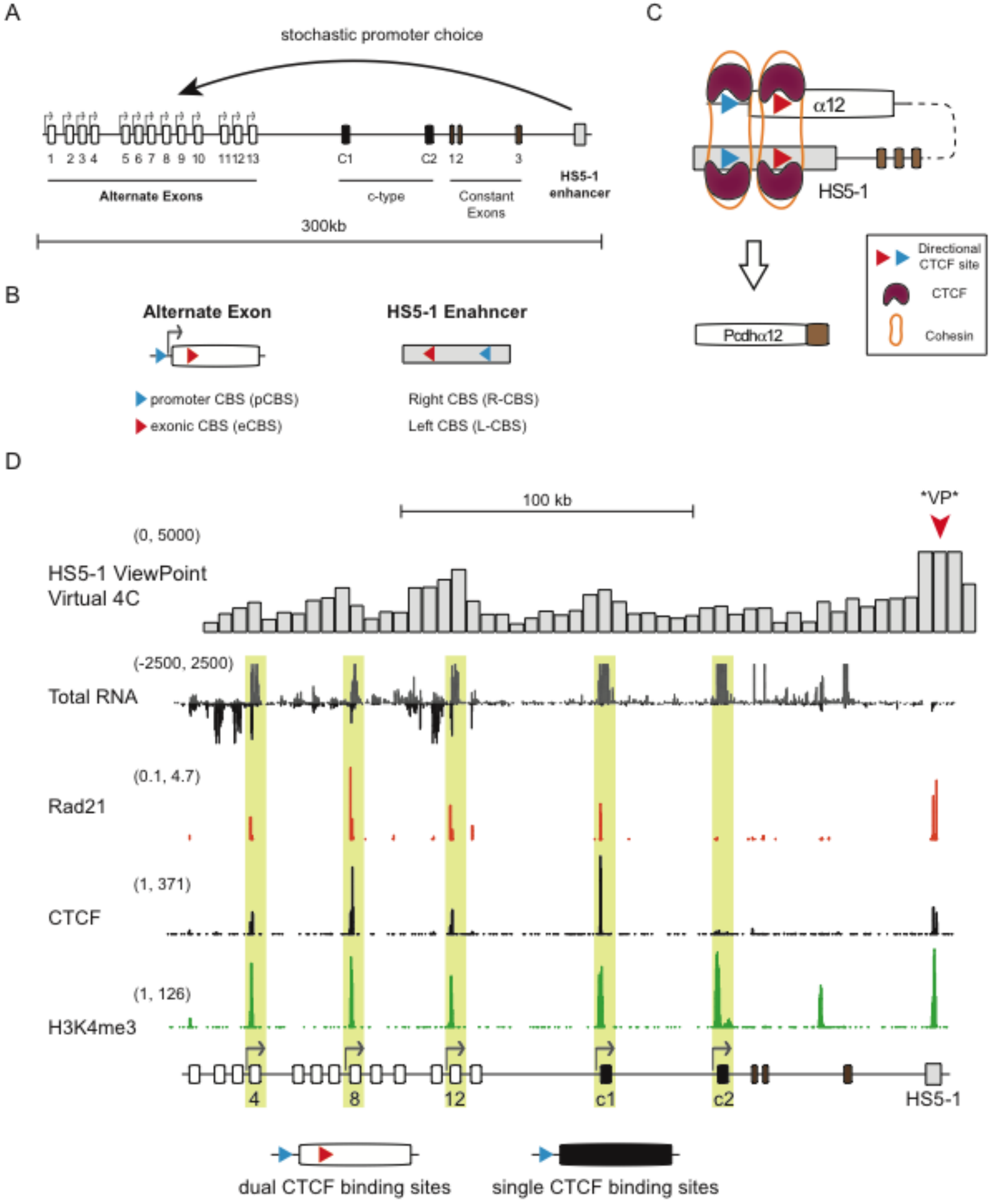

A
stochastic promoter choice
1 2 3 4 5 6 7 8 9 10 11 12 13
C1
C2
12
3
Alternate Exons
c-type
Constant Exons
HS5-1 enhancer
300kb
B
Alternate Exon
HS5-1 Enahncer
promoter CBS (pCBS)
exonic CBS (eCBS)
Right CBS (R-CBS)
Left CBS (L-CBS)
C
α12
HS5-1
Pcdhα12
Directional CTCF site
CTCF
Cohesin
D
100 kb
VP
(0, 5000)
HS5-1 ViewPoint Virtual 4C
(-2500, 2500)
Total RNA
(0.1, 4.7)
Rad21
(1, 371)
CTCF
(1, 126)
H3K4me3
4
8
12
c1
c2
HS5-1
dual CTCF binding sites
single CTCF binding sites

Figure 2

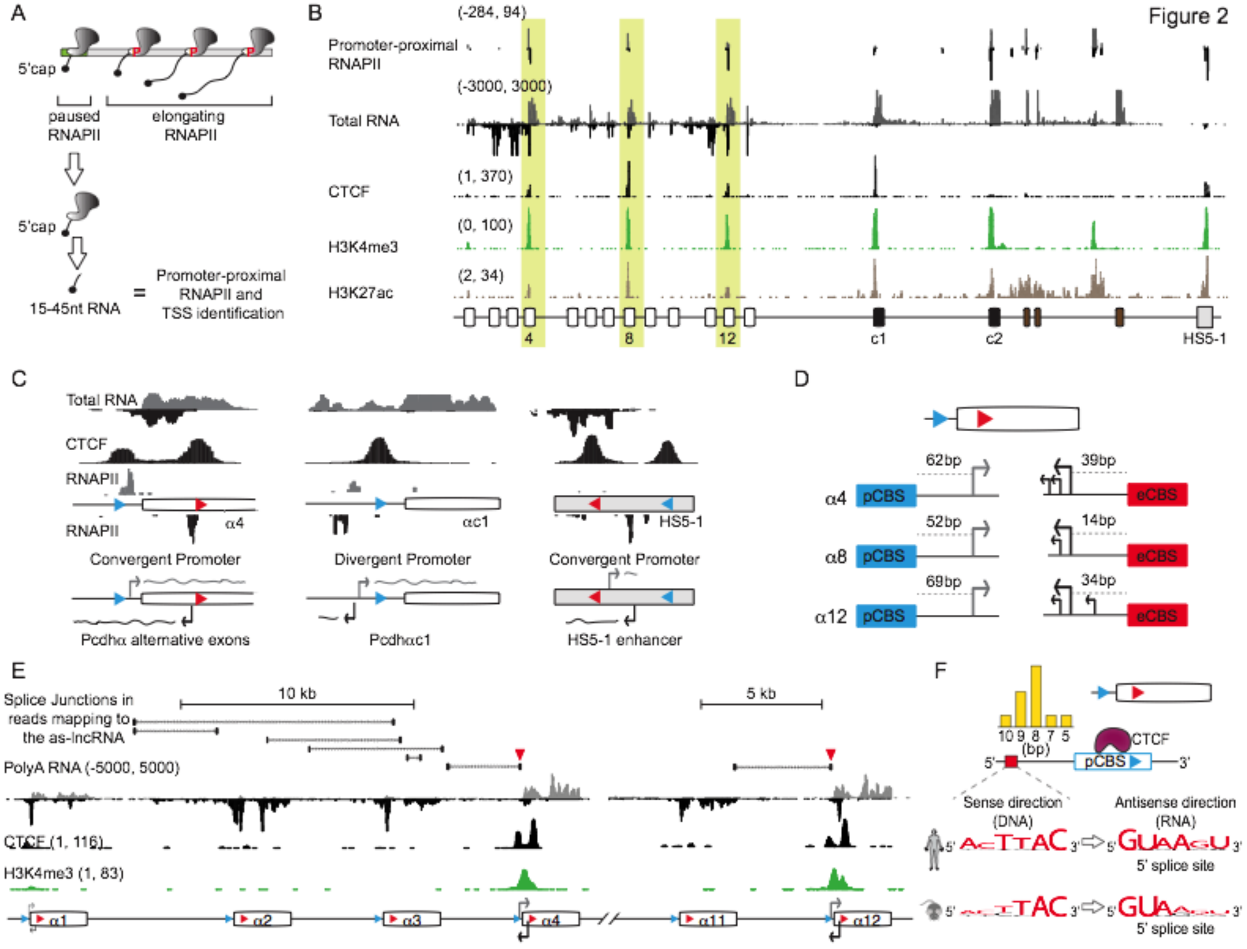

A
5'cap
paused RNAPII
elongating RNAPII
5'cap
15-45nt RNA
=
Promoter-proximal RNAPII and TSS identification
B
(-284, 94)
Promoter-proximal RNAPII
(-3000, 3000)
Total RNA
(1, 370)
CTCF
(0, 100)
H3K4me3
(2, 34)
H3K27ac
4
8
12
c1
c2
HS5-1
C
Total RNA
CTCF
RNAPII
RNAPII
α4
αc1
HS5-1
Convergent Promoter
Divergent Promoter
Convergent Promoter
Pcdhα alternative exons
Pcdhαc1
HS5-1 enhancer
D
α4
pCBS
62bp
39bp
eCBS
α8
pCBS
52bp
14bp
eCBS
α12
pCBS
69bp
34bp
eCBS
E
Splice Junctions in reads mapping to the as-lncRNA
10 kb
5 kb
PolyA RNA (-5000, 5000)
CTCF (1, 116)
H3K4me3 (1, 83)
α1
α2
α3
α4
α11
α12
F
10 9 8 7 5
(bp)
CTCF
5'
pCBS
3'
Sense direction (DNA)
Antisense direction (RNA)
5' AcTTAC 3'
5' GUAAGU 3'
5' splice site
5' TAC 3'
5' GUA 3'
5' splice site

Figure 3

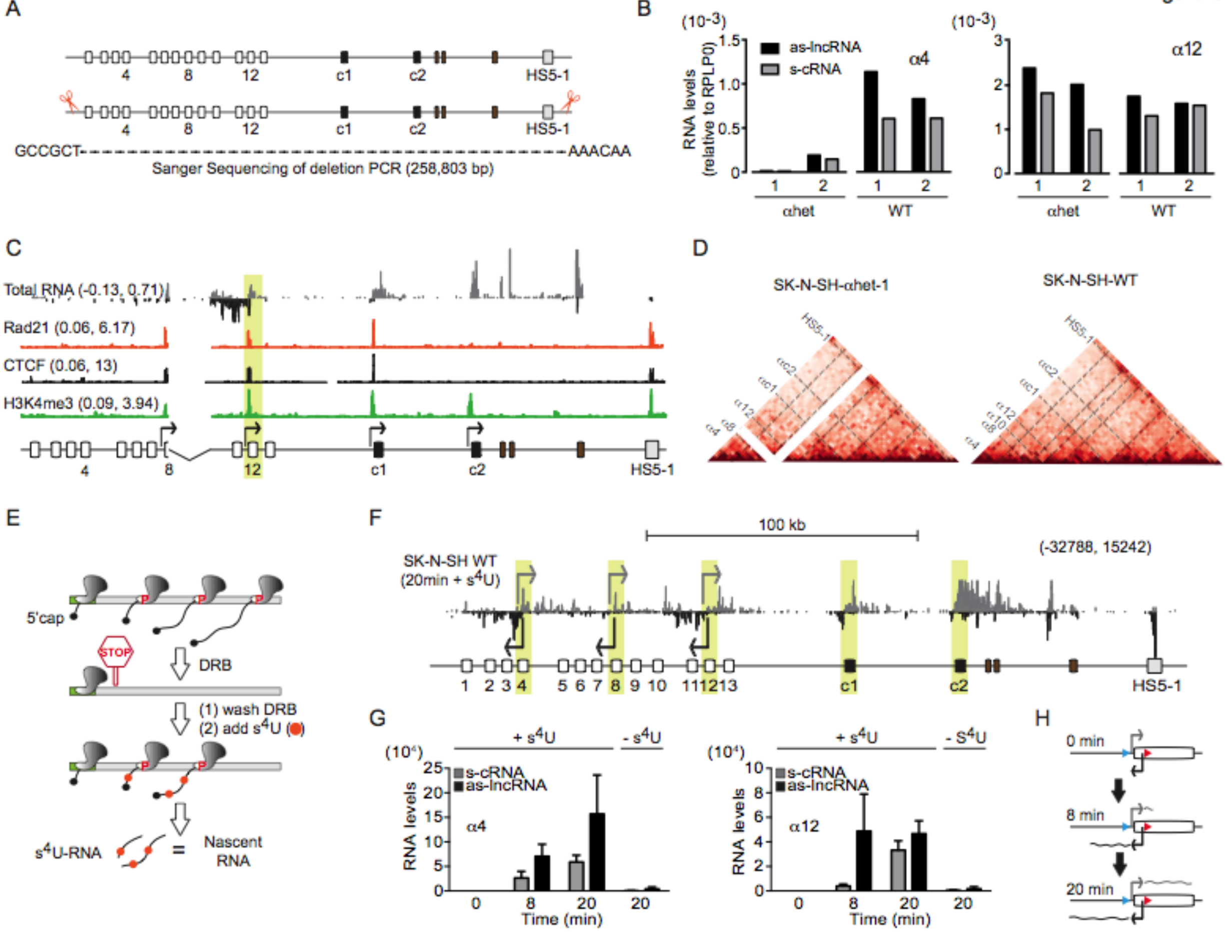

A
4
8
12
c1
c2
HS5-1
GCCGCT
AAACAA
Sanger Sequencing of deletion PCR (258,803 bp)
B
(10-3)
1.5
1.0
0.5
0
RNA levels
(relative to RPLP0)
as-lncRNA
s-cRNA
α4
1
2
αhet
WT
(10-3)
3
2
1
0
α12
C
Total RNA (-0.13, 0.71)
Rad21 (0.06, 6.17)
CTCF (0.06, 13)
H3K4me3 (0.09, 3.94)
4
8
12
c1
c2
HS5-1
D
SK-N-SH-αhet-1
SK-N-SH-WT
HS5-1
αc2
αc1
α12
α8
α4
α10
E
5'cap
STOP
DRB
(1) wash DRB
(2) add s4U
s4U-RNA
=
Nascent RNA
F
100 kb
(-32788, 15242)
SK-N-SH WT
(20min + s4U)
1
2
3
4
5
6
7
8
9
10
11
12
13
c1
c2
HS5-1
G
(10-4)
+ s4U
- s4U
25
20
15
10
5
0
s-cRNA
as-lncRNA
RNA levels
α4
0
8
20
20
Time (min)
(10-4)
+ s4U
- S4U
10
8
6
4
2
0
α12
H
0 min
8 min
20 min

Figure 4

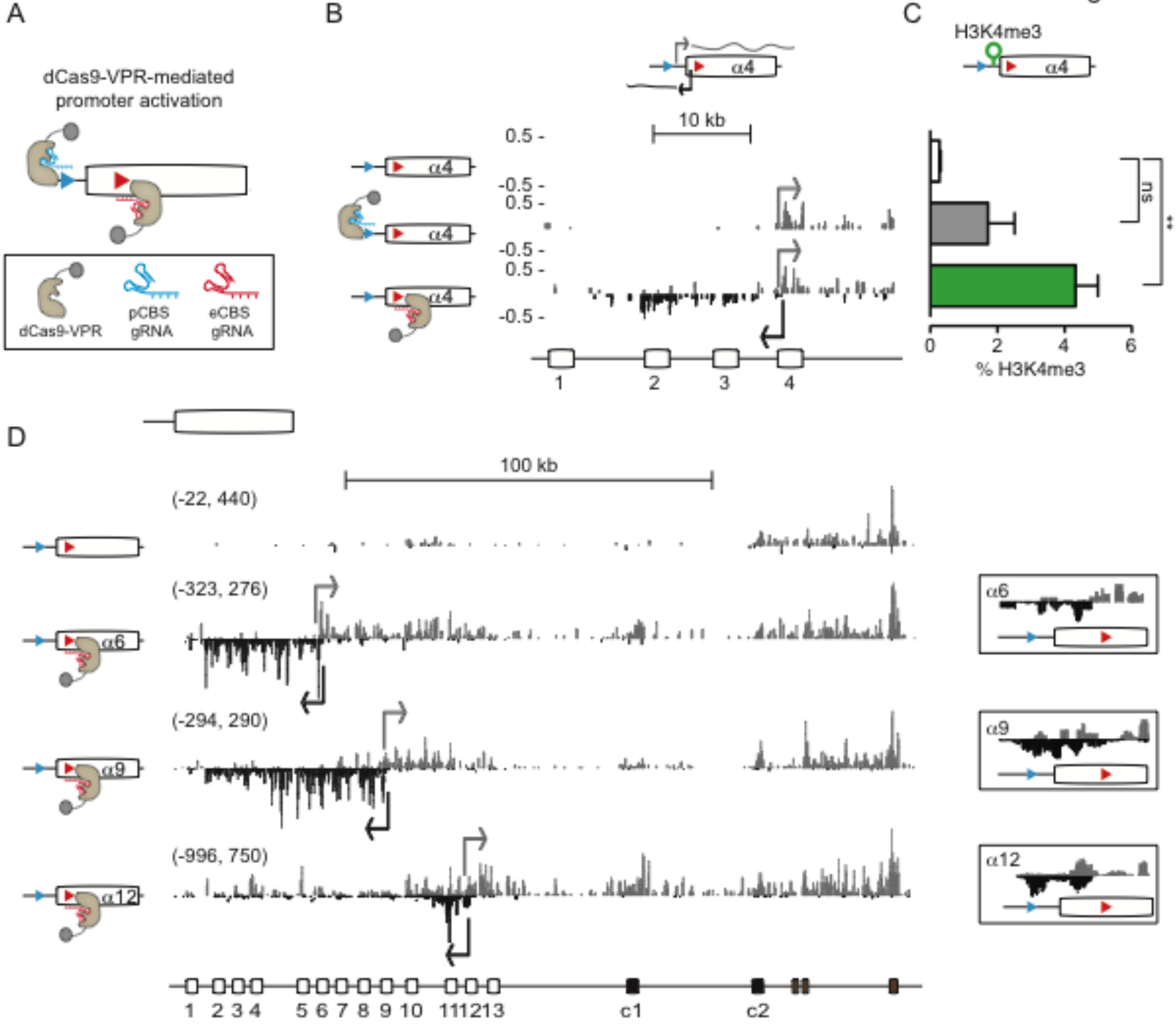

A
dCas9-VPR-mediated
promoter activation
dCas9-VPR
pCBS
gRNA
eCBS
gRNA
B
α4
10 kb
0.5 -
-0.5 -
0.5 -
-0.5 -
0.5 -
-0.5 -
1
2
3
4
C
H3K4me3
α4
ns
**
0
2
4
6
% H3K4me3
D
100 kb
(-22, 440)
(-323, 276)
α6
(-294, 290)
α9
(-996, 750)
α12
1
2 3 4
5 6 7 8 9 10
111213
c1
c2

Figure 5

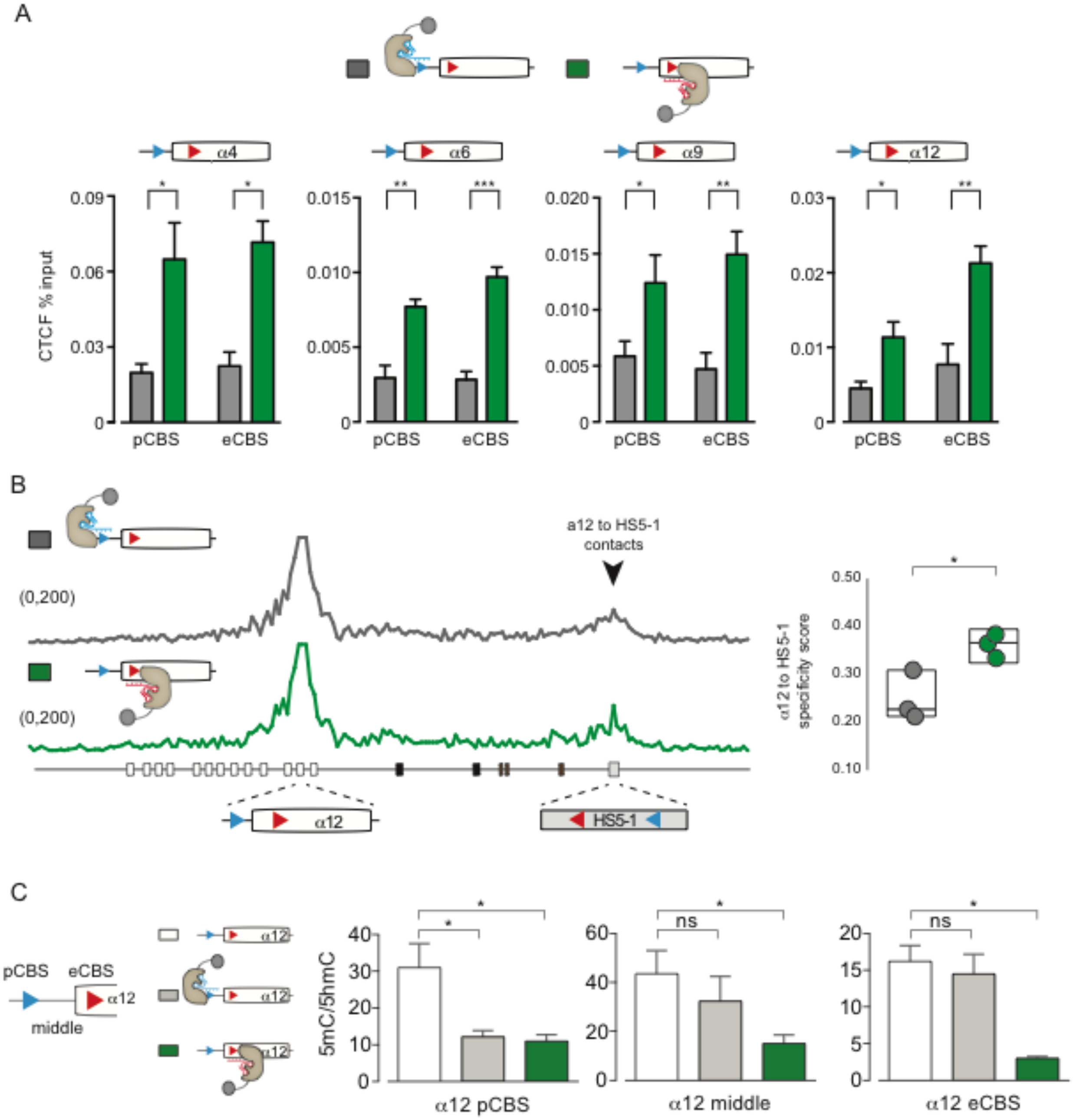

A
α4
α6
α9
α12
CTCF % input
0.09
0.06
0.03
0
0.015
0.010
0.005
0.020
0.03
0.02
0.01
*
**
***
pCBS
eCBS
B
a12 to HS5-1 contacts
(0,200)
(0,200)
α12
HS5-1
α12 to HS5-1 specificity score
0.50
0.40
0.30
0.20
0.10
*
C
pCBS
eCBS
middle
α12
5mC/5hmC
40
30
20
10
0
60
15
5
ns
*
α12 pCBS
α12 middle
α12 eCBS

Figure 6

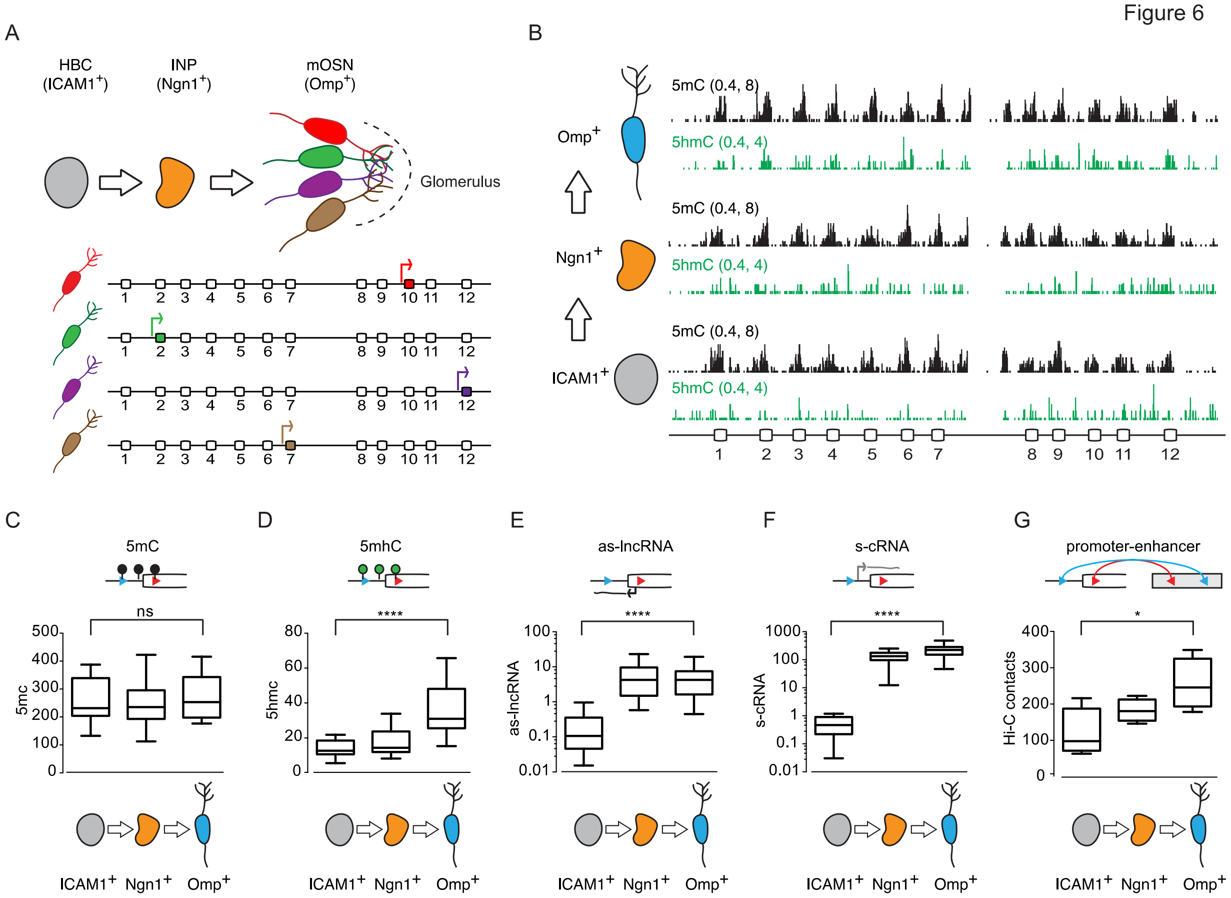

A
HBC
(ICAM1$^+$)
INP
(Ngn1$^+$)
mOSN
(Omp$^+$)
Glomerulus
1 2 3 4 5 6 7 8 9 10 11 12
B
Omp$^+$
5mC (0.4, 8)
5hmC (0.4, 4)
Ngn1$^+$
5mC (0.4, 8)
5hmC (0.4, 4)
ICAM1$^+$
5mC (0.4, 8)
5hmC (0.4, 4)
1 2 3 4 5 6 7 8 9 10 11 12
C
5mC
ns
5mc
0 100 200 300 400 500
ICAM1$^+$ Ngn1$^+$ Omp$^+$
D
5mhC
****
5hmc
0 20 40 60 80
ICAM1$^+$ Ngn1$^+$ Omp$^+$
E
as-lncRNA
****
as-lncRNA
0.01 0.1 1 10 100
ICAM1$^+$ Ngn1$^+$ Omp$^+$
F
s-cRNA
****
s-cRNA
0.01 0.1 1 10 100 1000
ICAM1$^+$ Ngn1$^+$ Omp$^+$
G
promoter-enhancer
*
Hi-C contacts
0 100 200 300 400
ICAM1$^+$ Ngn1$^+$ Omp$^+$

Figure 7

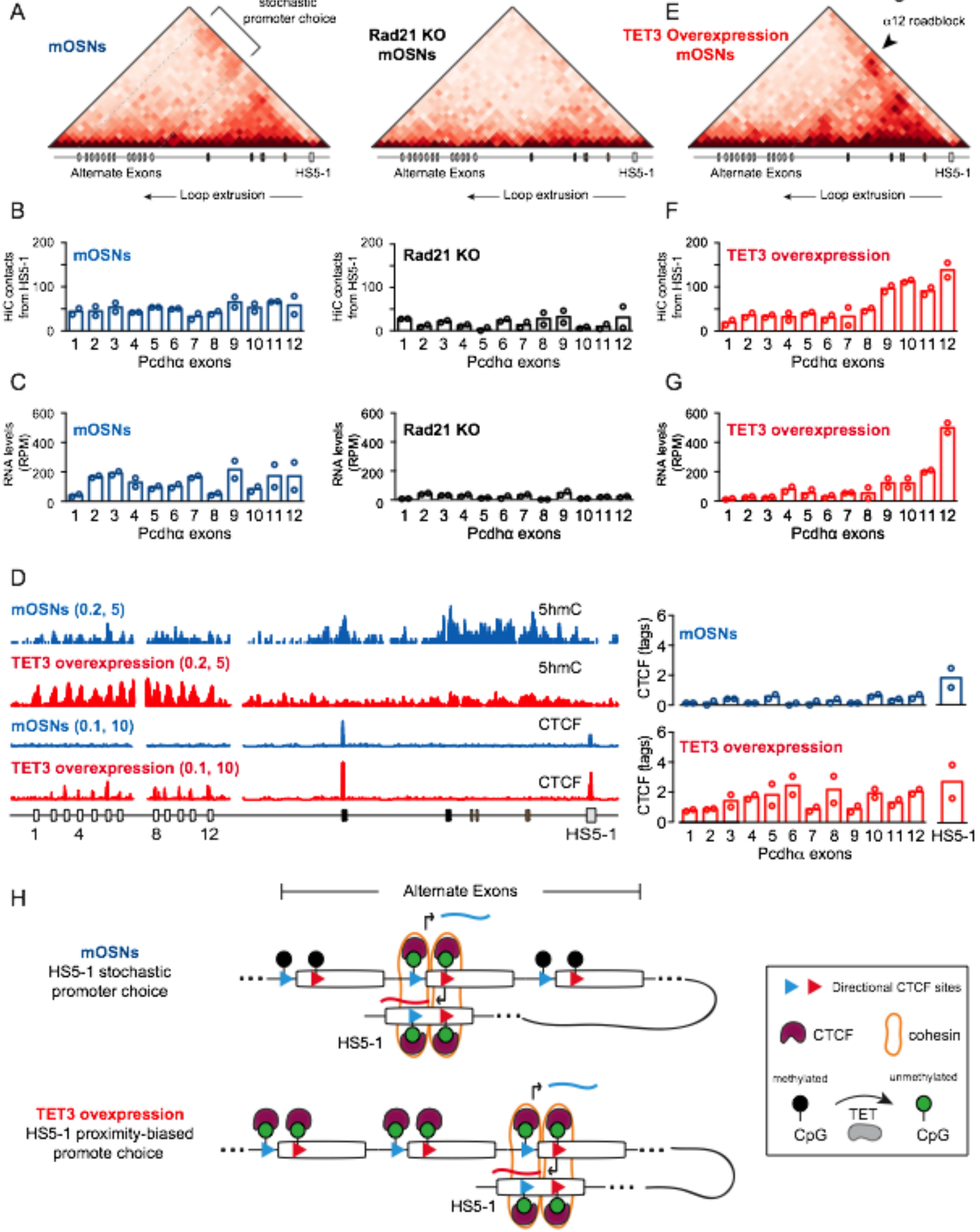
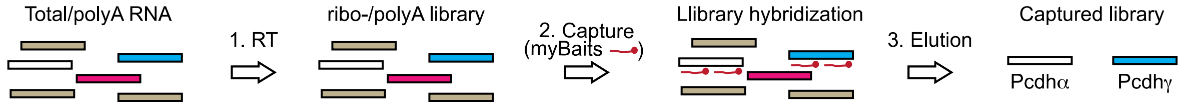
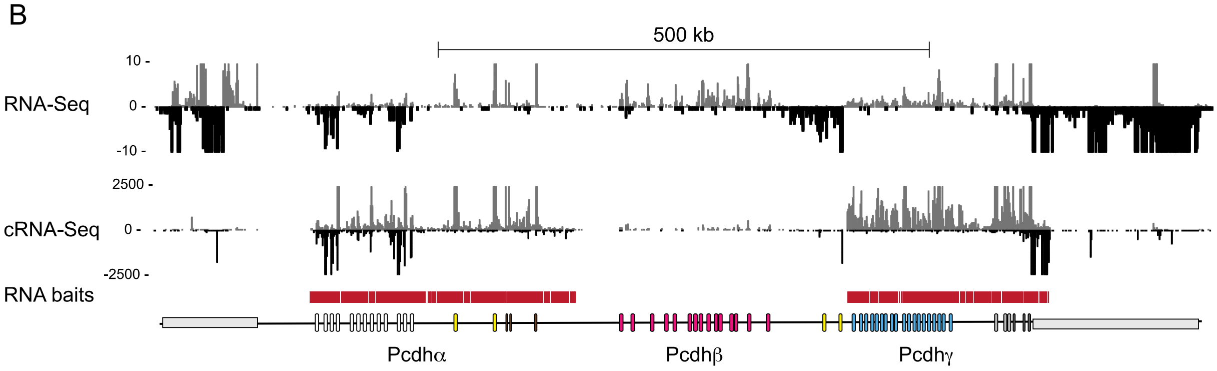

A
mOSNs
stochastic promoter choice
Alternate Exons
HS5-1
Loop extrusion
Rad21 KO mOSNs
Alternate Exons
HS5-1
Loop extrusion
E
TET3 Overexpression mOSNs
α12 roadblock
Alternate Exons
HS5-1
Loop extrusion
B
HiC contacts from HS5-1
200
100
0
mOSNs
1 2 3 4 5 6 7 8 9 10 11 12
Pcdhα exons
Rad21 KO
F
TET3 overexpression
C
RNA levels (RPM)
600
400
200
0
mOSNs
Rad21 KO
G
TET3 overexpression
D
mOSNs (0.2, 5)
5hmC
TET3 overexpression (0.2, 5)
5hmC
mOSNs (0.1, 10)
CTCF
TET3 overexpression (0.1, 10)
CTCF
1 4 8 12
HS5-1
CTCF (tags)
6
4
2
0
mOSNs
TET3 overexpression
1 2 3 4 5 6 7 8 9 10 11 12 HS5-1
Pcdhα exons
H
Alternate Exons
mOSNs
HS5-1 stochastic promoter choice
HS5-1
TET3 ovexpression
HS5-1 proximity-biased promote choice
HS5-1
Directional CTCF sites
CTCF
cohesin
methylated
unmethylated
CpG
TET
CpG

Figure S1

A



B



C

	Uniquely Mapped Reads	Pcdhα Mapped Reads	% Pcdhα Mapped Reads	Pcdhγ Mapped Reads	% Pcdhγ Mapped Reads	CBX5 Mapped Reads	% CBX5 Mapped Reads
RNA-Seq	35721051	3530	0.01	8259	0.02	1796	0.005
cRNA-Seq	3398107	161360	4.7	371737	10.9	96433	2.8
fold enrichment		46		45		54	

D

	Pcdhα Mapped Reads	(+) Pcdhα Mapped Reads	(+) Pcdhα Mapped Reads (%)	(-) Pcdhα Mapped Reads	(-) Pcdhα Mapped Reads (%)
RNA-Seq	3530	3010	85	520	15
cRNA-Seq	161360	132350	82	29010	18

E

	as-lncRNA	sc-RNA
Pcdhα4	508	424
Pcdhα12	1255	1556

Figure S2

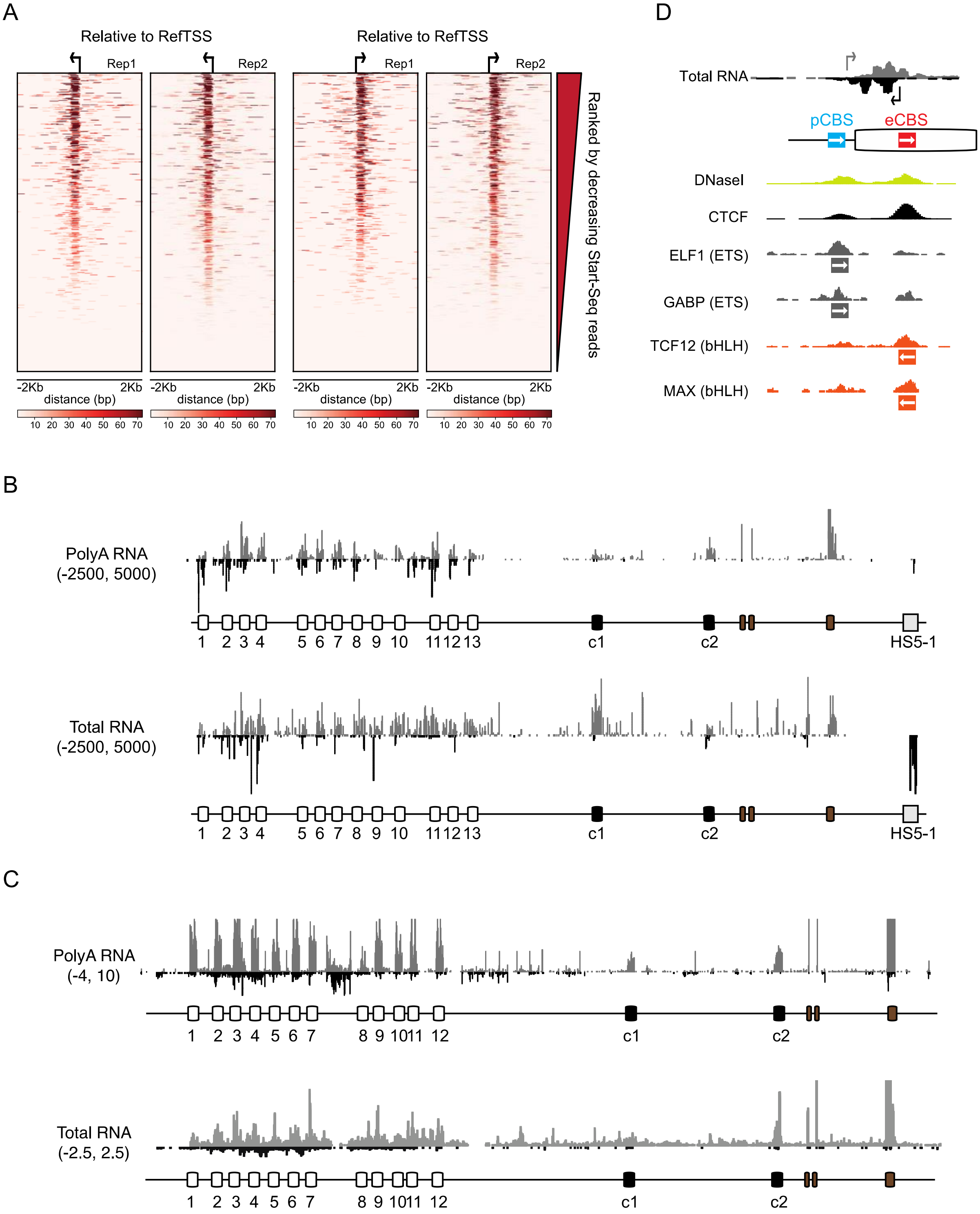

A
Relative to RefTSS
Rep1
Rep2
Relative to RefTSS
Rep1
Rep2
Ranked by decreasing Start-Seq reads
-2Kb
2Kb
distance (bp)
10 20 30 40 50 60 70
D
Total RNA
pCBS
eCBS
DNaseI
CTCF
ELF1 (ETS)
GABP (ETS)
TCF12 (bHLH)
MAX (bHLH)
B
PolyA RNA
(-2500, 5000)
1 2 3 4 5 6 7 8 9 10 11 12 13 c1 c2 HS5-1
Total RNA
(-2500, 5000)
1 2 3 4 5 6 7 8 9 10 11 12 13 c1 c2 HS5-1
C
PolyA RNA
(-4, 10)
1 2 3 4 5 6 7 8 9 10 11 12 c1 c2
Total RNA
(-2.5, 2.5)
1 2 3 4 5 6 7 8 9 10 11 12 c1 c2

Figure S3

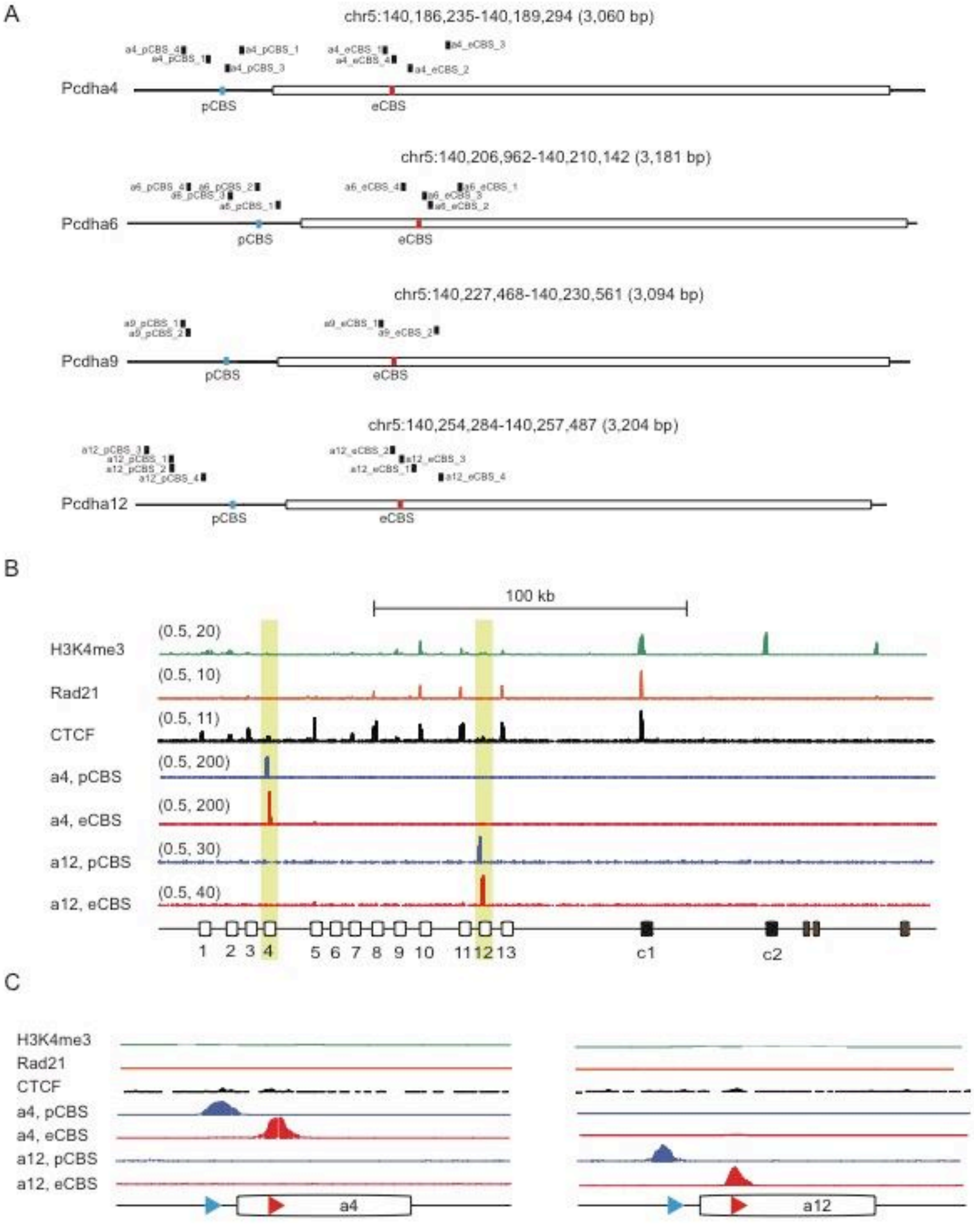

A
chr5:140,186,235-140,189,294 (3,060 bp)
a4_pCBS_4
a4_pCBS_1
a4_pCBS_1
a4_pCBS_3
a4_eCBS_1
a4_eCBS_4
a4_eCBS_3
a4_eCBS_2
Pcdha4
pCBS
eCBS
chr5:140,206,962-140,210,142 (3,181 bp)
a6_pCBS_4
a6_pCBS_2
a6_pCBS_3
a6_pCBS_1
a6_eCBS_4
a6_eCBS_1
a6_eCBS_3
a6_eCBS_2
Pcdha6
pCBS
eCBS
chr5:140,227,468-140,230,561 (3,094 bp)
a9_pCBS_1
a9_pCBS_2
a9_eCBS_1
a9_eCBS_2
Pcdha9
pCBS
eCBS
chr5:140,254,284-140,257,487 (3,204 bp)
a12_pCBS_3
a12_pCBS_1
a12_pCBS_2
a12_pCBS_4
a12_eCBS_2
a12_eCBS_3
a12_eCBS_1
a12_eCBS_4
Pcdha12
pCBS
eCBS
B
100 kb
H3K4me3
(0.5, 20)
Rad21
(0.5, 10)
CTCF
(0.5, 11)
a4, pCBS
(0.5, 200)
a4, eCBS
(0.5, 200)
a12, pCBS
(0.5, 30)
a12, eCBS
(0.5, 40)
1
2
3
4
5
6
7
8
9
10
11
12
13
c1
c2
C
H3K4me3
Rad21
CTCF
a4, pCBS
a4, eCBS
a12, pCBS
a12, eCBS
a4
a12

Figure S4

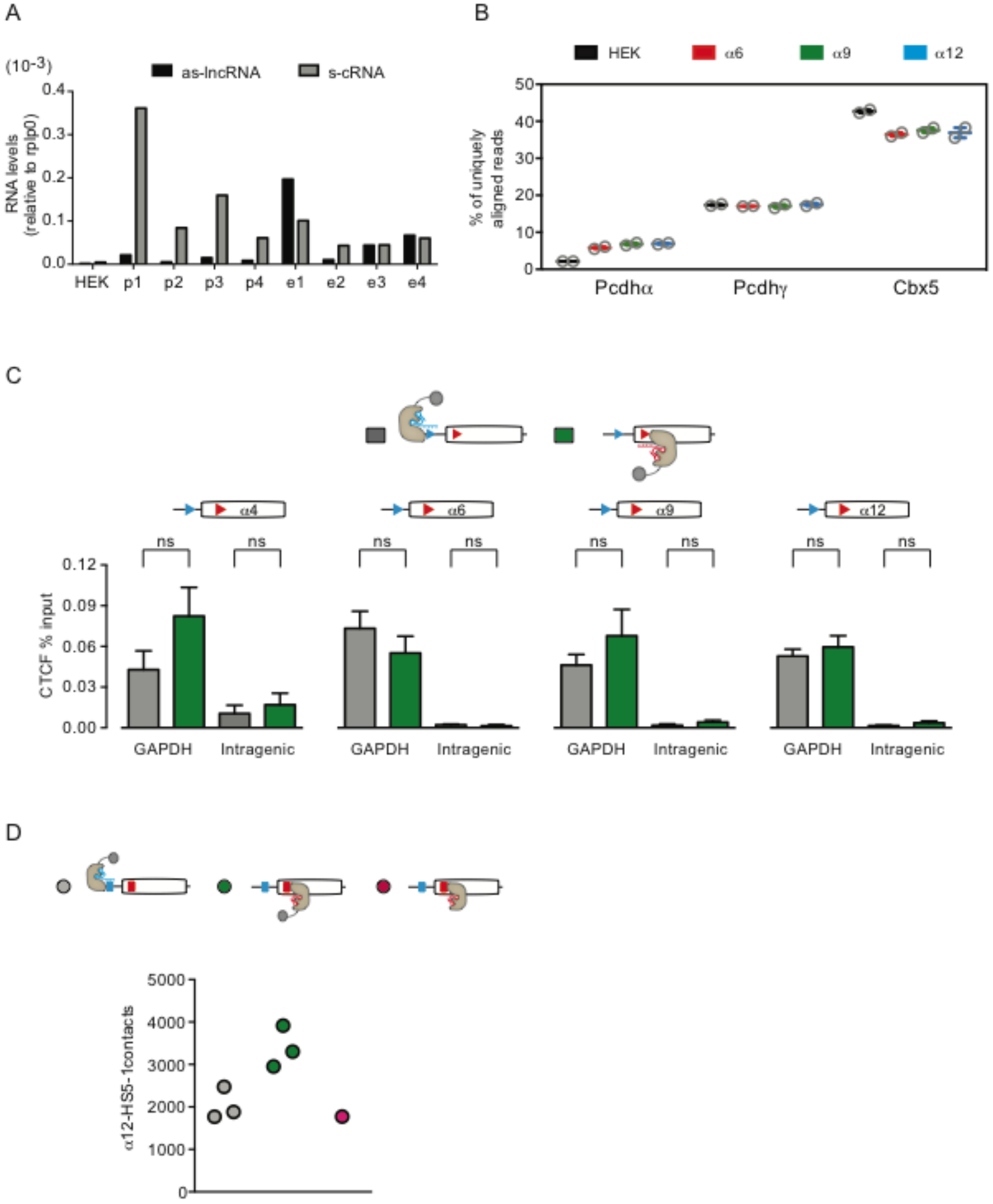

A
(10-3)
as-lncRNA
s-cRNA
RNA levels
(relative to rplp0)
0.4
0.3
0.2
0.1
0.0
HEK
p1
p2
p3
p4
e1
e2
e3
e4
B
HEK
α6
α9
α12
% of uniquely
aligned reads
50
40
30
20
10
0
Pcdhα
Pcdhγ
Cbx5
C
α4
α6
α9
α12
ns
CTCF % input
0.12
0.09
0.06
0.03
0.00
GAPDH
Intragenic
D
α12-HS5-1contacts
5000
4000
3000
2000
1000
0

Figure S5

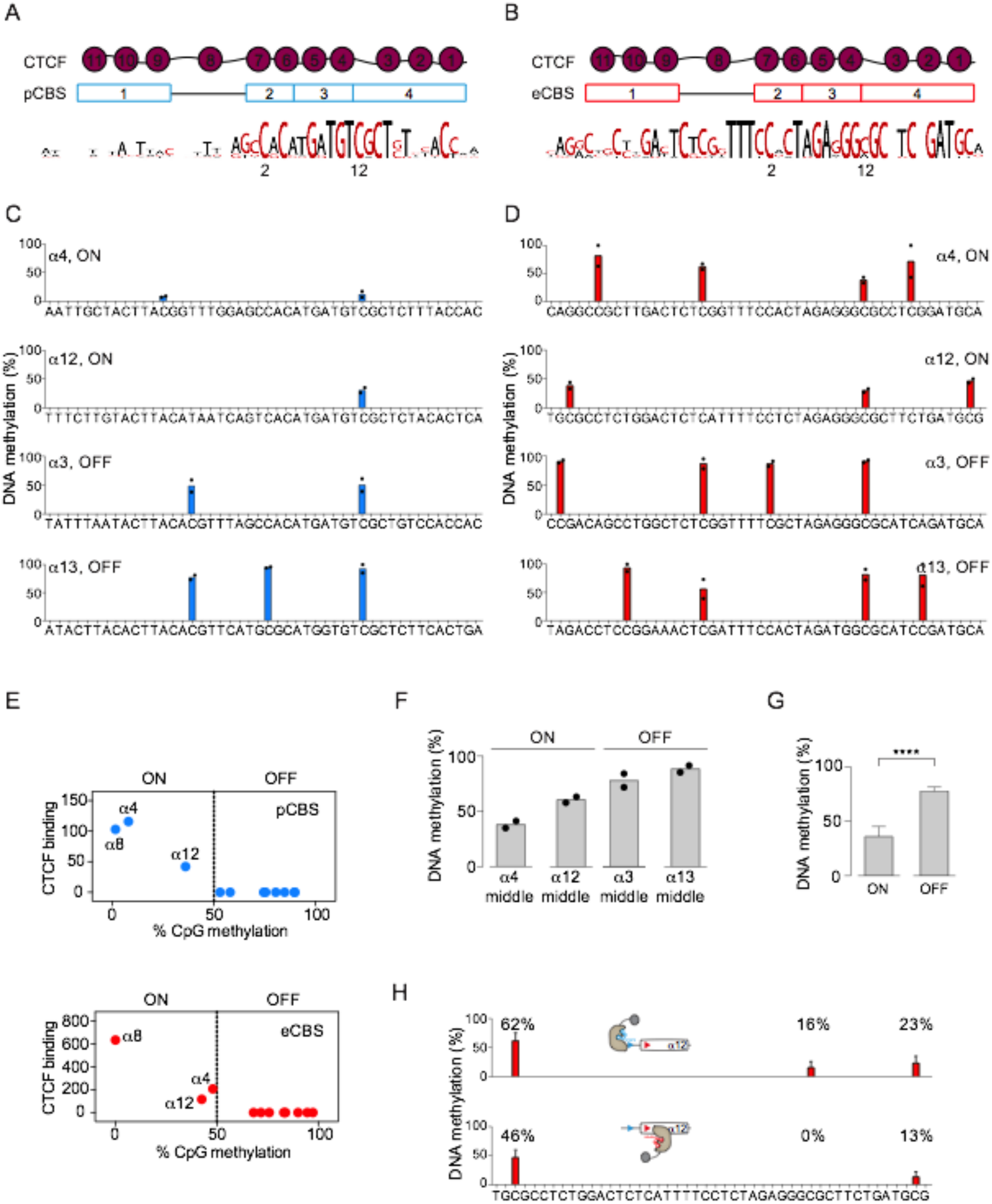

A
CTCF
11 10 9 8 7 6 5 4 3 2 1
pCBS
1 2 3 4
2
12
B
CTCF
11 10 9 8 7 6 5 4 3 2 1
eCBS
1 2 3 4
2
12
C
DNA methylation (%)
100
50
0
α4, ON
AATTGCTACTTACGGTTTGGAGCCACATGATGTCGCTCTTTACCAC
α12, ON
TTTCTTGTACTTACATAATCAGTCACATGATGTCGCTCTACACTCA
α3, OFF
TATTTAATACTTACACGTTTAGCCACATGATGTCGCTGTCCACCAC
α13, OFF
ATACTTACACTTACACGTTCATGCGCATGGTGTCGCTCTTCACTGA
D
DNA methylation (%)
100
50
0
α4, ON
CAGGCCGCTTGACTCTCGGTTTCCACTAGAGGGCGCCTCGGATGCA
α12, ON
TGCGCCTCTGGACTCTCATTTTCCTCTAGAGGGCGCTTCTGATGCG
α3, OFF
CCGACAGCCTGGCTCTCGGTTTTCGCTAGAGGGCGCATCAGATGCA
α13, OFF
TAGACCTCCGGAAACTCGATTTCCACTAGATGGCGCATCCGATGCA
E
ON
OFF
pCBS
CTCF binding
150
100
50
0
α4
α8
α12
0
50
100
% CpG methylation
ON
OFF
eCBS
CTCF binding
800
600
400
200
0
α8
α4
α12
0
50
100
% CpG methylation
F
ON
OFF
DNA methylation (%)
100
50
0
α4 middle
α12 middle
α3 middle
α13 middle
G
****
DNA methylation (%)
100
50
0
ON
OFF
H
DNA methylation (%)
100
50
0
62%
α12
16%
23%
46%
α12
0%
13%
TGCGCCTCTGGACTCTCATTTTCCTCTAGAGGGCGCTTCTGATGCG

Figure S6

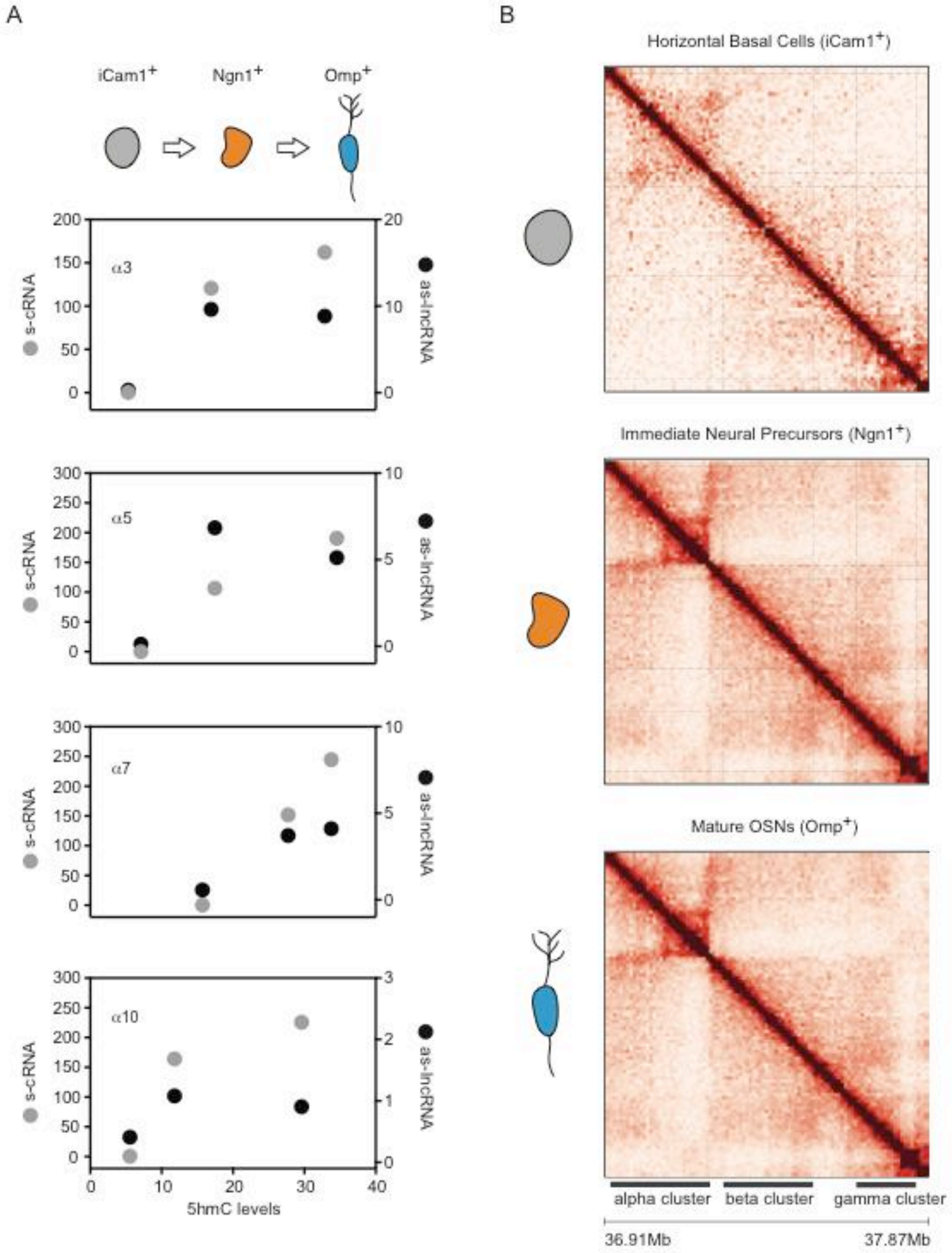
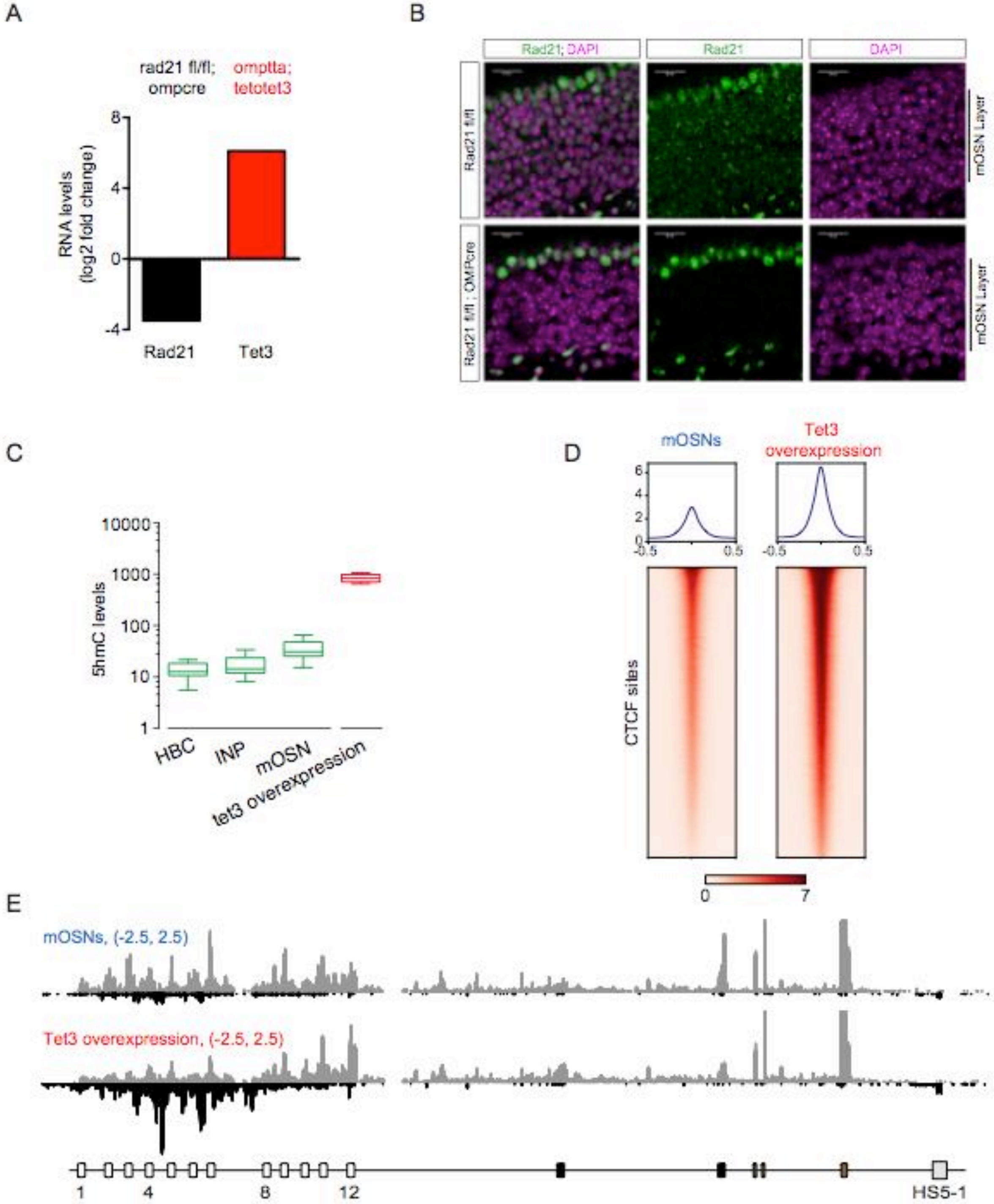

A
iCam1+
Ngn1+
Omp+
s-cRNA
as-lncRNA
α3
0
50
100
150
200
0
10
20
α5
0
50
100
150
200
250
300
0
5
10
α7
0
50
100
150
200
250
300
0
5
10
α10
0
50
100
150
200
250
300
0
1
2
3
0
10
20
30
40
5hmC levels
B
Horizontal Basal Cells (iCam1+)
Immediate Neural Precursors (Ngn1+)
Mature OSNs (Omp+)
alpha cluster
beta cluster
gamma cluster
36.91Mb
37.87Mb

Figure S7


A
rad21 fl/fl;
ompcre
omptta;
tetotet3
RNA levels
(log2 fold change)
8
4
0
-4
Rad21
Tet3
B
Rad21; DAPI
Rad21
DAPI
Rad21 fl/fl
Rad21 fl/fl ; OMPcre
mOSN Layer
mOSN Layer
C
5hmC levels
10000
1000
100
10
1
HBC
INP
mOSN
tet3 overexpression
D
mOSNs
Tet3
overexpression
6
4
2
0
-0.5
0.5
-0.5
0.5
CTCF sites
0
7
E
mOSNs, (-2.5, 2.5)
Tet3 overexpression, (-2.5, 2.5)
1
4
8
12
HS5-1