Scuola Internazionale Superiore di Studi Avanzati
PhD course in Functional and Structural Genomics

The scent of genome complexity: exploring genomic instability in mouse Olfactory Epithelium

Thesis submitted for the degree of *“Philosophiae Doctor”*

Candidate:
Alice Urzì

Supervisor:
Prof. Stefano Gustincich
Co-Supervisor:
Dr. Remo Sanges

Academic Year 2015/2016

This page intentionally left blank

This page intentionally left blank

Abstract

In the olfactory epithelium (OE) the detection of volatile compounds (odors) is accomplished by a large family of olfactory receptors (ORs), located on the surface of the cilia of olfactory sensory neurons (OSNs). These represent the major sensory component of the OE and reside in the nasal cavity. The extraordinary chemical diversity of olfactory ligands is matched in the mouse genome by a collection of more than 1200 mouse and 350 human active OR genes encoding for G-protein-coupled receptors (GPCRs).

Each mature OSN in the OE is thought to express only one allele of a single OR gene (monoallelic and monogenic expression). A given OR gene is expressed in a mosaic or punctate pattern of OSNs within a characteristic zone of the OE.

The transcriptional mechanisms that underlie this extraordinarily tight regulation of gene expression remain unclear.

I hypothesize that OR expression choice can be influenced by somatic LINE-1-associated genomic variations. Indeed, it is now well established that active LINE-1s can create genomic rearrangements at insertional and post-insertional stages. Besides promoting genome plasticity and diversification during evolution, somatic variations can contribute to gene expression regulation for those genes that are characterized by a stochastic and monoallelic expression.

Under this hypothesis, I expect the genomic sequence around the expressed ORs to be different with respect to that around the same ORs in non-expressing cells, for the presence of variations able to activate chromatin and promote ORs transcription.

I first showed high LINE-1 expression and retrotransposition in OE. Then I investigated the presence and involvement of LINE-1-associated variations with OR expression, comparing the genomic sequence around an active and an inactive OR locus.

In particular, I analyzed a genomic region of 50 kb around the *Olfr2* TSS taking advantage of a GFP knock-in mouse. In these mice, the OSNs naturally expressing *Olfr2* co-express also GFP. Targeted sequencing of *Olfr2* locus revealed hundreds of heterozygous structural variants (insertions, deletions, inversions and duplications) in the vicinity of the locus. Deletions were the most abundant variation category.

By end point PCR I validated six LINE-1 associated deletions potentially involved in *Olfr2* expression. Nevertheless, functional validation experiments *in vivo* will be performed to prove their effective role in *Olfr2* choice.

Looking at the putative mechanisms supporting the deletions, I started investigating a possible involvement of DSBs. With this aim, I performed a chromatin immunoprecipitation and sequencing (ChIP-Seq) analysis for endogenous γ-H2AX (an early response marker for DNA-DSBs) in mouse OE and liver. I performed a general characterization of endogenous γ-H2AX in normal tissues. In both tissues analyzed, γ-H2AX signal was not randomly distributed in the genome but preferentially localized within transcribed and regulatory regions.

Overall, γ-H2AX peaks were depleted in the OR clusters. Interestingly, an exception was given by a peak located within the *Olfr2* locus, in close proximity to two validated deleted regions.

Abbreviations

CDS	Coding Sequence
Cer	Cerebellum
ChIP-Seq	Chromatin ImmunoPrecipitation sequencing
CNV	Copy Number Variation
DRS	Direct Repeat Site
DSB	Double Strand Break
EN	Endonuclease
GBC	Global Basal cell
gDNA	Genomic DNA
GFP	Green fluorescent protein
γH2AX	Phosphorylated-H2AX
GPCR	G-protein coupled receptor
HBC	Horizontal Basal cell
H	Hippocampus
HR	Homologous recombination
IHC	Immunohistochemestry
IP	Immunoprecipitation
IR	Inverted repeat
IRES	Internal-ribosomal entry site
ITR	Inverted terminal repeat
K	Kidney
L	Liver
LCM	Laser capture microdissector
LCR	Locus control region
LINE-1	Long interspersed nuclear element-1
LSD1	Lysin demethylase 1
LTR	Long terminal repeat
MDA	Multiple Displacement Amplification
NAHR	Non-Allelic Homologous Recombination
NanoCAGE	Cap analysis of gene expression

NHEJ	Non-homologous end joining
NPC	Neural precursor cell
O/N	Over night
OB	Olfactory bulb
OE	Olfactory epithelium
Olfr2	Olfactory receptor 2
OR	Olfactory Receptor
ORF	Open Reading Frame
OSN	Olfactory Sensory Neuron
pA	Poly-A tail
PB	Pac Bio
PCR	Polymerase chain reaction
Poll II	RNA Polymerase II
RNP	Ribonucleoprotein
RT	Retrotranscriptase
RT	Room Temperature
RT-qPCR	Real Time quantitative PCR
SC	Supporting Cell
SINE	Short interspersed nuclear element
SSA	Single strand annealing
SV	Structural variation
TE	Transposable element
Topo II	Topoisomarase II
TPRT	Target primed reverse transcription
TSD	Target site duplication
TSS	Transcription start site
ub-H2AX	Ubiquitinatinated-H2AX
UPR	Unfolded protein response
UTR	Untranslated region
VNO	Vomeronasal organ
VR	Vomeronasal receptor
WB	Western blot
WGA	Whole genome amplification

1 Introduction

Here I present the main biological concepts at the basis of this work.

In the first section I give a general overview of the mouse olfactory system, focusing on the problem of OR transcriptional regulation in mouse OE.

In the second section I describe the general features of transposable elements with particular attention to the effects of LINE-1 retrotransposition on the genome stability.

In the third section I describe how transposable elements can modulate gene expression.

Finally, in the fourth section I introduce the relationship between transposable elements and DNA-double strand breaks repair systems.

1.1 The olfactory system

One of the most fascinating questions in neurobiology is how myriads of external stimuli can be rapidly perceived, processed by the brain and finally drive mammalian behavior. The detection of an external stimulus is accomplished by different sensory systems, among which the one responsible for the sense of smell is still functionally poorly characterized. In this work we focused our investigation on the olfactory system of mice, since they are a well-characterized and representative experimental model to study molecular mechanisms involved in mammalian olfaction.

Even though olfaction in mice has a principal role, the belief that olfaction is a secondary sense in human, compared to other species that rely on smell to detect food, predators and mates, is overly simplistic. Emerging evidence, like the discovery that hundreds of thousands of distinct odors can be discriminated by human nose and the discovery of the early impairment of olfaction in many neurodegenerative diseases (Barresi et al. 2012), highlighted the importance of studying this chemosensory system even in humans.

During the last thirty years the development of new technologies allowed to start answering some fundamental questions about olfaction, from the molecular identity of ORs to the interpretation of how olfactory information is coded in the brain (Nagai, Armelin-Correa, and Malnic 2016). While these discoveries represent a solid starting point, there are still open questions about olfaction. One of these regards the molecular mechanisms underlying OR expression.

1.1.1 Mouse Olfactory Epithelium (OE)

In mouse, the olfactory system consists of two main organs: the main OE and the olfactory bulb (OB). Olfactory perception begins when odorous ligands activate ORs expressed in olfactory sensory neurons (OSNs) of the OE. The OE is located in the posterior upper part of the nasal cavity where it is arranged over multiple cartilaginous structures called turbinates (Figure 1.1). It is composed by at least four different cell populations and each of these cell types occupies a specific position in the neuroepithelium (Figure 1.2).

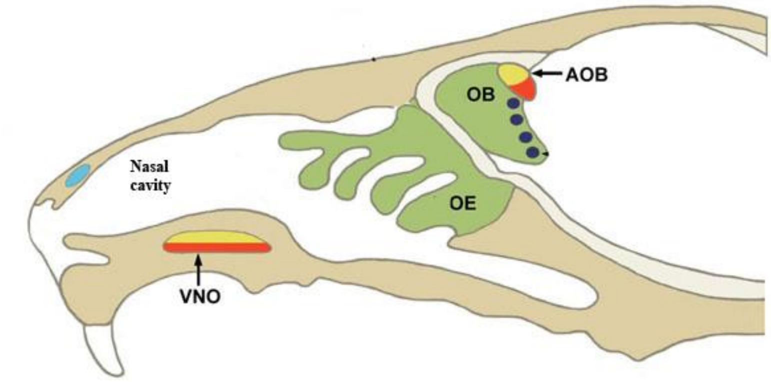


Figure 1.1. Mouse olfactory system. Sagittal cross section through the nasal region of the head: main olfactory epithelium (OE) and main olfactory bulb (OB) are shown in green, and the two layers of the vomeronasal organ (VNO) and corresponding connection regions in the accessory olfactory bulb (AOB) are shown in yellow and red (Brennan and Zufall 2006)

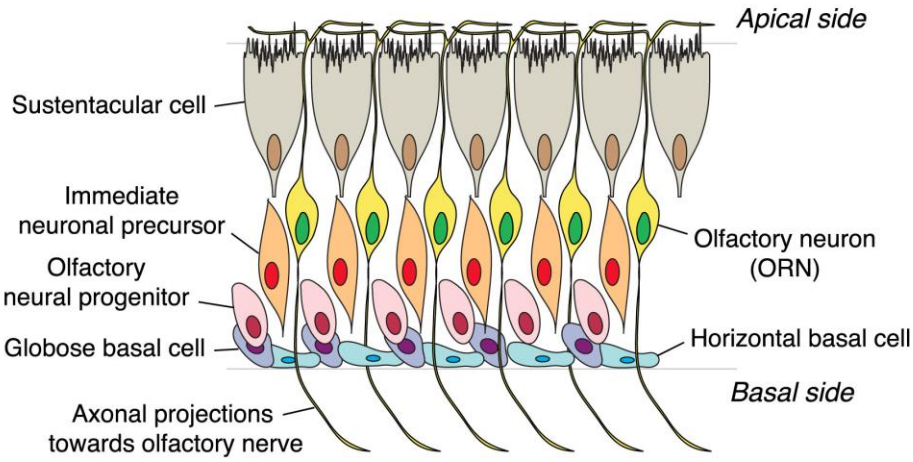


Figure 1.2. Cellular architecture of OE. A schematic representation of mouse OE is shown, the apical layer is populated by sustentacular cells; the intermediate layer is composed by mature and immature olfactory sensory neurons and the basal layer is characterized by the presence of globose and horizontal basal cells, olfactory neural progenitors and immediate neuronal precursors.

The basal layer of the OE is composed by globose basal cells (GBCs), horizontal basal cells (HBCs) and olfactory neural progenitors. GBCs and HBCs have been characterized as multipotent stem cells, able to generate OSN progenitors during both embryonic and adult neurogenesis (Jessica H. Brann and Firestein 2014).

OSNs, which occupy the middle layer of the OE, are bipolar neurons projecting a single dendrite with a thickened ending (the olfactory knob) that extends to the epithelial surface. The olfactory knob contains non-motile sensory cilia where ORs are expressed to detect odors. A unique axon penetrates the skull through the cribriform plate and synapses in the olfactory bulb. OSNs have a short average life-time (about 70 days) and are replaced periodically throughout the lifespan of an individual (Sultan-Styne et al. 2009).

Sustentacular cells (SCs) compose the apical layer of OE that is in contact with the air flow circulating through the nasal cavities. SCs are functionally similar to glial cells in the central nervous system.

1.1.2 OE neurogenesis

Different works have demonstrated that neurogenesis is not restricted to embryonic development, but persists in specific areas into adulthood (Graziadei and Graziadei 1979; Altman 1962).

There are three main neurogenic areas in the nervous system: the subgranular zone, which generates new granule cells to the dentate gyrus of the hippocampus; the subventricular zone (SVZ), which provides new interneurons to the OB; and the OE, which supplies new OSNs that innervate OB. In the OE, the renewal ability is likely a protective mechanism to preserve the sense of smell over time, given that the OE is directly and continuously exposed to insults coming from the outer environment (Hurtt et al. 1988).

Embryonic origin of OE

In OE the first signs of cellular differentiation can be observed at embryonic day 10 (E10), when first embryonic stem cells and developing OSNs are detected (Murdoch and Roskams 2008). OSN dendrites are observed at E11 when the first OE and OB connections are reported. Indeed, OE-OB synaptogenesis takes place between E11 and

E15. Interestingly, the first OR expression predates OE-OB synaptogenesis, since it has been reported as early as E11 (López-Mascaraque and de Castro 2002).

Adult neurogenesis

GBs and HBCs are responsible for generating OSNs during the late embryonic to postnatal stages. Stem cell populations generate Ascl1+ progenitors which in turn give rise to Neurogenin-1 and NeuroD1+ immediate neuronal precursors (INPs) (Packard et al. 2011; Suárez, García-González, and de Castro 2012). INPs differentiate in GAP-43+ immature OSNs that finally reach maturity once expressing olfactory marker protein (OMP) (Figure 1.3). An essential requirement for OSNs maturation is the stable expression of ORs. Indeed, ORs are expressed not only on the cilia of OSNs, where they are involved in detecting odors, and at the axon termini, where they direct the neuronal innervation pathway towards the OB.

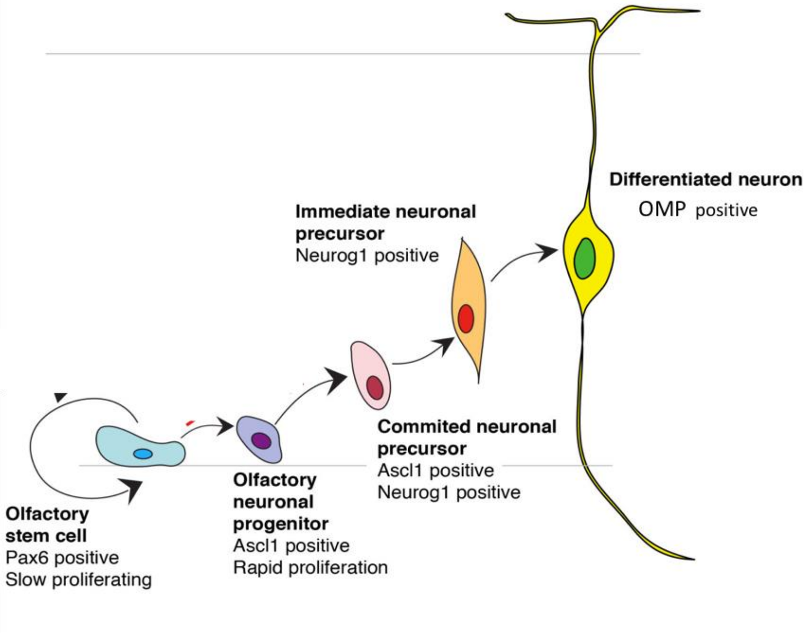


Figure 1.3. Schematic representation of OSNs differentiation. OSN differentiation process is shown. For each cell type, peculiar expressed genes are shown.

1.1.3 OR genes: structure and distribution in the genome

In 1991 Linda Buck and Richard Axel identified in rats an extremely large multigene family of transmembrane proteins that were hypothesized to be ORs on the basis of their typical expression pattern, restricted to the OE. Further studies confirmed that OR genes form the largest multigene family ever found in vertebrates (X. Zhang and Firestein 2002).

Each OR gene includes an intronless coding region of about 1 kb, preceded by upstream regulatory exons and terminating with a polyadenylation signal. The number of 5’ non-coding exons ranges between 1 and 4.

In mouse there are about 1200 OR genes, organized into 27 gene clusters dispersed throughout the genome and represented on all chromosomes except chromosomes 12 and Y. 20% of the total mouse OR genes are pseudogenes, which do not code for functional OR proteins. In humans there are 950 ORs, distributed in 100 locations throughout the genome, and 60% of them are pseudogenes (X. Zhang and Firestein 2002). This size difference between mouse and human repertories can be due to human development of trichromatic color vision that is very powerful in detecting environmental signals. It is likely that olfaction has become less important for primates confronted to other dichromatic mammalian species.

1.1.4 OR gene choice: monogenic and monoallelic expression

Each mature OSN expresses only one OR gene from the entire genome repertoire, according to a peculiar genomic feature which can be summarized as “one-receptor-one-neuron rule”. Moreover, in a given OSN a specific OR is expressed only from a single allele, and OSNs expressing the maternal or paternal allele coexist mosaically in the OE.

Expression of one OR gene per OSN does not mean that each OR gene is expressed in the same number of neurons. Conversely, counts of OSNs expressing a given OR gene in mouse or rat give variable results from gene to gene (Mombaerts 2004; Rodriguez-Gil et al. 2010). Moreover, unequal tissue RNA levels across the OR repertoire can vary by almost 300-fold (Young et al. 2008).

Interestingly, once selected by the OSN, the same OR is expressed both on the dendritic cilia and on the axon termini of the neuron. Axons from different OSNs that express the same OR gene coalesce into one or a few glomeruli of the OB. Therefore, each selected OR is involved in sensing odors and a main actor in directing OSN innervation pattern towards the OB. Therefore, a stable OR expression is essential for both the differentiation and integration of the mature OSN into the functional olfactory network.

The OR choice is supposed to occur at the level of the olfactory neural progenitor cells which then differentiate, first in an immature neuron and finally in a mature OSNs which innervates the OB (Figure 1.3).

So far the molecular mechanisms underlying the OR transcriptional regulation are almost completely unknown. Many hypotheses and models have been proposed in the last twenty years to explain how each OSN selects the right OR allele.

As already mentioned, OR expression has been reported as early as E11, before OE-OB synaptogenesis. Therefore, even at the earliest embryonic ages, OSNs expressing the same OR have restricted zonal/regional expression patterns, suggesting that retrograde signals from the OB do not influence OR gene choice in the OE (López-Mascaraque and de Castro 2002).

It is well accepted that different elements and events are involved in the selection of a given OR, including the gene coding sequence, short DNA sequences upstream of OR coding sequences, locus control regions (LCR)-like conserved elements, and feedback signals given by the expression of a full-length OR protein (Serizawa et al. 2003).

Here, we summarize the state of the art about the molecular mechanisms involved in ORs transcriptional regulation.

Different DNA sequences were described acting as proximal and distal OR regulatory sites. Homeodomain and O/E-like binding sites were identified in the proximal upstream regions of some ORs (Hirota and Mombaerts, 2004) (Rothman et al. 2005).

Additional sequences, called P and H elements, were described to act as distal enhancer regulating the probability of OR gene choice differentially across their cluster (Khan, Vaes, and Mombaerts 2011).

Moreover, the chromatin state has been demonstrated to have central role in OR regulation. In particular, singular OR expression can be initiated by lysine demethylase 1 (LSD1), which catalyzes an epigenetic switch that allows localization of the active OR allele to a specific nuclear subdomain distinct from inactive ORs (Clowney et al. 2012) (Lyons et al. 2013) (Tan, Zong, and Xie 2013).

Recently, Dalton and colleagues hypothesized that, following the OR gene choice, the unfolded protein response (UPR) could serve as a molecular signal that triggers a negative feedback to down regulate LSD1(Dalton, Lyons, and Lomvardas 2013). Therefore, the so-called “one OSN-one OR rule” is enforced by a negative feedback signal that requires functional OR expression.

Historically, one of the major limitations to the understanding of transcriptional regulation for OR genes has been the lack of a clear characterization of OR promoters and transcription start sites (TSSs). In 2012 a NanoCAGE (a cap analysis of gene expression protocol adapted for low amounts of input RNA) experiment on RNA from mouse OE identified TSSs of 87% of the OR genes giving an important contribution to reveal genomic features which can participate to OR expression choice (Plessy et al. 2012).

Nevertheless, the molecular mechanisms able to determine which OR allele is chosen, among the 2400 different ones available, remain unclear.

In this work we hypothesized that transposable elements could be involved in the choice of the OR gene, in particular LINE-1-associated structural variations in the locus of active OR genes could have a role in determining the receptor expression.

1.2 Transposable elements

Transposable elements (TEs) comprise a multitude of repetitive DNA sequences with the ability to mobilize and change locations in the genome (Kazazian 2004). They were first discovered in maize plants in the mid-1940 but they are generally present in the genomes of prokaryotes and eukaryotes. Initially they were considered *junk DNA* and *selfish DNA parasites* due to their lack of a clear functional role and ability to replicate independently and therefore being a potential threat towards genomic integrity (Fedoroff 2012).

Given the discovery that TEs occupy around 40% of mouse genome (Bourc’his and Bestor 2004) (Walsh, Chaillet, and Bestor 1998) (Mouse Genome Sequencing Consortium et al. 2002), the concept of junk DNA is not acceptable and the idea has shifted towards understanding their functional role. Even if the modern view of TEs is still controversial it is now evident that they still participate in shaping genomes, influencing gene expression and contributing to tissue-specific transcriptional programs, in particular as enhancer-like elements and regulator of chromatin structure (John L. Goodier and Kazazian 2008) (Bodega and Orlando 2014).

According to their mechanism of mobilization, TEs can be divided in two main groups: class II TEs, or DNA transposons, that move throughout the genome using a “cut and paste” mechanism, and class I TEs, or retrotransposons, that multiply themselves in the genome via a RNA intermediate, with a “copy and paste” mechanism (Figure 1.4) (Wicker et al. 2007).

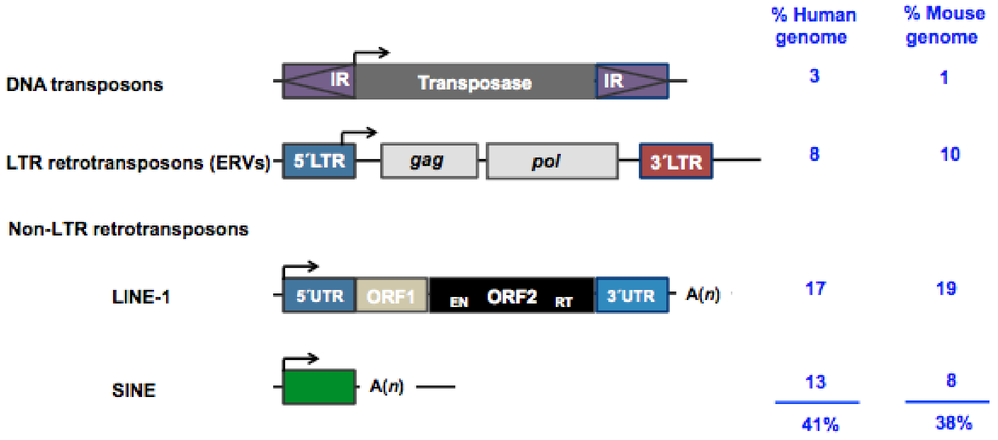


Figure 1.4. Classes of TEs found in mammalian genomes. Different types of TEs (DNA transposons; ERV, endogenous retrovirus; LINE-1, long interspersed element class 1; SINE, short interspersed element) are shown. Fractions of the human and mouse genomes occupied by each TE type is represented as percentage on the right. IR, inverted repeat; UTR, untranslated region; EN, endonuclease; RT, reverse transcriptase; LTR, long terminal repeat; ORF, open reading frame (Garcia-Perez, Widmann, and Adams 2016).

1.2.1 DNA transposons

During mobilization DNA transposons are first excised from their original location as double-stranded DNA and then reinserted elsewhere in the genome. DNA transposons structure is characterized by a transposase encoding gene included between the inverted terminal repeats (ITRs) (Figure 1.4). The autonomously encoded transposase is able to catalyze both the excision and integration of the repetitive element: it recognizes ITRs and binds them, then it “cuts” the DNA transposon and “pastes” it into a new genomic site, giving rise to a non-replicative mechanism where the DNA transposon is moved to a new location but no transposon copies are generated. During the insertion, the target site DNA is duplicated at both ends of the transposable element, forming the so called target site duplications (TSD), which are unique for each different DNA transposon (Munoz-Lopez and Garcia-Perez 2010).

The classification of DNA transposons is commonly based on the sequence, the ITRs and /or the TSDs. Among the subclass I there are: Tc1/mariner, PIF/Harbinger, hAT, Mutator, Merlin, Transib, P, piggyBac and CACTA. In the Subclass II there are Helitron and Maverick transposons. The most widespread TE family in nature is the Tc1/mariner that is present in diverse taxa as rotifers, fungi, plants, fish and mammal (Lopez and Garcìa, 2010). Currently there are no active DNA transposons in mammals, and recent computational analyses indicated that their activity ceased in the primate lineage at least 37 million years ago (Pace and Feschotte 2007). At present they comprise approximately the 3% of the human reference genome (Beck et al. 2011) and the 4% of the mouse genome (Keane et al. 2014).

1.2.2 Retrotransposons

Retrotransposons constitute 42% of the genome and duplicate via a RNA intermediate using a “copy-and-paste” mechanism. Retrotransposons are first transcribed into RNA, then retrotranscribed and inserted in a new genomic location. With this mechanism, a second novel insertion is created, while the original transposon is preserved.

Depending on whether they are able to encode the transposition machinery or not, retrotransposons can be classified in autonomous and non autonomous. Autonomous retrotransposons encode for the proteins necessary to their own retrotransposition.

Non-autonomous elements include processed pseudogenes and short interspersed nuclear elements (SINE), which overall constitute 13% of human genome. In order to be mobilized they need retrotransposition proteins to be encoded by the autonomous elements.

Among autonomous elements a further classification can be made into elements which possess Long Terminal Repeats (LTR) and elements without LTRs (non-LTR retrotransposons). The most prominent members of the first category are the endogenous retroviruses (ERVs) which comprise approximately 8% of the genome and the mouse intracisternal A-particles (IAPs).

Non-LTR retrotransposons account for 21% of the genome and comprise Long Interspersed Nuclear Elements (LINE), which can be further divided into three subfamilies LINE1, LINE2 and LINE3. Among these three, only LINE1 or LINE-1 elements are currently active in humans and mice (Lee et al. 2012). For this reason, we will focus only on this subfamily in the following paragraphs.

Long interspersed nuclear elements (LINE-1)

LINE-1s are highly abundant and constitute approximately 17% of the human genome and 19% of the mouse genomes, respectively (Mandal and Kazazian, 2008).

Most LINE-1s are retrotransposition-defective due to 5’ truncation, internal rearrangements or point mutations that disrupt their open reading frames. Among the 5000 LINE-1 full length copies present in the human genome, roughly 80-100 elements contain two intact open reading frames and are considered retrotransposition-competent (Beck et al., 2011).

By comparison, the mouse genome is estimated to contain about 11,000 LINE-1 full length copies, among which at least 3000 are still active (J. L. Goodier 2001).

Retrotransposition-competent LINE-1s encode the machinery necessary to support their own replication through an RNA intermediate. They are 6 kb in length and contain a 5’ untranslated region (UTR), two open reading frames (ORF1 and ORF2) and a 3’UTR followed by a polyadenilation signal (AATAAAA) (Figure 1.5).

The 5’UTR region contains an internal RNA polymerase II promoter and several binding sites for regulatory elements such as a CpG island, two SRY box binding domains, a YY-box transcriptional initiation start site and a RUNX binding site. Antisense promoter activity (ASP) located at the LINE-1 5’UTR has been extensively demonstrated (Speek 2001)

The ORF1 and ORF2 encode for proteins necessary for retrotransposition activity. ORF1 encodes a 40kDa protein (ORF1p) with RNA-binding activity. It contains a highly-conserved C-terminal region, an RNA binding motif and a less-conserved N-terminal α-helical domain. On the other hand, ORF2 encodes a 150 kDa protein (ORF2p) with three conserved domains, an N- terminal endonuclease (EN) domain, a central reverse-transcriptase domain (RT) and a C-terminal zinc knuckle-domain. The LINE-1 endonuclease (EN) domain is responsible for the DNA double strand break (DSB) at the insertion target site, whereas the RT activity generates the cDNA copy of LINE-1 to be inserted into the new genomic location (Beck et al., 2011) (Ostertag and Kazazian 2001).

Human and mice share the same structure of the LINE-1 gene, apart from the promoter region at the 5’UTR. Indeed, the 5’ UTR region of rodents is characterized by the presence of repeated conserved monomers of about 200bp followed by a short non monomeric region. The variability at the 5’UTR region can reside in differences in both number and sequence of the monomers.

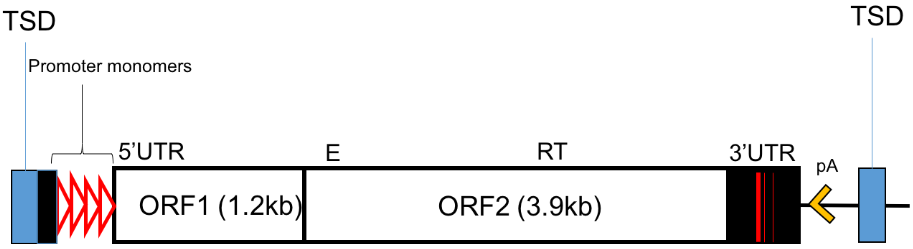


Figure 1.5. Structure of the mouse LINE-1. The 5'UTR region is composed of a variable number of monomers (~ 200 bp, red triangles) having promoter activity. Polymorphisms are present at the 3'UTR, here indicated with red vertical axis. ORF1 and ORF2 open reading frame are indicated. E =endonuclease domain; RT= retrotranscriptase domain; pA= polyadenylation signal; TSD= target site duplication. Adapted from (Mears and Hutchison 2001).

Depending on the 5'UTR monomeric organization murine LINE1s are divided into different sub-classes. Phylogenetically all mouse LINE-1 elements seem to derive from a common ancestor which has spawned several subfamilies differing in their 5' end region: *V*, *A*, *Tf* and *Gf* subtypes (J. L. Goodier 2001)

The *V* family has no identifiable monomers and is supposed to be the oldest and inactive subfamily.

The *A* family members have a monomeric structure at their 5' region and contain about 6500 full length elements. A subset (900) of *A* elements is supposed to be retrotransposition-competent (or active) because they contain intact ORF1 and ORF2 and are transcribed, although their retrotransposition capacity has never been directly assayed.

The *Tf* and *Gf* families, derived from a common ancestor (*F*-family), have different monomeric structure at the 5'UTR and contain a large number of transposable elements. The *Tf* type has 1800 active elements among 3000 full length members, whereas the most recently discovered *Gf* type includes 400 active elements among 1500 full length members (Goodier et al., 2001). *In vitro* experiments which tested the activity of various regions of the mouse *Tf* 5'UTR have revealed that the promoter activity lies within the monomers and therefore the promoter activity seems to be enhanced proportionally to the number of monomers (Ostertag and Kazazian, 2001).

1.2.3 Molecular mechanism of LINE-1 retrotransposition

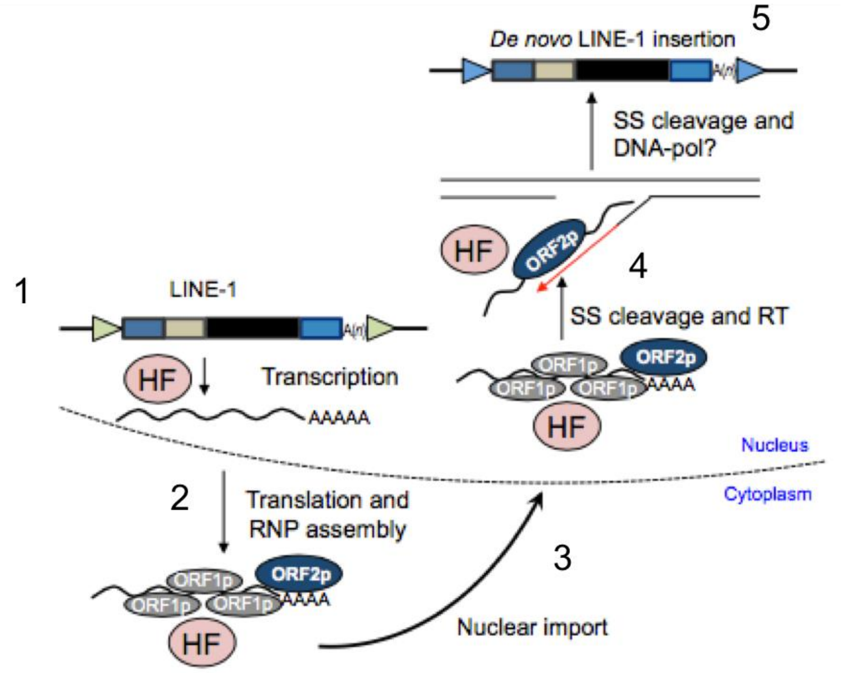


Figure 1.6. LINE-1 retrotransposition cycle. The full-length active LINE-1 RNA is first transcribed in the nucleusby cell host factors (HF) (1) and transported to the cytoplasm where ORF1p and ORF2p are translated (2). These proteins preferentially bind *in cis* to their encoding mRNA, generating a ribonucleoprotein particle (RNP), which is shuffled back into the nucleus (3). The EN activity of ORF2 protein generates a single strand (SS) break in genomic DNA that is used by the ORF2-RT activity to generate the first-strand cDNA (4) (red arrow). How second strand synthesis occurs is not well understood. New LINE-1 insertions are often flanked by TSDs (blue or green arrowheads) and are also often 5′-truncated (not shown). HF, host factor involved in the retrotransposition process. (Adapted from Garcìa et al., 2016).

During retrotransposition, the LINE-1 sequence is transcribed by RNA polymerase II starting from its own internal promoter, leading to the generation of a bicistronic mRNA (Figure 1.6-step1). This mRNA molecule has a polyA tail that can be encoded by its own weak but functional polyadenylation signal, or by a signal present in the downstream genomic sequence, in this case leading to the so called LINE-1-mediated 3’- transduction. It is still not known whether a 7-methylguanosine cap is added to LINE-1 transcripts, while it is clear they do not contain introns (Ostertag and Kazazian, 2001).

The next step involves the transport of the LINE-1 RNA molecule to the cytoplasm, where the ORF1 and ORF2 sequences are translated into proteins (Figure 1.6-step2). The shuttling mechanism involved in the process is still unknown.

In the cytoplasm, multiple copies of ORF1p and only few copies of ORF2p bind with the LINE-1 RNA molecule, creating a stable ribonucleoprotein complex (RNP) (Figure 1.6-step3) (Kulpa and Moran 2006)(Beck et al., 2011).

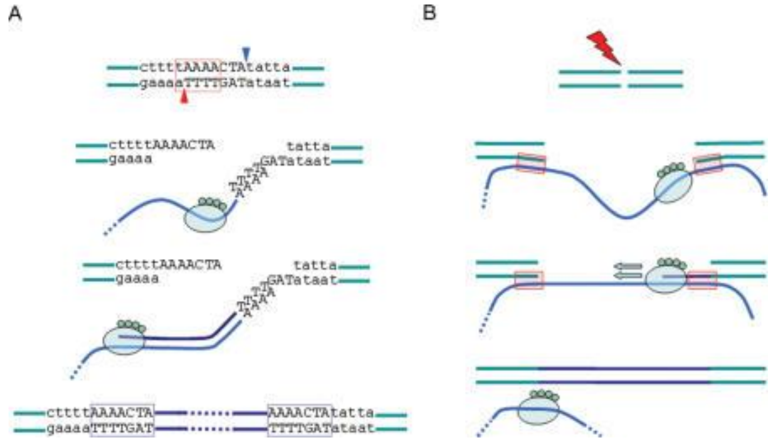


Figure 1.7. Comparison between TPRT and ENi LINE-1 insertions. (A) TPRT-mediated LINE-1 insertion. LINE-1 endonuclease (red arrowhead) cut first strand DNA at the 5'-TTTT/A-3' consensus (red box) and allows LINE-1 mRNA (blue line) to anneal to genomic DNA using its poly(A) tail. Reverse transcriptase activity of LINE-1 ORF2 (oval) synthesizes LINE-1 cDNA (purple line) using LINE-1 mRNA as template and 3'OH from nicked genomic DNA as primer. Second-strand cleavage (blue arrowhead) occurs 7–20 bp downstream from first cleavage site, creating nicks which are repaired to form TSDs (blue dotted boxes). The insertion process is completed by the attachment of the LINE-1 cDNA and synthesis of the second strand. (B) Schematic representation of ENi mechanism. The creation of a genomic double-strand break (red thunderbolt) is followed the attachment of free-floating LINE-1 mRNA (blue line) to newly separated ends using small complementary sequence. Finally the gap may be filled in by DNA synthesis by either the LINE-1 RT, cellular repair polymerases or both (S. K. Sen et al. 2007).

This complex, through a mechanism that has not yet been clarified is carried back into the nucleus (Figure 1.6-step3). Here two different LINE-1 integration mechanisms can be mediated (Figure 1.6-step4-5): the canonical target primed reverse transcription (TPRT) or the endonuclease-independent (ENi) retrotransposition (Viollet, Monot, and Cristofari 2014) (Figure 1.7).

The target primed reverse transcription (TPRT) consists of a coupled reverse transcription/integration process (Morrish et al. 2002). During TPRT, ORF2p endonuclease activity produces a single-strand nick in the genomic DNA preferentially at the consensus sequence 5′-TTAAAA/3′-AATTTT. The ORF2p reverse transcriptase activity, priming the reaction within the polyA tail, extends the free 3′-OH group using the LINE-1 RNA as a template (Viollet, Monot, and Cristofari 2014). After that, the second strand at the integration site is cleaved and used to prime the synthesis of the cDNA second strand. The typical hallmarks of this TPRT-derived integration mechanism include the target site duplications (TSDs), 7-20 bp sequences present at each end of the new LINE-1 element, and a dA-rich tail of variable length (Cordaux and Batzer 2009). Another feature of this integration process is the fact that the majority of the newly inserted LINE-1 elements are 5’ truncated, and therefore unable to retrotranspose any longer. This truncation may be caused by an inefficiency of the reverse transcriptase in the polymerization process of the new cDNA copy, or by the activity of a cellular RNAse H, reflecting a possible attempt by the host defense machinery to protect the genomic integrity (Ostertag and Kazazian, 2001; Beck et al., 2011).

In the endonuclease-independent (ENi) retrotransposition process, at the level of a pre- existing double-strand break, LINE-1 mRNA molecules can attach to the protruding ends using small stretches of complementary bases without the need for a further endonuclease cleavage. At this point the LINE-1 reverse transcriptase, the host repair polymerase or both synthetize the missing DNA bases, leading to a LINE-1 integration that lacks the structural features of the TPRT-mediated insertion. The typical hallmarks of this alternative process are unusual structures caused by LINE-1 integration at atypical target sequences, LINE-1 truncations predominantly at the 3′ ends and lack of TSDs (Morrish et al., 2002). LINE-1 integration mediated by the ENi mechanism has been observed at the level of telomeres although it is a significantly less efficient process, rarely found in vivo (Babushok 2005).

1.2.4 LINE1 mobilization effects

Host genomes use different mechanisms to minimize potential consequences of transposons mobilization among which DNA methylation is one example. Nevertheless, LINE-1 retrotransposition events are able to remodel the structure of the genome with different mechanisms.

Insertional mutagenesis

This is the first described genomic modification induced by LINE-1 insertions. In this mechanism, LINE-1 element inserts in the exon of a gene, inducing an interruption of the coding sequence (Kazazian et al., 1998).

3’ and 5’transduction

As previously mentioned, the polyadenylation signal present at the 3’UTR of the LINE-1 element is functional but weak, and therefore often substituted by downstream stronger signals. This mechanism, the so called “3’ transduction”, causes the transcription, and possibly the retrotransposition of a segment of genomic sequence present at the 3’ of the LINE-1 element. 3’ transduction events were reported in mouse and human, where 15 out of 66 uncharacterized LINE-1 sequences were demonstrated to carry 3’genomic sequences with an average length of 207 bp (J. L. Goodier 2000). Usually shorter and less common is the 5’ transduction that can be identified only at the level of full length LINE-1 elements: in this case the transduction occurs when the LINE-1 element is transcribed starting from an upstream promoter, causing the retrotransposition of a segment of genome upstream to the LINE-1 5’UTR (Beck et al., 2011).

Heterochromatinization

In 1998, after the description of chromosome X inactivation (also called lyonization process), it was hypothesized for the first time a possible involvement of LINE-1 elements in the heterochromatinization of the X chromosome (Lyon 1998).

Only recently it has been demonstrated that actually LINE-1 elements participate during the process in two steps: silent LINE-1s create a heterochromatic compartment in which genes are recruited, while a subset of active and expressed LINE-1s help in X-chromosome inactivation propagation to those genes that are prone to escape (Chow et al. 2010).

LINE-1-associated genomic deletions

Genomic deletions can be generated by LINE-1s at 2 distinct stages of the retrotransposon life cycle: (*i*) at the time of insertion of the element at a new genomic locus via either classical endonuclease-dependent or non-classical endonuclease-independent retrotransposition, and (*ii*) at a post-insertional stage, by re-combination between LINE-1 elements potentially inserted in the genome for a long time. This latter mechanism is part of the interaction between TEs and DNA-DSB repair systems which will be discuss in detail in paragraph 1.4.1.

Deletion generated at the insertion site

When a new LINE-1 element inserts in the genome according to the canonical endonuclease-dependent TPRT, it can generate either small duplications or small deletions of the target site nucleotides depending on endonuclease cleavage.

Although there is a clear LINE-1 EN consensus “bottom strand” cleavage site (5′-TTAAAA/3′-AATTTT) and variants of the sequence (Jurka 1997), it has been demonstrated that there is little or no target site preference for top strand cleavage. According to this model, top strand cleavages downstream of the initial endonucleolytic nick ultimately will lead to the generation of TSDs. By contrast, top strand cleavages in direct opposition or upstream of the initial endonucleolytic nick will result either in conservative insertions or the generation of small deletions.

Simple modifications of this model can produce additional rearrangements like LINE-1 chimeric elements and large deletions (Gilbert, Lutz-Prigge, and Moran 2002).

In cultured human cells it has been observed that the 10% of the integrations mediated by an engineered LINE-1 element were characterized by genomic loss, as also observed in the human and chimpanzee genomes (Beck et al., 2011).

1.2.5 Where and When do LINE-1s mobilize?

LINE-1 retrotransposition in germline and during embryo development

Where and when LINE-1 elements are expressed and retrotransposed during the life of an organism is a fundamental question to address.

Previously LINE-1 retrotransposition was believed to occur only in the germline during gametogenesis and the accumulation of LINE-1 elements in the genome during evolution was considered a clear sign of heritable retrotransposition. LINE-1 mobilization during these developmental phases can be linked to the typical pattern of DNA hypomethylation observed in the cells of the germline, fundamental for the epigenetic reprogramming that occurs during germ cell specification (Smallwood and Kelsey 2012). Indeed, it has been demonstrated that germ cell populations of mice lacking de novo methyltransferase 3-like (Dnmt3L) present higher concentrations of LINE-1 transcripts (Bourchis and Bestor, 2004). Recently, the impact of LINE-1 retrotransposition on viability and quality of fetal oocytes in mice or fetal oocyte attrition (FOA), has been described (Malki et al. 2014).

However, this idea is slowly changing since several lines of evidence, collected using different models of LINE-1-transgenic mice and rats, are showing that a burst of LINE-1 retrotransposition events might occur also in the soma, especially during early stages of embryogenesis (Kano et al. 2009). These somatic retrotransposition events are not incorporated into germ cells, so they are not heritable and will not accumulate in the genome of all cells. Clearly, these events provide sources of genomic diversity within distinct somatic cells of an individual, generating somatic mosaicism in a particular tissue (Vitullo et al. 2012).

High concentrations of LINE-1-mRNA was also present in mouse embryos (Kano et al. 2009. It was demonstrated that probably LINE-1 mobilization and integration occur probably more often during embryogenesis than in the germline. In particular, by using an LINE-1 transgenic mouse model, high levels of LINE-1-mRNA expression were detected, both in germ cells and in embryos, particularly at pre-implantation stages and later on at E10.5 (Kano et al. 2009).

Moreover, it has been recently demonstrated that the RT protein encoded by LINE-1 has a fundamental role during the early embryonic development: by incubating mouse zygotes with the non-nucleoside RT inhibitor nevirapine or microinjecting murine zygotes with morpholino-modified antisense oligonucleotides against the LINE-1 5'end region, they observed an arrest of development at the two- and four-cell stage (Pittoggi et al., 2003).

However, the frequency and specific timing of retrotransposition events in early embryos and during gametogenesis remain to be determined.

Interestingly, further evidence demonstrated that LINE-1 retrotransposition is not limited to germ line and embryogenesis but new LINE-1 insertions can occur also in tumors (Moran et al. 1996) (Ostertag 2000) and in the brain (Muotri et al. 2005; Muotri et al. 2010; Coufal et al. 2009; Erwin et al. 2016).

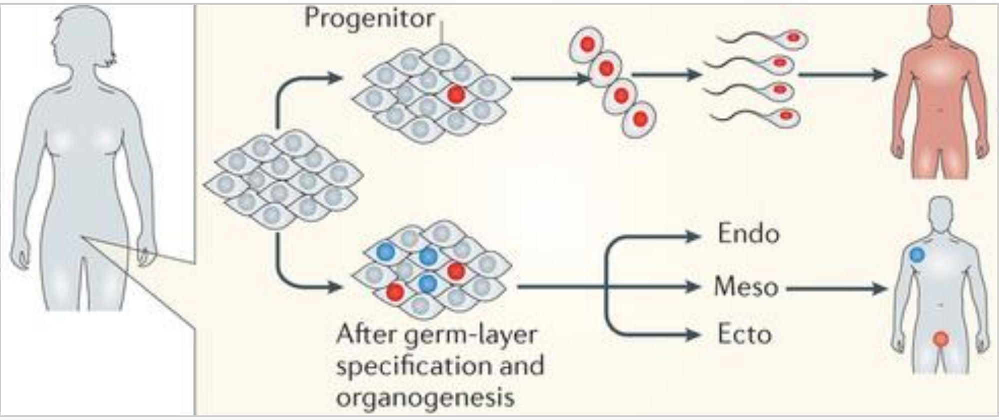


Figure 1.8. Consequences of somatic retrotransposition events during embryogenesis. Somatic retrotransposition can happen at any time during embryogenesis. Retrotransposition events that occur in early pluripotent progenitor cells will result in somatic mosaicism: these unique cells will contribute to all tissues of the body of the individual, including the germ line. Somatic retrotransposition that happens after germ-layer specification and organogenesis, however, results in tissue-specific insertions that are not hereditary. Adapted from (Erwin et al., 2014).

Somatic LINE-1 retrotransposition in the nervous system

Multiple levels of complexity characterize the intricate cellular network in the nervous system. Here, besides cells of different subtypes there are many cells that, although belonging to the same subtype, display different structural and functional features.

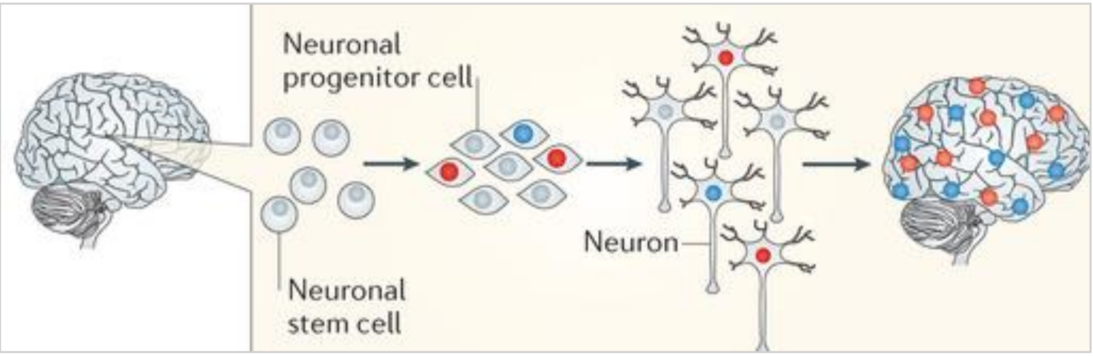


Figure 1.9. Consequences of somatic retrotransposition events in the brain. Blue and red nuclei in represent unique genomes as result of somatic retrotransposition in neuronal progenitor cells. Adapted from (Erwin et al., 2014).

Epigenetic regulation, alternative splicing and post-translational modifications are some of the factors involved in the determination of these diversities. The discovery of neurons with different genotypes, the so called somatic mosaicism, makes the nervous system more complex than ever thought (Figure 1.9) (Erwin, Marchetto, and Gage 2014).

Muotri and colleagues in 2005 demonstrated that both endogenous and engineered LINE-1 elements can retrotranspose in the mammalian brain (Muotri et al., 2005).

LINE-1 transcripts were detected in neural progenitor cells (NPCs) and new LINE-1 insertions can accumulate in NPCs in mouse models of human LINE-1 retrotransposition and in human NPCs in culture (Coufal et al., 2009) (Muotri et al., 2005). This selective expression of LINE-1 in NPCs compared to other tissues seems to be explained by a change in methylation level of LINE-1 promoter in brain cells (Coufal et al; 2009).

Moreover, the transcription factor SOX2 seems to repress LINE-1 transcription in rat adult hippocampal NPCs. Indeed, during neuronal differentiation, the low expression of SOX2 corresponds to a higher LINE-1 transcription and retrotransposition (Muotri et al., 2005).

Recently, single-cell genomics-based studies coupled with next-generation DNA sequencing have allowed researchers to demonstrate that mosaic genomes are a peculiarity of the human brain (Baillie et al. 2011) (Erwin et al. 2016) (Evrony et al. 2012)(Upton et al. 2015), although there is an ongoing debate about the frequency of retrotransposition in this tissue (Evrony et al. 2015)(Richardson et al. 2014).

In human brain LINE-1 retrotransposition has been clearly characterized only in hippocampus (Upton et al., 2015) but little is known about the level of retrotransposition in other brain areas. Recently, Macia and colleagues demonstrated that both LINE-1 expression and retrotransposition can occur even in post-mitotic differentiated neurons (Angela Macia et al. 2016).

1.3 TEs and gene expression

During the last 15 years, several works have been investigating how retrotransposition can modulate gene expression. In particular, great attention has been posed on the nervous system, where LINE-1s have been shown to be able to mobilize and create somatic mosaicism.

1.3.1 LINE-1s function as portable promoters

Newly inserted LINE-1 elements can affect somatic genomes by modulating RNA abundance changing the expression of messenger RNA (mRNA) and/or non-coding RNA (ncRNA). Gene expression levels are modulated differently depending on where, and in which orientation, a LINE-1 element inserts into the gene (Viollet et al., 2014). For example, no effect on transcript levels is reported if LINE-1 elements are inserted into an intron in the antisense orientation, whereas a substantial decrease in transcript abundance is observed when a LINE-1 element is inserted in sense orientation (Viollet et al., 2014).

LINE-1 elements contain both a sense and antisense Pol II promoter in their 5′UTR and ORF1 sequence, respectively, as well as a recently discovered Pol II promoter in their untranslated 3′UTR. Bidirectional transcription from the sense and antisense promoter can produce chimeric transcripts, non-coding RNA, antisense mRNA or double stranded RNA (dsRNA), which can affect gene expression in distinct ways. For example, in vitro experiments on human embryonic stem cells demonstrated that the antisense promoter can be used as alternative promoter in driving the transcription of the genomic sequence upstream to the 5'UTR of the LINE-1 element also in a tissue-specific way, inducing tissue-specific gene expression of specific genes (Mätlik, Redik, and Speek 2006) (A. Macia et al. 2011).

The promoter in the 3′UTR is in the sense orientation and since the majority of new somatic LINE-1 insertions are 5'truncated, it is likely that many newly inserted LINE-1 sequences still contain the 3′ UTR promoter and may initiate transcription of downstream genes (Faulkner et al.,2009).

DNA methylation and histone modifications are well characterized epigenetic mechanisms able to keep LINE-1 transcription under control: when these repressive marks are relieved, LINE-1s are activated. Activation of previously silenced retrotransposons can cause expression changes at neighboring genes by altering the timing or tissue specificity of gene expression. Conversely, retrotransposons are targets for DNA methylation, which induces a repressive chromatin structure that can spread to coding sequences in their proximity, effectively silencing them (Viollet et al., 2014).

1.3.2 TEs as tissue specific enhancers

TEs can regulate gene expression also harboring transcription factor binding sites that can act as host gene enhancers in specific tissues or developmental stages. In particular, among LTR elements, ERVs are present in 5-25% of the genomic regions bound by the pluripotency-associated transcription factors OCT4 or NANOG in human and mouse embryonic stem cells (ESCs). Moreover, in mouse ESCs, ERV elements contain the binding sites for the pluripotency-associated transcription factor SOX2 (Bourque et al., 2008) (Kunarso et al., 2010).

During brain development SINE elements can act as enhancers for host genes. SINE insertions are associated with both ISL1 and Fgf8 genes, involved in motor neuron development and brain development respectively (Bejerano et al. 2006) (Sasaki and Matsui 2008).

DNA transposons, even though they no longer mobilize in most mammals, can act as enhancer to influence host gene expression. For example, some DNA transposons have been shown to act as neocortical enhancers for genes involved in neuronal development (Notwell et al. 2015).

1.3.3 SINEs as regulators of chromosome organization

TEs can play an important role in influencing the organization of mammalian chromosomes, and they are indeed enriched within regions of mammalian genomes that bind CTCF. Moreover, in mice SINE B2 elements carry a CTCF-binding motif (Bourque et al. 2008)(Schmidt et al. 2012)(Sundaram et al. 2014).

CTCF is a well characterized protein which can have multiple functions, acting as an insulator to block the interaction between an enhancer and a promoter, or to prevent the spreading of chromatin domains and as an anchor that assembles chromatin into loops or domains allowing the interaction of regulatory elements (Merkenschlager and Nora 2016). Together with cohesin, CTCF allows developmental long-range enhancers to regulate gene expression (Merkenschlager and Odom 2013).

One example of a SINE insertion that can influence developmental gene regulation through effects on chromatin domains occurs during the expression of growth hormone (GH) in the developing pituitary. Here the SINE element, in collaboration with cohesin elements, is necessary and sufficient to prevent the spreading of repressive chromatin in the GH domain (Lunyak et al. 2007).

1.3.4 LINE-1 in the healthy brain: functional role?

It is still an open question whether LINE-1 retrotransposition has a function in the brain. LINE-1 elements could act as enhancers or promoters for host genes in the brain but at the same time it is also possible that LINE-1 elements are on the way to be domesticated by the host genome, like the domestication of DNA transposons in the immune system. It has been hypothesized that the RAG1 and RAG2 genes, necessary for V(D)J recombination in developing lymphocytes were derived from the Transib superfamily of DNA transposons (Jurka et al. 2005).

LINE-1 retrotransposition in the brain is potentially mutagenic but LINE-1 *de novo* insertions affecting brain development or function in patients remains to be demonstrated.

Supporting studies show how LINE-1 activity in the healthy brain induce genomic rearrangements that could delete genomic regions proximal to genes, but again any functional meaning has to be proved (Erwin et al. 2016). In particular Erwin and colleagues, using a single cell sequencing approach, observed that somatic LINE-1-associated variants are composed of two classes: <u>LINE-1 retrotransposition-dependent</u> insertions and <u>retrotransposition-independent</u> LINE-1-associated variants (Figure 1.10).

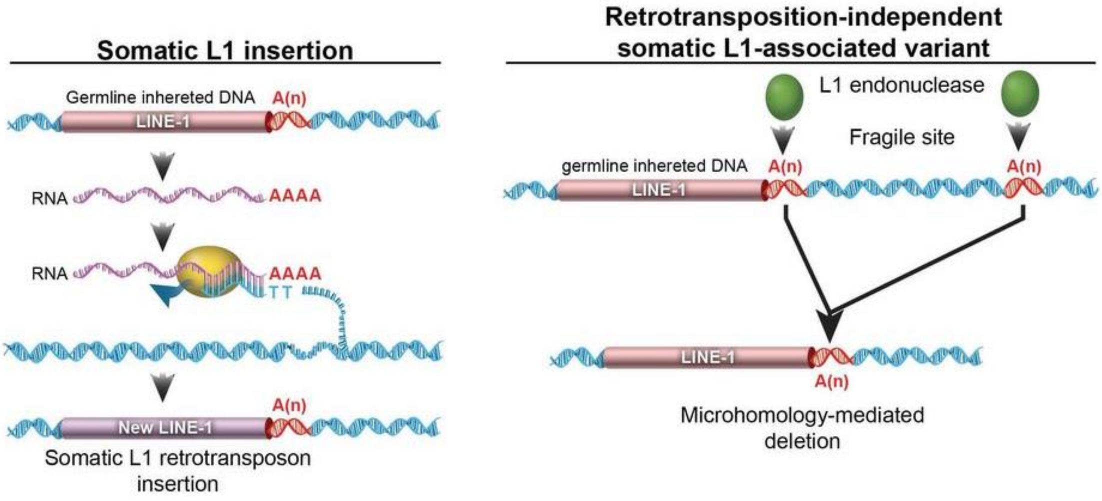


Figure 1.10. Representation of two possible LINE-1 associated variants. On the right: a germline-inherited LINE-1 sequence is transcribed into RNA. The L1 endonuclease and reverse transcriptase protein nicks the genomic DNA and reverse transcribes the L1 RNA, resulting in the insertion of a new copy of LINE-1 sequence. On the left: LINE-1 endonuclease preferentially cuts a a germline-inherited LINE-1 sequence and recombination with a downstream A microsatellite results in a microhomology-mediated deletion. The A microsatellite region may be nicked by the L1 endonuclease or a fragile site within the genome of neural progenitor cells (Erwin et al. 2016).

The first class comprises LINE-1 retrotransposition-dependent insertions. The second class comprises deletions mediated by LINE-1 endonuclease or other mechanism but independent of LINE-1 retrotransposition, indeed no additional inserted sequence or TSD flanking regions are found in the deleted sequence.

In order to discover the mechanism supporting for retrotransposition-independent LINE-1 associated deletions the authors first demonstrated that LINE-1 sequences are prone to instability since they contain the preferential sequence motif recognized by the endonuclease they encode. In particular, the increased expression of LINE-1 endonuclease during neural differentiation induced DBSs preferentially at LINE-1 loci and microhomology-mediated repair system could support the formation of these deletions (Erwin et al. 2016). Therefore, LINE-1-associated genomic regions are predisposed to somatic CNV in the neurogenic brain areas and this could be one of the mechanism explaining the huge neuronal diversity in the nervous system (more details about LINE-1 creating DSBs are described in paragraph 1.4).

1.3.5 Evidence of somatic LINE-1 expression in mouse OE (background results)

The laboratory of S.Gustincich, in collaboration with the group of P. Carninci added an important piece of information about LINE-1 expression in neuronal cells types outside the central nervous system (Pascarella et al.,2014). A NanoCAGE expression analysis in mouse OE revealed massive LINE-1 transcription in OR and VR loci. Both full-length and truncated LINE-1 elements were transcribed in OE (Pascarella 2014). Several mechanisms have been proposed to explain how LINE-1 could modulate transcription of proximal VR and OR genes, including regulation of gene expression via LINE-1 5’ and 3’ promoters and LINE-1 as substrate from chromatin modifications. Seeding and spreading of heterochromatin could be triggered by the GC-rich region of the LINE-1 5’ sense promoter which is a target for methylation (Y. Zhang et al. 2012). Conversely, similarly to a mechanism already described for the mouse growth hormone locus, demethylation of LINE-1 promoter can drive transcription and induce functional chromatin domains which contrast the influence of repressive chromatin modifications (Lunyak et al., 2007). LINE-1 transcription can also generate small non-coding RNAs and these could be involved in the regulation of local chromatin structure (Olovnikov, Aravin, and Fejes Toth 2012). Alternatively, transcriptional activation of LINE-1 in

ORs and VRs loci may be involved in the differentiation of olfactory and vomeronasal neurons. Both OE and the vomeronasal organ (VNO) are characterized by the ability to constantly regenerate throughout the lifespan of an organism (J. H. Brann and Firestein 2010) and interestingly LINE-1 elements are known to be active during adult neurogenesis (Muotri et al. 2005) (Kuwabara et al. 2009).

We consider NanoCAGE results about LINE-1 expression in OE important preliminary data for the project presented in this thesis. NanoCAGE analysis was a starting point to go further investigating a possible involvement of LINE-1s in OR transcriptional regulation. In particular, we hypothesized that LINE-1-associated structural variations (SVs) could have a role in the activation of ORs transcription.

1.4 TEs, DSBs and SVs: a combination to genomic instability

Mammalian genomes are characterized by a high density of repetitive elements; therefore, it is likely to find them as substrates for genomic SVs at a post-insertional stage.

In fact, the basis of the involvement of repetitive sequences in genomic instability is not limited to their abundance in the genomes. If we consider them as pieces of homologous sequences it is clear that they have the ability to alter those DNA repair processes which rely on homologous recombination, thus resulting in genomic rearrangements.

1.4.1 TEs and DNA-DSBs repair systems

The mechanisms according to which TEs, along with DSBs repair systems, can lead to genomic structural variations are reviewed in detail in the work of Hedges and Deininger (Hedges and Deininger 2007). Here we report some representative examples.

TEs can cause genome instability, serving as by-product of those double-strand break (DSB) repair processes which rely on sequence homology. The DNA-damage response (DDR) enables the cells to sense DSBs, propagate DNA damage signals, and activate signaling cascades that subsequently activate a multitude of cellular responses, until the resolution of the lesions. DDR is characterized by the early phosphorylation of the H2AX histone at the site of DSB, which can recruit different repair proteins.

Two DNA-damage repair mechanisms which need tracts of sequence homology are homologous recombination repair (HRR) and single strand annealing (SSA).

HRR involves the use of hundreds of base pairs of sequence homology, resulting error-free repair (Johnson 2000). This pathway is mostly active during the late-S and G2 phases where sister chromatids are available (Haber 2000) (Figure 1.11). Although sister chromatids are the most common template, interspersed TE sequences can offer alternative non-allelic homologous sequence on which the invading strand can anneal (Shurjo K. Sen et al. 2006). In this particular case the mechanism is called non allelic homologous repair (NAHR) (Figure 1.11). Different structural variations, deletions, insertions, and inversions, usually referred to as copy-number variations (CNVs) can be created as result of NAHR, depending upon the relative position of the non-allelic sequences involved. Some possible TE-associated CNVs are represented in Figure 1.11.

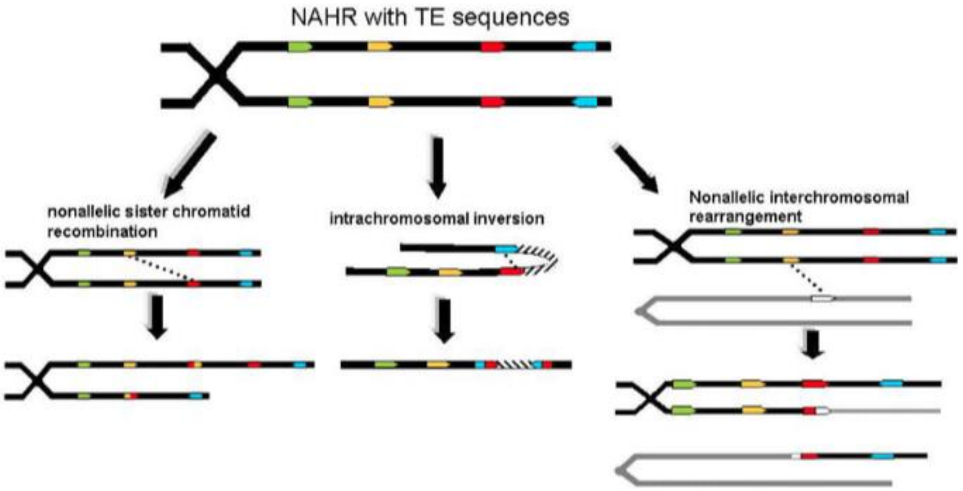

NAHR with TE sequences
nonallelic sister chromatid recombination
intrachromosomal inversion
Nonallelic interchromosomal rearrangement

Figure 1.11. Non allelic Homologous Recombination events among TEs. Three of several possible CNV rearrangements resulting from non allelic recombination are shown. In the first instance, misalignment of sister chromatids during recombination or DNA repair yield insertion and deletion mutations. In the second instance, recombination between two homologous TEs in direct orientation on the same physical chromosome results in an inversion of the intervening sequence. In the third image, alignment with a non homologous chromosome results in the translocation of chromosome arms.

The other repair mechanism based on homologous recombination is SSA. This pathway is active when a DSB occurs within a pair of repetitive sequences with the same strand orientation (direct repeats). During SSA, sequences on both sides of the DSB are cleaved allowing homologous sequences close to the exposed ends to associate and ligate (Figure 1.12) (Pâques and Haber 1999). Usually the homologous sites are located close to the break, in order to reduce the genetic loss. Usually such homology between sequences in close proximity is found in regions enriched for repetitive elements.

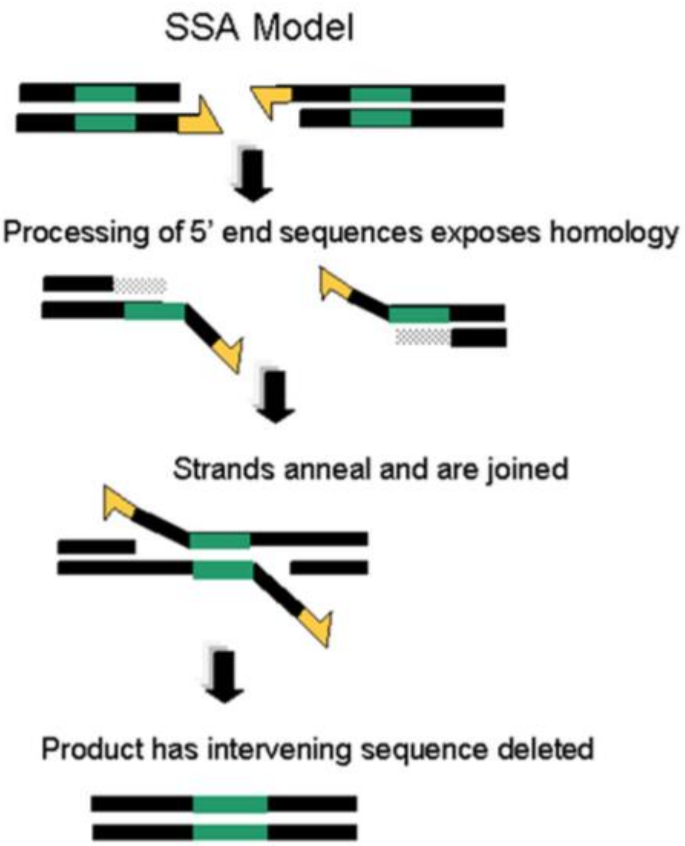

SSA Model
Processing of 5' end sequences exposes homology
Strands anneal and are joined
Product has intervening sequence deleted

Figure 1.12. Single Strand Annealing (SSA). In the SSA model, the 5' ends surrounding the breakage region are resected, exposing adjacent homology. The homologous strands anneal with each other and are ligated, resulted in the deletion of resulted sequence.

1.4.2 Interspersed repeats are hotspot for LINE-1 endonuclease (LINE1-EN) cleavage

Recent evidence suggests TEs involvement in genomic instability mediated by LINE1-EN at a pre-insertional stage: overexpression of LINE-1 endonuclease has been demonstrated to play a role in generating endogenous DSBs in human cells (Gasior et al. 2006). Moreover, typical TSDs flanking the new inserted LINE-1 contain part of LINE1-EN consensus. They can therefore represent preferred hotspots for new EN cleavage generating genomic rearrangements possibly mediated by the recombination repair mechanisms previously described (Tremblay, Jasin, and Chartrand 2000).

Finally, the protein kinase ATM, a crucial protein in many DNA repair signaling processes, was demonstrated to be involved in LINE-1 retrotransposition mechanism (Gasior et al. 2006), thus suggesting an additional correlation between LINE-1 and DNA-DSBs repair systems.

1.4.3 The histone γ-H2AX as marker of DSBs

Genome stability is a condition required for the fitness of all living organisms. One threaten to this stability can be identified in DNA DSBs which are among the most cytotoxic DNA lesions (E. P. Rogakou et al. 1998).

When DNA damages are not properly repaired they can lead to genetic mutations, chromosome rearrangements, or even to cell apoptosis. The phosphorylation of H2AX histone (variant of the H2A histone), on the C-terminal residue (Ser139), is one of the earliest molecular events in response to DNA double-strand breaks (DSBs). The phosphorylated H2AX histone (γ-H2AX) is a crucial cellular signal for the subsequent recruitment of the DNA-damage repair protein machinery (E. P. Rogakou et al. 1998). Therefore, γ-H2AX can be a suitable marker for investigating the presence and the genomics localization of DSBs in different organisms and in different tissues of the same organism.

1.4.4 H2AX-mediated response to DNA-DSB

Histone variant H2AX is a key DDR component. DDR involves many proteins: *sensors* as MRE11–RAD50–NBS1 (MNR) complex and ataxia-telangiectasia mutated (ATM); *mediators* as mediator of DNA damage check-point 1 (MDC1); *effectors* as check-point kinase 1 (CHK1 and CHK2). In response to DSBs, different protein kinases such as ATM, ATR or DNA-PKcs can rapidly phosphorylate H2AX. Phosphorylation is not the only kind of modification occurring on H2AX. Recent evidence showed that during DDR H2AX can also be subject to different modifications such as ubiquitination and acetylation. In particular, H2AX monoubiquitination at Lys119/Lys120 residues has been demonstrated to be critical for the formation of γ-H2AX and the recruitment of MDC1 to DNA damage sites (Pan et al. 2011).

After H2AX phosphorylation, the ATM kinase, which is retained at the damage site, promotes γ-H2AX signal amplification and propagation up to several Mb on each side of the DSBs in mammals (Emmy P. Rogakou et al. 1999).

Therefore, once phosphorylated, H2AX is able to trigger chromatin structural alterations at the DNA damage site to promote DNA repair. Overall, it seems to have both structural and functional roles cooperating in chromatin decondensation and retention of specific factors close to the DSB (Kruhlak et al. 2006; Celeste et al. 2003).

An alternative and intriguing DSB repair mechanism has been recently described in the work of Onozawa and colleagues. Both induced and spontaneous DNA DSBs were demonstrated to be repaired by the insertion of 50 to 1000 bp sequences termed "template-sequence insertions" (TSIs) derived from distant regions of the genome (Onozawa et al. 2014) . These TSIs were derived from genic, retrotransposon, or telomere sequences and were not deleted from the donor site in the genome, leading to the hypothesis that they were derived from reverse-transcribed RNA.

1.4.5 Investigating γ-H2AX and DSBs in physiological conditions

So far, the majority of the experimental efforts have focused on studying exogenous DSBs in different cellular systems after ionizing radiation (IR) exposure or after treatment with chemical compounds able to artificially induce DSBs.

Although DSBs are mostly stochastic and pathological in nature, they can also be intermediates of physiological processes such as antigen receptor diversification in lymphocytes (Cui and Meek 2007), retrotransposition (Gasior et al. 2006) and transcription (Seo et al. 2012a; Madabhushi et al. 2015).

Frequency of endogenous DSBs

Replicating cancerous cells are characterized by high levels of naturally occurring DSBs (Seo et al. 2012a) but little is known about the frequency of DSBs occurring in healthy cells or tissues under physiological conditions.

Are endogenous DSBs too rare to be biologically relevant in normal cells?

One answer came from a work published in 2003 from Vilenchik and Knudson in which they demonstrated that endogenous DSBs produced at sites of single strand lesions (SSLs) during cell S-phase have a rate of 50 events per cell cycle in mammalian cells. Moreover, the rate of DSBs production was estimated to be approximately equal to that produced by 1.5 Gy of IR (Vilenchik and Knudson 2003).

Interestingly, DSBs frequency can vary among cell types differing in specific exposure and response to endogenous free radicals, environmental stresses and in genomic replication rates (Hedges and Deininger 2007).

Endogenous DSBs distribution in the genome

Important information about genome distribution of endogenous DSBs was obtained with the development of the DSB-Capture technique. DSB-Capture is a sequencing-based method able to detect DSBs in situ and directly map these at single-nucleotide resolution (Lensing et al. 2016). Endogenous DSBs profiling was generated for normal human epidermal keratynocytes (NHEK), providing the most comprehensive DSB landscape in a normal human cell line.

Lensing and colleagues showed that the vast majority (76%) of DSBs overlap with DNAseI sensible sites, suggesting a link between regulatory chromatin and genome instability. Moreover, DSBs also correlated with markers of active genes, enhancer regions, CTCF binding sites and the transcription start factor p63. Interestingly, 38% of DSBs overlapped with RNA PolII sites linking DSBs to transcription. Overall genic regions, particularly 5’ UTRs and promoters, were enriched compared to intergenic ones. Finally, increased gene expression was correlated with increased DNA damage around TSS, whereas damage within gene bodies showed little association with gene expression. Finally, no enrichment was found in heterochromatic regions (Lensing et al., 2016).

Interestingly an earlier work investigating endogenous DSBs distribution with a completely different approach (chromatin immunoprecipitation and sequencing experiment (Chip-seq) for γ-H2AX) obtained concordant results (Seo et al., 2012). Chip-seq data from human replicating cells (Jurkat cells) demonstrated γ-H2AX preferential distribution in euchromatic portions of the genome in particular at the active TSSs (Seo et al., 2012).

1.4.6 Sources of naturally occurring DSBs

Endogenous DSBs can arise spontaneously during DNA replication as a result of oncogenic stress. Naturally-occurring DSBs as intermediates of LINE-1 retrotransposition have been extensively described in the previous paragraphs. Nevertheless, other physiological processes can be correlated with the formation of endogenous DSBs. Here we report some examples:

DSBs in VDJ recombination

A well characterized role for naturally occurring DSBs comes from the immune system, where RAG1/RAG2-induced DNA-DSBs are key mediators of V(D)J recombination, (Cui and Meek, 2007). Moreover, it has been demonstrated that γ-H2AX, which is highly phosphorylated in response to V(D)J recombination, has a role in coordinating DSB repair with cellular proliferation and survival to prevent translocations and suppress lymphomagenesis (Yin and Bassing 2008).

Controversial role of DSBs in transcription

Transcription can be a source of genomic instability due to possible interference with the DNA replication machinery, causing replication fork stalling eventually associated with DNA damage and recombination (Fong, Cattoglio, and Tjian 2013) (Branzei and Foiani 2010) (Aguilera 2002). Moreover, it has been reported that DSBs occurring within a gene body inhibit the transcription of the gene, while the transcription of adjacent genes is not altered (Pankotai et al. 2012).

Other works reported that actively transcribed genes are usually hotspots for DNA DSBs but at the same time they are repaired faster than non-transcribed genes. This suggests a role for transcription in the repair of DSBs (Chaurasia et al. 2012).

Possible involvement of endogenous DSBs in transcription has been described in two recent works showing that DNA DSBs is coupled and necessary for transcriptional activation and elongation of stimulus inducible genes in mice and human (Madabhushi et al., 2015) (Bunch et al. 2015).

In particular, Madabhushi and colleagues, using both molecular and genome-wide next-generation sequencing methods, demonstrated that neuronal activity NMDA stimulation causes the formation of DSBs in the promoters of a subset of early-response genes in mouse primary cortical neurons (Madabhushi et al., 2015). Surprisingly, the specific chemical induction of DSBs within the promoter of selected early response genes induces their expression even in absence of an external stimulus (Madabhushi et al., 2015). Moreover, activity-dependent DSB formation is mediated by type II topoisomerase (Topo II) (Madabhushi et al., 2015). Overall DSB formation is shown to be a physiological event that rapidly resolves topological constraints to early-response gene expression in neurons (Madabhushi et al., 2015). In the same year, the group of Calderwood demonstrated that γ-H2AX is accumulated during RNA Pol II pausing release in TSSs of stimulus-inducible protein coding genes in humans (Bunch et al., 2015).

DSBs as mediator of DNA topological stress

Additional data suggesting possible roles of DSBs in physiological conditions come from different works published in 2013 by the groups of Kravatsky and Kretova that developed a method for precise genome-wide mapping of DSBs in human chromosomes (Tchurikov et al., 2013). DSBs hotspots distribute preferentially within genomic regions characterized by H3K27ac and H3K4me3 marks and CTCF binding sites (Tchurikov et al. 2013). Moreover, DSBs hot spots are scattered along chromosomes and delimit 50–250 kb DNA domains. Interestingly, 30% of the domains possess coordinately expressed genes, therefore DSBs distributed outside both silenced and expressed gene clusters together with the insulator protein CTCF. These results are consistent with the view that DSBs are involved in reducing topological stress imposed by long regions of uniform chromatin states (Tchurikov et al., 2013).

Supporting observations come from an analysis of CTCF in mammalian cells, showing that CTCF can orchestrate long-range chromosomal interactions, suggesting a mechanism by which insulators establish regulatory domains (Kurukuti et al. 2006).

Reimand and colleagues demonstrated that type II Topoisomerase β bound CTCF binding sites flanking transcriptionally associated domains (TADs) where are involved in solving topological constrains (Uusküla-Reimand et al. 2016).

Moreover, a ChIP-seq analysis of CTCF-binding sites distribution in the genome of primary human fibroblasts demonstrated that the majority of CTCF signal was depleted within gene clusters (Kim et al. 2007).

Finally, two recent works demonstrated that CTCF orientation is important for regulation of HOX (Narendra et al. 2015) and protocadherin clusters (Guo et al. 2015).

γ-H2AX involvement in additional biological processes

Recently, increasing experimental evidence supported the involvement of DSB-induced γ-H2AX in several non-canonical physiological processes besides DNA-damage response. X-chromosome inactivation in somatic cells, neural stem cell development and cellular senescence maintenance are some of the numerous biological processes in which γ-H2AX can perform both functional and structural roles. A detailed description of γ-H2AX involvement in all the cited biological process is present in (Turinetto and Giachino 2015).

Non-canonical biological roles for γ-H2AX open the possibility that it can be involved in specialized functions in different cell types. Possibly, for all the γ-H2AX supported-biological processes, the occurrence of both induced and endogenous DSBs promotes the initial H2AX phosphorylation. Once phosphorylated, H2AX becomes central to several biological functions possibly unrelated to the DNA DSB response in strict sense (Turinetto and Giachino, 2015).

1.4.7 Investigation of endogenous γ-H2AX in mouse tissues

The first evidence of endogenous γ-H2AX expression in mouse tissues comes from the work of Sedelnikova and colleagues in which five mouse organs (liver, testis, kidney, lung and brain) were shown to display H2AX phosphorylation (Sedelnikova et al. 2004). A further exhaustive analysis of endogenous γ-H2AX distribution in mouse brain revealed that γ-H2AX is expressed from embryonic life to senescence in the absence of experimentally evoked damage to cellular DNA. In particular, neurogenic areas of the forebrain and cerebellum showed both focal and non-focal phosphorylation of H2AX that was shown to be linked to cell proliferation. On the other hand, in cerebral cortex of senescent mice γ-H2AX foci are most likely related to DSB occurrence and repair (Barral et al. 2014).

1.5 Objective

With this work we aim to assess whether genome instability could be involved in olfaction.

The overall objective of the project consists in understanding the possible role of somatic genomic variations associated to LINE-1 in the regulation of OR choice in the mouse OE. To achieve this final goal, we organized the research work around four intermediate objectives:

- Investigation of LINE-1 expression and mobilization in mouse OE (3.1)
- Identification of LINE-1 associated structural variations in *Olfr2* locus (3.2)
- Validation of identified deletions (3.3)
- Genome-wide analysis of endogenous γH2AX in mouse OE and Liver (L) (3.4)

2 Methods and Materials

2.1 Animals

C57BL/6J (Harlan) mice were obtained from SISSA animal facility; B6;129P2*Olfr2*^{tm1Mom}/MomJ (Jackson) were kindly provided by the professor Anna Menini (SISSA). All animal experiments were performed in accordance with European guidelines for animal care and following SISSA Ethical Committee permissions. Mice were housed and bred in SISSA non-SPF animal facility, with 12 hour dark/light cycles and controlled temperature and humidity. Mice had ad libitum access to food and water.

2.1.1 B6;129P2*Olfr2*^{tm1Mom}/MomJ

B6;129P2*Olfr2*^{tm1Mom}/MomJ are characterized by OSNs expressing the mutated locus co-express tauGFP by virtue of internal ribosomal entry site (IRES)-mediated co-translation (Bozza et al. 2002). *Olfr2*/GFP-expressing OSNs are visible under fluorescent microscope due to GFP auto-fluorescence.

2.1.2 Summary of all mouse samples for each experiment

Mouse strain	Experiment	Tissue	Age
C57BL/6J	RT-qPCR (LINE-1 expression)	OE; K; Cer	p21
C57BL/6J	RT-qPCR (LINE-1 CNV)	OE; K; L; H	3m
C57BL/6J	IHC anti-ORF2	OE; K; L; Cer	p6; 1m
B6;129P2*Olfr2*^{tm1Mom}/MomJ	Olfr2 locus analysis	OE	p6
C57BL/6J	WB anti-γH2AX	OE; K; L; Cer	p6;1m;12m
C57BL/6J	ChIP-seq γH2AX	OE; L	p6;1m

Table 2.1. Summary of all mouse samples. For each experiment performed in this thesis, mouse strain, tissues and ages are indicated. OE, olfactory epithelium; K, kidney; Cer, cerebellum; L, liver; H, hippocampus. 1m/3m/12m, one /three/twelve months after birth; p6/p21, six days/twentyone days after birth.

2.2 Genomic DNA extraction

Mouse tissues were dissected from C57BL/6J. Tissues were homogenized at room temperature in 1ml of lysis buffer (Tris pH 8.0 100mM; EDTA pH 8.0 5mM; SDS 0.5%; NaCl 150mM) using a glass-Teflon potter. RNA was digested by incubation at 37°C for 1 hour with Rnase A (40 µg/mL) (Sigma). After having added the Proteinase K (Roche) with a final concentration of 10 µg/mL, samples were incubated O/N at 37°C. The day after, genomic DNA was extracted using the phenol/chloroform/isoamyl alcohol method: 1 volume of phenol (water-saturated, pH 8.0) (Sigma) was added to the samples, followed by a centrifugation at 10000 rpm for 20 minutes. The aqueous upper phase was collected in a new tube, and 1 volume of phenol:(chloroform-isoamyl alcohol (24:1)) was added, followed by a centrifugation at 10000 rpm for 10 minutes. The upper aqueous phase was again collected in a new tube, 1 volume of chloroform:isoamyl alcohol (24:1) was added and centrifuged at 10000 rpm for 10 minutes. The resulting upper aqueous phase was finally collected in a new tube and DNA was precipitated by adding two volumes of 100% ethanol. DNA white flakes were transferred into fresh tubes containing 200 µL of 70% ethanol and centrifuged at 12000 rpm for 15 minutes at 4°C in order to wash and gradually hydrate them. After ethanol removal, the DNA pellets were air dried and dissolved in 300 µL of Tris 10mM pH 8.0 O/N. The DNA quality was finally assessed by gel electrophoresis using a 0.9% ethidium bromide agarose gel.

2.2.1 Genomic DNA quantification using Quant-iTTM PicoGreen® dsDNA kit (Invitrogen)

The copy number variation analysis performed in this study needs to detect very small differences in a high number of copies of the target DNA (LINE-1) between different samples. For this reason, it is necessary to use a precisely quantified small amount of genomic DNA. According to the literature (Coufal et al., 2009), the proper amount of genomic DNA for this analysis is 80 picograms corresponding approximately to 12 genomes. Therefore, the Quant-iTTM PicoGreen® dsDNA kit (Invitrogen) was used as a highly sensitive DNA quantification system. Quant-iTTM PicoGreen® dsDNA kit (Invitrogen) is an ultrasensitive double strand DNA (dsDNA) quantification method which allows to detect and quantify extremely small amounts of DNA in solution, from 25 pg/mL up to 1000 ng/mL. In this study, DNA concentrations raging from 0,9 ng/mL and 1,1 ng/mL were accepted. Quantification of gDNA for LINE-1 CNV assay was performed by Lavinia Floreani.

2.3 Total RNA extraction

Mouse tissues were dissected from C57BL/6J mice at the age of p21 and immediately snap frozen in liquid nitrogen. Total RNA was extracted using Trizol (Invitrogen) according to manufacture's instructions. RNA samples were treated with DNase I (Ambion). cDNA was prepared from 1 µg of RNA using the iSCRIPT™ cDNA Synthesis Kit (Bio-Rad) according to manufacturer's instructions.

2.4 Quantitative Real-Time PCR (RT-qPCR) with Taqman probes.

In this work, a Taqman multiplex RT-qPCR with relative quantification was performed. Quantitative PCR experiments were performed using iQ5 real-time PCR detection system (BIORAD). For each sample, calibration curves of cDNA or genomic DNA were assayed to verify both amplification efficiencies and whether the dilution point chosen for the analysis was within the linear range of the reaction. Data obtained by qPCR co-amplifications of the target DNA sequence (LINE-1) and the internal invariable control were analyzed using the $2^{-\Delta\Delta Ct}$ (Livak and Schmittgen, 2001) method. Standardization was performed considering the highest ΔCt value as calibrator. Statistical analysis of the data obtained by a minimum of three qPCR replica was performed by the paired, two-tailed, Student's t-test. Values were considered statistically significant when p-value resulted < 0.05.

2.4.1 LINE-1 expression and copy number variation (CNV) Taqman assays

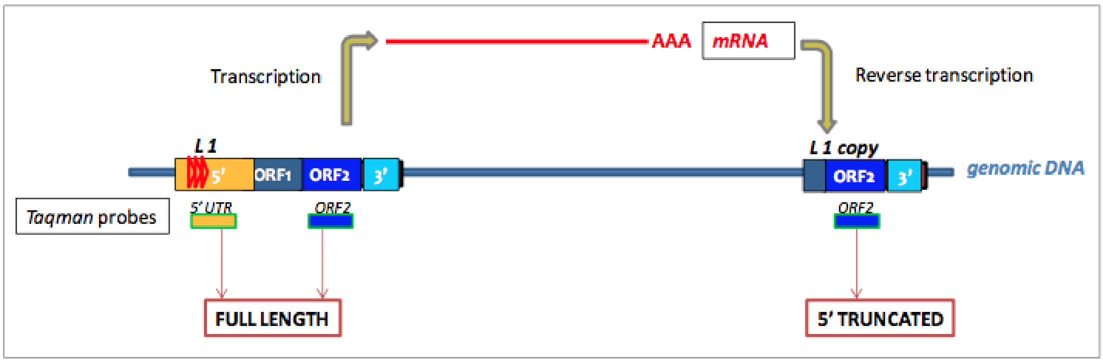


Figure 2.1. **Multiplex quantitative PCR assay for mouse LINE1 (L1) expression and copy number variation analysis.** A cartoon of the rationale of the multiplex qPCR used in this study is shown. After retrotransposition most of LINE-1 elements are 5' truncated and inactive. Taking advantage of this peculiarity, specific Taqman probes were designed to discriminate intact full length L1 (5'UTR probe, yellow box) from retrotransposed copies, which are 5' truncated (ORF2 probe, blue box).

Previously in our lab a new assay to assess LINE-1 retrotransposition in mouse tissues was developed, based on the protocol published by Coufal and colleagues in humans (Coufal et al., 2009). This technique uses Taqman RT-qPCR to evaluate LINE-1 copy number but can be suitable also for evaluate LINE-1 mRNA expression. Taqman qPCR strategy was planned considering that LINE-1 5' UTR probes detect LINE-1 full length forms, corresponding to original LINE-1 elements. Probes complementary to the ORF2 region at the LINE-1 3'end detect the entire LINE-1 repertoire, including both full length and truncated forms, which correspond to both original and retrotransposed LINE-1 elements. Since murine 5'UTR region of LINE-1 is composed of different repetitive monomers which characterize different LINE-1 subfamilies, we designed specific probes for each active mouse LINE-1 type (A, Tf and Gf) (Figure 2.1) (Goodier et al., 2001). In this way, we were able to discriminate between the different mouse L1 subfamilies. In order to perform a relative quantification of the truncated and the full length LINE1 copies, we used, as internal controls, Taqman probes designed against glycerhaldeyde 3-phosphate dehydrogenase or GAPDH for expression assay, and MicSAT for CNV assay (Table 2.2). RT-qPCR cyclying conditions are shown in Table 2.3.

RT-qPCR assay for LINE-1 CNV was performed by Lavinia Floreani.

	Primer forward 5’→3’	Primer reverse 5’→3’	Taqman probe 5’→3’
ORF2	CCCTCAACAGAGGAATGGAT	CCATCCATTTGGCTAGGAAT	AAATGTGGTACATCTACACAATGGA
A	TGAGCACTGAAACTCAGAGGAG	GATTGTTCTTCTGGTGATTCTGTTA	GAATCTGTCTCCCAGGTCTG
Tf	CCAAACACCAGATAACTGTACACC	CGTGGGAGACAAGCTCTCTT	TGAAAGAGGAGAGCTTGCCT
Gf	TGCCCACTGAAACTAAGGAGA	GCTTGTTCTTCAGGTGACTCTGT	TGCTACCCTCCAGGTCTGCT
GAPDH	CGACCCCTTCATTGACCTC	CTCCACGACATACTCAGCACC	CTCCACTCACGGCAAATTC
MicSAT	GAACATATTAGATGAGTGAGTTAC	GTTCTACAAATCCCGTTTCCAA	ACTGAAAAACACATTCG

Table 2.2. List of primer used for RT-qPCR Taqman assays.

	RT-qPCR program	
Initial activation step	95°C 20 sec	
denaturation	95°C 10 sec	40 cycles
annealing/extension	59°C 30 sec	
end of reaction	4°C forever	

Table 2.3. RT-qPCR cycling conditions

2.5 Western Blot

Total protein lysates from mouse tissues were prepared by homogenization in Sample Buffer 2× (4% SDS, 20% Glycerol, 0.12M Tris pH 6.8, and 10% BME). Equal amount of proteins was separated in 15% SDS-polyacrylamide gel transferred to nitrocellulose membrane. Immunoblotting was performed with the following primary antibodies: Anti-phospho-Histone H2A.X 1:1000 (Ser139) (clone JBW301, Millipore), anti-Histone H2A.X 1:1000 (ab11175, Abcam), anti-βactin 1:10000 (A5441, Sigma). All primary antibodies were incubated overnight at 4°C.

Signals were revealed by using ECL (Amersham) after one-hour incubation at room temperature with secondary antibodies conjugated with horseradish peroxidase.

2.6 Immunohistochemistry

C57BL/6J mice were sacrificed by decapitation and cervical dislocation.

L, K and Cer were dissected from C57BL/6J mice at p6 and left overnight at 4°C in 4% paraformaldehyde (PFA). For OE, after mouse decapitation, the skin and the jaw were removed from the heads. In parallel the entire liver was dissected, then the samples were left overnight at 4°C in 4% paraformaldehyde (PFA). Only for 1 month old mice, the head was incubated overnight in EDTA solution 0.5M pH 8 for an additional decalcification step. All tissues, after a 4-h cryoprotection step in a 30% sucrose/1× PBS solution at 4°C, were included in Frozen section medium Neg-50 (Richard Allan Scientific) and snap frozen in liquid nitrogen. Frozen blocks were brought into a cryostat (Microm International) and left for 60 min at −21°C, if not used, they were stored at -80°C.

Immunohistochemistry was performed 16 μm-thick cryo-slices prepared with Vibratome 1000s (Leica), both prepared from 6 days (p6) and 1 month old C57BL/6J mice (n = 3). Slides were blocked with PBS, 10% FBS, 1% BSA, and 1% fish gelatine (filtered) for 1 h at room temperature, and the primary (anti-Line1 1:100 M300, Santa Cruz, anti-Olfactory Marker Protein 1:2000 544-10001, Wako) and secondary antibodies were diluted in PBS, 1% BSA, 0.1% fish gelatine, and 0.3% Triton X-100. Incubation with primary antibodies was performed for 16 h at room temperature;

Incubation with conjugated-secondary antibodies was performed for 2 h at room temperature. Nuclei were labelled with DAPI. Slides were mounted with mounting medium for fluorescence Vectashield (Vector Lab) and observed with a confocal microscope.

2.7 Methods for gene clusters identification, random regions selection, repeats annotation and coverage calculation

Data collection – mouse genome (assembly GRCm38) was downloaded from Ensembl (release 85)1 FTP (ftp://ftp.ensembl.org/pub/release-85/fasta/mus_musculus/dna/Mus_musculus.GRCm38.dna.primary_assembly.fa.gz). Pfam2 gene domain annotations were downloaded from BioMart.

Identification of gene clusters – gene clusters table was generated based on physical gene location using python scripts. In particular, two consecutive genes sharing the same domain were considered to belong to a cluster if the distance between them is

≤ 500 Kbp regardless of the presence of other genes within the cluster, a threshold already used in other gene family studies.

Random regions selection – random regions were selected relying on gene cluster composition in order to reflect their length. Each gene cluster was randomized 100 times using the shuffleBed program from BEDTools5 denying shuffled intervals to overlap each other and to fall inside gene cluster intervals with the next parameters: -noOverlapping, -excl.

Repeats annotation – repetitive elements were annotated using RepeatMasker (v4.0.6) against Repbase database (v20.05 – Release 20150807) with the next parameters: -species mouse, -s, -gff.

Coverage calculation – a simple python script was used to divide repetitive elements in families and the coverage was calculated for each kind of gene cluster and its respective randomizations using coverageBed from BEDTools with default parameters. Finally, the mean and the standard deviation were obtained from the hundred randomization ad plotted together with the cluster coverage in the final barplot using R (v3.3.1) and ggplot2 (2.1.0).
Analysis was performed by Massimiliano Volpe.

2.8 Sample preparation for Laser Capture Microdissector

OE samples were prepared from 6 days old B6;129P2-*Olfr2*^{tm1Mom}/MomJ (*Olfr2*-GFP mice). After decapitation, the skin and the jaw were removed from the heads, and the samples were left overnight in 1× ZincFix fixative (BD Biosciences) diluted in DEPC-treated water. After a 4-h cryoprotection step in a 30% sucrose/1× ZincFix solution, heads were included in Frozen section medium Neg-50 (Richard Allan Scientific) and snap frozen in liquid nitrogen. Frozen blocks were brought into a cryostat (Microm International) and left for 60 min at −21°C. Serial coronal sections of mouse heads (14 µm) were cut with a clean blade, transferred on PEN-coated P.A.L.M. MembraneSlides (P.A.L.M. Microlaser Technologies), and immediately stored at −80°C. Before usage, the slides were brought to room temperature and air-dried for 2 min. The MOE was morphologically identified and different pools of GFP-positive and GFP negative OSNs were selected with fluorescent microscope, microdissected, and collected with a Zeiss P.A.L.M. LCM microscope (Carl Zeiss Inc.) in P.A.L.M. tubes with adhesive caps and immediately used for subsequent whole genome amplification (Figure 2.2).

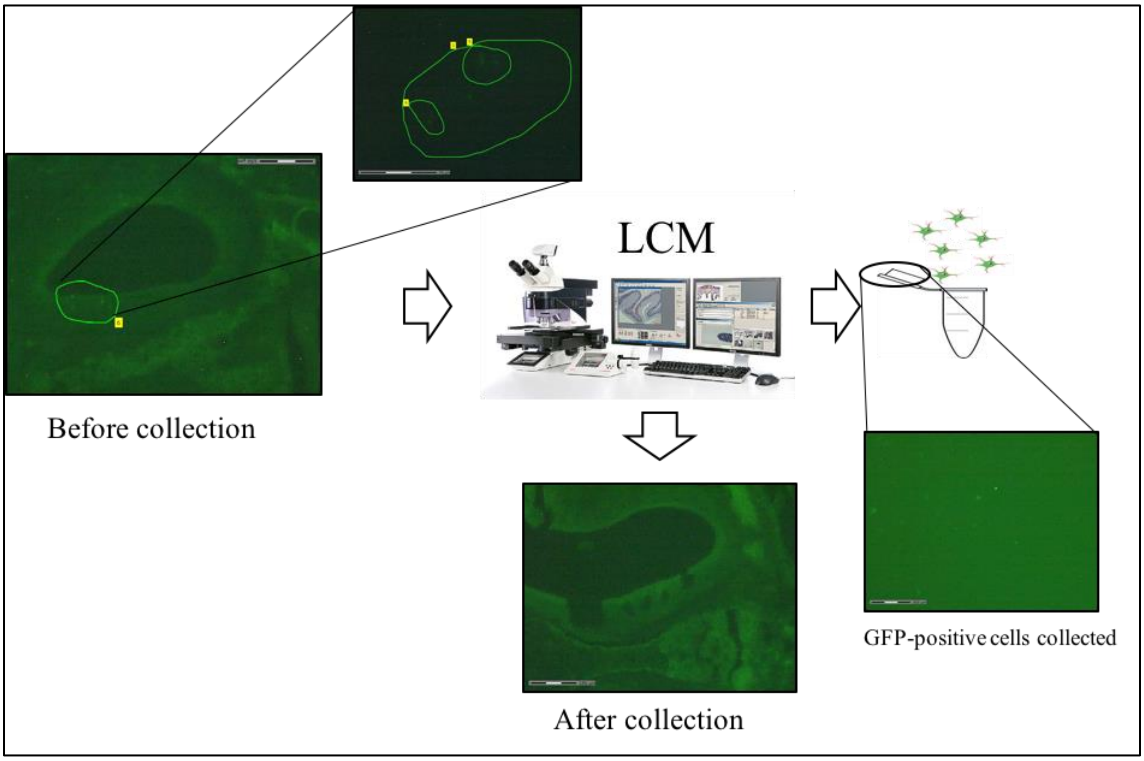


Figure 2.2. Laser capture microdissection, capitulation and collection of GFP-positive cells from OE.

2.9 Whole genome amplification (WGA)

2.9.1 Multiple Displacement Amplification (MDA)

Multiple displacement amplification is a non-PCR based DNA amplification technique. This method can rapidly amplify minute amounts of DNA samples to a reasonable quantity for genomic analysis. The reaction starts by annealing random hexamer primers to the template: DNA synthesis is carried out by a high fidelity enzyme, called ϕ29 DNA polymerase, at a constant temperature. In this work we used Repli-g Single Cell kit (QIAGEN), a commercially available MDA kit specialized for single cells starting material. We followed the manufacture's instructions and we incubated the samples for amplification 16 hours at 30°C. After amplification, MDA products were checked on 0.8% agarose gel before and after column purification with QIAquick PCR purification kit. Compared with conventional PCR amplification techniques, MDA generates larger sized products (5-10 kb) without PCR amplification biases: for this reason, we chose MDA amplification as definitive method to produce starting DNA material to use in downstream analysis. Nevertheless, we were aware of possible MDA amplification artifacts, in fact we included a non-MDA control sample (OE sample) in the experiment.

2.10 Long Range PCR (LR-PCR) amplification of 50 kb *Olfr2* locus

Purified MDA products and bulk genomic DNA (without any MDA amplification) were used as template for LR-PCR amplification of 50 kb genomic sequence around *Olfr2* TSS. For a first Pac Bio sequencing the 50 kb around *Olfr2* gene were divided into 11 amplicons of about 5 kb each. For the subsequent Illumina sequencing the 50 kb sequence locus was divided into 13 amplicons with a size ranging from about 400 bp to about 5 kb (amplicon 2 was divided into three sub-amplicons for Illumina sequencing). We performed the PCR amplification using a LR-PCR amplification kit (QIAGEN) following the manufacture's instruction. For each PCR reaction we used about 100 ng of purified MDA product or bulk genomic DNA. PCR products were check on 0.9% agarose gel and purified with QIAquick PCR purification kit (QIAGEN) following the manufacture's. Purified products were quantified with Nano Drop (ThermoScientific).

LR-PCR program and primer used are shown in Table 2.4. and Table 2.5, respectively.

Long Range PCR program

Initial activation step	93°C 3min
denaturation	93°C 15 sec
annealing	60°C 30 sec
extension	68°C 6 min
end of reaction	4°C forever

Table 2.4. Long Range PCR reaction cycling conditions. Denaturation, annealing and extension steps were repeated for 35 cycles.

Amplicon	Forward Sequence (5'→3')	Reverse Sequence (5'→3')
1	CTATCCAAAGCCACTGATAAGG	GTGAATTTCATGCAGTCATTG
2.1	CAACCTGGCATCTAAATAAAG	GAGTTAGCAGAAAAAGCAAAGTCT
2.2	AGACTTTGCTTTTTCTGCTAACTC	CCAGGTGTTCAGAAATGAAAGTTATG
2.3	CATAACTTTCATTTCTGAACACCTGG	CTACTTGAACTTGAGCTTTGTGGTGG
3	GGTTGAGCCACCACAAAGCTC	CTTGATGGTGATAGATGCACTG
4	CAGTGCATCTATCACCATCAAG	GCACTAAGTACCCAGGGAAAG
5	CTTTCCCTGGGTACTTAGTGC	CCACCTTCATCCTCATTCGT
6	CAGTGATGGACTATGGTGGAC	CTACCTAACCATCTCTGCTGG
7	GACCGCTATGTGGCCATCTG	CTGTCCAATCACAAGCCTCCC
8	TTGATTTCAATGGCTTTCAGA	CCAAATCGGATAGGGGACTA
9	GACTGGAGTGAAGTGGAATC	GTGTTCAGTACCAAGCACTC
10	CCAATGCCTGTGAAGTTCAGAAG	GTTTCCCTCTTAGGAATGCTTTC
11	CCAAGCAAATGGACTGAAG	GTGAGTGAGCCTGTCCTATC

Table 2.5. List of primers used for Long Range PCR amplification of 50kb *Olfr2* locus. Primers were designed on the reverse strand.

2.11 Sequencing techniques

2.11.1Pac Bio RS II sequencing

Sequencing was performed at GATC Biotech (Germany). Pac Bio RS II is a single molecule real time sequencing technique based on the properties of zero-mode waveguides, developed by Pacific Biosciences of California, Inc. This technique is characterized by the ability to sequence very long reads with high accuracy and coverage without any PCR amplification bias. For these reasons it is particularly suitable for the detection of structural variations.

LR-PCR products of each amplicon were pooled together in equimolar ratio to reach 2µg of total DNA and sent for sequencing. A total of two Pac Bio libraries were sequenced.

2.11.2 MiSeq Illumina Sequencing

Sequencing was performed at IGA Technology Services (Udine). MiSeq sequencing technology can produce 2 x 300 paired-end reads in a single run, allowing detection of target variants with unmatched accuracy, especially within homopolymer regions. Illumina sequencing DNA-libraries were built for each of DNA amplicons for the three samples, a total of 39 libraries were sequenced.

2.11.3 Bioinformatic analysis and structural variation discovery

Pac Bio and Illumina reads were mapped with BWA MEM (http://bio-bwa.sourceforge.net/) on reference mouse genome (GCRm38/mm10 assembly).

Variation discovery on Illumina reads was performed with Pindel bioinformatics tool.

Here we report some Pindel parameters which have to be taken into account for the interpretation of the results.

1) Threshold coverage: minimum coverage of 5 Illumina reads was chosen, meaning that each variation has to be covered by at least 5 reads in at least one sample to be included in Pindel output. Therefore, when a variation is associated to a specific sample (with at least 5 supporting reads), it could be present also in the other samples with a coverage under the 5 reads threshold.

2) Length range: we considered only variations longer than 50 bp and shorter than the length of the correspondent PCR amplicon. An exception was done for insertions since Pindel algorithm can identify only very short insertions (500 bp as maximum).

3) Variation identity: Pindel recognizes two variations as different if their coordinates diverged for at least 1 bp.

Bionformatics analyses were performed by Aurora Maurizio.

2.12 Deletions repeat enrichment

Repeat description in correspondence of the deletions was based on repeatmasker annotation (rmsk) (http://www.repeatmasker.org/). Bedtools intersect were used (Quinlan and Hall 2010).

Analyses were performed by Aurora Maurizio

2.13 Pac Bio validation

High-identity Illumina sequencing dataset was combined with a complementary Pac Bio data set. Pac Bio long reads reads were employed to validate Pindel deletions. Illumina split reads supporting the deletions were aligned over PB reads using blastn, regardless the sample (NB: reads supporting the deletions found in the OE sample were aligned also over Pac Bio reads coming from the MDA sample and viceversa).
blastn parameters (word_size 20, perc_identity 85, evalue 1e-10) (Altschul et al. 1990).

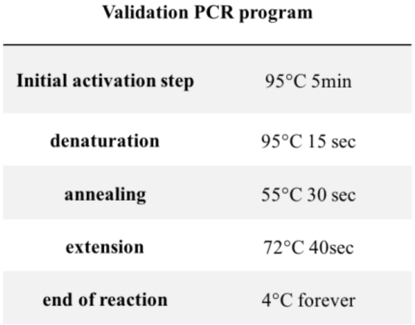
85 % minimum identity was imposed because of the 14% error rate affecting PacBio reads.

Each query Illumina fasta sequence should align for its entire length (+- 10 bp) over the corresponding Pac Bio read in order to consider the deletion supported by the Illumina read validated.

Bionformatics analyses were performed by Aurora Maurizio

2.14 End-point PCR validation assay

Validation PCR assays were performed with ExTaq DNA-Polymerase (Takara) following the manufacture's protocol.

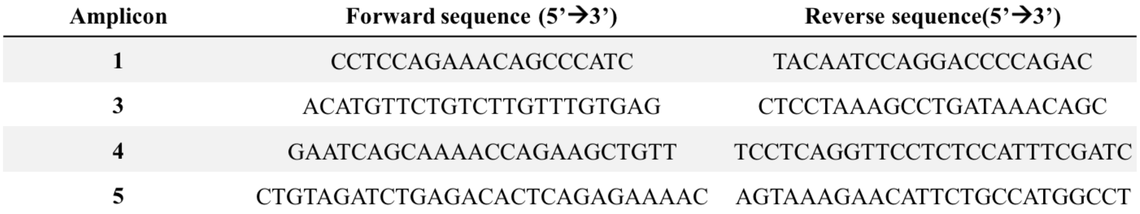


Validation PCR program

Initial activation step	95°C 5min
denaturation	95°C 15 sec
annealing	55°C 30 sec
extension	72°C 40sec
end of reaction	4°C forever

Table 2.6. Validation PCR reaction cycling conditions. Denaturation, annealing and extension steps were repeated for 40 cycles.

When necessary, we adapted melting temperatures and DNA polymerase extension step depending on the specific primers features for each validation assay. A list of validation primers used is shown in Table 2.7.



Amplicon	Forward sequence (5'→3')	Reverse sequence(5'→3')
1	CCTCCAGAAACAGCCCATC	TACAATCCAGGACCCCAGAC
3	ACATGTTCTGTCTTGTTTGTGAG	CTCCTAAAGCCTGATAAACAGC
4	GAATCAGCAAAACCAGAAGCTGTT	TCCTCAGGTTCCTCTCCATTTCGATC
5	CTGTAGATCTGAGACACTCAGAGAAAAC	AGTAAAGAACATTCTGCCATGGCCT

Table 2.7. List of primers used for Pindel deletion validation PCR assays. Primers were designed on the forward strand.

Sanger sequences were analyzed with UCSC Blat tool and with NCBI Blast tool in order to verify whether they were supporting any Pindel deletion.

Representation of Sanger sequences, Illumina reads and Pindel deletions are shown in the results as screenshots of uploaded tracks on USCS Genome Browser tool.

2.15 Chromatin Immunoprecipitation (ChIP)

Mouse tissues were lysed and cross-linked in freshly prepared 1% formaldehyde solution. The crosslinking reaction was stopped by adding Glycine (0.125 M), then the tissue was homogenized using a Dounce homogenizer and sonicated.

100 μg of chromatin sample was immuno-precipitated O/N with 2 ug of anti-phospho-Histone H2A.X (Ser139) Antibody (clone JBW301, Millipore) or with IgG-conjugated magnetic beads. DNA was de-crosslinked at 65°C O/N and extracted with standard phenol/clorophorm protocol (see genomic DNA extraction). Finally, extracted DNA was quantified with Picogreen. For each sample 10 ng of IP DNA and input DNA were sent for Illumina sequencing libraries construction.

2.15.1 ChIP samples sequencing and bioinformatics analysis

ChIP samples were sequenced with Illumina High Seq paired-end sequencing at Deep Seq facility of School of Life Sciences, Queen's Medical Centre at Nottingham University.

A filtering pipeline was used to filter reads with low sequencing score and reads aligning to adaptor sequences. First, raw reads were trimmed against adaptors using scythe (https://github.com/vsbuffalo/scythe). The remaining reads were quality trimmed using sickle (https://github.com/najoshi/sickle). Reads passing the filters were mapped to the mouse reference genome (build mm10/GRCm38) using bwa (http://bio-bwa.sourceforge.net/). In the bam files obtained, duplicates were marked using picard and filtering was performed in order to remove reads with mapping quality below 60, duplicates and improper pairs. The filtered data was further sorted and mate fixed. Filtering, sorting and mate fixing were performed using samtools (version 0.1.19). Bam files were then converted to paired-end bed format using bamToBed utility from bedtools suite (version 2.25.0). Peak calling was performed on the filtered data using epic (version 0.1.18), a peak caller based on SICER suitable to identify diffused domains of enrichment, which is the pattern expected for gamma-H2AX signal.

A simple perl script was used to convert epic outputs to bed format (conversion includes: ordering the columns as appropriate for bed format, substitute spaces with tabs, assign an ID to each peak, defined as chromosome_start). From each of the resulting bed files peaks overlapping blacklisted genomic regions were removed using intersectBed from bedtools suite. Blacklisted regions for mm9 were downloaded from https://sites.google.com/site/anshulkundaje/projects/blacklists and lifted to mm10 using the UCSC Genome Browser liftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). To obtain a representative set of ChIP-seq peaks for each biological condition (liver at P6, OE at P6 and OE at 1 month) we considered the intersection of the peak sets obtained in the two replicates and these were used in subsequent analyses.

The bioinformatics methods described above were performed by Margherita Francescatto.

Peak genomic distribution

Peak annotation was performed with ChIP-seq NEBULA online-tool specific for ChIP experiments on histone modifications (Boeva et al. 2012). Default parameters were used for the analysis.

Gene Ontology (GO) functional annotation

GO enrichment analysis was performed using GREAT online tool. For each peak the nearest TSS was annotated within 1 Mb. Each sample dataset (foreground dataset) was analyzed using all the other sample datasets as background dataset. Only GO terms with FDR<0.01 were included in the output (McLean et al. 2010).

Peak annotation with respect to mouse CpG islands

The annotation of the peaks identified with respect to CpG islands was performed using the AnnotatePeak.pl function of the HOMER suite of tools (Heinz et al. 2010).

Analyses performed by Margherita Francescatto.

Peak annotation with respect to different class of repeats

Analyses were performed by Massimiliano Volpe.

Comparison of ChIP-seq peaks with L and OE expression data

Liver CAGE expression data

Liver expression data was derived from the mouse tissue catalogue of FANTOM5 consortium. The table containing normalized expression values across all mouse samples profiled within phases I and II of the FANTOM5 project (Arner et al. 2015) was downloaded from FANTOM5 website (http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/mm9.cage_peak_phase 1and2combined_tpm_ann.osc.txt.gz). The data corresponding to liver neonatal samples closer to the age of mice for which we have ChIP-seq data (N6, N7, N10, N20, N25 and N30) was extracted using a custom R script. Data was filtered in order to retain only CAGE peaks with at least 1tpm (tpm=tags per million) in all samples. A file in bed format was created reporting genomic coordinates of CAGE peaks, peak annotation and average expression. Since FANTOM5 expression data is natively annotated in mm9, we lifted the bed file to mm10 using the liftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver).

OE expression data

Expression data from the work of Ibarra-Soria and colleaues (Ibarra-Soria et al. 2014) was downloaded, converted to tab separated format and used to create a bed file containing TSS coordinates (transcription start site +40 bp to make it generally comparable to CAGE peaks), corresponding annotation and average expression across the 6 replicates.

Expression correlation analyses were performed by Margherita Francescatto.

Comparison with chromatin segmentation of the L mouse genome

We downloaded 11 ChIP-seq datasets in bigWig format:

- 7 liver histone marks (H3k27ac, H3k27me3, H3k36me3, H3k4me1, H3k4me3, H3k79me2, H3k9ac) and corresponding input were downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31039
- liver CTCF and Pol2 ChIP-seq and corresponding input were downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29184

All bigWig files were converted to bedgraph using the bigWigToBedGraph tool (downloaded from UCSC genome browser, hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/bigWigToBedGraph). The data was then binned using the chromhmm-tools (https://github.com/daler/chromhmm-tools/blob/master/chromhmm_signalize.py). The file config.txt was created following instructions from chromhmm-tools manual.

The signal was then binarized using ChromHMM (Ernst and Kellis 2012)function "BinarizeSignal" and the HMM chromatin state model was built using the function "LearnModel", specifying 10 states and the genome build of interest (mm9).

The relative enrichment of the states belonging to the segmentation so created with respect to the γ-H2AX peaks identified in the three conditions was calculated using the function "OverlapEnrichment". To identify peaks corresponding to enhancer regions as characterized by ChromHMM model we intersected (intersectBed) each of the three peak sets with the bed file containing the 10-state segmentation of the genome and extracted peaks overlapping state 3.

Comparison with CTCF, Pol II and DNase data.

Because the peak calling is not an exact process, we accepted two features to overlap if they were located with 1kb of each other (in other words, intersections were performed using windowBed from bedTool suite with w parameter set to 500). Intersection analyses were performed by Aurora Maurizio.

Statistical analysis and plotting

All statistical analysis and plots in sections X, Y, Z were performed using R.

3 Results

3.1 LINE-1 expression and mobilization in mouse OE

3.1.1 Validation of NanoCAGE data with qPCR Taqman assay

We started our analysis validating recently published NanoCAGE data about LINE-1 expression in mouse OE (Pascarella 2014), performing RT-qPCR with Taqman probes. Real time experiments were performed on C57BL/6J wild type mice at the age of p21, a total number of six mice were used. For each mouse we compared LINE-1 expression among three tissues: OE and cerebellum (Cer) as neuronal tissues and kidney (K) as non-neuronal one.

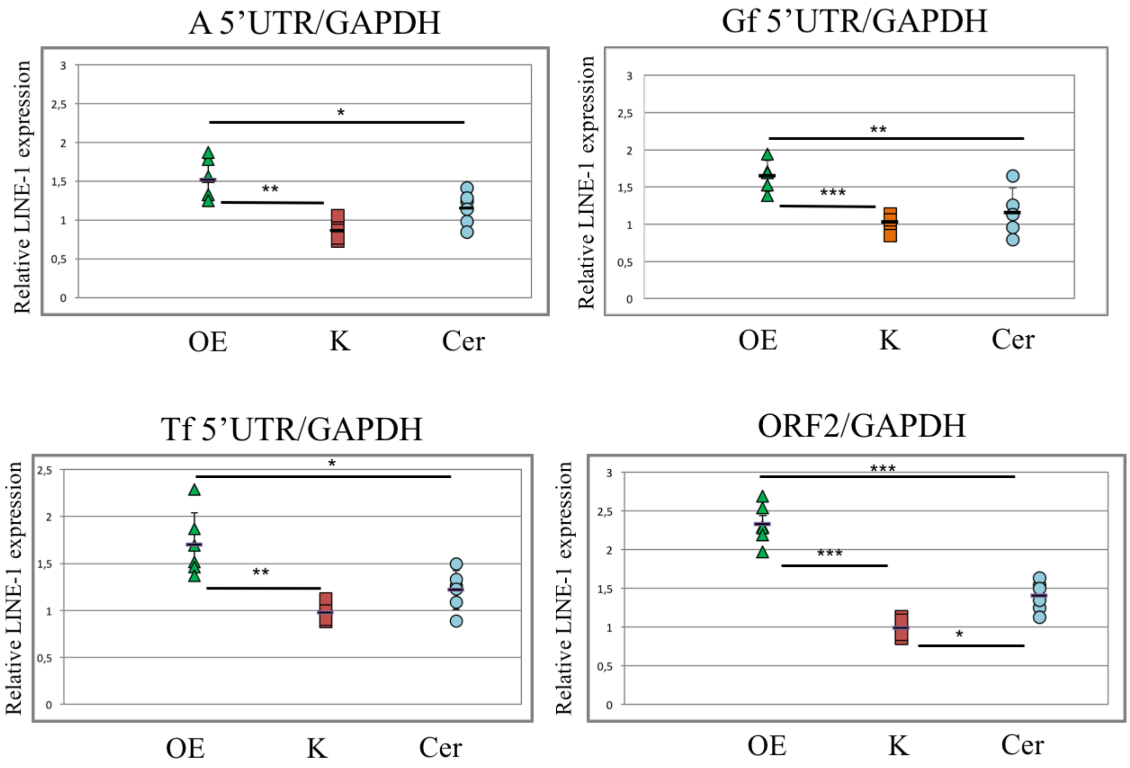


Figure 3.1. RT-qPCR analysis for LINE-1 expression in mouse tissues in p21 wild type mice. Different RT-qPCR Taqman probes were designed for each of the three mouse LINE-1 subfamilies (A,Gf,Tf) and for the ORF2 coding sequence, common to all mouse LINE-1s. A probe on GAPDH was used as internal control. OE, olfactory epithelium; K, kidney; Cer, Cerebellum. A minimum of three independent replicates were performed for each assay. * p-values<0.05; **p-values < 0.01; *** p-values <0.001 derived from t-Student paired test.

For all RT-qPCR assays (A, Tf, Gf and ORF2) we observed a higher expression of LINE-1 elements in OE compared with K and Cer. No significant differences were observed comparing Cer with K except for ORF2 assay, where LINE-1 expression in cerebellum was significantly higher than in kidney (Figure 3.1). Overall these results confirmed the expression of both full length and 5'truncated LINE-1 transcripts in OE.

3.1.2 LINE-1 ORF2 protein expression in mouse tissues

To further confirm LINE-1 expression in OE we performed fluorescent-IHC experiments to compare LINE-1 ORF2 protein expression in different mouse tissues. We analyzed LINE-ORF2 protein signal in OE, Cer, K and liver (L) of C57BL/6J wild type mice at the age of p6. In agreement with RT-qPCR data we detected ORF2 protein signal mainly in OE and Cer compared with K, where no ORF2 signal was detected. Few positive cells were detected also in L.

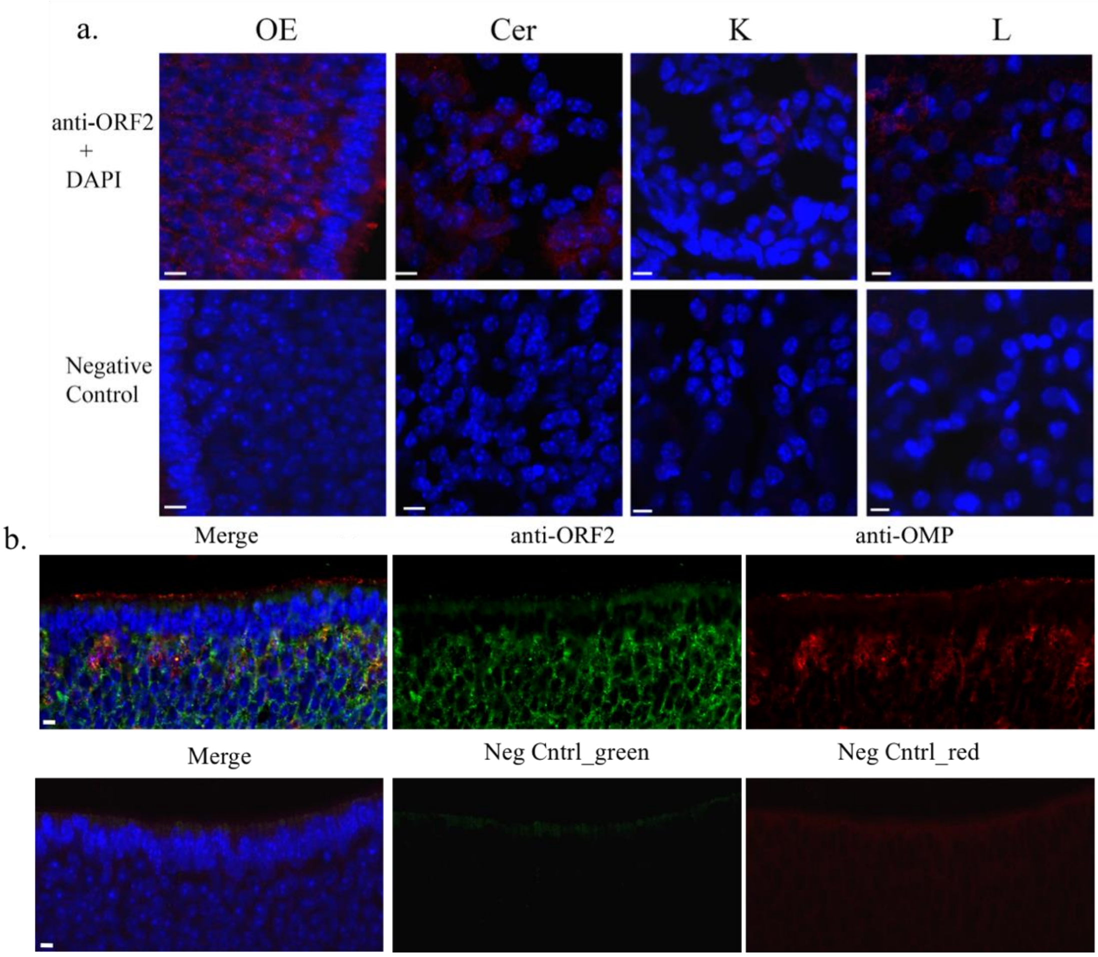


Figure 3.2. IHC anti LINE-1 ORF2 protein in C57BL/6J wild type mice. a. Merge signal (ORF2+ DAPI) were shown in 16 µm frozen slices from each mouse tissue analyzed and for negative controls. OE, Olfactory epithelium; Cer, Cerebellum; K, kidney; L, liver. Nikon confocal 40x oil, zoom 4x. b. Anti-ORF2 signal in 16 mm frozen slices from 1m OE was shown in green, OMP signal was shown in red and DAPI in blue. Nikon confocal 40x oil, zoom 2x. For all the figures, Neg Cntrl = sample incubated only with secondary Ab. White bars correspond to 10 µm.

Then, we looked in detail at OE. To verify whether ORF2 signal was localized only in mature OSNs, we performed a double fluorescent-IHC for ORF2 and for OMP, a cytoplasmic protein which labels only mature OSNs. We were able to observe a co-localization of the two signals in the intermediate level of OE where mature OSNs are present. Interestingly ORF2 signal was diffused also in the basal layer of the epithelium were basal cells and OSN progenitors are located (Figure 3.2, b).

Overall, ORF2 protein presented a punctuate expression pattern. It seemed to be localized preferentially in the cytoplasm but some positive nuclei were also observed in OE, Cer and L.

3.1.3 LINE-1 CNV in mouse tissues

We further investigated LINE-1 copy number variation (CNV) in different mouse tissues. We used the RT-qPCR CNV assay described in methods, able to detect both full length and truncated LINE-1elements.

The CNV assay was performed in five C57BL/6J at 3 months of age for different neuronal and non-neuronal tissues: OE, hippocampus (H), K and liver (L). As previously observed (Coufal et al. 2009), we were able to detect a high number of LINE-1 sequences in H compared to K and L. Moreover, we surprisingly observed that LINE-1 mobilization was increased in OE compared with all other tissues (Figure 3.3).

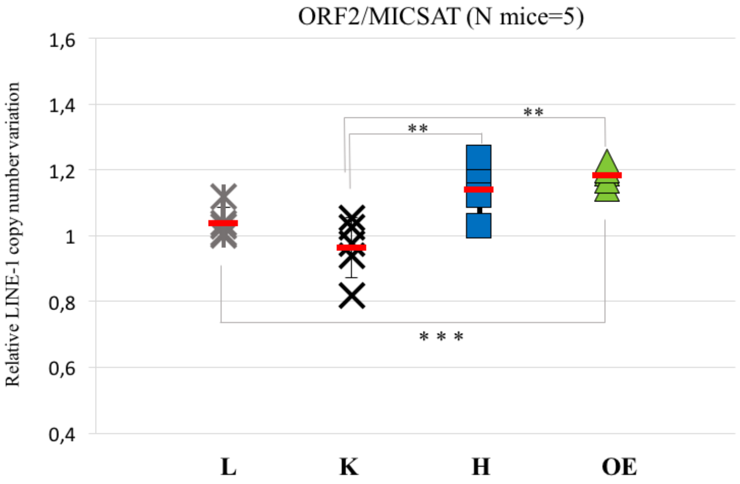


Figure 3.3. qPCR analysis of total LINE1 copy number in mouse tissues. Relative quantification of total LINE-1 obtained by qPCR using the ORF2-MICSAT Taqman probes. OE, olfactory epithelium; L, liver; K, kidney; H, Hippocampus. L1 ORF2 number is significantly higher in all brain regions respect to liver and kidneys. Each symbol represents the average of 3 qPCR independent replica. Red lines represent the average of all samples. Standard deviations are indicated. *P<0.05; **P<0.01; ***P<0.001 resulting from *t-student* paired test.

These results are consistent given the ability of OE, like the hippocampus, to constantly regenerate during the adult life on an organism thanks to the presence of neuronal stem cell populations (Hurtt et al., 1988).

3.1.4 Analysis of repeats enrichment in OR clusters

We then focused our attention on OR gene clusters, investigating the distribution of different classes of repeats within them (detailed method information in paragraph 2.7). Interestingly the OR-clusters showed a specific enrichment for LINE-1 elements compared with other classes of repeats. Surprisingly, LINE-1 enrichment was a peculiarity of OR genes. Indeed, other gene clusters analyzed presented enrichment in different classes of repeats (i.e: Zinc Finger clusters were enriched for LTRs) or no enrichment at all (i.e:Trypsin clusters) (Figure 3.4).

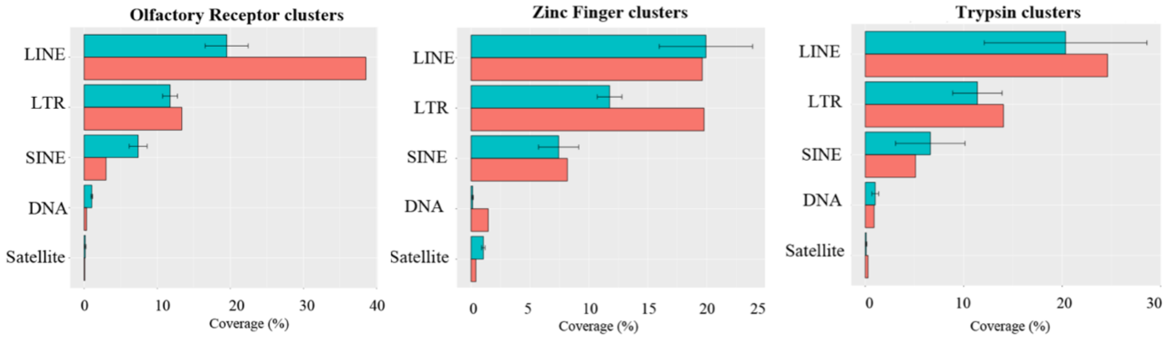


Figure 3.4. **Repeat enrichment analysis for different mouse gene clusters.** Real clusters are shown in red and random clusters are shown in green. Black bars represent standard deviation for each repeat class enrichment.

3.2 Structural variation (SV) discovery in *Olfr2* locus

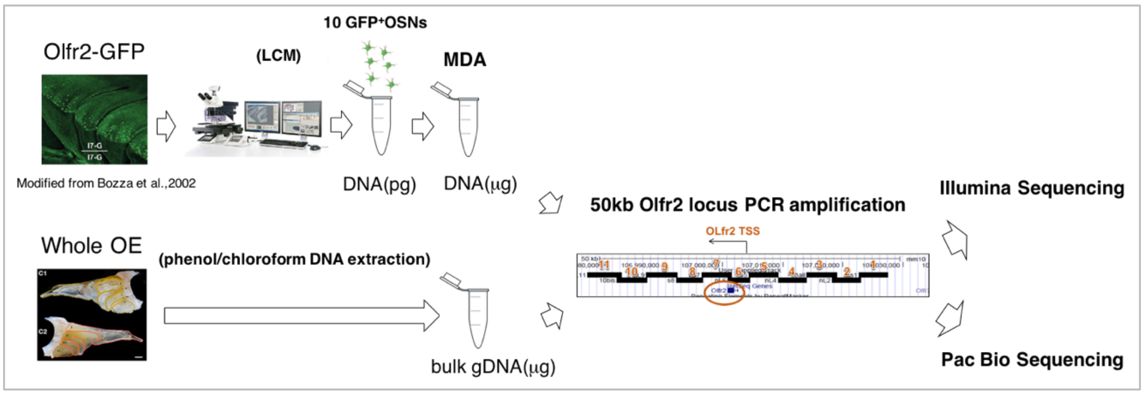
Given the relatively high LINE-1 expression and increase in CNV in OE compared with other mouse tissues and given the peculiar LINE-1 enrichment in OR clusters, we decided to further investigate possible LINE-1 effects in a specific OR locus.

We started to investigate whether LINE-1-associated non-annotated structural variations (LINE-1-SVs) are present in a selected OR locus. Moreover, to assess a link between LINE-1-SV and OR expression we performed the analysis in parallel on an active and an inactive locus of the same OR.

To this aim we took advantage of B6;129-*Olfr2*-GFP mice in which OSNs expressing *Olfr2* receptor also express GFP protein. With this tool we were able to collect different pools of 10 GFP-positive cells expressing the receptor *Olfr2* and carrying the active *Olfr2* locus (see details in methods and material). In parallel bulk genomic DNA was extracted from total OE of an age-matched mouse from the same litter.

We performed *Olfr2* locus targeted-sequencing combining two different sequencing techniques, Pac Bio and Illumina systems.

Experimental flowchart:



3.2.1 Multiple displacement amplification (MDA) of 10 cells

Different pools of 10 GFP-positive cells were collected by LCM from B6;129-*Olfr2*-GFP mice at the age of p6 and immediately amplified with MDA, a whole genome amplification technique (see details 2.9.1). For all MDA replicas we obtained long DNA fragment output ranging from 1 to 10 kb (Figure 3.5).

Random primer amplification in MDA negative control was expected from the MDA kit protocol (Figure 3.5). To verify successful *Olfr2* locus amplification, we checked by PCR the presence of *Olfr2* coding sequence (CDS) in each MDA amplified product (data shown in Appendix, Figure-A1.

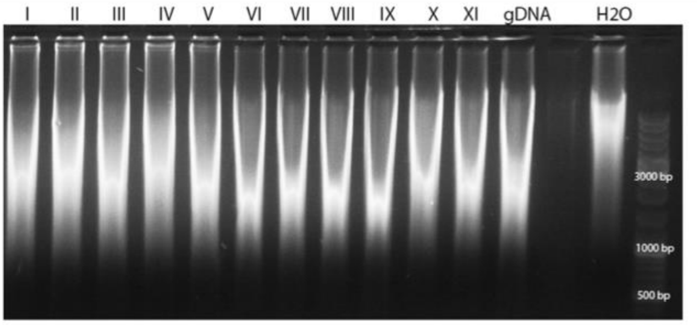


Figure 3.5.**Whole genome MDA amplification of 10 cells DNA.** MDA amplification was performed on 11 independent pools of 10 cells, previously collected by LCM. A positive control with 20 ng of gDNA as starting material and a negative control with no input material is shown. The presence of DNA in the MDA-negative control is expected due to random primer amplification.

3.2.2 Long Range PCR (LR-PCR) amplification of the *Olfr2* locus

LR-PCR amplification of *Olfr2* locus for Pac Bio sequencing

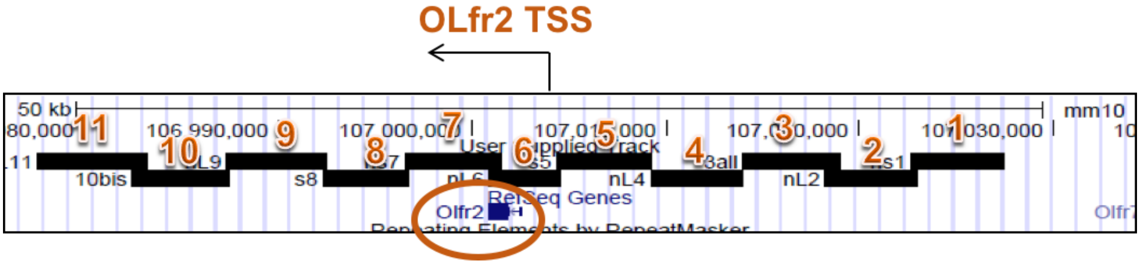


Figure 3.6. Schematic representaion of 50 kb region around *Olfr2* TSS locus. UCSC Genome Browser represetation of 50 kb on mouse chromosome 7 in *Olfr2* locus. Black rectangles and red numbers indicated the PCR amplicons upstream (1-5) and downstream (6-11) the TSS (indicated with a black arrow). Red cyrcle indicates *Olfr2* trascript. *Olfr2* trascritpion is indicated on the reverse strand.

In order to amplify 50 kb around *Olfr2* transcription start site (TSS) (25 kb upstream and downstream the TSS) we divided the locus in 11 amplicons of about 5 kb each (Figure 3.6). We performed locus amplification on 11 MDA biological replicates (MDA I-XI), each derived from a pool of 10 GFP-positive cells collected by LCM. In parallel, we performed *Olfr2* locus amplification from bulk genomic DNA, extracted from OE (gDNA-OE) and not amplified by MDA. Bulk OE DNA sample can be considered a “technical positive control” in order to be aware of possible MDA amplification artifacts formation which can compromise downstream PCR amplification. At the same time, it is a “negative biological control” because it consists of whole OE cell population among which *Olfr2*-expressing cells (GFP positive) represented less than the 0.1% of all OSNs which are about ten million neurons.

Therefore, we expected that if a putative genomic variation is involved in *Olfr2* expression it will be underrepresented or even undetectable in OE compared with MDA sample.

The best PCR products from different MDA biological replicas and from bulk OE gDNA were purified (all PCR reaction products ware purified) and pooled together for Pac Bio sequencing (Figure 3.7).

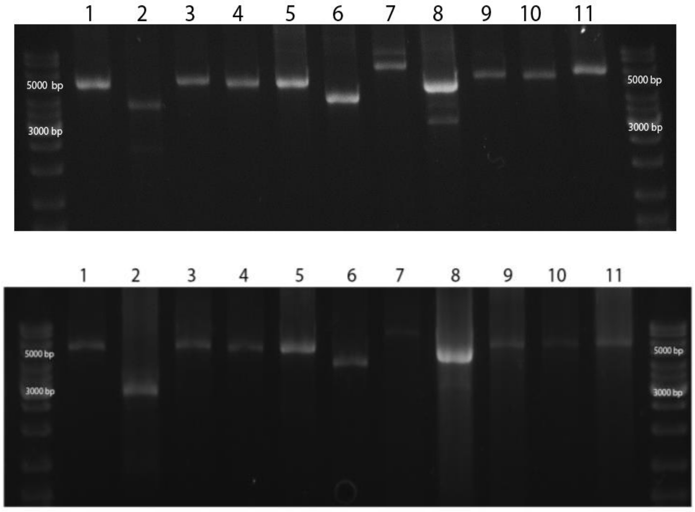


Figure 3.7. **LR-PCR amplification of 50 kb *Olfr2* locus.** Both panels show purified PCR products for 11 amplicons covering 50 kb of *Olfr2* locus in MDA sample (top panel) and OE sample (bottom panel), respectively. Each PCR was perform in parallel to negative control with no DNA input (data not shown).

Pac Bio reads from MDA and OE samples were firstly mapped on the reference genome (MouseGRCm38/mm10 assembly) to verify the 50kb *Olfr2* locus coverage with PCR amplification. For both the samples we were able to cover almost of the targeted locus, 82% and 72% of reads were mapped for MDA sample and gDNA-OE sample respectively, demonstrating that both MDA amplification and LR-PCR amplification techniques were successful. Unfortunately, LR-PCR amplification failed for amplicon 7, due to the presence of the long IRES-GFP construct, and for amplicon 2 due to high density of repetitive sequences which made very difficult primer design. Pac Bio coverage is reported in Appendix (Figure-A 3).

LR-PCR amplification of *Olfr2* locus for Illumina sequencing

The same LR-PCR amplification of 50kb *Olfr2* locus was performed for Illumina sequencing. Here we were able to optimize amplicon 2 PCR dividing it in three smaller sub-amplicons (called 2.1, 2.2 and 2.7). Therefore, for amplicon 2 we had only Illumina supporting reads. Amplicon 7 was excluded from the amplification due to the impossibility to cover by PCR IRES-GFP region.

For Illumina sequencing we performed LR-PCR amplification of *Olfr2* locus on two different MDA biological replicates out of 11 (MDA-V and MDA-XI) and on the same bulk OE gDNA (OE sample), used for Pac Bio. Finally, 13 PCR amplicons for each sample were sequenced for a total of 39 Illumina libraries.

After sequencing, Illumina reads were mapped on the reference mouse genome (MouseGRCm38/mm10 assembly) to check for *Olfr2* locus coverage among the three samples (MDA V, MDA XI and OE) (Figure 3.8). As for Pac Bio sequencing we obtained good coverage of the locus that was comparable between MDA and OE samples. Looking at the reads coverage comparing the 5’ and 3’ regions with respect to *Olfr2* TSS, we observed that the most covered amplicons were located upstream the *Olfr2* gene.

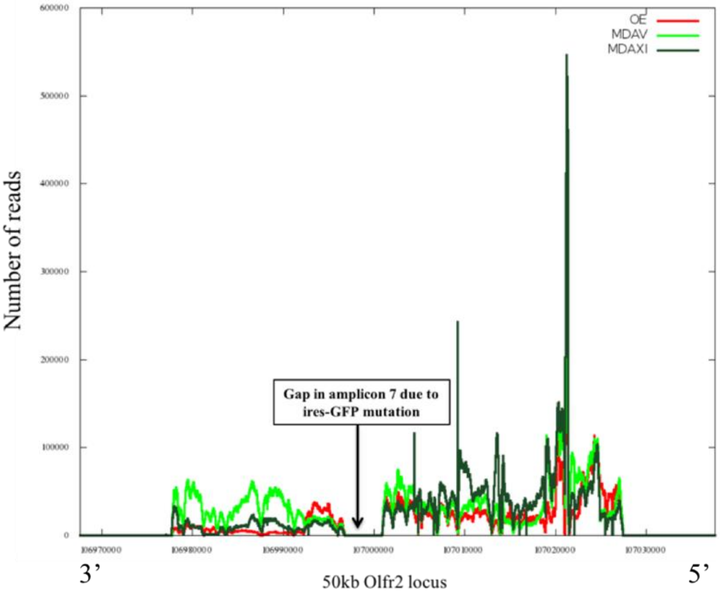


Figure 3.8. Illumina sequencing coverage of 50kb *Olfr2* locus. Coverage for OE (red), MDAV (light green) and MDA XI (dark green) sample is shown. Number of reads per *Olfr2* 50 kb locus coordinates are plotted. The arrow shows the amplification gap at the level of amplicon 7 due to IRES-GFP insertion. Minus strand is shown.

3.2.3 Identification of non-annotated SVs

Since OR loci are characterized by high density of annotated TEs which contribute to high genomic complexity, our initial goal was to identified the presence of non-annotated genomic variations in *Olfr2* locus.

Variation discovery was performed with Pindel tool on Illumina reads for each PCR amplicon, indeed amplicons were sequenced as single units of the *Olfr2* locus.

Overall in the 50 kb locus, we identified 2400 deletions, 34 inversions, 407 tandem duplications and 806 insertions (Table 3.1).

Interestingly all the variations were found in a "heterozygous condition", meaning that for each variation we found only a small percentage of reads ("alternate reads") supporting the non-annotated variant, while a majority of the reads ("reference reads") supported the invariant reference sequence.

	DELETIONS	INVERSIONS	TANDEM DUPLICATIONS	INSERTIONS
MDAV	755	12	116	200
MDAXI	826	15	100	424
MDAV_and_MDAXI	242	2	76	65
MDAV_and_OE	62	2	15	10
MDAXI_and_OE	55	0	6	14
MDAV_and_MDAXI_and_OE	240	1	92	58
OE	220	2	2	35
Total	2400	34	407	806

Table 3.1. Number of Pindel variations identified in Illumina samples. A complete list of CNVs indentified with Pindel in Illumina reads per each sample is shown. "MDAV_and_MDAXI" shows variations shared by both MDA biological replicates; "MDAV_and_MDAXI_and_OE" shows variation shared by all the three samples.

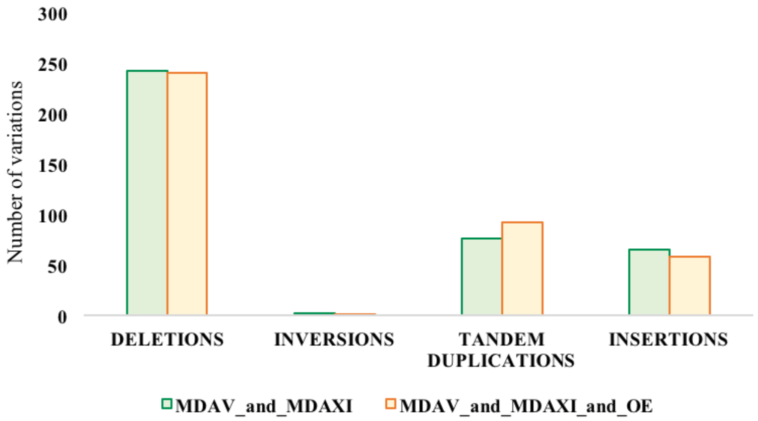
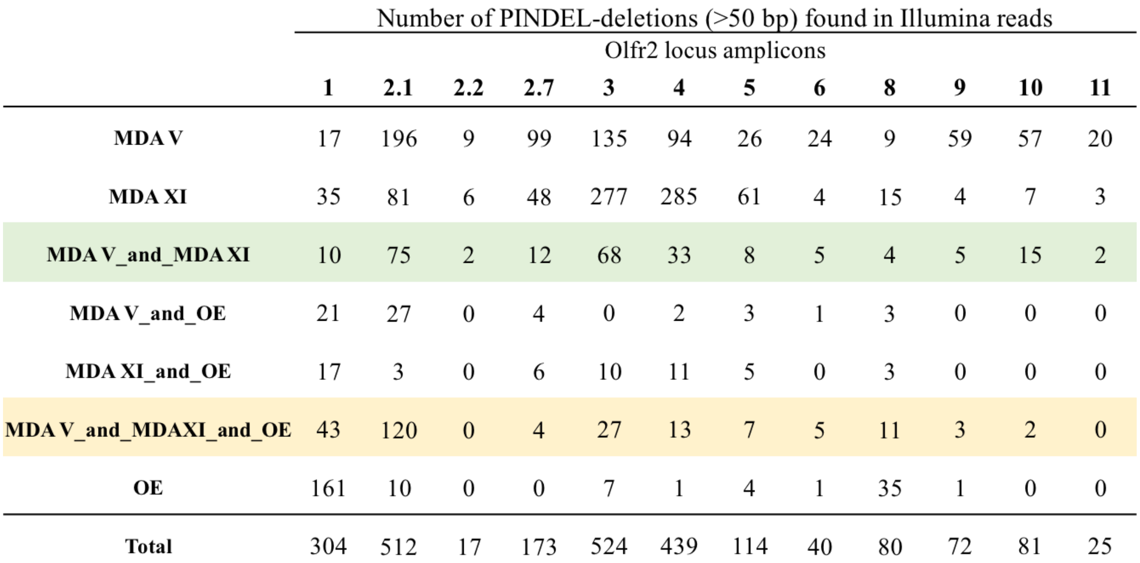


Figure 3.9. Number of Pindel variations per sample. Number of Pindel variations found in both MDA samples (green) and those found in common between MDA and OE sample (yellow) are compared.

We identified the number of variants for each sample combination (Table 3.1). In order to reduce false positives due to possible MDA artifacts we filter starting Pindel dataset considering only those variants found in both MDA replicates (MDA V and MDA XI). We then compared the number of variants present only in both MDA replicates with those common between MDA and OE samples. For each category, we observed how the number of MDA-specific (present only in both MDA V and XI) and common variations (present in both MDA and OE samples) are comparable (Figure 3.9).

3.3 Genomic deletions in *Olfr2* locus

Since deletions were the most abundant variation found in the *Olfr2* locus, we decided to start performing validation analysis on this category. In Table 3.2, the distribution of deletions among different samples is shown.



	Number of PINDEL-deletions (>50 bp) found in Illumina reads											
	Olfr2 locus amplicons											
	1	**2.1**	**2.2**	**2.7**	**3**	**4**	**5**	**6**	**8**	**9**	**10**	**11**
MDA V	17	196	9	99	135	94	26	24	9	59	57	20
MDA XI	35	81	6	48	277	285	61	4	15	4	7	3
MDA V_and_MDA XI	10	75	2	12	68	33	8	5	4	5	15	2
MDA V_and_OE	21	27	0	4	0	2	3	1	3	0	0	0
MDA XI_and_OE	17	3	0	6	10	11	5	0	3	0	0	0
MDA V_and_MDAXI_and_OE	43	120	0	4	27	13	7	5	11	3	2	0
OE	161	10	0	0	7	1	4	1	35	1	0	0
Total	304	512	17	173	524	439	114	40	80	72	81	25

Table 3.2. Number of Pindel deletions in Illumina samples. A complete list of Pindel deletions found in different Illumina samples, per each amplicon, is shown. “MDAV_and_MDAXI” shows variations shared by both MDA biological replicates; “MDAV_and_MDAXI_and_OE” shows variation shared by all the three samples.

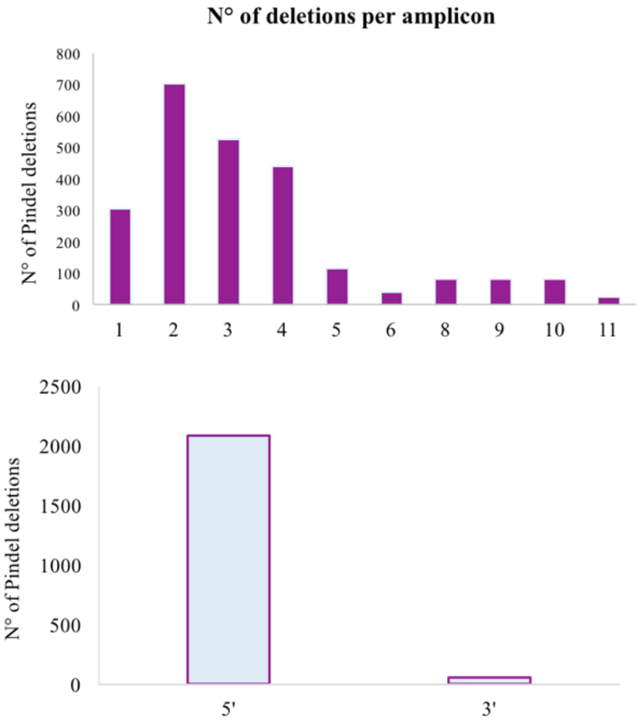


Figure 3.10. Number of Pindel deletions per amplicon. Number of Pindel deletions for each *Olfr2* locus amplicon are shown (Top panel). Average number of Pindel deletions in 5' amplicons (1-5) is compared with average number of Pindel deletions in 3' amplicons (6-11) (Bottom panel).

In agreement with a general higher coverage for the amplicons upstream the *Olfr2* TSS (1-5) compared with the coverage of those downstream (6-11) (data shown in Appendix), the highest number of deletions was found at the 5' of the *Olfr2* coding sequence (Figure 3.10, bottom panel).

As expected amplicon 6, which contains the *Olfr2* coding region, presented a very low number of deleted sequences, suggesting that the most conservative part of the locus is characterized by a low genomic variability (Figure 3.10, top panel).

General length frequency distribution of all the deletions found in *Olfr2* locus is represented in Figure 3.11. As anticipated, deletion length is overall distributed in a range within 50 and 5000 bp. Moreover, we observed four frequency peaks, the first within 50 and 300 bp, two peaks around 2000bp, and finally a peak around 4500 bp (Figure 3.11, left panel).

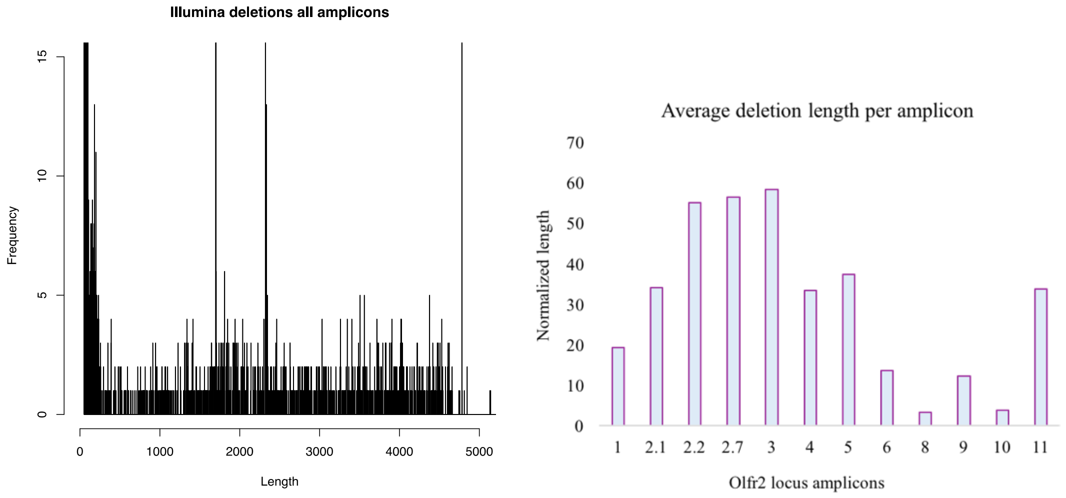


Figure 3.11. Frequency of deletion length distribution. Frequency of length distribution for all Pindel deletions is shown (Left panel). Normalized average length of deletions for each amplicon is shown (Right panel).

Interestingly, looking at single amplicons we can observe that, on average, the longer deletions are located in the 5’ amplicons while the deletions in the 3’ amplicons are shorter (Figure 3.11, right panel). All the amplicons have a similar length ranging from 3500 bp to 5500bp except for amplicon 2.2 which was much shorter than the others, about 400 bp.

3.3.1 Validation of Pindel deletions with Pac Bio long reads

The first validation of deletions found in Illumina reads was performed taking advantage of Pac Bio reads. Pac Bio sequencing is characterized by higher percentage of sequencing error (about 14 %) compared with Illumina, but it is not affected by sequencing amplification bias, since it is based on single molecule sequencing.

Each Pindel deletion supported by Illumina reads was checked for the presence of additional supporting Pac Bio reads (Figure 3.12).

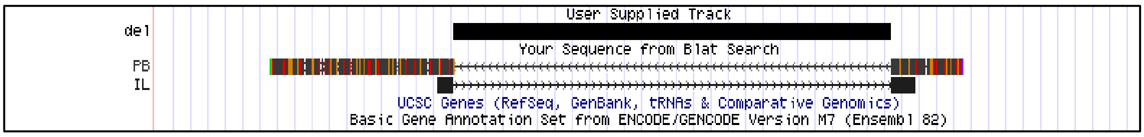

User Supplied Track
del
Your Sequence from Blat Search
PB
IL
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
Basic Gene Annotation Set from ENCODE/GENCODE Version M7 (Ensembl 82)

Figure 3.12. Representative example of a Pac Bio read supporting a Illumina deletion. UCSC Genome Browser 8 kb screenshot of a 4kb Pindel deletion found in Illumina reads and supported by PacBio read IL, Illumina read; PB, Pac Bio read; del, deleted region.

Deletions in amplicon 2 (for technical reason previously explained), 6 and 8 did not have any Pac Bio supporting reads.

For all the other amplicons, different Pindel deletions supported by the same PacBio read were considered as a same unique deletion, therefore the original number of total deletions found in Illumna was drastically reduced (Table 3.3). Overall, only 2% (55 out of 2400) of all the Pindel deletions found in Illumina reads was still supported by Pac Bio reads. This result could find two possible explanations. One possibility is that a high fraction of Illumina deletions was sequencing artifacts, a second option is that Pac Bio sequencing did not reach saturation.

	Number of PINDEL-deletions (>50 bp) supported by PB reads						
	Olfr2 locus amplicons						
	1	**3**	**4**	**5**	**9**	**10**	**11**
MDA V	0	2	4	4	1	0	0
MDA XI	0	7	6	0	0	0	1
MDA V_and_MDA XI	1	12	5	0	0	0	1
MDA V_and_OE	0	0	0	0	0	0	0
MDA XI_and_OE	0	0	0	0	0	0	0
MDA V_and_MDA XI_and_OE	4	1	2	1	2	1	0
OE	0	0	0	0	0	0	0
Total	5	22	17	5	3	1	2

Table 3.3. Number of Pindel deletions supported by Pac Bio (PB) reads. Number of PB-supported deletions in Illumina samples per each amplicon is shown. MDAV_and_MDAXI shows variations shared by both MDA biological replicates; MDAV_and_MDAXI_and_OE shows variation shared by all the three samples.

In order to have an overview of the Pac Bio supported deletion distribution throughout the 50kb *Olfr2* locus, a condensed representation of the deletions pattern for each amplicon is shown in Figure 3.13.

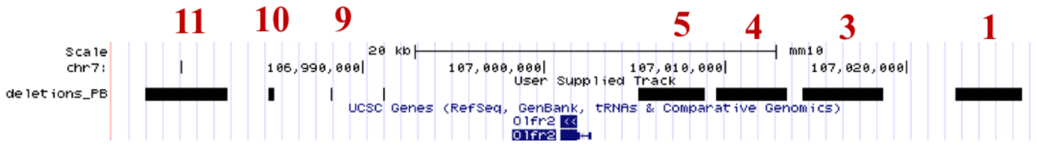


Figure 3.13. Pac Bio-supported deletions distribution in *Olfr2* locus. UCSC Genome Browser representation of Pac Bio-supported deletions in 50 kb *Olfr2* locus. Number of amplicons harboring the deletions are indicated in red. For each deletion, a "dense" representation is indicated with black rectangles. *Olfr2* coding sequence is highlighted in blue.

Looking at the frequency of deletion length distribution, after Pac Bio filter, we note that the number of very small and very long Illumina deletions (<100 bp and >4000 bp) was reduced, the majority of deletions having a length ranging from ca. 2000 to 4000 bp.

Focusing on single amplicons, interestingly we can observe that the longer deletions were maintained in the 5’ amplicons compared with the 3’ ones (Figure 3.14, right panel).

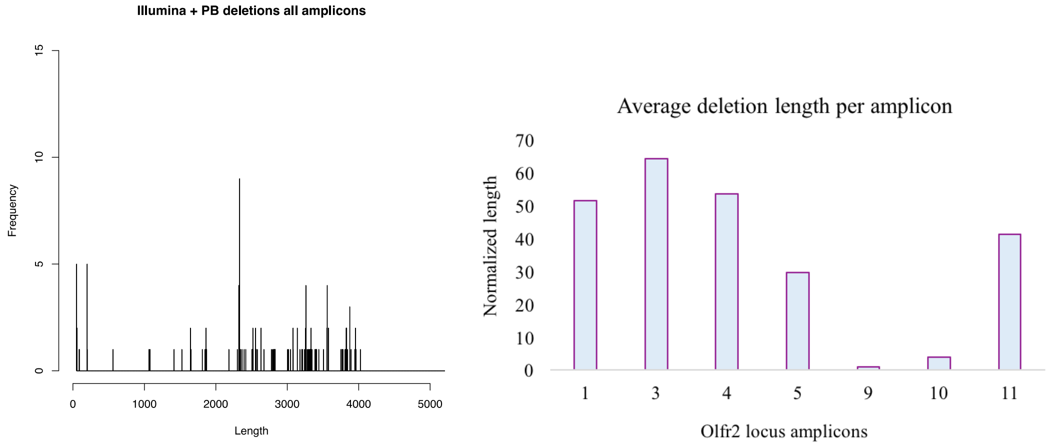


Figure 3.14. Frequency of length distribution for Pac Bio validated deletions. Frequency of length distribution for all Pac Bio-supported deletions is shown (Left panel). Normalized average length of deletions for each amplicon is shown (Right panel).

Comparing deletions distribution among samples, we observed that 19 out of 55 deletions were present in both MDA replicates but not in OE samples, suggesting they could be those implicated in *Olfr2* expression. The majority of MDA-specific deletions were located in amplicon 3, while amplicon 1 presents the highest number of common deletions (Figure 3.15). Nevertheless, we cannot exclude that MDA specific deletions are artifacts since we did not find any OE read supporting them. Importantly, the risk of MDA artifacts can be excluded for the 11 deletions shared by both MDA and OE samples.

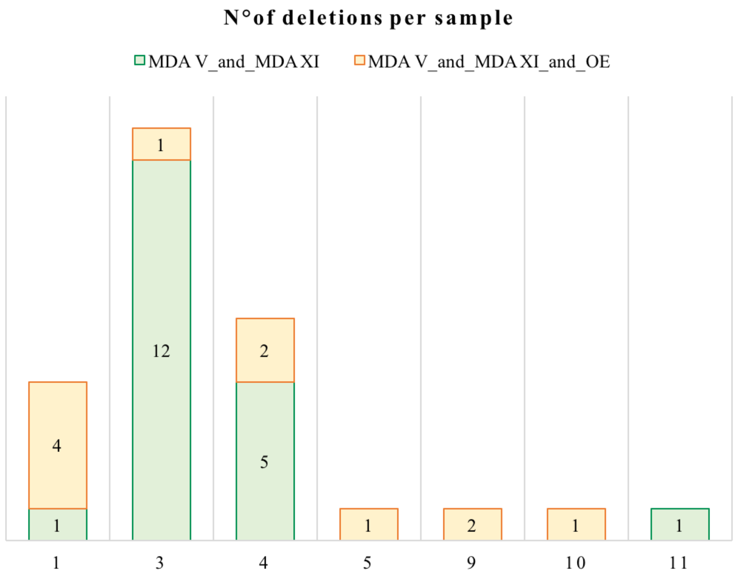


Figure 3.15. Number of Pac Bio-supported deletions per sample. Number of Pac Bio-supported deletions in MDA samples (green) and in MDA and OE samples (yellow) is shown for each amplicon**.**

Finally, for all the Pac Bio supported deletions which were shared by both MDA replicates and OE sample (common deletions), we calculated the Illumina coverage (“reference reads”/ “alternate reads” x 100) comparing the three samples (OE, MDA V and MDA XI) (Figure 3.16). Interestingly, for the majority of Pac Bio-supported common deletions, MDA coverage (in at least one MDA replicate) was higher than OE coverage. In particular, the deletion “2898_3887” in amplicon 3 was characterized by a very high coverage in one MDA replicate (MDA XI) (Figure 3.16).

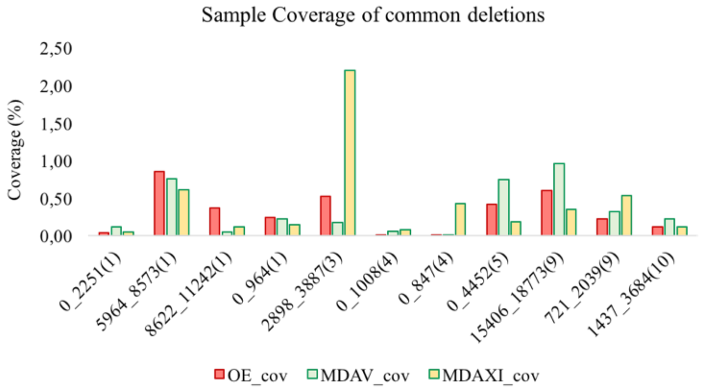


Figure 3.16. Illumina coverage for Pac Bio-supported common deletions. Percentage of Illumina sample coverage is shown for each deletion shared by MDA and OE. Number on x-axis indicated the ID of each deletion for each amplicon.

3.3.2 Independent technical PCR validation of Pindel deletions

In parallel to Pac Bio validation, Pindel deletions were validated independently by qualitative end-point PCR. We decided to perform PCR validation as independent approach in order to overcome the problem of potential lack of saturation for Pac Bio sequencing. In other words, if we considered as true only the deletions supported by both Illumina and Pac Bio reads we could have missed the portion of putative true Pindel deletions not detected by Pac Bio sequencing. Therefore, we decided to select the deletions to be validated among those called by Pindel in the MDA Illumina data regardless of their presence in the Pac Bio ones.

However, one technical problem for designing PCR assay was the high number of Pindel deleted sequences, often differing for just 1 nucleotide and often overlapping each other (Figure 3.17, top panel). Therefore, given the complex deletion pattern for each amplicon, we designed validation primers on a consensus of Illumina reads at the 5’ and 3’ of each deletion (Figure 3.17, bottom panel). With this approach we aimed at detecting by PCR at least the most covered deletions.

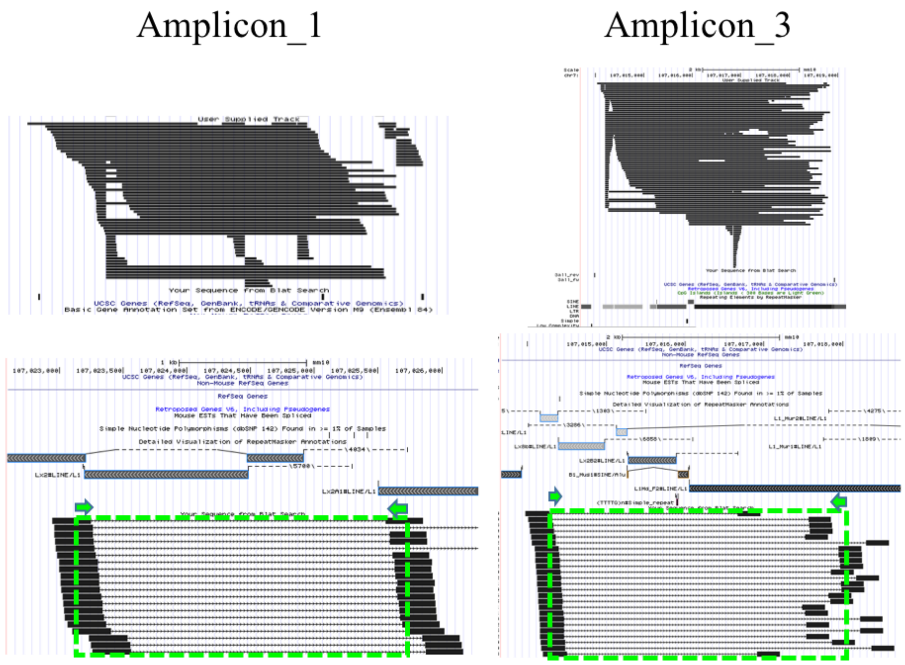

Amplicon_1
Amplicon_3
User Supplied Track
Your Sequence from Blat Search
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
RefSeq Genes
Retroposed Genes V6, Including Pseudogenes
Detailed Visualization of RepeatMasker Annotations

Figure 3.17. Pindel deletions and Illumina consensus reads. Top panels: two representative examples of complex Pindel deletion patterns in amplicon 1 and in amplicon 3. Bottom panels: Illumina consensus reads supporting Pindel deletions in amplicon 1 and amplicon 3. Green arrows indicate validation primers; green rectangle indicates the region amplified with validation assay, consisting of different Pindel deletions. Detailed Repeat Masker annotation for different repetitive elements is shown above the Illumina consensus reads.

Additional challenge in designing specific validation primers was due to the high density of repetitive elements overlapping the deleted sequence (Figure 3.17, bottom panels). Nevertheless, we could design at least one specific PCR validation assay for amplicon 1, amplicon 3, amplicon 4 and amplicon 5.

Technical PCR validations were performed on amplified PCR products preserved from the initial long PCR amplicons (MDA and OE) which were sent for Pac Bio and Illumina sequencing (Figure 3.18).

Putative "deleted" PCR bands were extracted from agarose gel and re-sequenced by Sanger sequencing. Interestingly, for each PCR reaction we were able to amplify both the deleted and the non-deleted sequence, although with variable efficiency (Figure 3.18).

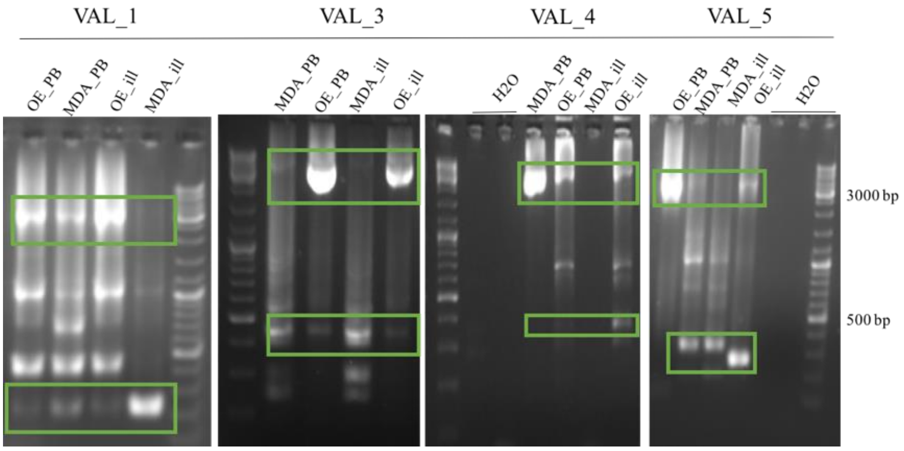


Figure 3.18. Deletion PCR validation results. Green rectangles indicated putative deleted and non-deleted bands for each validation assay. "MDA_PB" and "OE_PB"= MDA and OE PCR amplicon sample sequenced by Pac Bio; "MDA_ill" and "OE_ill"= MDA and OE PCR amplicon samples sequenced by Illumina.

An example of Sanger sequence of an independent PCR product, supporting the deletions according to the Illumina reads pattern, is shown in Figure 3.19.

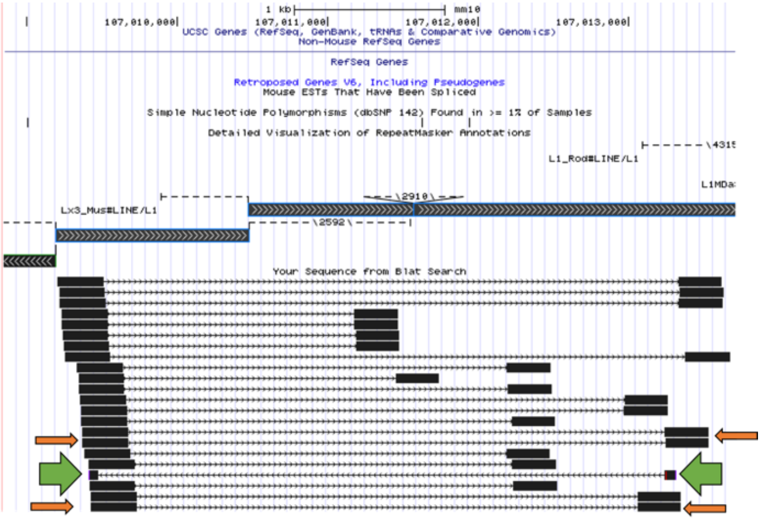

1 kb
mm10
107,010,000
107,011,000
107,012,000
107,013,000
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
Non-Mouse RefSeq Genes
RefSeq Genes
Retroposed Genes V6, Including Pseudogenes
Mouse ESTs That Have Been Spliced
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Detailed Visualization of RepeatMasker Annotations
\4315
L1_Rod#LINE/L1
L1MDa
\2910\
Lx3_Mus#LINE/L1
\2592\
Your Sequence from Blat Search

Figure 3.19. Example of Sanger sequence supporting Pindel deletions. UCSC Genome Browser 5kb screenshot: representative Sanger sequence is indicated by green arrows; Illumina reads supported by Sanger sequence are indicated by orange arrows.

Confirmed Sanger sequences were then intersected with the coordinates of Pindel deletions. From this intersection we could identify the deleted sequences matching with the Sanger sequence of the DNA band amplified by validation primer assays (Figure 3.20).

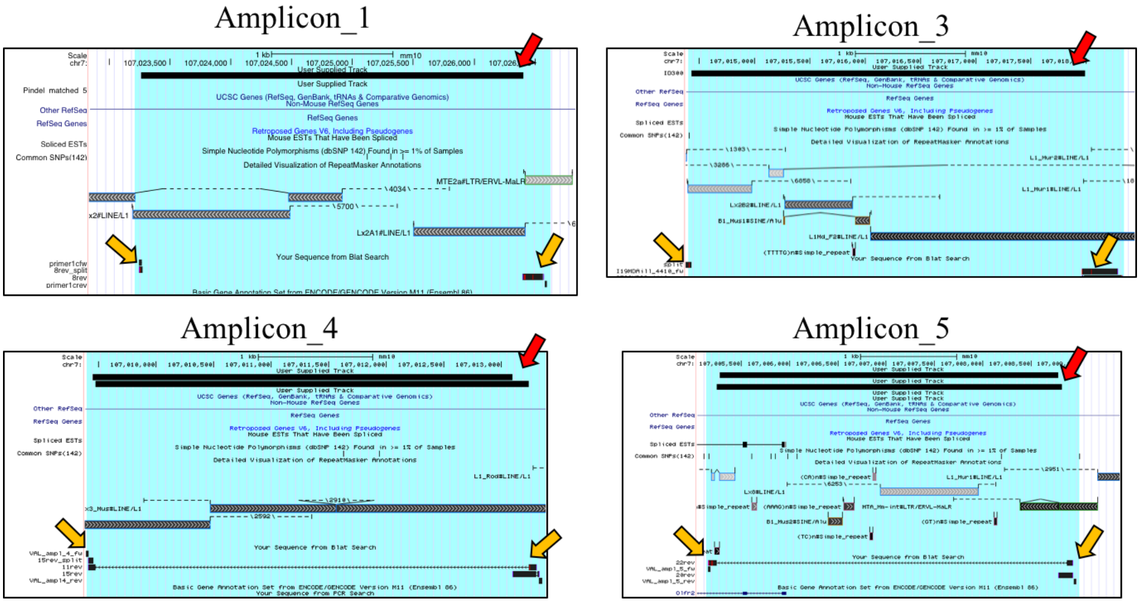

Amplicon_1
Amplicon_3
Amplicon_4
Amplicon_5
User Supplied Track
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
Non-Mouse RefSeq Genes
RefSeq Genes
Retroposed Genes V6, Including Pseudogenes
Mouse ESTs That Have Been Spliced
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Detailed Visualization of RepeatMasker Annotations
Your Sequence from Blat Search
Basic Gene Annotation Set from ENCODE/GENCODE Version M11 (Ensembl 86)

Figure 3.20. Sanger sequences supporting selected Pindel deletions. Sanger sequences for amplicons 1, 3, 4 and 5 are shown with black rectangles and indicated by the yellow arrows. Red arrows indicate the Pindel deletions supported by Sanger sequences. Detailed Repeat Masker annotation is activated.

So far we were able to validated 6 different deletions in about 20 kb upstream the *Olfr2* coding sequence. Validated deletions were found in amplicon 1, 3, 4 and 5 (Table 3.4).

Amplicon 3 deletion was supported in OE sample by only 1 read (under the Pindel coverage threshold of 5 reads), for this reason it was initially classified as MDA-specific deletion.

Interestingly, 3 out of 6 PCR validated deletions were also supported by Pac Bio reads. Moreover, deletions in amplicon 1 and amplicon 3, which were shared between OE and MDA samples, had a higher coverage in MDA compared with OE. Thus suggesting they could be involved in *Olfr2* expression without the risk to MDA artifacts. For these reasons we focused on these deletions for further independent validations.

	Amplicon_1	Amplicon_3	Amplicon_4		Amplicon_5	
PCR_validated	MDA+OE	MDA+OE	OE+MDA	MDA	MDA	MDA
ID_Pindel (best match)	199	300	159	149	260	256
deletion_bp	3141 bp	3555 bp	3766 bp	3646 bp	3507 bp	3571 bp
NON-del_band_bp	3356 bp	3957 bp	3965 bp	3965 bp	3815 bp	3815 bp
expected_del_band_bp	215 bp	402 bp	199 bp	319 bp	308 bp	244 bp
cov_MDA V % (Ref/Alt)	5918/7 **(0,11%)**	51816,5/293 **(0,5%)**	7198/0 **(0%)**	5139/6 **(0,1%)**	14782.5/121 **(0,8%)**	11377/0 **(0%)**
cov_MDA XI % (Ref/Alt)	13924,5/7 **(0,05%)**	173767,5/534 **(0,3%)**	25172.5/46 **(0,2%)**	18846/0 **(0%)**	25803/0 **(0%)**	17390/34 **(0,2%)**
cov OE % (Ref/Alt)	36589/16 **(0,04%)**	118457/1 **(0,0008%)**	8802/0 **(0%)**	6727/0 **(0%)**	14061/0 **(0%)**	7111.5/0 **(0%)**
MDA-PB reads	1	1	0	0	9	0
OE-PB reads	0	0	0	0	0	0

Table 3.4. PCR validated deletions summary information. For each deletion validated by PCR all the related information is summarized. PCR_validated= sample where deletion was validated by PCR; ID_Pindel (best match) = ID of Pindel deletion which is supported by Sanger sequence with the highest nucleotide precision; deletion_bp= length of deleted sequence; NON-del_band_bp= length of non-deleted sequence; expected_del_band= length of expected deleted PCR band; cov_MDA V% , MDAXI%, OE (Ref/Alt)= Illumina reads coverage for each sample calculated by the number of alternate reads (Alt), supporting the deletion and the number of reference reads (Ref), supporting the reference sequence; MDA and OE-PB reads= number of supporting Pac Bio reads for sample.

3.3.3 Amplicon 3-deletion additional validations

So far, we further validate amplicon 3 deletion from bulk OE gDNA and from total MDA amplified DNA (starting material before PCR amplification). With this approach we could exclude even the possibility of being looking at PCR artifacts, possibly produced during the initial *Olfr2* locus PCR amplification (Figure 3.21).

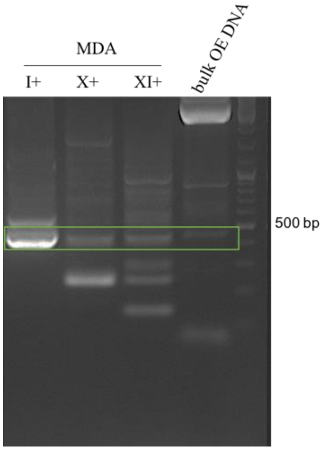


Figure 3.21. Additional PCR validation for amplicon3-deletion. Validated amplicon 3-deletion bands are shown for three MDA replicates (I, X, XI) and from total gDNA extracted from OE (bulk OE DNA) in the green rectangle.

3.3.4 Deletions distribution with respect to *Olfr2* locus repetitive elements

In order to assess if LINE-1 elements could be involved in deletions formation, we start investigating how deletions are distributed with respect to different classes of annotated repetitive elements in 50 kb of *Olfr2* locus. Within region analyzed, 80 different repetitive elements are annotated by Repeat Masker, covering a total of 34805 bp. The remaining 15191 bp consist of genomic DNA. Among these repeats, 12 LTR (3944 bp), 11 SINE (1594 bp) and 29 LINE (27341 bp) fragments are annotated (Figure 3.22).

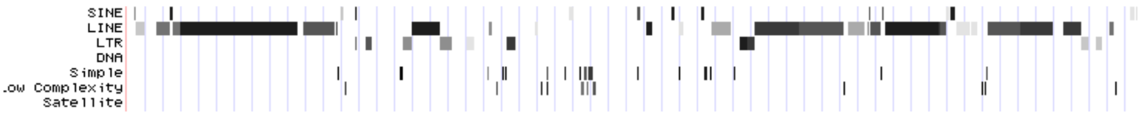

50kb Olfr2 locus

Figure 3.22. Representation of repetitive elements annotated in 50kb *Olfr2* locus. UCSC Genome Browser screenshot of 50 kb *Olfr2* locus. Black rectangles indicate a "dense" UCSC representation of annotated repeats.

In order to normalize for differences in locus occupancy by different classes of repeats, we calculated percentage of bases covered by deletions. Pindel deletions covered 98 % of LINE, 85 % of SINE and 79 % LTR occupied regions (Figure 3.23). Interestingly, when considering the deletions supported by both Illumina and PacBio reads, the loss of LINE sequences (from 98 to 77% of bp covered) was less pronounced compared with SINEs and LTRs coverage (from 85 to 46% and from 79 to 36% of bp covered, respectively).

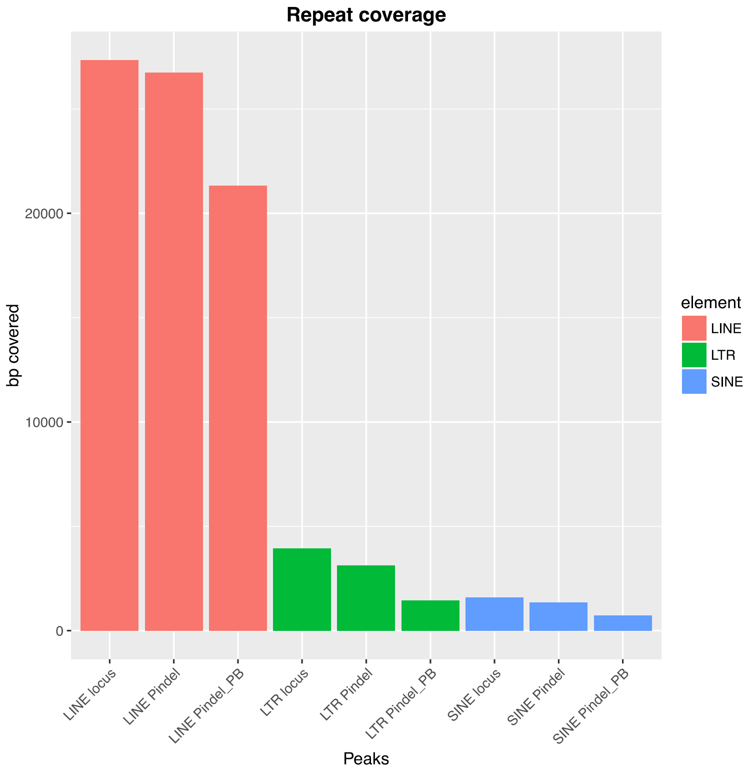


Figure 3.23. Repeat coverage for *Olfr2* deletions. Deletions bp coverage for different classes of repetitive elements is shown. LINE coverage is represented in red, LTR coverage in green and SINE coverage in blue. For each class, the number of bp covered in the locus (first column for each group), the number of bp covered by Pindel deletions (second column), and the number of bp covered by Pindel deletions supported by Pac Bio (PB) reads (third column), is shown.

3.3.5 Direct repetitive sites (DRSs) delimiting validated deletions

Looking more in details at the six validated deletions we noticed that they were all characterized by the presence of two annotated short direct repetitive sequences at the 5' and 3' of the deletion, respectively. Interestingly, the resulting deleted sequence harbors only one of the two repeat sites (Figure 3.24).

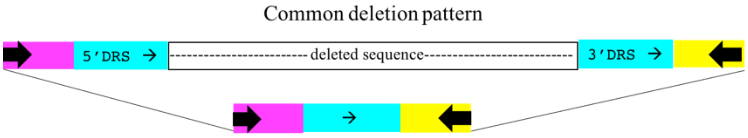

Common deletion pattern
5'DRS →
------------------------ deleted sequence ------------------------
3'DRS →
→

Figure 3.24. Common structural pattern of validated deletions. Non-deleted and resulting deleted sequences are shown. Black filled arrows represent validation primers, violet rectangles and yellow rectangles represent sequences at 5' and 3'of the deletion, respectively. DRSs are highlighted in light blue.

We referred to these peculiar repeats as “Direct Repetitive Sites” (DRSs), interestingly they were annotated on different LINE-1 elements located at the deletion borders.

A representative example of amplicon 3 deletion pattern and DRSs sites is shown in (Figure 3.25).

Looking at the amplicon 3-deletion, we noticed that the 3555 bp deleted sequence was characterized by several annotated LINE-1 elements harboring different alternative LINE-1 EN cutting sites. In particular, “5’-TTAGAA-3’” and “5’-CTAAAA-3’” sites were repeated once, “5’-TTGAAA-3’” and “5’-TTAAAG-3’” were repeated three times in the deleted regions.

Amplicon_3 reference sequence

ACATGTTCTGTCTTGTTTGTGAGGTTTACCCTAAGCTTTTT
ATTAGACATATTGAATTTTAAATATTCTTCTTATTTCTTTG
AAAAATGTCATGCAGTGTATTTTGAT

-----------------3555 bp ----------------------------

ATTTTGATGGGGATTGCATTGAATCTGTAGATTGCTTTTGG
CAAGATAGCCATTTTTACAATGTTGATCCTGCCAATCCATG
AGCATGGGAGATCTTTCCATCTTCTGAGATCTTCTTTAATT
TCTTTCTTCAGAGACTTGAAGTTCTTGTCATACAGATCTTT
CACTTCCTTAGTTAGAGTCACGCCAAAGTATTTTATATTAT
TTGTGACTATTGAGAAGGGTGTTGTTTCCCTAATTTCTTTC
TCAGCCTGTTTATCCTTTGTGTACAGAAAGGCCATCGACTT
GTTTGAGTTAATTTTATATCCAGCTACTTCATTGAAGC**TGT**
TTATCAGGCTTTAGGAG

Amplicon_3 deleted sequence

ACATGTTCTGTCTTGTTTGTGAGGTTTACCCTAAGCTTTTT
ATTAGACATATTGAATTTTAAATATTCTTCTTATTTCTTTG
AAAAATGTCATGCAGTGTATTTTGATGGGGATTGCATTGAA
TCTGTAGATTGCTTTTGGCAAGATAGCCATTTTTACAATGT
TGATCCTGCCAATCCATGAGCATGGGAGATCTTTCCATCTT
CTGAGATCTTCTTTAATTTCTTTCTTCAGAGACTTGAAGTT
CTTGTCATACAGATCTTTCACTTCCTTAGTTAGAGTCACGC
CAAAGTATTTTATATTATTTGTGACTATTGAGAAGGGTGTT
GTTTCCCTAATTTCTTTCTCAGCCTGTTTATCCTTTGTGTA
CAGAAAGGCCATCGACTTGTTTGAGTTAATTTTATATCCAG
CTACTTCATTGAAGC**TGTTTATCAGGCTTTAGGAG**

Figure 3.25. Structural conformation of amplicon 3-deletion. Left panel: reference, non-deleted sequence. Sequences at the 5’ and 3’ of the deletion are highlighted in violet and yellow, respectively. Lentgh of deleted sequence is indicated. DRSs are highlighted in light blue. Forward and reverse validation primers are indicated in bold font. Right panel: non-annotated deleted sequence. Portion at the 5’ and 3’ of the deletion are highlighted in violet and yellow, respectively. The left DRS is highlighted in light blue. Forward and reverse validation primers are indicated in bold font.

We hypothesized a deletion supporting mechanism mediated by the presence of DRSs, flanking the deletion, and by LINE-1 EN cutting sites within the deletion. According to this mechanism, deletions could be generated after LINE1-EN-dependent DSB repair, mediated by homologous recombination between DRSs.

3.4 ChIP-Seq analysis of endogenous γ-H2AX in mouse OE and L

Given the potential involvement of endogenous LINE1-EN-dependent DSBs in the deletions formation, we start investigating the expression and the genomic localization of a DSBs marker (γ-H2AX) in mouse tissues, under physiological conditions.

3.4.1 Preliminary results: endogenous γ-H2AX expression in mouse tissues

Before investigating endogenous γ-H2AX genomic distribution in mouse OE we needed to assess whether the phosphorylated form of the histone was detectable in different mouse tissues under physiological conditions. Therefore, we performed western blot for γ-H2AX in two neural tissues (OE and cerebellum (Cer)) and two non neural ones (liver (L) and kidney (K)). Each tissue was analyzed at the two different ages of 6 days and 30 days after birth (Figure 3.26).

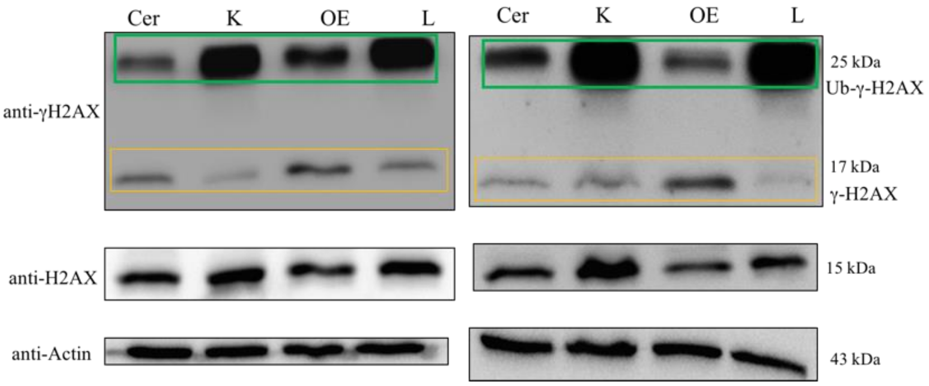


Figure 3.26. WB anti-γH2AX in different mouse tissues. Whole tissue protein lysates from C57BL/6J mice at p6 (left panel) and 1m (right panel) were analyzed for endogenous γ-H2AX expression. Cer, cerebellum; K, kidney;OE, olfactory epithelium; L, liver. Phosphorylated form of H2AX (γ-H2AX) is shown at 17 kDa (yellow square) and phosphorylated plus monoubiquitinated H2AX (Ub-γH2AX) is shown at 25 kDa (in the green square). Total H2AX and β-actin were developed on the same membrane.

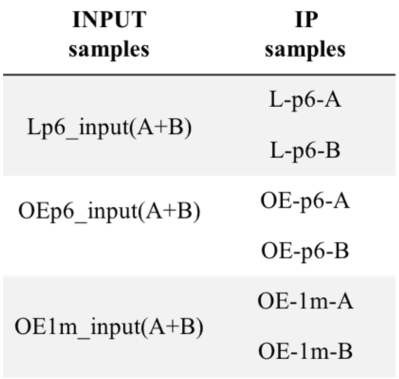
At both the ages analyzed, with the same anti-γ-H2AX antibody we were able to detect two different and specific signals, the expected phosphorylated H2AX (γ-H2AX) at 17 kDa and the mono-ubiquitinated and phopshorylated form of the histone (Ub-γ-H2AX) at 25kDa. In general, all the samples in all the conditions showed a stronger Ub-γ-H2AX signal compared with γ-H2AX. Interestingly, the Ub-γ-H2AX was more intense in non-neural tissues compared with neural ones, while the γ-H2AX signal presented an opposite trend.

Additional replicas and further experiments should be performed to better understand the biological meaning which could underlie tissue differences in Ub-γ-H2AX/γ-H2AX ratio.

3.4.2 General characterization of γ-H2AX peaks

In order to investigate how γ-H2AX distributes in the mouse genome we performed a chromatin immuno-precipitation (IP) and sequencing experiment in C57BL/6J mice, analyzing OE at two ages (6 days (p6), 1 month (1m), after birth) and liver (L) at p6. For each IP experiment, OE and L were pooled together from about 10 mice in order to get a suitable quantity of chromatin.

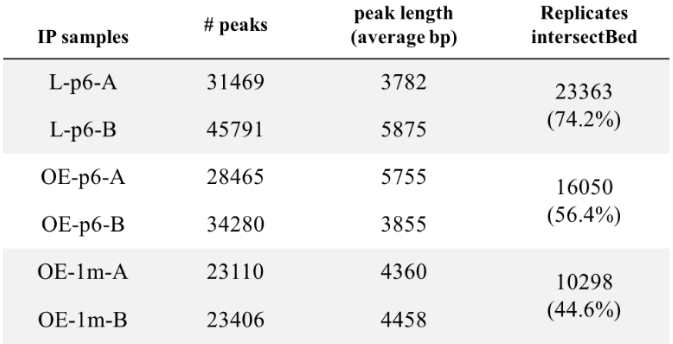
For each condition, we sequenced two different biological IP replicates each derived from different pools of mice. In parallel we sequenced a same quantity of INPUT sample (total starting chromatin) as control (Table 3.5).



INPUT samples	IP samples
Lp6_input(A+B)	L-p6-A
	L-p6-B
OEp6_input(A+B)	OE-p6-A
	OE-p6-B
OE1m_input(A+B)	OE-1m-A
	OE-1m-B

Table 3.5. ChIP-Seq samples sequenced by Illumina sequencing. For each sample two biological replicates (A and B) were sequenced. Input chromatin for each sample was sequenced as background control.

Peak calling was performed by SICER tool (see details in Methods and Materials). To obtain a representative set of ChIP-seq peaks for each biological condition (L at p6, OE at p6 and OE at 1 month) we considered the intersection of the peak sets obtained in the two replicates and these were used in subsequent analyses. The number of peaks called for each sample, their length and the number of intersection-peaks are reported in Table 3.6.



IP samples	# peaks	peak length (average bp)	Replicates intersectBed
L-p6-A	31469	3782	23363 (74.2%)
L-p6-B	45791	5875	
OE-p6-A	28465	5755	16050 (56.4%)
OE-p6-B	34280	3855	
OE-1m-A	23110	4360	10298 (44.6%)
OE-1m-B	23406	4458	

Table 3.6. **ChIP-seq peak calling output.** Peaks were called with EPIC tool, number of peaks and peak lengths for each biological replicate are indicated. For each sample peaks from two biological replicates were intersected and intersected sample datasets were used in following analysis.

From an intersection among the peak datasets resulted that almost 6000 peaks were shared among the three samples, 2000 peaks were found only in OEp6, 6000 only in OE1m and almost 12000 peaks were exclusive for Lp6.

For all the following analysis we used peak datasets resulted from peaks intersection between two biological replicates for each sample.

A subset of shuffled peaks was generated as background control dataset to use in all the analysis.

γ-H2AX peaks genome distribution

To visualize γ-H2AX peaks distribution in mouse genome we used NEBULA, a tool able to calculate for each sample the proportion of peaks falling within different genomic features (e.g. within genes, intergenic, 5’UTRs ect).

For both OE and L datasets about the 90% of peaks fall within the regulatory and transcribed portion of the genome (gene bodies, promoters, 5’UTRs and enhancers), while only a 10% fall in intergenic regions (Figure 3.27).

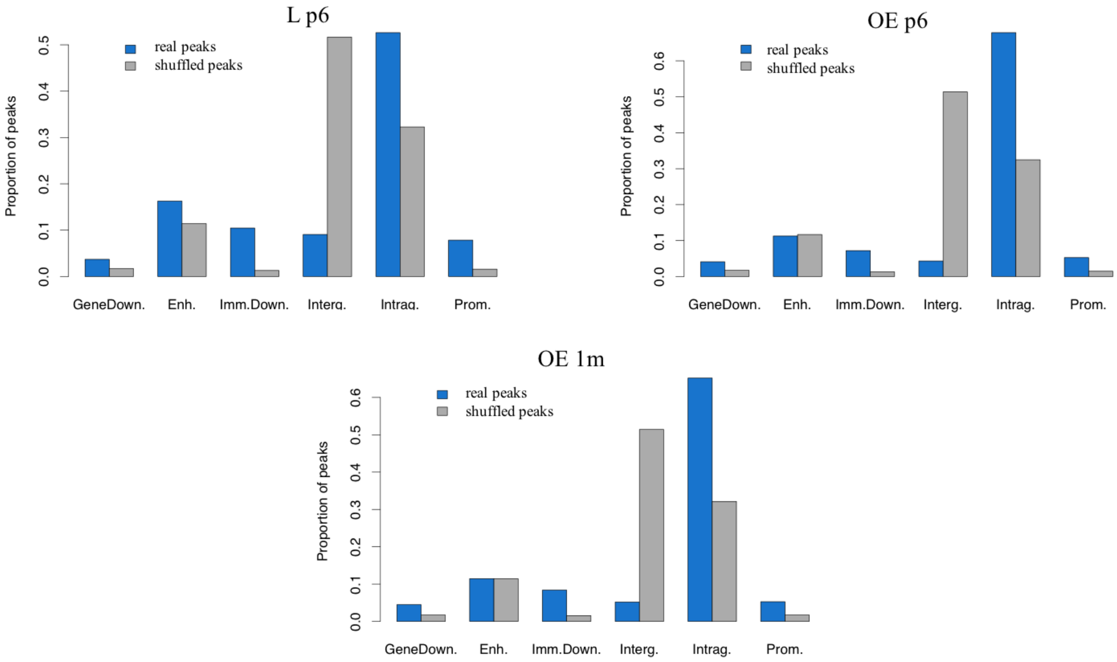


Figure 3.27. ChIP-seq peaks genome annotation. Peaks genome distribution was performed with the bioinformatics tool NEBULA; the genomic regions were considered with respect to gene start site (TSS). For each sample the proportion of peaks falling in each genomic region is shown. NEBULA legend: Gene Down=gene downstream (3'UTR), Ehn= enhancer, Imm.Down.=Immediate downstream (5'UTR), Interg.=intergenic, Intrag.=intragenic, Prom=promoter. Real sample peaks were represented in blue, shuffled peaks were represented in grey.

γ-H2AX peaks distribution with respect to Transcription Start Sites (TSSs)

Using NEBULA, we were able to investigate peaks distribution with respect to annotated TSSs. Shuffled peaks did not show any enrichment at TSSs, while IP sample peaks were enriched within 0 and 10kb downstream the TSS (Figure 3.28).

L sample seems to have sharper distribution at TSSs than OE samples, some differences between OE and L peaks distribution at TSSs could be due to poorer TSSs annotation in OE tissue compared with L.

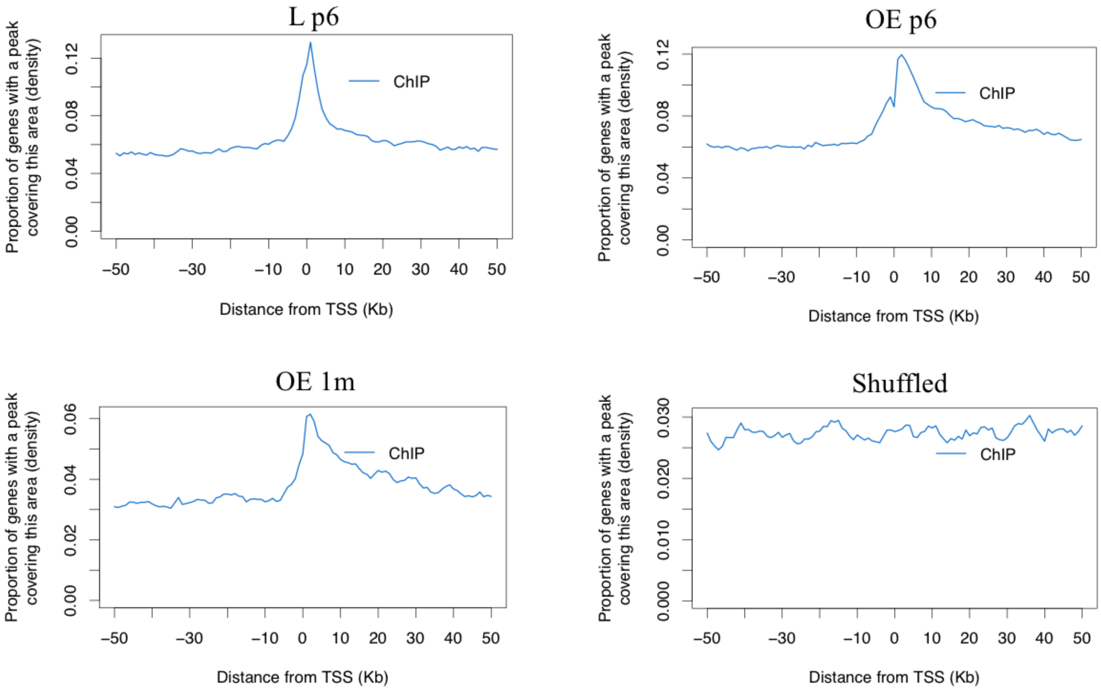


Figure 3.28. ChIP-Seq peaks distribution around TSS. Peaks genome distribution with respect to annotated TSS was performed with the bioinformatics tool NEBULA. For real and shuffled peak datasets peak density was plotted with the distance from annotated TSSs.

γ-H2AX peaks localization at GpC islands

Given the poor characterization of γ-H2AX in physiological conditions, we wondered whether identified sites of γ-H2AX enrichment identified co-localize with CpG islands, as it could be a feature of γ-H2AX peaks to prefer being deposed on CpG islands or avoid them.

Sample_dataset	CpG-overlapping peaks (%)
Lp6	21.66
OEp6	24.21
OE1m	20.88
shuffle	3.01

Table 3.7. Percentage of peaks overlapping CpG islands. For each sample the percentage of peaks overlapping CpG islands is shown.

We observe that about 20 to 24% of the peaks identified overlap at least one annotated CpG island (Table 3.7). This is in sharp contrast with the set of randomly created ChIP-seq peaks, only about 3% of which overlaps a CpG island. This suggests that at least part of the γH2AX-bound DNA in physiological conditions is associated to CG rich regions.

3.4.3 γ-H2AX peaks functional annotation

Gene Ontology enrichment analysis

In order to assess any functional enrichment for γ-H2AX peaks we took advantage of the software GREAT, a bioinformatics tool specific for ChIP-Seq peak annotation and gene ontology (GO) enrichment. In particular, we looked for "Biological process" enrichment of each sample dataset (foreground dataset) against total number of sample peaks (background dataset). Surprisingly, L and OE sample peaks turned out enriched for different biological processes (Figure 3.29).

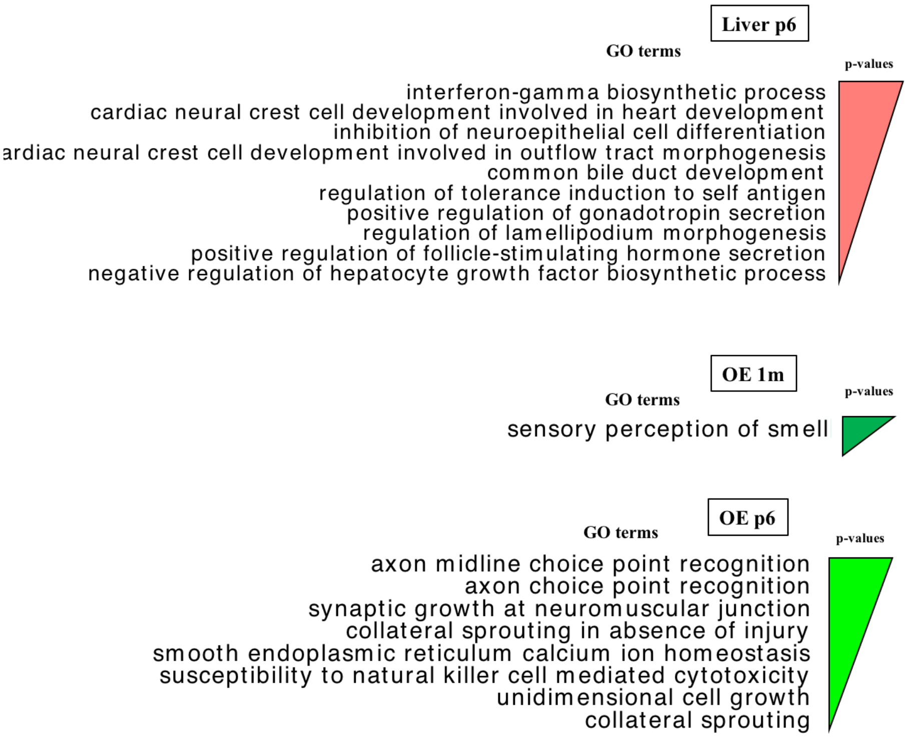


Figure 3.29. ChIP-seq peaks enrichment for GO Biological Process. "Biological Process" GO enrichment analysis was performed by GREAT tool. Top 10 GO categories were shown for each sample. For each peaks the nearest TSS was annotated. Only p-values < 0.0001 were considered in the output results.

Some interesting processes enriched in L were "interferon-gamma biosynthetic process", "common bile duct development" and "negative regulation of hepatocyte growth biosynthetic process"; OE p6 was enriched for biological processes linked to “axon choice point recognition” and “collateral sprouting”; surprisingly “sensory perception of smell” was the only biological function enriched in OE 1m.

Overall GO results suggested some tissue specificity for peaks functional annotation.

3.4.4 γ-H2AX peaks distribution within different regulatory sites

To compare peaks distribution with respect to different regulatory sites, we took advantage of publicly available datasets, looking for the best possible match with OE and L tissues. While L datasets were readily available, we couldn’t find much data for the OE: we used instead olfactory bulb, cortex and whole brain tissues, as the closer tissue types we could find data for.

CTCF, RNA Pol II and DNAseI regulatory sites

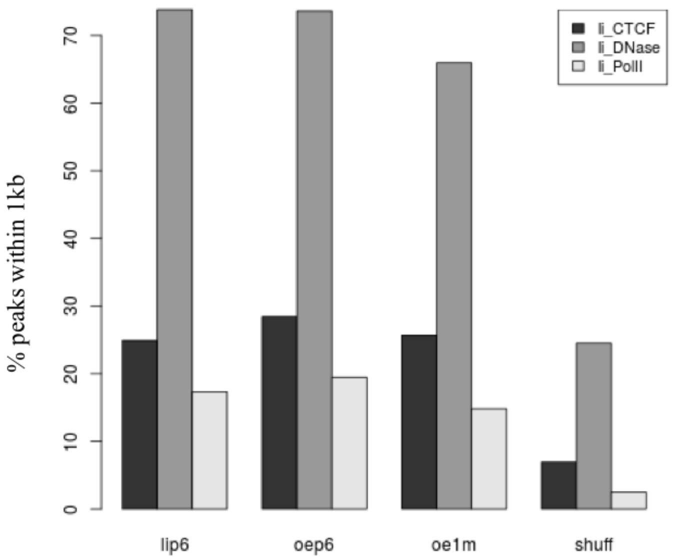


Figure 3.30.**ChIP-Seq peaks overlap within DNAse,CTCF and PolIII regulatory sites.** Results obtained from mouse liver datasets are shown. lip6= liver p6 peaks; shuff= shullfe peaks. For each sample the percentage of peaks falling within 1 kb from a regulatory site is shown.

The majority of the peaks (about 70-75%) for each sample fell within 1kb of a DNAseI hypersensitive site (Figure 3.30), confirming what obtained in the work of Lensing and colleagues (Lensing et al., 2016). This result suggests a link between regulatory chromatin and genome instability. Looking at shuffled dataset, this was true only for 25% of peaks.

A similar analysis was performed to investigate peak distribution with respect to CTCF binding sites. CTCF is a regulatory protein responsible for the formation, together with cohesion factors, of chromatin loops between different parts of the genome. Moreover, CTCF can be involved in different gene regulatory pathways acting as repressor or activator of gene transcription. Interestingly we observed that about 30% of the peaks fall within CTCF binding sites, for each sample (Figure 3.30).

We further investigated γ-H2AX peaks distribution with respect to RNA Pol II binding sites. About 10-15% of peaks for each sample fall within Pol II binding sites suggesting that a subset of γ–H2AX peaks could be involved in transcription (Figure 3.30).

Similar results were obtained comparing the peaks with brain and olfactory bulb datasets as presented in the Appendix (Figure-B 5).

Chromatin segmentation analysis

To gain further insight about the localization of γ-H2AX with respect to other chromatin marks, we used publicly available ChIP-seq data on PolII, CTCF and 7 histone modifications to create a 10-state segmentation of the mouse genome based on combinatorial re-occurrence of subsets of the marks under analysis (see details in Methods and Materials).

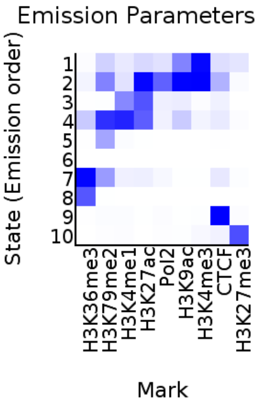
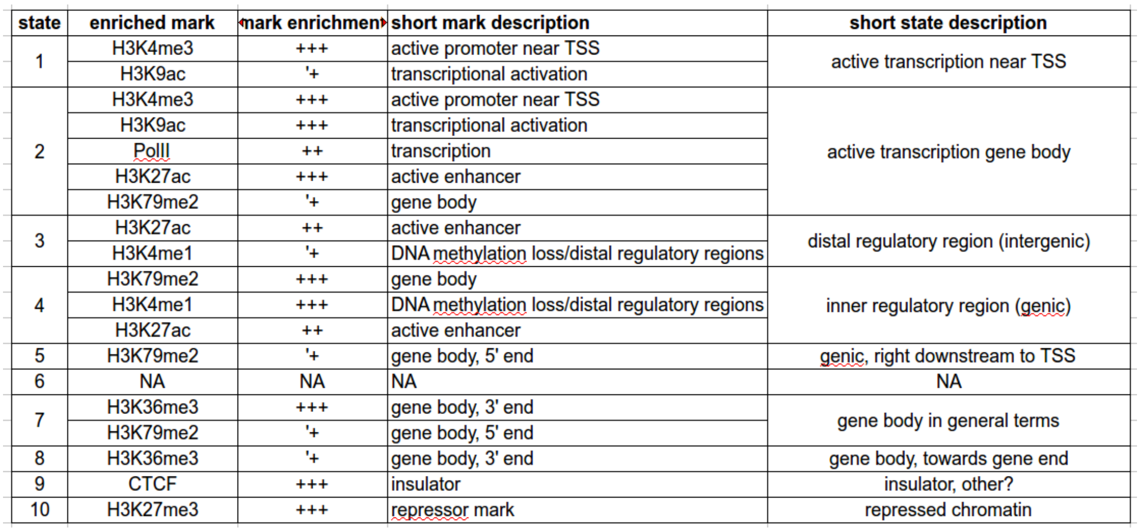


Figure 3.31. Chromatin segmentation of mouse genome. 10-state segmentation of mouse genome. Publicly available ChIP-seq data on PolII, CTCF and 7 histone modifications were used. White and blue colors represent, respectively, no enrichment and tha maximum enrichment.

As summarized above, this means in practice that each position within the chromatin is assigned to one of 10 possible states, each one characterized by enrichment in one or more chromatin marks (with the exclusion of state 6, for which none of the chromatin marks is enriched) (Figure 3.31).

To create a human readable annotation of the identified states and an interpretation of their meaning, we created a summary of the marks enriched for each state and corresponding function and used these to formulate a short description for the state itself (Table 3.8).



state	enriched mark	mark enrichmen	short mark description	short state description
1	H3K4me3	+++	active promoter near TSS	active transcription near TSS
	H3K9ac	'+	transcriptional activation	
2	H3K4me3	+++	active promoter near TSS	active transcription gene body
	H3K9ac	+++	transcriptional activation	
	PolII	++	transcription	
	H3K27ac	+++	active enhancer	
	H3K79me2	'+	gene body	
3	H3K27ac	++	active enhancer	distal regulatory region (intergenic)
	H3K4me1	'+	DNA methylation loss/distal regulatory regions	
4	H3K79me2	+++	gene body	inner regulatory region (genic)
	H3K4me1	+++	DNA methylation loss/distal regulatory regions	
	H3K27ac	++	active enhancer	
5	H3K79me2	'+	gene body, 5' end	genic, right downstream to TSS
6	NA	NA	NA	NA
7	H3K36me3	+++	gene body, 3' end	gene body in general terms
	H3K79me2	'+	gene body, 5' end	
8	H3K36me3	'+	gene body, 3' end	gene body, towards gene end
9	CTCF	+++	insulator	insulator, other?
10	H3K27me3	+++	repressor mark	repressed chromatin

Table 3.8. Summary of marks enriched for each state. For each chromatin state, enriched mark, level of enrichment, mark description and short state description are indicated. +++= maximum enrichment; += minimum enrichment; NA= no enrichment.

We used built-in features of the ChromHMM framework (see details in Methods and Materials) to annotate the γ-H2AX peak sets. As shown in Figure 3.32, the three real peak sets show similar enrichment patterns, with states 4, 2 and 7 being the ones with the strongest enrichment. This is consistent with the observation that a very large proportion of the peaks is located in gene bodies and suggests that this deposition pattern is linked to a regulatory function of γ-H2AX within the gene bodies in physiological conditions. As expected, and consistently with previous observations, for the set of randomly distributed peaks no enrichment is detected for all states.

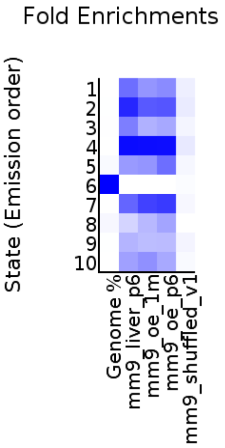


Figure 3.32. Sample fold enrichment among different chromatin states. White and blue colored squares represent, respectively, no enrichment and tha maximum enrichment.

Peaks overlapping active enhancers

Given that the combination of H3K27ac and H3K4me1 marks active enhancers (Calo and Wysocka 2013) we extracted the set of peaks overlapping chromatin segments in state 3 (Table 3.9**).**

Sample_dataset	State-3-overlapping peaks
Lp6	20.76%
OEp6	19.57%
OE1m	16.36%
shuffle	3.73%

Table 3.9. Percentage of peaks overlapping active enhancers. For each sample, the percentage of peaks in overla with state-3 marks (active enhancers) is indicated.

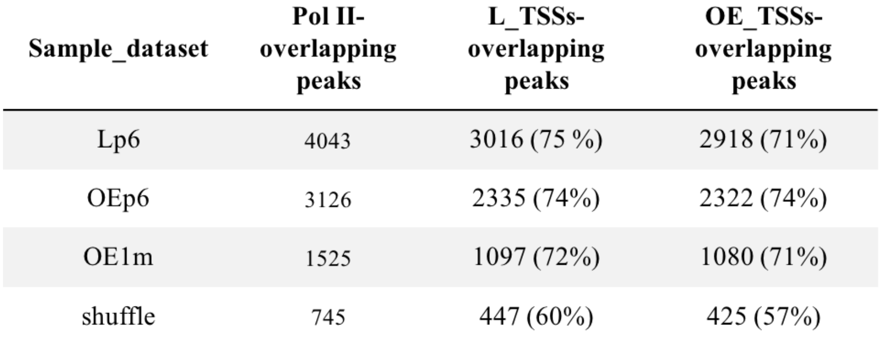
This suggests that at least part of the γ-H2AX signal co-localizes with putative enhancers.

3.4.5 γ-H2AX peaks correlation with expression

Pol II-overlapping γ-H2AX peaks correlation with OE and L active TSSs

We first investigated whether the ca. 15% of γ-H2AX peaks overlapping Pol II binding sites were also in overlap with an active TSS, in OE or L datasets (see details in

Methods and Materials for information about the expression datasets used). For each sample, the number of Pol II-overlapping peaks associated to an active TSS in L and OE datasets is shown in Table 3.10.



Sample_dataset	Pol II-overlapping peaks	L_TSSs-overlapping peaks	OE_TSSs-overlapping peaks
Lp6	4043	3016 (75 %)	2918 (71%)
OEp6	3126	2335 (74%)	2322 (74%)
OE1m	1525	1097 (72%)	1080 (71%)
shuffle	745	447 (60%)	425 (57%)

Table 3.10. Pol II-overlapping peak correlation with active TSSs. The column "Pol II-overlapping peaks" indicates the number of γ-H2AX peaks overlapping PolII binidng sites; the column "L_TSSs-overlapping peaks" indicates the Pol2-overlapping peaks that were also associated to a TSS active according to the FANTOM5 liver dataset (numbers in brackets indicate the corresponding percentage out of the total number of Pol2-associated peaks); similarly the column "OE_TSSs-overlapping peaks" indicates the number of PolII-associated peaks overlapping one active TSS according to the OE dataset.

As expected there is a very high concordance between the peaks overlapping Pol II signal and peaks overlapping an active TSS (about 71 to 75 %). This is remarkably high given that the techniques and the samples used are fairly different.

γ-H2AX peaks correlation with OE and L active TSSs

Comparison with TSSs active in the OE

OE-active TSSs overlapping γ-H2AX peaks belonging to any of the three considered peak sets have higher expression levels with respect to OE-active TSSs that do not overlap γ-H2AX peaks. We do not observe the same effect for OE-active TSSs overlapping a mock-up set of peaks randomly distributed along the genome (Figure 3.33, left panel).

Even though there are no major differences between expression levels of OE-active TSSs within/outside the three peak sets considered, we observe a trend further suggesting at least a certain degree of tissue specificity (Figure 3.33, right panel).

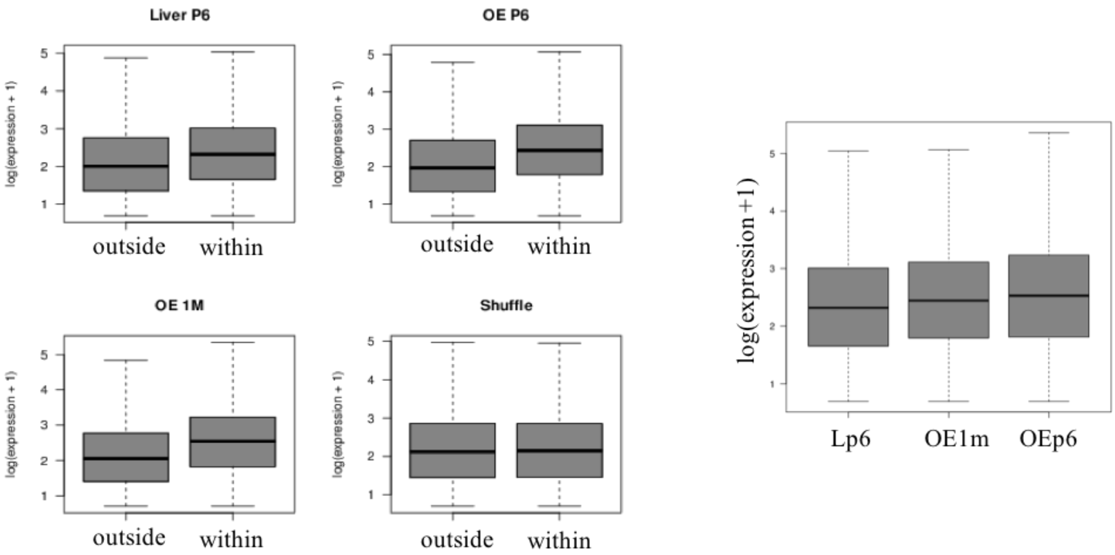


Figure 3.33. **ChIP-seq peaks comparison with active OE-TSS.** Left panel: per each sample, the expression values of OE-active TSSs falling within and without the peaks are compared. Right panel: fold induction of OE-active TSSs expression values is shown for each sample.

Comparison with TSSs active in L

A similar comparison performed in L tissue doesn't give as a stark evidence that in general TSSs active in L overlapping γ-H2AX peaks have higher expression than those located more far away from the peaks (Figure 3.34). We cannot exclude that the weaker effect might be due to technical reasons. Possible explanations include the fact that: the nature of CAGE data is very different from RNAseq data, there are less samples and therefore the signal is likely less robust, the age of the mice from which the liver was derived is fairly different from what was used for ChIP-seq library preparation.

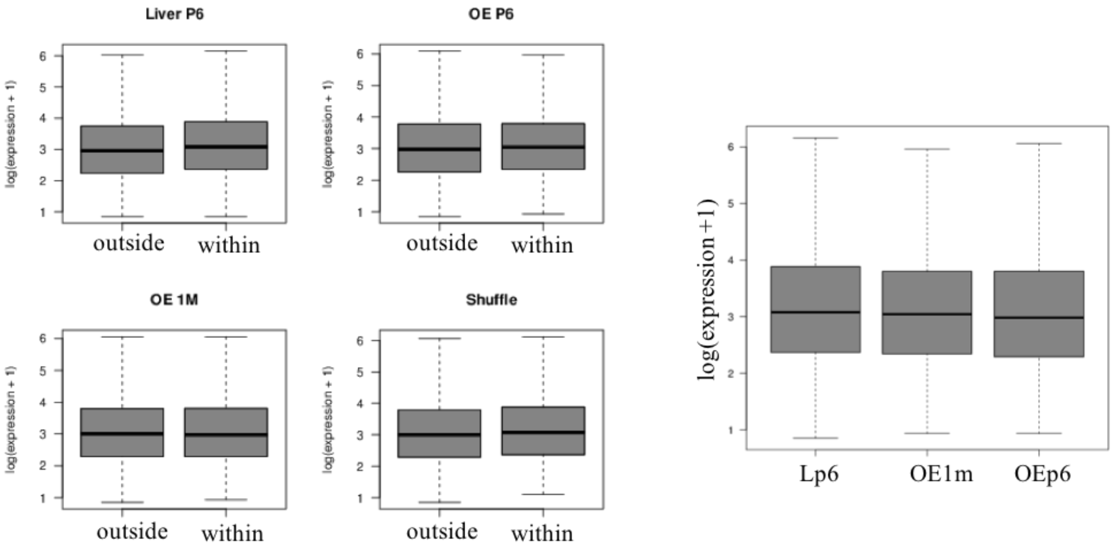


Figure 3.34. ChIP-seq peaks comparison with active L-TSS. Left panel: per each sample, the expression values of L-active TSSs falling within and without the peaks are compared. Right panel: fold induction of L-active TSSs expression values is shown for each sample.

ChIP-seq signal is stronger for peaks overlapping active TSSs in both L and OE

In addition, we evaluated peaks correlation with active TSSs considering peak p-values, which give a quantification of the strength of the corresponding ChIP-seq peaks. We first compared the p-values of ChIP-seq peaks overlapping active TSSs in both L (Figure 3.35,top panels) and OE (Figure 3.35, bottom panels). It is evident that peaks overlapping active TSSs have more significant p-values associated to them.

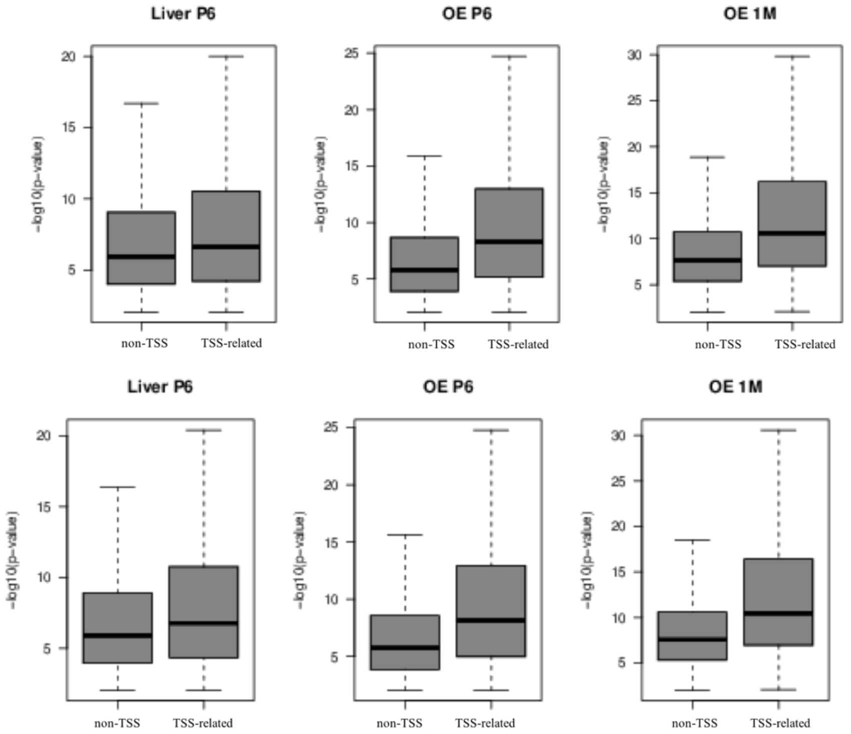


Figure 3.35.Comparison of peaks p-values with respect to active-TSSs. Top panels are referred to TSSs active in L and bottom panels to TSSs active in OE. For each sample the p-values of peaks associated to a TSS is compared with the p-values of peaks not associated to a TSS.

ChIP-seq peaks overlapping active TSSs in both L and OE

Since there is a notably similar proportion of TSSs overlapped in the comparison with the two distinct expression datasets, we investigated the number of γ-H2AX peaks overlapping active TSSs, in both OE and L datasets.

Checking the expression data, we can say that there is in general a high overlap between peaks overlapping active TSSs in the two datasets (Table 3.11). That is, most γ-H2AX peaks that overlap an active TSS in the L dataset overlap an active TSS in the OE dataset. Given that γ-H2AX peaks are not extremely long, this means in most cases that the same expressed gene is overlapped.

	L_ TSSs-overlapping peaks	OE_ TSSs-overlapping peaks	L and OE- TSSs-overlapping peaks	L and OE-TSSs-overlapping peaks (percentage)
Lp6	4040	4490	3541	87.6
OEp6	3119	3630	2802	89.8
OE1m	1524	1935	1352	88.7

Table 3.11. **Peak overlap with active TSSs.** The column “L_TSSs-overlapping peaks” indicates the number of γ-H2AXpeaks that were associated to a TSS active according to the FANTOM5 liver dataset; similarly the column “OE_TSSs-overlapping peaks” indicates the number of γ-H2AXpeaks that were associated to a TSS active according to the RNA-seq OE dataset; the column “L and OE_TSSs-overlapping peaks” indicates the number of γ-H2AXpeaks that were associated to a TSS active in both L and OE datasets. The last column expresses the percentage of γ-H2AXpeaks that were associated to a TSS active in both L and OE datasets.

3.4.6 γ-H2AX peaks enrichment for different classes of repeats

In agreement with the previously observed γ-H2AX signal distribution within the transcribed portion of the genome, we observed a clear enrichment of γ-H2AX peaks for SINE elements. Among different classes of repetitive elements, SINEs are indeed those associated with actively transcribed genes (Figure 3.36).

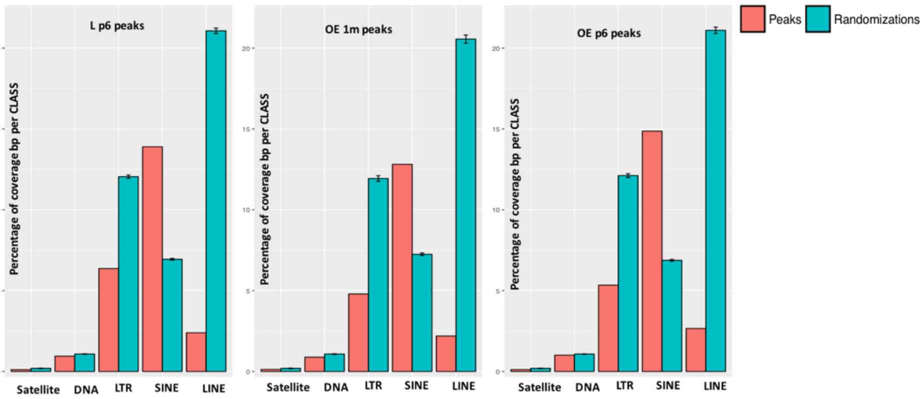


Figure 3.36. ChIP-seq peaks enrichment for different classes of repetitive elements. For all the samples the percentage of peaks coverage (bp) for each class of repeats is shown. Real peaks are shown in red and random peaks are shown in green. Black lines represent standard deviation.

3.4.7 γ-H2AX peaks within gene clusters

Chip-seq analysis of CTCF binding sites in human fibroblasts showed that CTCF-depleted domains included clusters of related genes that are transcriptionally co-regulated, supporting a role for CTCF-binding sites acting as insulators. Gene clusters, among which OR-clusters, are surrounded by a pair of consecutive CTCF-binding sites in the genome (Kim at al., 2006).

Interestingly, we previously observed that 30% of γ–H2AX peaks overlapped with CTCF binding sites (Figure 3.30). Therefore, we wanted to verify if 30% of CTCF-overlapping γ-H2AX peaks follow the same CTCF localization with respect to olfactory clusters. We compared the distribution of CTCF peaks (public dataset) and CTCF-overlapping γH2AX peaks, with respect to OR gene clusters considering different intervals, within and outside the clusters.

Overall, as expected, CTCF-overlapping γH2AX peaks from both OE and L were depleted inside olfactory clusters but preferentially distributed outside the co-regulated regions, thus confirming their distribution according to CTCF signal (Figure 3.37).

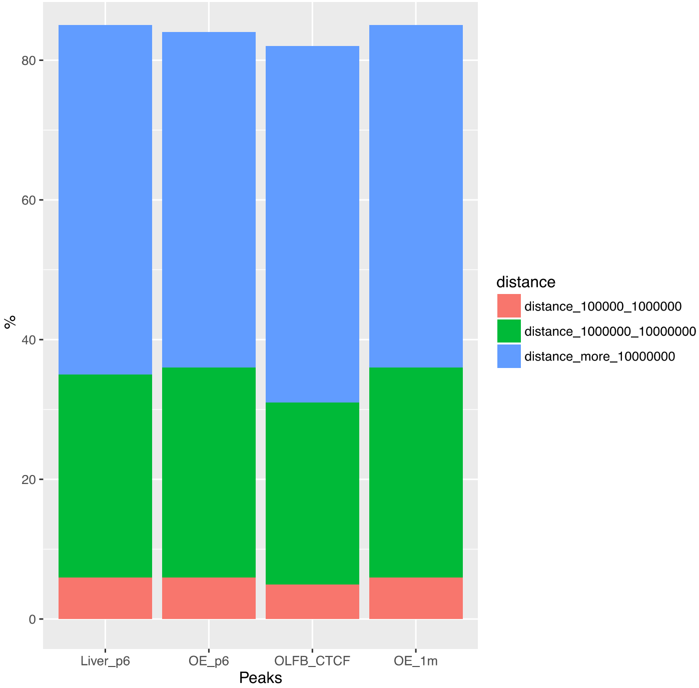


Figure 3.37. CTCF-overlapping γ-H2AX peaks distribution with respect to OR-gene clusters. CTCF-overlapping peaks were intersected with olfactory clusters to see the % of γH2AX peaks falling inside the clusters and within different range of intervals outside the clusters. Too few peaks fall within 0 and 100 kb to be visible in the graphic. A parallel analysis was performed with CTCF peaks from OB public dataset (OLFB_CTCF).

γ-H2AX peaks within *Olfr2* cluster

Finally, we investigated further how all γ–H2AX peaks were distributed with respect to the OR cluster where *Olfr2* is located on chromosome 7 and which was extensively studied in the structural variation analysis. As expected the peaks were depleted inside the 1Mb *Olfr2* cluster but preferentially distributed outside it (Figure 3.38). Interestingly, the only exception was one OE1m γ–H2AX peak which falls inside the *Olfr2* locus. In particular, it was located between amplicon 3 and amplicon 1 validated deletions, overlapping amplicon 2 coordinates. (Figure 3.39). The peak is characterized by a high number of supporting reads, suggesting it represents a real signal.

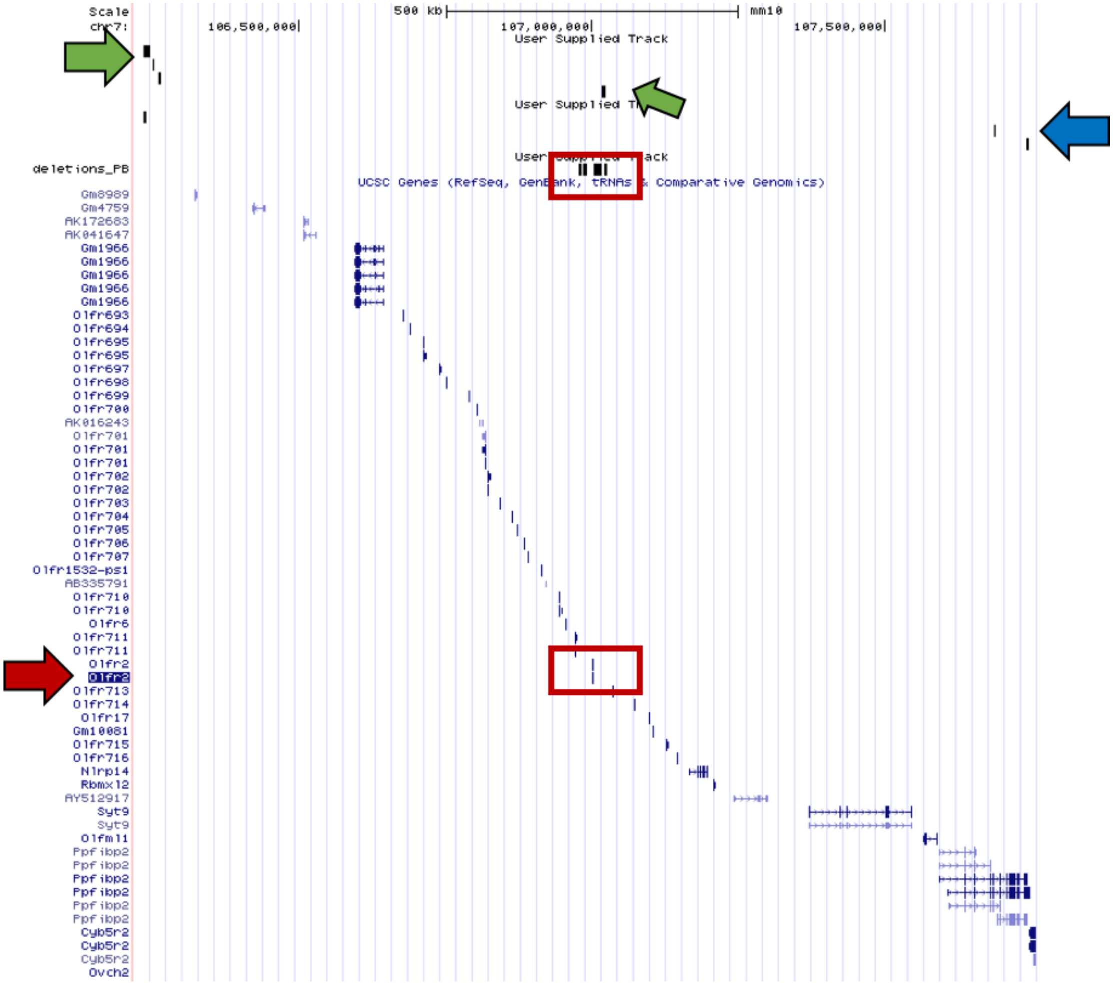


Figure 3.38. H2AX peaks distribution with respect to *Olfr2* cluster. UCSC Genome Browser 1Mb screenshot showing the entire *Olfr2* cluster on chromosome 7. OE γ-H2AX peaks are indicated by the blue (OEp6) and green (OE1m) arrows. No L peaks were detected in the same region. Red rectangles indicated the *Olfr2* validated deletions and red arrow indicated *Olfr2* coding region.

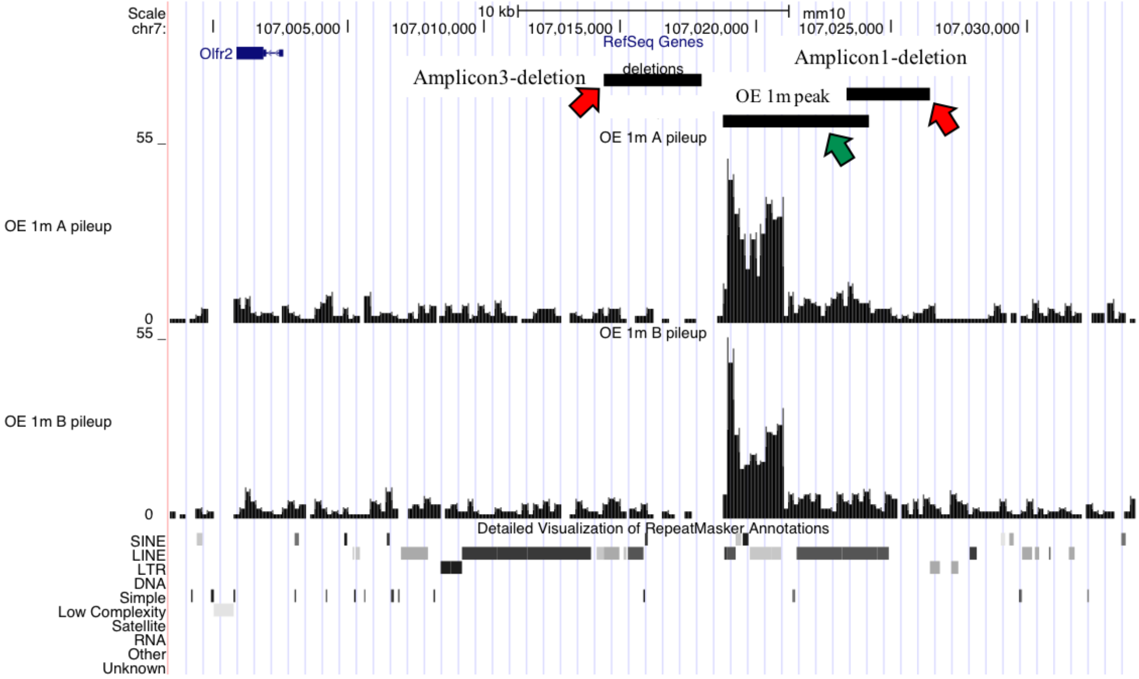


Figure 3.39. OE γ-H2AX peak localization with respect to validated deletions. Amplicon 1 and amplicon 3-deletions are indicated by red arrows; OE γ-H2AX peak is indicated by green arrow. ChOE1m A and B pileups show the peak-supporting reads for each ChIP-sample replicate. Pileup of reads in the two sample replicates (OE 1m A and OE 1m B) are indicated.

4 Discussion

Here we discuss our results. In particular, we will focus our attention on LINE-1 retrotransposition in OE, LINE-1 enrichment in OR loci, analysis of genomic deletions in *Olfr2* locus and finally about endogenous DSBs in OE and L.

4.1 LINE-1 expression and CNV in OE

It has been demonstrated that both full length and truncated LINE-1 transcripts are barely expressed but detectable in several somatic human tissues like liver, prostate, heart muscle, and stomach (Faulkner et al. 2009). Conversely, in other somatic tissues as kidney, spleen and adrenal gland, LINE-1 expression is considered below the detection limit (Faulkner et al. 2009; Belancio et al. 2010).

Recently, LINE-1 expression and engineered LINE-1 retrotransposition in mature neuronal cells have been investigated, demonstrating that engineered LINE-1 retrotransposition can take place in non-dividing mature neuronal cells (Angela Macia et al. 2016). In 2014, Pascarella and colleagues demonstrated, with NanoCAGE analysis from OE of C57BL/6J mice at p21, LINE-1 transcription at vomeronasal (VR) and OR loci (Pascarella 2014).

In agreement with these evidence, by using RT-qPCR assays, we demonstrated for the first time that both full length and 5’truncated endogenous LINE-1 transcripts are highly expressed in mouse OE compared with K and Cer from C57BL/6J mice at p21. RT-qPCR results were comparable for each of the three full length LINE-1 subfamilies analyzed (Gf, Tf and A). Consistently, looking at the total number of LINE-1 transcripts (ORF2 probe), their relative expression in OE was even higher. Indeed, using the ORF2 probe we were able to detect significantly higher LINE-1 expression also in Cer compared with K (3.1.1).

RT-qPCR expression results for LINE-transcripts were confirmed also by immunofluorescent-IHC experiments for LINE-1 ORF2 protein, comparing OE with Cer, K and L tissues from C57BL/6J mice at p6. The strongest ORF2 signal was detected in OE, where ORF2 presented a punctuate and cytoplasmatic/perinuclear localization in the majority of cells. Nevertheless, in some OE cells positive nuclei were detected. In Cer only few cells were positive for ORF2, showing a punctuate pattern localized both in the cytoplasm and in the nucleus. In L we detected a similar ORF2 pattern but with a higher background signal in the secondary antibody control. No signal was detected in K.

Looking in detail at OE, ORF2 was expressed in the basal layers of OE consisting of stem cells and OSNs progenitor cells. Moreover, in the OE intermediate layer, where OSNs are located, ORF2 co-localization with OMP protein was observed by double IHC. ORF2/OMP co-localization suggests that the expression of LINE-ORF2 protein can be also supported by mature and differentiated OSNs, in agreement with results from the work of Macia and colleagues (Macia et al., 2016). Remarkably, ORF2 pattern in mouse OE resembles the ORF2 expression pattern observed in murine breast-cancer tissues at the early stage of tumor progression (Gualtieri et al., 2013). Overall LINE-1 expression results were in agreement with data published by Belancio and colleagues. The authors, indeed, demonstrated variable expression of endogenous LINE-1 in somatic human tissues, with the amount of LINE-1 level in some tissues comparable to those detected in cancer cells (Belancio et al., 2010).

The detection of LINE-1 transcripts can have different consequences on OE gene expression. Indeed, LINE-1 sequences harbor two endogenous sense promoters, at 5' and 3'UTRs, respectively, and one antisense 5'promoter. In particular, Faulkner and colleagues showed how 5'-truncated LINE-1 sequences can function as portable promoters for protein coding genes thanks to the presence of endogenous promoter in the 3'UTR (Faulkner et al., 2009). Moreover, functional expression of LINE-1 ORF2, carrying a conserved EN domain, can create DSBs, thus increasing genomic instability in OE.

Another cause of genomic instability derives from potential LINE-1 transcripts integration in new genomic locations (LINE-1 retrotransposition). Several works, using cell-based engineered LINE-1-retrotransposition assay (Muotri et al. 2005; Muotri et al. 2010; Coufal et al. 2009) and next-generations sequencing approaches (Baillie et al. 2011; Evrony et al. 2012; Erwin, Marchetto, and Gage 2014; Erwin et al. 2016) demonstrated that LINE-1 retrotransposition is not restricted to germ line and embryogenesis but LINE-1 elements are active and can mobilize in neuronal precursor cells (NPCs) during adult neurogenesis. Intriguingly, recent evidence confirmed that engineered LINE-1 can also mobilize in post-mitotic differentiated neurons (Angela Macia et al., 2016).

In particular, in order to assess the copy-number variation (CNV) of L1 elements in human tissues, Coufal and colleagues also set up a protocol of Taqman qPCR and applied it to different human tissues and brain regions. They estimated the presence of approximately 80 more L1 copies in the hippocampus (H) compared to other organs such as heart and liver. Overall, they observed that there is a substantial variability between individuals, and also that the hippocampus, probably because of its neurogenic niche, seems to harbor a higher copy number of LINE-1 elements compared to other brain regions (Coufal et al. 2009)(Richardson et al. 2014).

In our work, we investigated LINE-1 CNV in OE, H, K and L from adult C57BL/6J mice. In agreement with results obtained by Coufal in human tissues (Coufal et al.,2009), we were able to detect an increase in both full length and 5'truncated LINE-1 copies in H compared with L and K. Additionally, we detected the highest rate of LINE-1 retrotransposition in OE and not in H. This result can be explained if we consider that OE, like H, is characterized by proliferating neuronal precursors, able to constantly regenerate mature OSNs. Moreover, another intriguing possibility is that part of the observed LINE-1 retrotransposition could be supported by differentiated OSNs as suggested by Macia and colleagues (Angela Macia et al., 2016) (3.1.3).

To further investigate LINE-expression and CNV in OE, it would be interesting to focus on single cells, distinguishing the contribution of LINE-1retrotransposition in progenitors from the ones in post-mitotic differentiated cells.

4.2 LINE-1 enrichment at OR-loci

Zooming to OE genomic sequence, we investigated how different classes of annotated repetitive elements are distributed within OR-clusters.

In 2009, Kambere and Lane performed a genomic analysis of LINE repeat content in OR and VR loci from five mammalian species (Kambere and Lane 2009). The authors demonstrated that VR and OR loci have higher LINE content than other clustered gene families. A LINE regulatory role in these clusters was suggested, since OR and VR gene clusters are not homologous and have different evolutionary origin. They showed that VR and OR loci resemble X-chromosome for LINE composition and segmental duplications.

These results suggest that LINE could have a role in shaping the epigenetic state in OR and VR loci as they do favoring long-range allelic inactivation on X chromosome (Bailey et al., 2000 and Lyon et al., 2003). Indeed, a peculiar characteristic of OR genes is the monogenic OR expression. In details, in each OSN all the OR alleles are inactivated except for the one which is randomly chosen (Reinsborough and Chess 2013; Herrada and Dulac 1997). In agreement with these observations, a study of mouse and human monoallelically expressed genes, demonstrated they are enriched for LINE elements (Allen, Schmidt, and Bridle 2003).

Concordantly with evidence obtained by Kamber and Lane, we observed that OR clusters are selectively enriched for LINE-1 repeats. To prove that LINE-1 enrichment is peculiar for OR-clusters we repeated the same analysis also for other gene clusters (Trypsin and Zinc Finger clusters), confirming that OR-clusters present a specific enrichment for LINE-1s.

4.3 Target sequencing of 50 kb in *Olfr2* locus

We took advantage of very long MDA-amplified DNA fragments to amplify ~50 kb within *Olfr2* locus. In particular, we selected as target sequence 25kb upstream and downstream *Olfr2* TSS as target sequence. For sequencing we used two different but complementary technologies. Pac Bio RSII is a single molecule sequencing technique able to sequence very long reads suitable for large variations discovery. However, it present two limiting disadvantages: a high error rate (about 14 %) and the lack of well characterized Pac Bio bioinformatics tools for variation discovery. Therefore, we combined Pac Bio sequencing with Illumina MySeq paired-end sequencing technique. Illuimina has the advantages to achieve a higher coverage with a lower percentage of sequencing errors. Furthermore, different bioinformatics tools have been developed to perform variation discovery in Illumina short reads.

4.4 Structural deletions in *Olfr2* locus

Somatic variations are defined as those genomic variations that occur in the soma, outside the germ line, and are not heritable by the following generations.

Looking at somatic tissues, when a somatic variation occurs in a cell progenitor it will be inherited by all the cells of the tissue deriving from that progenitor. On the contrary, if a variation occurs in a post-mitotic differentiated cell it will be a unique feature of that cell.

As previously mentioned, somatic LINE-1 retrotransposition occurs mainly in NPCs but it can also be supported in post-mitotic differentiated neurons (Coufal et al.,2009; Angela Macia et al.,2016).

The OE is characterized by heterogeneous cell populations. The basal layer consists of stem cells and OSN progenitor cells, while in the intermediate layer reside differentiated OSNs. Whether all OSNs expressing the same OR derive from the same neuronal progenitor is not well understood.

Nevertheless, here we hypothesized that OR choice occurs in OSN progenitors, suggesting that, if a LINE-1-associated genomic variation is involved in the choice, it will be present in all the differentiated neurons expressing the selected OR.

Part of the originality of this work lies in the idea to investigate the involvement of LINE-1-associated variations in OR-choice, taking advantage of knock-in (KI) *Olfr2*-GFP mice.

Olfr2-GFP mice, indeed, gave us the possibility to investigate structural variations comparing both transcriptionally active and inactive *Olfr2* loci at the same time, in GFP-positive and in bulk OE gDNA, respectively. In order to decrease the genomic complexity and increase the sequencing coverage, we decided to work with the lowest number of cells compatible with LCM collection.

Working with low number of cells requires whole-genome amplification (WGA) techniques to increase a very low DNA starting quantity. WGA techniques, in turn, can introduce amplification artifacts, increasing noise and false discoveries in sequencing analysis.

In this work we overcome the problem of MDA artifacts using bulk OE gDNA as sample representative of the inactive *Olfr2* loci in parallel to GFP-positive cells. Bulk genomic DNA extracted from OE was indeed directly used as input for long range PCR amplification of *Olfr2* locus without any MDA amplification step.

At the same time, bulk OE DNA represents a biological negative control since it consists of all OE cellular populations among which *Olfr2*-expressing cells represent about the 0.1%.

Pindel results showed out hundreds of deletions ranging from 50 bp to about 4000 bp, among which a high proportion was specific for MDA samples. It has to be mentioned that a possible limitation of our approach could be due to the relatively short length of *Olfr2* PCR amplicons (around 5 kb) which do not allow us to identify very large CNVs.

To limit the risk of artifacts we considered as more relevant those deletions that were found in both MDA replicates and supported by both Illumina and Pac Bio reads. Independent PCR validations confirmed six deletions at the 5’ of *Olfr2* TSS, among which two were supported by Pac Bio reads and shared by two MDA replicates.

We focused on deletion on amplicon 3 because it was represented in MDA samples by the 0.5 and 0.3% of the Illumina reads in MDA V and MDA XI, respectively. Conversely, amplicon 3 deletion was found only in the 0.0008% of OE reads, suggesting a possible link with *Olfr2* expression.

Importantly, validation of this deletion in OE excluded the possibility of MDA artifacts. Remarkably, we further validated this deletion on total MDA and bulk OE samples, both samples which preceded *Olfr2* locus-targeted PCR amplification.

Although the high coverage in MDA compared with OE samples suggested that the deletion could be involved in *Olfr2* expression we will need to demonstrate its functional role *in vivo*. To this purpose, in the near future we are going to set up a functional validation experiment. Taking advantage of crispr-cas9 technology we will reproduce the deletion in the genomic sequence of *Olfr2* locus in OSN progenitors of a transient anosmic mouse. After OE regeneration, we expect to see a change in the number of GFP-positive *Olfr2* expressing cells.

So far, we just started thinking about a possible supporting mechanism for deletions found in *Olfr2* locus. Remarkably, all the 6 PCR validated deletions in amplicon 1, 3, 4 and 5 were characterized by the presence of two directed repetitive sites (DRSs) of few nucleotides at the immediate 5’ and 3’ sequence surrounding the deletion. Interestingly, only one DRS was found in the deleted sequence.

Comparing different DRSs, we did not find any consensus motif, so further analysis is needed to characterized in detail each deletion structural pattern.

Until now, we focused on amplicon 3 deletion and we noticed that the two DRSs (5’-ATTTGAT-3’) map respectively on two different, annotated, 5’truncated LINE-1 elements, distant 3555 bp from each other. Remarkably, they were both annotated on minus strand, which is the same strand from where *Olfr2* is transcribed.

Surprisingly, we found different alternative LINE-1 endonuclease cutting sites (Jurka 1997) within the deleted sequence which mainly consist of annotated LINE-1 elements. In particular, “5’-TTAGAA-3’” and “5’-CTAAAA-3’” sites were present one time, “5’-TTGAAA-3’” and “5’-TTAAAG-3’” were present three times (3.3).

Recently, it has been shown that overexpressed LINE-EN can create DNA DSBs preferentially on LINE-1 loci which harbor EN consensus sites themselves (Gasior et al.,2006; Erwin et al., 2016).

While we are conscious that alternative mechanisms could be involved, we speculated that DSRs could support a micro-homology recombination mechanism, necessary for DSBs repair and possibly resulting in sequence deletion. A model for a supporting mechanism is represented by the SSA previously described in Figure 1.12. Interestingly, this pathway is active when a DSB occurs within a pair of directly repeated sequences (Pâques and Haber 1999).

Similar observations were reported in the recent work of Gage and colleagues (Erwin et al., 2016). They identified several retrotransposition independent LINE-1 associated deletions in human brain (Erwin et al., 2016). In particular, they showed that the deletions were created by micro-homology recombination between A(n) microsatellite or fragile sites, triggered by LINE-1 EN cleavage at those sites. Indeed, LINE-1-EN-induced DSBs preferentially occurred in LINE-1 sequences harboring themselves several EN cutting sites. This observation led to the hypothesis that the increase of LINE-1 retrotransposition during neuronal differentiation generates DSBs preferentially in LINE-1 loci.

To prove that LINE-EN preferentially cuts at LINE-1 loci, they overexpressed LINE-1 construct in HEK T cells and then performed ChIP-seq experiment for γ-H2AX. As expected, LINE-1 induced γ-H2AX peaks preferentially mapped on LINE-1 sequences.

In our work, to investigate the possible involvement of DSB repair mechanisms in the formation of *Olfr2* deletions, we performed a ChIP-Seq experiment for endogenous γ-H2AX (DSBs marker) in mouse OE and L. From this analysis we expected to see a preferential distribution of endogenous γ-H2AX within LINE-1 sequences in OR-loci in OE tissue compared with L.

Interestingly, preliminary WB analysis for γ-H2AX in OE and L showed high expression of the endogenous DSBs marker in both tissues under physiological conditions. Unexpectedly, when looking at Chip-seq data, we found that both OE and L γ-H2AX peaks were depleted within LINE-1 sequences and were preferentially distributed outside the OR clusters.

Depletion of γ-H2AX peaks within LINE-1 sequences suggested that the contribution of endogenous LINE-1-EN to DSBs in a normal tissues is probably too low to be detectable by ChIP-seq experiment or is occurring at a certain time only in a very limited number of cells.

Looking then at *Olfr2* cluster, we confirmed that OE γ-H2AX signal was depleted within the cluster except for one single peak. Indeed, one γ-H2AX peak (from OE 1m sample) was found within a region located about 15 kb upstream *Olfr2* TSS. Intriguingly, that specific γ-H2AX peak was located between amplicon 1- and amplicon 3-validated deletions, suggesting the presence of fragile sites in those regions. So far we can say that the deletion-associated γ-H2AX peak is characterized by a high number of supporting reads, suggesting it is a real signal. Further analyses are ongoing to outline a possible link to the generation of genomic deletions.

4.5 Characterization of endogenous γ-H2AX in OE

Even though ChIP-seq analysis of endogenous γ-H2AX in OE could not exhaustively suggest DSBs involvement in deletions, we decided to benefit from ChIP-Seq data to characterize endogenous γ-H2AX genomic distribution in OE and L tissues under physiological conditions. Indeed, several works investigated DNA-DSBs in cell lines upon treatment with IR or chemical compounds but little is know about DSBs naturally occurring in normal tissues.

We performed ChIP-Seq experiment on OE at two ages, p6 and 1m, according to the mouse ages used for LINE-1 and SVs analyses. We chose L as “control” tissue for ChIP-seq analysis since L is a very well characterized homogenous tissue with a wide range of public datasets available. Moreover, OE and L are characterized by different cell types with different expression pattern. Interestingly, it was demonstrated that L is among the tissues with the fewest number of ectopically expressed ORs (Flegel et al. 2013)

Nevertheless, we cannot exclude that OE and L could share common biological features. Interestingly, one example is represented by the cytochrome P450, which is expresses specifically by the these two tissues (Reed, Lock, and De Matteis 1986).

We started the analysis of ChIP-Seq data looking at the genomic distribution of γ-H2AX peaks. Overall, γ-H2AX signal in OE and L presented very similar distribution patterns. Peaks were enriched in the regulatory (promoters and enhancers) and transcribed (gene bodies) portion of the genome. Consistently, the majority of peaks fall within 0 and 10 kb downstream the annotated TSSs. In addition, 75% of γ-H2AX peaks were localized within a DNAseI sensible site, suggesting a link between genomic instability and regulatory chromatin. Moreover, about 20-25% of peaks fall within CpG islands, suggesting that at least part of the γH2AX-bound DNA in physiological conditions is associated to CG rich regions.

CG content has been previously related to function: for example, it has been reported that transcription start sites located within CpG islands have distinct features with respect to those located outside (Carninci et al. 2006; Sandelin et al. 2007). Interestingly, it was shown that GC-rich regions could be particularly vulnerable to certain types of DNA damage (Ma et al. 2011). Finally, GC-rich extracellular DNA has been shown to promote DSB in adipose-derived mesenchymal stem cells (Kostyuk et al. 2015).

We additionally performed a chromatin segmentation analysis. Overall, the three sample datasets showed very similar enriched patterns. γ-H2AX peaks were preferentially enriched for Pol II binding sites and histone modifications found on active promoters (H3K4me3, H3K9ac) and gene bodies (H3K36me3, H3K79me2). This is consistent with the observation that a very large proportion of the peaks is located in gene bodies and suggests that this deposition pattern is linked to a regulatory function of γ-H2AX within the gene bodies in physiological conditions. Moreover, chromatin segmentation analysis showed that about 20% of γ-H2AX peaks were enriched for histone modifications labeling active enhancers and distal regulatory regions (H3K27ac and H3K4me1, respectively) . This result further confirm preferentially distribution of γ-H2AX peaks for enhancers, suggesting DSBs. Recently, it has been demonstrated that Topoisomerase I nicking is required for ligand-dependent enhancer activation (Puc et al. 2015). These results showed a link between transcription and DNA-damage repair response, where nicking is necessary to relieve topological stress due to DNA torsion, promoting enhancer transcription.

Overall the general characterization of γ-H2AX showed that endogenous DSBs are not randomly distributed in the genome. Interestingly they are strongly associated with genes, promoters and active enhancers.

In agreement, it has been previously shown that endogenous γ-H2AX signal was enriched in transcribed regions and at Pol II pausing sites, in human Jurkat cells (Seo et al. 2012b). Recently, endogenous DSBs have been characterized in normal human keratinocytes showing their preferentially distribution in correspondence of DNAseI sensible sites and PolII binding sites (Lensing et al. 2016).

To investigate further peaks association with genes we performed GO functional annotation. Surprisingly the biological processes associated with γ-H2AX peaks were very different for OE and L. Intriguingly, the only GO term enriched in OE 1m sample was clearly linked to olfaction, suggesting a link to tissue specific gene expression.

Interestingly, about the 75 % of Pol II-overlapping peaks was associated to an active TSS in both OE and L datasets, confirming γ-H2AX link with transcription. We speculated that the left 25% of Poll II overlapping peaks, which did not overlap an active TSS, could represent signal associated to Pol II elongation sites. Indeed, Pol II elongation sites are located within the gene rather than in the TSS. Anyway, to confirm this hypothesis further analyses are needed.

Moreover, we looked in details at peaks correlation with active TSSs in OE and L expression datasets. OE-active TSSs overlapping γ-H2AX peaks belonging to either of the three considered peak sets had higher expression levels with respect to OE-active TSSs not overlapping γ-H2AX peaks. We didn’t observe the same effect for OE-active TSSs overlapping a mock-up set of peaks randomly distributed along the genome. Even though there are no major differences between expression levels of OE-active TSSs within/outside the three peak sets considered, we noted that for both OE samples the expression levels of OE-active TSS within the peaks was higher than in L set.

When we looked at L expression dataset we did not find the same evidence that TSSs active in liver overlapping γ-H2AX peaks have higher expression than those located more far away from the peaks. We cannot exclude that the weaker effect might be due to technical reasons related to the liver expression data.

Finally, since we found a notably similar proportion of TSSs overlapped in the comparison with the two distinct expression datasets, we investigated the number of γ-H2AX peaks overlapping active TSSs, in both OE and L datasets. As expected, for each sample, the large majority of peaks overlapping an active TSS in L, overlapped also an active TSS in OE. Given the moderate peak lengths, we can deduce that the overlapped TSSs are the same in both the two tissue datasets.

Overall, expression-correlation results suggested that a portion of endogenous γ-H2AX signal can be related to gene expression under physiological conditions. In particular, γ-H2AX peaks seem to be related to genes expressed by both OE and L tissues rather than to tissue-specific genes. Further analyses are ongoing to investigate this issue.

Moreover, RNA-seq of OE and L from the very same ChIP samples analyzed would be needed to increase the significance of our analysis.

Finally, we investigated the γ-H2AX distribution with respect to CTCF insulator protein to assess whether DSBs could co-operate with CTCF in orchestrating long-range chromosomal interactions. Indeed, previous analysis of human CTCF genome distribution demonstrated that CTCF-depleted domains included clusters of related genes that are transcriptionally co-regulated, supporting a role for CTCF-binding sites acting as insulators. Gene clusters, among which OR clusters, are surrounded by a pair of consecutive CTCF-binding sites in the genome (Kim at al., 2006). In addition, Tchurikov and colleaugues showed localization of DSBs and CTCF at the border of both expressed and silenced gene clusters (Tchurikov et al., 2013). Additional observations came from the work of Reimand and colleagues, where they demonstrated that type II Topoisomerase β (TopoIIβ) binds CTCF binding sites flanking transcriptionally associated domains (TADs). Here, TopoIIβ creating DNA-DSBs is involved in solving topological constrains (Uusküla-Reimand et al. 2016). In agreement with these observations, we first observed that 30% of γ-H2AX peaks fall inside CTCF binding sites, this was true for both OE and L. We then looked at 30% CTCF-overlapping peaks distribution with respect to olfactory gene clusters. Interestingly, γ-H2AX peaks were depleted inside the clusters, but preferentially distributed outside of them together with CTCF, suggesting a possible involvement for DBSs in mediating topological stress facilitating long range chromosomal interactions. The same distribution was observed for all 10 different gene clusters analyzed (3.4).

Overall 77.6% (54.7%, 63.1%) of the Lp6 (OEp6 and OE1m respectively) peaks is associated to one among all the features we investigated (CTCF, DNAseI, Pol II, CpG and active TSSs) supporting functional roles previously associated to those features. Interestingly, the percentage is higher for the Lp6 peak set, consistently with the fact that most of the supporting data used in the comparison is liver-derived (in other words, for liver we have additional data which is fundamentally matching in terms of both tissue and age and for liver we have the highest “annotation coverage”). The remaining portions of the peaks results to be independent of those features, suggesting that they could have a functional role that make them independent of the functional marks we investigated.

4.6 Conclusion

Taken together, the results of this work showed that endogenous LINE-1 elements are active in mouse OE under physiological conditions. Moreover, the high density of LINE-1 elements in OR-loci let us speculate that they have a possible regulatory function for the expression of OR genes. Genomic analysis of a particular OR locus showed a high degree of genomic complexity, characterized by high number of SVs among which LINE-1-associated deletions are the predominant variant. Finally, genome-wide analysis of endogenous DSBs in OE and L showed DNA damage is detectable in normal tissues, and are linked to transcribed and regulatory genomic regions.

Overall, we conclude that the ability for LINE-1s to create DSBs, together with interspersed homology, suggested a role of TEs in increasing genetic instability in OR loci. Here, LINE-1 dependent structural variations could contribute to influence gene regulation of OR genes in a monoallelic and stochastic fashion.

Bibliography

Aguilera, Andrés. 2002. “The Connection between Transcription and Genomic Instability.” *The EMBO Journal* 21 (3): 195–201. doi:10.1093/emboj/21.3.195.

Allen, S. W., R. W. Schmidt, and S. L. Bridle. 2003. “A Preference for a Non-Zero Neutrino Mass from Cosmological Data: A Non-Zero Neutrino Mass from Cosmological Data.” *Monthly Notices of the Royal Astronomical Society* 346 (2): 593–600. doi:10.1046/j.1365-2966.2003.07022.x.

Altman, J. 1962. “Are New Neurons Formed in the Brains of Adult Mammals?” *Science (New York, N.Y.)* 135 (3509): 1127–28.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10. doi:10.1016/S0022-2836(05)80360-2.

Arner, Erik, Carsten O. Daub, Kristoffer Vitting-Seerup, Robin Andersson, Berit Lilje, Finn Drabløs, Andreas Lennartsson, et al. 2015. “Transcribed Enhancers Lead Waves of Coordinated Transcription in Transitioning Mammalian Cells.” *Science (New York, N.Y.)* 347 (6225): 1010–14. doi:10.1126/science.1259418.

Babushok, D. V. 2005. “L1 Integration in a Transgenic Mouse Model.” *Genome Research* 16 (2): 240–50. doi:10.1101/gr.4571606.

Baillie, J. Kenneth, Mark W. Barnett, Kyle R. Upton, Daniel J. Gerhardt, Todd A. Richmond, Fioravante De Sapio, Paul M. Brennan, et al. 2011. “Somatic Retrotransposition Alters the Genetic Landscape of the Human Brain.” *Nature* 479 (7374): 534–37. doi:10.1038/nature10531.

Barral, Serena, Riccardo Beltramo, Chiara Salio, Patrizia Aimar, Laura Lossi, and Adalberto Merighi. 2014. “Phosphorylation of Histone H2AX in the Mouse Brain from Development to Senescence.” *International Journal of Molecular Sciences* 15 (1): 1554–73. doi:10.3390/ijms15011554.

Barresi, Marina, Rosella Ciurleo, Sabrina Giacoppo, Valeria Foti Cuzzola, Debora Celi, Placido Bramanti, and Silvia Marino. 2012. “Evaluation of Olfactory Dysfunction in Neurodegenerative Diseases.” *Journal of the Neurological Sciences* 323 (1–2): 16–24. doi:10.1016/j.jns.2012.08.028.

Beck, Christine R., José Luis Garcia-Perez, Richard M. Badge, and John V. Moran. 2011. “LINE-1 Elements in Structural Variation and Disease.” *Annual Review of*

Genomics and Human Genetics 12 (1): 187–215. doi:10.1146/annurev-genom-082509-141802.

Bejerano, Gill, Craig B. Lowe, Nadav Ahituv, Bryan King, Adam Siepel, Sofie R. Salama, Edward M. Rubin, W. James Kent, and David Haussler. 2006. "A Distal Enhancer and an Ultraconserved Exon Are Derived from a Novel Retroposon." *Nature* 441 (7089): 87–90. doi:10.1038/nature04696.

Belancio, V. P., A. M. Roy-Engel, R. R. Pochampally, and P. Deininger. 2010. "Somatic Expression of LINE-1 Elements in Human Tissues." *Nucleic Acids Research* 38 (12): 3909–22. doi:10.1093/nar/gkq132.

Bodega, Beatrice, and Valerio Orlando. 2014. "Repetitive Elements Dynamics in Cell Identity Programming, Maintenance and Disease." *Current Opinion in Cell Biology* 31 (December): 67–73. doi:10.1016/j.ceb.2014.09.002.

Boeva, V., A. Lermine, C. Barette, C. Guillouf, and E. Barillot. 2012. "Nebula--a Web-Server for Advanced ChIP-Seq Data Analysis." *Bioinformatics* 28 (19): 2517–19. doi:10.1093/bioinformatics/bts463.

Bourc'his, Déborah, and Timothy H. Bestor. 2004. "Meiotic Catastrophe and Retrotransposon Reactivation in Male Germ Cells Lacking Dnmt3L." *Nature* 431 (7004): 96–99. doi:10.1038/nature02886.

Bourque, G., B. Leong, V. B. Vega, X. Chen, Y. L. Lee, K. G. Srinivasan, J.-L. Chew, et al. 2008. "Evolution of the Mammalian Transcription Factor Binding Repertoire via Transposable Elements." *Genome Research* 18 (11): 1752–62. doi:10.1101/gr.080663.108.

Bozza, Thomas, Paul Feinstein, Chen Zheng, and Peter Mombaerts. 2002. "Odorant Receptor Expression Defines Functional Units in the Mouse Olfactory System." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 22 (8): 3033–43. doi:20026321.

Brann, J. H., and S. Firestein. 2010. "Regeneration of New Neurons Is Preserved in Aged Vomeronasal Epithelia." *Journal of Neuroscience* 30 (46): 15686–94. doi:10.1523/JNEUROSCI.4316-10.2010.

Brann, Jessica H., and Stuart J. Firestein. 2014. "A Lifetime of Neurogenesis in the Olfactory System." *Frontiers in Neuroscience* 8 (June). doi:10.3389/fnins.2014.00182.

Branzei, Dana, and Marco Foiani. 2010. “Maintaining Genome Stability at the Replication Fork.” *Nature Reviews Molecular Cell Biology* 11 (3): 208–19. doi:10.1038/nrm2852.

Brennan, Peter A., and Frank Zufall. 2006. “Pheromonal Communication in Vertebrates.” *Nature* 444 (7117): 308–15. doi:10.1038/nature05404.

Bunch, Heeyoun, Brian P. Lawney, Yu-Fen Lin, Aroumougame Asaithamby, Ayesha Murshid, Yaoyu E. Wang, Benjamin P. C. Chen, and Stuart K. Calderwood. 2015. “Transcriptional Elongation Requires DNA Break-Induced Signalling.” *Nature Communications* 6 (December): 10191. doi:10.1038/ncomms10191.

Calo, Eliezer, and Joanna Wysocka. 2013. “Modification of Enhancer Chromatin: What, How, and Why?” *Molecular Cell* 49 (5): 825–37. doi:10.1016/j.molcel.2013.01.038.

Carninci, Piero, Albin Sandelin, Boris Lenhard, Shintaro Katayama, Kazuro Shimokawa, Jasmina Ponjavic, Colin A M Semple, et al. 2006. “Genome-Wide Analysis of Mammalian Promoter Architecture and Evolution.” *Nature Genetics* 38 (6): 626–35. doi:10.1038/ng1789.

Celeste, Arkady, Oscar Fernandez-Capetillo, Michael J. Kruhlak, Duane R. Pilch, David W. Staudt, Alicia Lee, Robert F. Bonner, William M. Bonner, and André Nussenzweig. 2003. “Histone H2AX Phosphorylation Is Dispensable for the Initial Recognition of DNA Breaks.” *Nature Cell Biology* 5 (7): 675–79. doi:10.1038/ncb1004.

Chaurasia, P., R. Sen, T. K. Pandita, and S. R. Bhaumik. 2012. “Preferential Repair of DNA Double-Strand Break at the Active Gene in Vivo.” *Journal of Biological Chemistry* 287 (43): 36414–22. doi:10.1074/jbc.M112.364661.

Chow, Jennifer C., Constance Ciaudo, Melissa J. Fazzari, Nathan Mise, Nicolas Servant, Jacob L. Glass, Matthew Attreed, et al. 2010. “LINE-1 Activity in Facultative Heterochromatin Formation during X Chromosome Inactivation.” *Cell* 141 (6): 956–69. doi:10.1016/j.cell.2010.04.042.

Clowney, E. Josephine, Mark A. LeGros, Colleen P. Mosley, Fiona G. Clowney, Eirene C. Markenskoff-Papadimitriou, Markko Myllys, Gilad Barnea, Carolyn A. Larabell, and Stavros Lomvardas. 2012. “Nuclear Aggregation of Olfactory Receptor Genes Governs Their Monogenic Expression.” *Cell* 151 (4): 724–37. doi:10.1016/j.cell.2012.09.043.

Cordaux, Richard, and Mark A. Batzer. 2009. "The Impact of Retrotransposons on Human Genome Evolution." *Nature Reviews Genetics* 10 (10): 691–703. doi:10.1038/nrg2640.

Coufal, Nicole G., José L. Garcia-Perez, Grace E. Peng, Gene W. Yeo, Yangling Mu, Michael T. Lovci, Maria Morell, K. Sue O'Shea, John V. Moran, and Fred H. Gage. 2009. "L1 Retrotransposition in Human Neural Progenitor Cells." *Nature* 460 (7259): 1127–31. doi:10.1038/nature08248.

Cui, X., and K. Meek. 2007. "Linking Double-Stranded DNA Breaks to the Recombination Activating Gene Complex Directs Repair to the Nonhomologous End-Joining Pathway." *Proceedings of the National Academy of Sciences* 104 (43): 17046–51. doi:10.1073/pnas.0610928104.

Dalton, Ryan P., David B. Lyons, and Stavros Lomvardas. 2013. "Co-Opting the Unfolded Protein Response to Elicit Olfactory Receptor Feedback." *Cell* 155 (2): 321–32. doi:10.1016/j.cell.2013.09.033.

Ernst, Jason, and Manolis Kellis. 2012. "ChromHMM: Automating Chromatin-State Discovery and Characterization." *Nature Methods* 9 (3): 215–16. doi:10.1038/nmeth.1906.

Erwin, Jennifer A., Maria C. Marchetto, and Fred H. Gage. 2014. "Mobile DNA Elements in the Generation of Diversity and Complexity in the Brain." *Nature Reviews Neuroscience* 15 (8): 497–506. doi:10.1038/nrn3730.

Erwin, Jennifer A, Apuã C M Paquola, Tatjana Singer, Iryna Gallina, Mark Novotny, Carolina Quayle, Tracy A Bedrosian, et al. 2016. "L1-Associated Genomic Regions Are Deleted in Somatic Cells of the Healthy Human Brain." *Nature Neuroscience* 19 (12): 1583–91. doi:10.1038/nn.4388.

Evrony, Gilad D., Xuyu Cai, Eunjung Lee, L. Benjamin Hills, Princess C. Elhosary, Hillel S. Lehmann, J.J. Parker, et al. 2012. "Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain." *Cell* 151 (3): 483–96. doi:10.1016/j.cell.2012.09.035.

Evrony, Gilad D., Eunjung Lee, Bhaven K. Mehta, Yuval Benjamini, Robert M. Johnson, Xuyu Cai, Lixing Yang, et al. 2015. "Cell Lineage Analysis in Human Brain Using Endogenous Retroelements." *Neuron* 85 (1): 49–59. doi:10.1016/j.neuron.2014.12.028.

Faulkner, Geoffrey J, Yasumasa Kimura, Carsten O Daub, Shivangi Wani, Charles Plessy, Katharine M Irvine, Kate Schroder, et al. 2009. "The Regulated

Retrotransposon Transcriptome of Mammalian Cells." *Nature Genetics* 41 (5): 563–71. doi:10.1038/ng.368.

Fedoroff, N. V. 2012. "Transposable Elements, Epigenetics, and Genome Evolution." *Science* 338 (6108): 758–67. doi:10.1126/science.338.6108.758.

Flegel, Caroline, Stavros Manteniotis, Sandra Osthold, Hanns Hatt, and Günter Gisselmann. 2013. "Expression Profile of Ectopic Olfactory Receptors Determined by Deep Sequencing." Edited by Johannes Reisert. *PLoS ONE* 8 (2): e55368. doi:10.1371/journal.pone.0055368.

Fong, Yick W., Claudia Cattoglio, and Robert Tjian. 2013. "The Intertwined Roles of Transcription and Repair Proteins." *Molecular Cell* 52 (3): 291–302. doi:10.1016/j.molcel.2013.10.018.

Garcia-Perez, Jose L., Thomas J. Widmann, and Ian R. Adams. 2016. "The Impact of Transposable Elements on Mammalian Development." *Development* 143 (22): 4101–14. doi:10.1242/dev.132639.

Gasior, Stephen L., Timothy P. Wakeman, Bo Xu, and Prescott L. Deininger. 2006. "The Human LINE-1 Retrotransposon Creates DNA Double-Strand Breaks." *Journal of Molecular Biology* 357 (5): 1383–93. doi:10.1016/j.jmb.2006.01.089.

Gilbert, Nicolas, Sheila Lutz-Prigge, and John V. Moran. 2002. "Genomic Deletions Created upon LINE-1 Retrotransposition." *Cell* 110 (3): 315–25.

Goodier, J. L. 2000. "Transduction of 3'-flanking Sequences Is Common in L1 Retrotransposition." *Human Molecular Genetics* 9 (4): 653–57. doi:10.1093/hmg/9.4.653.

———. 2001. "A Novel Active L1 Retrotransposon Subfamily in the Mouse." *Genome Research* 11 (10): 1677–85. doi:10.1101/gr.198301.

Goodier, John L., and Haig H. Kazazian. 2008. "Retrotransposons Revisited: The Restraint and Rehabilitation of Parasites." *Cell* 135 (1): 23–35. doi:10.1016/j.cell.2008.09.022.

Graziadei, G. A., and P. P. Graziadei. 1979. "Neurogenesis and Neuron Regeneration in the Olfactory System of Mammals. II. Degeneration and Reconstitution of the Olfactory Sensory Neurons after Axotomy." *Journal of Neurocytology* 8 (2): 197–213.

Guo, Ya, Quan Xu, Daniele Canzio, Jia Shou, Jinhuan Li, David U. Gorkin, Inkyung Jung, et al. 2015. "CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function." *Cell* 162 (4): 900–910. doi:10.1016/j.cell.2015.07.038.

Haber, James E. 2000. "Partners and Pathways." *Trends in Genetics* 16 (6): 259–64. doi:10.1016/S0168-9525(00)02022-9.

Hedges, D.J., and P.L. Deininger. 2007. "Inviting Instability: Transposable Elements, Double-Strand Breaks, and the Maintenance of Genome Integrity." *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 616 (1–2): 46–59. doi:10.1016/j.mrfmmm.2006.11.021.

Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 2010. "Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities." *Molecular Cell* 38 (4): 576–89. doi:10.1016/j.molcel.2010.05.004.

Herrada, Gilles, and Catherine Dulac. 1997. "A Novel Family of Putative Pheromone Receptors in Mammals with a Topographically Organized and Sexually Dimorphic Distribution." *Cell* 90 (4): 763–73. doi:10.1016/S0092-8674(00)80536-X.

Hurtt, M. E., D. A. Thomas, P. K. Working, T. M. Monticello, and K. T. Morgan. 1988. "Degeneration and Regeneration of the Olfactory Epithelium Following Inhalation Exposure to Methyl Bromide: Pathology, Cell Kinetics, and Olfactory Function." *Toxicology and Applied Pharmacology* 94 (2): 311–28.

Ibarra-Soria, Ximena, Maria O. Levitin, Luis R. Saraiva, and Darren W. Logan. 2014. "The Olfactory Transcriptomes of Mice." *PLoS Genet* 10 (9): e1004593.

Johnson, R. D. 2000. "Sister Chromatid Gene Conversion Is a Prominent Double-Strand Break Repair Pathway in Mammalian Cells." *The EMBO Journal* 19 (13): 3398–3407. doi:10.1093/emboj/19.13.3398.

Jurka, J. 1997. "Sequence Patterns Indicate an Enzymatic Involvement in Integration of Mammalian Retroposons." *Proceedings of the National Academy of Sciences of the United States of America* 94 (5): 1872–77.

Jurka, J., V.V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. 2005. "Repbase Update, a Database of Eukaryotic Repetitive Elements." *Cytogenetic and Genome Research* 110 (1–4): 462–67. doi:10.1159/000084979.

Kambere, Marijo B., and Robert P. Lane. 2009. "Exceptional LINE Density at V1R Loci: The Lyon Repeat Hypothesis Revisited on Autosomes." *Journal of Molecular Evolution* 68 (2): 145–59. doi:10.1007/s00239-008-9195-0.

Kano, H., I. Godoy, C. Courtney, M. R. Vetter, G. L. Gerton, E. M. Ostertag, and H. H. Kazazian. 2009. "L1 Retrotransposition Occurs Mainly in Embryogenesis and Creates Somatic Mosaicism." *Genes & Development* 23 (11): 1303–12. doi:10.1101/gad.1803909.

Kazazian, Haig H. 2004. "Mobile Elements: Drivers of Genome Evolution." *Science (New York, N.Y.)* 303 (5664): 1626–32. doi:10.1126/science.1089670.

Keane, Thomas M., Kim Wong, David J. Adams, Jonathan Flint, Alexandre Reymond, and Binnaz Yalcin. 2014. "Identification of Structural Variation in Mouse Genomes." *Frontiers in Genetics* 5 (July). doi:10.3389/fgene.2014.00192.

Khan, Mona, Evelien Vaes, and Peter Mombaerts. 2011. "Regulation of the Probability of Mouse Odorant Receptor Gene Choice." *Cell* 147 (4): 907–21. doi:10.1016/j.cell.2011.09.049.

Kim, Tae Hoon, Ziedulla K. Abdullaev, Andrew D. Smith, Keith A. Ching, Dmitri I. Loukinov, Roland D. Green, Michael Q. Zhang, Victor V. Lobanenkov, and Bing Ren. 2007. "Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome." *Cell* 128 (6): 1231–45. doi:10.1016/j.cell.2006.12.048.

Kostyuk, Svetlana, Tatiana Smirnova, Larisa Kameneva, Lev Porokhovnik, Anatolij Speranskij, Elizaveta Ershova, Sergey Stukalov, Vera Izevskaya, and Natalia Veiko. 2015. "GC-Rich Extracellular DNA Induces Oxidative Stress, Double-Strand DNA Breaks, and DNA Damage Response in Human Adipose-Derived Mesenchymal Stem Cells." *Oxidative Medicine and Cellular Longevity* 2015: 1–15. doi:10.1155/2015/782123.

Kruhlak, Michael J., Arkady Celeste, Graham Dellaire, Oscar Fernandez-Capetillo, Waltraud G. Müller, James G. McNally, David P. Bazett-Jones, and André Nussenzweig. 2006. "Changes in Chromatin Structure and Mobility in Living Cells at Sites of DNA Double-Strand Breaks." *The Journal of Cell Biology* 172 (6): 823–34. doi:10.1083/jcb.200510015.

Kulpa, Deanna A., and John V. Moran. 2006. "Cis-Preferential LINE-1 Reverse Transcriptase Activity in Ribonucleoprotein Particles." *Nature Structural & Molecular Biology* 13 (7): 655–60. doi:10.1038/nsmb1107.

Kurukuti, S., V. K. Tiwari, G. Tavoosidana, E. Pugacheva, A. Murrell, Z. Zhao, V. Lobanenkov, W. Reik, and R. Ohlsson. 2006. “CTCF Binding at the H19 Imprinting Control Region Mediates Maternally Inherited Higher-Order Chromatin Conformation to Restrict Enhancer Access to Igf2.” *Proceedings of the National Academy of Sciences* 103 (28): 10684–89. doi:10.1073/pnas.0600326103.

Kuwabara, Tomoko, Jenny Hsieh, Alysson Muotri, Gene Yeo, Masaki Warashina, Dieter Chichung Lie, Lynne Moore, Kinichi Nakashima, Makoto Asashima, and Fred H Gage. 2009. “Wnt-Mediated Activation of NeuroD1 and Retro-Elements during Adult Neurogenesis.” *Nature Neuroscience* 12 (9): 1097–1105. doi:10.1038/nn.2360.

Lee, E., R. Iskow, L. Yang, O. Gokcumen, P. Haseley, L. J. Luquette, J. G. Lohr, et al. 2012. “Landscape of Somatic Retrotransposition in Human Cancers.” *Science* 337 (6097): 967–71. doi:10.1126/science.1222077.

Lensing, Stefanie V, Giovanni Marsico, Robert Hänsel-Hertsch, Enid Y Lam, David Tannahill, and Shankar Balasubramanian. 2016. “DSBCapture: In Situ Capture and Sequencing of DNA Breaks.” *Nature Methods* 13 (10): 855–57. doi:10.1038/nmeth.3960.

López-Mascaraque, L., and F. de Castro. 2002. “The Olfactory Bulb as an Independent Developmental Domain.” *Cell Death and Differentiation* 9 (12): 1279–86. doi:10.1038/sj.cdd.4401076.

Lunyak, V. V., G. G. Prefontaine, E. Nunez, T. Cramer, B.-G. Ju, K. A. Ohgi, K. Hutt, et al. 2007. “Developmentally Regulated Activation of a SINE B2 Repeat as a Domain Boundary in Organogenesis.” *Science* 317 (5835): 248–51. doi:10.1126/science.1140871.

Lyon, M.F. 1998. “X-Chromosome Inactivation: A Repeat Hypothesis.” *Cytogenetic and Genome Research* 80 (1–4): 133–37. doi:10.1159/000014969.

Lyons, David B., William E. Allen, Tracie Goh, Lulu Tsai, Gilad Barnea, and Stavros Lomvardas. 2013. “An Epigenetic Trap Stabilizes Singular Olfactory Receptor Expression.” *Cell* 154 (2): 325–36. doi:10.1016/j.cell.2013.06.039.

Ma, Wenjian, Jim W. Westmoreland, Dmitry A. Gordenin, and Mike A. Resnick. 2011. “Alkylation Base Damage Is Converted into Repairable Double-Strand Breaks and Complex Intermediates in G2 Cells Lacking AP Endonuclease.” Edited by

Nancy Maizels. *PLoS Genetics* 7 (4): e1002059. doi:10.1371/journal.pgen.1002059.

Macia, A., M. Munoz-Lopez, J. L. Cortes, R. K. Hastings, S. Morell, G. Lucena-Aguilar, J. A. Marchal, R. M. Badge, and J. L. Garcia-Perez. 2011. "Epigenetic Control of Retrotransposon Expression in Human Embryonic Stem Cells." *Molecular and Cellular Biology* 31 (2): 300–316. doi:10.1128/MCB.00561-10.

Macia, Angela, Thomas J. Widmann, Sara R. Heras, Veronica Ayllon, Laura Sanchez, Meriem Benkaddour-Boumzaouad, Martin Munoz-Lopez, et al. 2016. "Engineered LINE-1 Retrotransposition in Non-Dividing Human Neurons." *Genome Research*, December. doi:10.1101/gr.206805.116.

Madabhushi, Ram, Fan Gao, Andreas R. Pfenning, Ling Pan, Satoko Yamakawa, Jinsoo Seo, Richard Rueda, et al. 2015. "Activity-Induced DNA Breaks Govern the Expression of Neuronal Early-Response Genes." *Cell* 161 (7): 1592–1605. doi:10.1016/j.cell.2015.05.032.

Malki, Safia, Godfried W. van der Heijden, Kathryn A. O'Donnell, Sandra L. Martin, and Alex Bortvin. 2014. "A Role for Retrotransposon LINE-1 in Fetal Oocyte Attrition in Mice." *Developmental Cell* 29 (5): 521–33. doi:10.1016/j.devcel.2014.04.027.

Mätlik, Kert, Kaja Redik, and Mart Speek. 2006. "L1 Antisense Promoter Drives Tissue-Specific Transcription of Human Genes." *Journal of Biomedicine and Biotechnology* 2006: 1–16. doi:10.1155/JBB/2006/71753.

McLean, Cory Y., Dave Bristor, Michael Hiller, Shoa L. Clarke, Bruce T. Schaar, Craig B. Lowe, Aaron M. Wenger, and Gill Bejerano. 2010. "GREAT Improves Functional Interpretation of Cis-Regulatory Regions." *Nature Biotechnology* 28 (5): 495–501. doi:10.1038/nbt.1630.

Mears, M. L., and C. A. Hutchison. 2001. "The Evolution of Modern Lineages of Mouse L1 Elements." *Journal of Molecular Evolution* 52 (1): 51–62.

Merkenschlager, Matthias, and Elphège P. Nora. 2016. "CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation." *Annual Review of Genomics and Human Genetics* 17 (1): 17–43. doi:10.1146/annurev-genom-083115-022339.

Merkenschlager, Matthias, and Duncan T. Odom. 2013. "CTCF and Cohesin: Linking Gene Regulatory Elements with Their Targets." *Cell* 152 (6): 1285–97. doi:10.1016/j.cell.2013.02.029.

Mombaerts, Peter. 2004. "Odorant Receptor Gene Choice in Olfactory Sensory Neurons: The One Receptor–one Neuron Hypothesis Revisited." *Current Opinion in Neurobiology* 14 (1): 31–36. doi:10.1016/j.conb.2004.01.014.

Moran, John V, Susan E Holmes, Thierry P Naas, Ralph J DeBerardinis, Jef D Boeke, and Haig H Kazazian. 1996. "High Frequency Retrotransposition in Cultured Mammalian Cells." *Cell* 87 (5): 917–27. doi:10.1016/S0092-8674(00)81998-4.

Morrish, Tammy A., Nicolas Gilbert, Jeremy S. Myers, Bethaney J. Vincent, Thomas D. Stamato, Guillermo E. Taccioli, Mark A. Batzer, and John V. Moran. 2002. "DNA Repair Mediated by Endonuclease-Independent LINE-1 Retrotransposition." *Nature Genetics* 31 (2): 159–65. doi:10.1038/ng898.

Mouse Genome Sequencing Consortium, Robert H. Waterston, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep F. Abril, Pankaj Agarwal, et al. 2002. "Initial Sequencing and Comparative Analysis of the Mouse Genome." *Nature* 420 (6915): 520–62. doi:10.1038/nature01262.

Munoz-Lopez, Martin, and Jose Garcia-Perez. 2010. "DNA Transposons: Nature and Applications in Genomics." *Current Genomics* 11 (2): 115–28. doi:10.2174/138920210790886871.

Muotri, Alysson R., Vi T. Chu, Maria C. N. Marchetto, Wei Deng, John V. Moran, and Fred H. Gage. 2005. "Somatic Mosaicism in Neuronal Precursor Cells Mediated by L1 Retrotransposition." *Nature* 435 (7044): 903–10. doi:10.1038/nature03663.

Muotri, Alysson R., Maria C. N. Marchetto, Nicole G. Coufal, Ruth Oefner, Gene Yeo, Kinichi Nakashima, and Fred H. Gage. 2010. "L1 Retrotransposition in Neurons Is Modulated by MeCP2." *Nature* 468 (7322): 443–46. doi:10.1038/nature09544.

Murdoch, B., and A. J. Roskams. 2008. "A Novel Embryonic Nestin-Expressing Radial Glia-Like Progenitor Gives Rise to Zonally Restricted Olfactory and Vomeronasal Neurons." *Journal of Neuroscience* 28 (16): 4271–82. doi:10.1523/JNEUROSCI.5566-07.2008.

Nagai, M. H., L. M. Armelin-Correa, and B. Malnic. 2016. "Monogenic and Monoallelic Expression of Odorant Receptors." *Molecular Pharmacology* 90 (5): 633–39. doi:10.1124/mol.116.104745.

Narendra, V., P. P. Rocha, D. An, R. Raviram, J. A. Skok, E. O. Mazzoni, and D. Reinberg. 2015. "CTCF Establishes Discrete Functional Chromatin Domains at the Hox Clusters during Differentiation.” *Science* 347 (6225): 1017–21. doi:10.1126/science.1262088.

Notwell, James H., Tisha Chung, Whitney Heavner, and Gill Bejerano. 2015. “A Family of Transposable Elements Co-Opted into Developmental Enhancers in the Mouse Neocortex.” *Nature Communications* 6 (March): 6644. doi:10.1038/ncomms7644.

Olovnikov, Ivan, Alexei A Aravin, and Katalin Fejes Toth. 2012. “Small RNA in the Nucleus: The RNA-Chromatin Ping-Pong.” *Current Opinion in Genetics & Development* 22 (2): 164–71. doi:10.1016/j.gde.2012.01.002.

Onozawa, M., Z. Zhang, Y. J. Kim, L. Goldberg, T. Varga, P. L. Bergsagel, W. M. Kuehl, and P. D. Aplan. 2014. “Repair of DNA Double-Strand Breaks by Templated Nucleotide Sequence Insertions Derived from Distant Regions of the Genome.” *Proceedings of the National Academy of Sciences* 111 (21): 7729–34. doi:10.1073/pnas.1321889111.

Ostertag, E. M. 2000. “Determination of L1 Retrotransposition Kinetics in Cultured Cells.” *Nucleic Acids Research* 28 (6): 1418–23. doi:10.1093/nar/28.6.1418.

Ostertag, E. M., and H. H. Kazazian. 2001. “Biology of Mammalian L1 Retrotransposons.” *Annual Review of Genetics* 35: 501–38. doi:10.1146/annurev.genet.35.102401.091032.

Pace, John K., and Cédric Feschotte. 2007. “The Evolutionary History of Human DNA Transposons: Evidence for Intense Activity in the Primate Lineage.” *Genome Research* 17 (4): 422–32. doi:10.1101/gr.5826307.

Packard, Adam, Maryann Giel-Moloney, Andrew Leiter, and James E. Schwob. 2011. “Progenitor Cell Capacity of NeuroD1-Expressing Globose Basal Cells in the Mouse Olfactory Epithelium.” *The Journal of Comparative Neurology* 519 (17): 3580–96. doi:10.1002/cne.22726.

Pan, Mei-Ren, Guang Peng, Wen-Chun Hung, and Shiaw-Yih Lin. 2011. “Monoubiquitination of H2AX Protein Regulates DNA Damage Response Signaling.” *Journal of Biological Chemistry* 286 (32): 28599–607. doi:10.1074/jbc.M111.256297.

Pankotai, Tibor, Céline Bonhomme, David Chen, and Evi Soutoglou. 2012. “DNAPKcs-Dependent Arrest of RNA Polymerase II Transcription in the Presence of DNA Breaks.” *Nature Structural & Molecular Biology* 19 (3): 276–82. doi:10.1038/nsmb.2224.

Pâques, F., and J. E. Haber. 1999. "Multiple Pathways of Recombination Induced by Double-Strand Breaks in Saccharomyces Cerevisiae." *Microbiology and Molecular Biology Reviews: MMBR* 63 (2): 349–404.

Pascarella, Giovanni. 2014. "NanoCAGE Analysis of the Mouse Olfactory Epithelium Identifies the Expression of Vomeronasal Receptors and of Proximal LINE Elements." *Frontiers in Cellular Neuroscience* 8. doi:10.3389/fncel.2014.00041.

Plessy, C., G. Pascarella, N. Bertin, A. Akalin, C. Carrieri, A. Vassalli, D. Lazarevic, et al. 2012. "Promoter Architecture of Mouse Olfactory Receptor Genes." *Genome Research* 22 (3): 486–97. doi:10.1101/gr.126201.111.

Puc, Janusz, Piotr Kozbial, Wenbo Li, Yuliang Tan, Zhijie Liu, Tom Suter, Kenneth A. Ohgi, Jie Zhang, Aneel K. Aggarwal, and Michael G. Rosenfeld. 2015. "Ligand-Dependent Enhancer Activation Regulated by Topoisomerase-I Activity." *Cell* 160 (3): 367–80. doi:10.1016/j.cell.2014.12.023.

Quinlan, A. R., and I. M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42. doi:10.1093/bioinformatics/btq033.

Reed, C. J., E. A. Lock, and F. De Matteis. 1986. "NADPH: Cytochrome P-450 Reductase in Olfactory Epithelium. Relevance to Cytochrome P-450-Dependent Reactions." *The Biochemical Journal* 240 (2): 585–92.

Reinsborough, Calder, and Andrew Chess. 2013. "An Epigenetic Trap Involved in Olfactory Receptor Gene Choice." *Developmental Cell* 26 (2): 120–21. doi:10.1016/j.devcel.2013.07.011.

Richardson, Sandra R, Iñigo Narvaiza, Randy A Planegger, Matthew D Weitzman, and John V Moran. 2014. "APOBEC3A Deaminates Transiently Exposed Single-Strand DNA during LINE-1 Retrotransposition." *eLife* 3 (April). doi:10.7554/eLife.02008.

Rodriguez-Gil, D. J., H. B. Treloar, X. Zhang, A. M. Miller, A. Two, C. Iwema, S. J. Firestein, and C. A. Greer. 2010. "Chromosomal Location-Dependent Nonstochastic Onset of Odor Receptor Expression." *Journal of Neuroscience* 30 (30): 10067–75. doi:10.1523/JNEUROSCI.1776-10.2010.

Rogakou, E. P., D. R. Pilch, A. H. Orr, V. S. Ivanova, and W. M. Bonner. 1998. "DNA Double-Stranded Breaks Induce Histone H2AX Phosphorylation on Serine 139." *Journal of Biological Chemistry* 273 (10): 5858–68. doi:10.1074/jbc.273.10.5858.

Rogakou, Emmy P., Chye Boon, Christophe Redon, and William M. Bonner. 1999. "Megabase Chromatin Domains Involved in DNA Double-Strand Breaks in Vivo." *The Journal of Cell Biology* 146 (5): 905–16. doi:10.1083/jcb.146.5.905.

Rothman, Andrea, Paul Feinstein, Junji Hirota, and Peter Mombaerts. 2005. "The Promoter of the Mouse Odorant Receptor Gene M71." *Molecular and Cellular Neuroscience* 28 (3): 535–46. doi:10.1016/j.mcn.2004.11.006.

Sandelin, Albin, Piero Carninci, Boris Lenhard, Jasmina Ponjavic, Yoshihide Hayashizaki, and David A. Hume. 2007. "Mammalian RNA Polymerase II Core Promoters: Insights from Genome-Wide Studies." *Nature Reviews Genetics* 8 (6): 424–36. doi:10.1038/nrg2026.

Sasaki, Hiroyuki, and Yasuhisa Matsui. 2008. "Epigenetic Events in Mammalian Germ-Cell Development: Reprogramming and beyond." *Nature Reviews Genetics* 2008 (2): 129–40. doi:10.1038/nrg2295.

Schmidt, Dominic, Petra C. Schwalie, Michael D. Wilson, Benoit Ballester, Ângela Gonçalves, Claudia Kutter, Gordon D. Brown, Aileen Marshall, Paul Flicek, and Duncan T. Odom. 2012. "Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages." *Cell* 148 (1–2): 335–48. doi:10.1016/j.cell.2011.11.058.

Sedelnikova, Olga A., Izumi Horikawa, Drazen B. Zimonjic, Nicholas C. Popescu, William M. Bonner, and J. Carl Barrett. 2004. "Senescing Human Cells and Ageing Mice Accumulate DNA Lesions with Unrepairable Double-Strand Breaks." *Nature Cell Biology* 6 (2): 168–70. doi:10.1038/ncb1095.

Sen, S. K., C. T. Huang, K. Han, and M. A. Batzer. 2007. "Endonuclease-Independent Insertion Provides an Alternative Pathway for L1 Retrotransposition in the Human Genome." *Nucleic Acids Research* 35 (11): 3741–51. doi:10.1093/nar/gkm317.

Sen, Shurjo K., Kyudong Han, Jianxin Wang, Jungnam Lee, Hui Wang, Pauline A. Callinan, Matthew Dyer, Richard Cordaux, Ping Liang, and Mark A. Batzer. 2006. "Human Genomic Deletions Mediated by Recombination between Alu Elements." *The American Journal of Human Genetics* 79 (1): 41–53. doi:10.1086/504600.

Seo, J., S. C. Kim, H.-S. Lee, J. K. Kim, H. J. Shon, N. L. M. Salleh, K. V. Desai, et al. 2012a. "Genome-Wide Profiles of H2AX and -H2AX Differentiate Endogenous and Exogenous DNA Damage Hotspots in Human Cells." *Nucleic Acids Research* 40 (13): 5965–74. doi:10.1093/nar/gks287.

———. 2012b. "Genome-Wide Profiles of H2AX and -H2AX Differentiate Endogenous and Exogenous DNA Damage Hotspots in Human Cells." *Nucleic Acids Research* 40 (13): 5965–74. doi:10.1093/nar/gks287.

Serizawa, Shou, Kazunari Miyamichi, Hiroko Nakatani, Misao Suzuki, Michiko Saito, Yoshihiro Yoshihara, and Hitoshi Sakano. 2003. "Negative Feedback Regulation Ensures the One Receptor-One Olfactory Neuron Rule in Mouse." *Science* 302 (5653): 2088–2094.

Smallwood, Sébastien A., and Gavin Kelsey. 2012. "De Novo DNA Methylation: A Germ Cell Perspective." *Trends in Genetics* 28 (1): 33–42. doi:10.1016/j.tig.2011.09.004.

Speek, M. 2001. "Antisense Promoter of Human L1 Retrotransposon Drives Transcription of Adjacent Cellular Genes." *Molecular and Cellular Biology* 21 (6): 1973–85. doi:10.1128/MCB.21.6.1973-1985.2001.

Suárez, Rodrigo, Diego García-González, and Fernando de Castro. 2012. "Mutual Influences between the Main Olfactory and Vomeronasal Systems in Development and Evolution." *Frontiers in Neuroanatomy* 6. doi:10.3389/fnana.2012.00050.

Sultan-Styne, Krista, Rafael Toledo, Christine Walker, Anna Kallkopf, Charles E. Ribak, and Kathleen M. Guthrie. 2009. "Long-Term Survival of Olfactory Sensory Neurons after Target Depletion." *The Journal of Comparative Neurology* 515 (6): 696–710. doi:10.1002/cne.22084.

Sundaram, Vasavi, Yong Cheng, Zhihai Ma, Daofeng Li, Xiaoyun Xing, Peter Edge, Michael P. Snyder, and Ting Wang. 2014. "Widespread Contribution of Transposable Elements to the Innovation of Gene Regulatory Networks." *Genome Research* 24 (12): 1963–76. doi:10.1101/gr.168872.113.

Tan, L., C. Zong, and X. S. Xie. 2013. "Rare Event of Histone Demethylation Can Initiate Singular Gene Expression of Olfactory Receptors." *Proceedings of the National Academy of Sciences* 110 (52): 21148–52. doi:10.1073/pnas.1321511111.

Tchurikov, Nickolai A., Olga V. Kretova, Daria M. Fedoseeva, Dmitri V. Sosin, Sergei A. Grachev, Marina V. Serebraykova, Svetlana A. Romanenko, Nadezhda V. Vorobieva, and Yuri V. Kravatsky. 2013. "DNA Double-Strand Breaks Coupled with PARP1 and HNRNPA2B1 Binding Sites Flank Coordinately Expressed Domains in Human Chromosomes." Edited by Nick Gilbert. *PLoS Genetics* 9 (4): e1003429. doi:10.1371/journal.pgen.1003429.

Tremblay, A., M. Jasin, and P. Chartrand. 2000. "A Double-Strand Break in a Chromosomal LINE Element Can Be Repaired by Gene Conversion with Various Endogenous LINE Elements in Mouse Cells." *Molecular and Cellular Biology* 20 (1): 54–60.

Turinetto, V., and C. Giachino. 2015. "Multiple Facets of Histone Variant H2AX: A DNA Double-Strand-Break Marker with Several Biological Functions." *Nucleic Acids Research* 43 (5): 2489–98. doi:10.1093/nar/gkv061.

Upton, Kyle R., Daniel J. Gerhardt, J. Samuel Jesuadian, Sandra R. Richardson, Francisco J. Sánchez-Luque, Gabriela O. Bodea, Adam D. Ewing, et al. 2015. "Ubiquitous L1 Mosaicism in Hippocampal Neurons." *Cell* 161 (2): 228–39. doi:10.1016/j.cell.2015.03.026.

Uusküla-Reimand, Liis, Huayun Hou, Payman Samavarchi-Tehrani, Matteo Vietri Rudan, Minggao Liang, Alejandra Medina-Rivera, Hisham Mohammed, et al. 2016. "Topoisomerase II Beta Interacts with Cohesin and CTCF at Topological Domain Borders." *Genome Biology* 17 (1): 182. doi:10.1186/s13059-016-1043-8.

Vilenchik, M. M., and A. G. Knudson. 2003. "Endogenous DNA Double-Strand Breaks: Production, Fidelity of Repair, and Induction of Cancer." *Proceedings of the National Academy of Sciences* 100 (22): 12871–76. doi:10.1073/pnas.2135498100.

Viollet, Sébastien, Clément Monot, and Gaël Cristofari. 2014. "L1 Retrotransposition: The Snap-Velcro Model and Its Consequences." *Mobile Genetic Elements* 4 (2): e28907. doi:10.4161/mge.28907.

Vitullo, Patrizia, Ilaria Sciamanna, Marta Baiocchi, Paola Sinibaldi-Vallebona, and Corrado Spadafora. 2012. "LINE-1 Retrotransposon Copies Are Amplified during Murine Early Embryo Development." *Molecular Reproduction and Development* 79 (2): 118–27. doi:10.1002/mrd.22003.

Walsh, C. P., J. R. Chaillet, and T. H. Bestor. 1998. "Transcription of IAP Endogenous Retroviruses Is Constrained by Cytosine Methylation." *Nature Genetics* 20 (2): 116–17. doi:10.1038/2413.

Wicker, Thomas, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, et al. 2007. “A Unified Classification System for Eukaryotic Transposable Elements.” *Nature Reviews. Genetics* 8 (12): 973–82. doi:10.1038/nrg2165.

Yin, Bu, and Craig H. Bassing. 2008. “The Sticky Business of Histone H2AX in V(D)J Recombination, Maintenance of Genomic Stability, and Suppression of Lymphoma.” *Immunologic Research* 42 (1–3): 29–40. doi:10.1007/s12026-008-8030-4.

Young, Janet M., RaeLynn M. Endicott, Sean S. Parghi, Megan Walker, Jeffrey M. Kidd, and Barbara J. Trask. 2008. “Extensive Copy-Number Variation of the Human Olfactory Receptor Gene Family.” *The American Journal of Human Genetics* 83 (2): 228–42. doi:10.1016/j.ajhg.2008.07.005.

Zhang, Xinmin, and Stuart Firestein. 2002. “The Olfactory Receptor Gene Superfamily of the Mouse.” *Nature Neuroscience* 5 (2): 124–133.

Zhang, Y., J. Shu, J. Si, L. Shen, M. R. H. Estecio, and J.-P. J. Issa. 2012. “Repetitive Elements and Enforced Transcriptional Repression Co-Operate to Enhance DNA Methylation Spreading into a Promoter CpG-Island.” *Nucleic Acids Research* 40 (15): 7257–68. doi:10.1093/nar/gks429.

Appendix A:

Supplementary information for SV analysis in Olfr2 locus

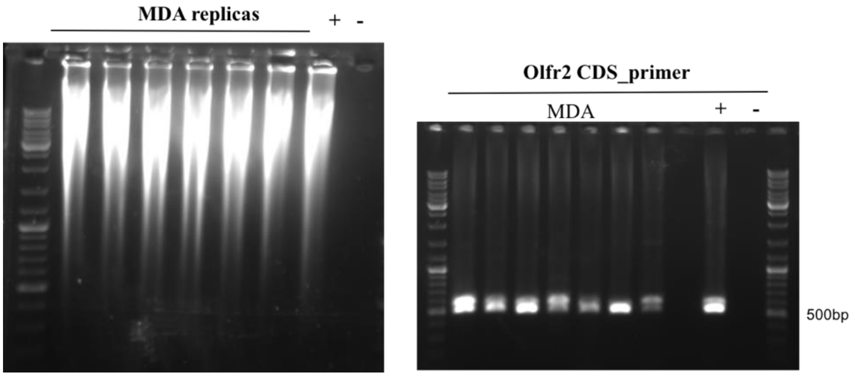


Figure-A1. MDA locus amplification PCR check. Right panel: 6 MDA amplification replicates from 10 single cells; + = MDA performed on control gDNA; - = MDA performed without input DNA (negative control). Left panel: control PCR for *Olfr2* coding sequence (CDS) on MDA amplifications.

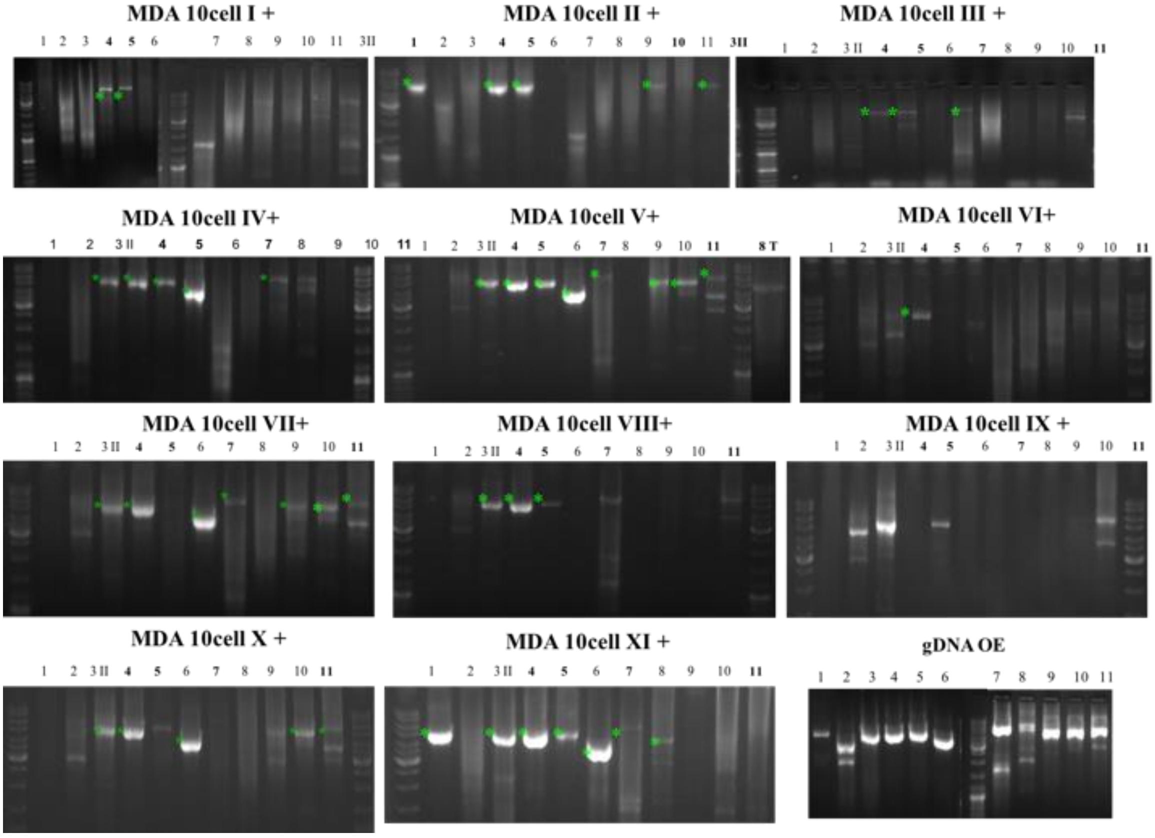

MDA 10cell I +
1 2 3 4 5 6 7 8 9 10 11 3II
MDA 10cell II +
1 2 3 4 5 6 7 8 9 10 11 3II
MDA 10cell III +
1 2 3II 4 5 6 7 8 9 10 11
MDA 10cell IV+
1 2 3II 4 5 6 7 8 9 10
MDA 10cell V+
11 1 2 3II 4 5 6 7 8 9 10 11 8T
MDA 10cell VI+
1 2 3II 4 5 6 7 8 9 10 11
MDA 10cell VII+
1 2 3II 4 5 6 7 8 9 10 11
MDA 10cell VIII+
1 2 3II 4 5 6 7 8 9 10 11
MDA 10cell IX +
1 2 3II 4 5 6 7 8 9 10 11
MDA 10cell X +
1 2 3II 4 5 6 7 8 9 10 11
MDA 10cell XI +
1 2 3II 4 5 6 7 8 9 10 11
gDNA OE
1 2 3 4 5 6 7 8 9 10 11

Figure-A2. Long range PCR amplification for Pac Bio sequencing. For each MDA replicate as for bulk OE gDNA, 50 kb of *Olfr2* locus amplification is shown. Green labels indicate examples of selected amplifications finally pooled together for sequencing.

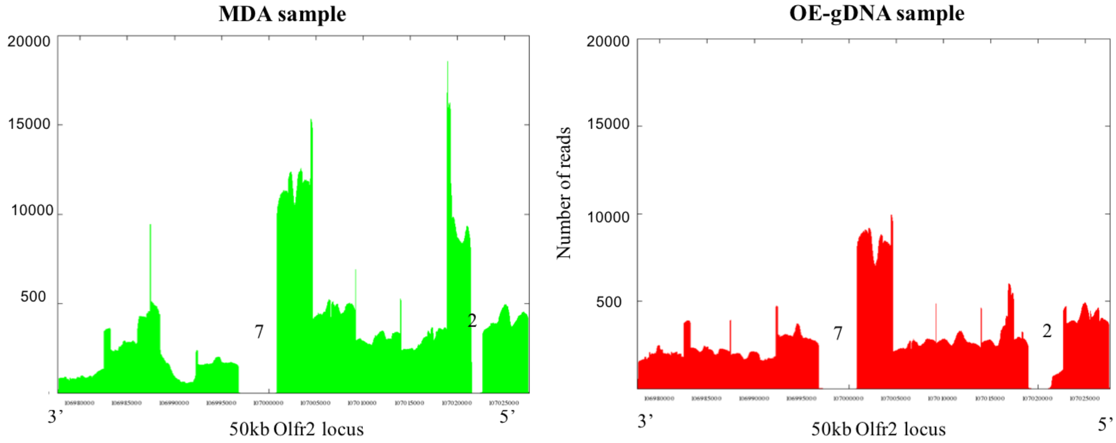


Figure-A 3. **Pac Bio sequencing coverage of 50kb *Olfr2* locus.** *Olfr2* 50 kb locus coverage in MDA sample (left) and OE sample (right) is represented plotting the number of reads for the 50 kb locus coordinates. Minus strand, which is the strand of *Olfr2* transcription, is shown. The numbers "7" and "2" indicate the amplicons which were not correctly amplified. Reads mapped in amplicon 2 coordinates are not specific.

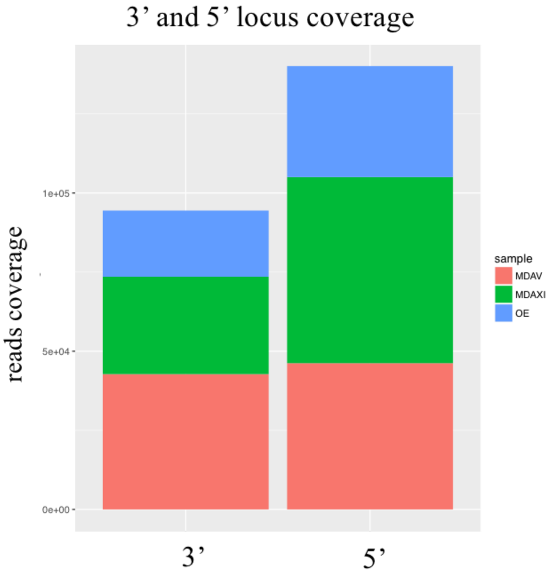


Figure-A 4. Illumina coverage at 5' and 3' amplicons. Illumina coverage for each sample comparing 5' and 3' amplicons with respect to *Olfr2* TSS. MDAV (red); MDAXI (green); OE (blue).

Appendix B:

Supplementary information for ChIP-Seq analysis

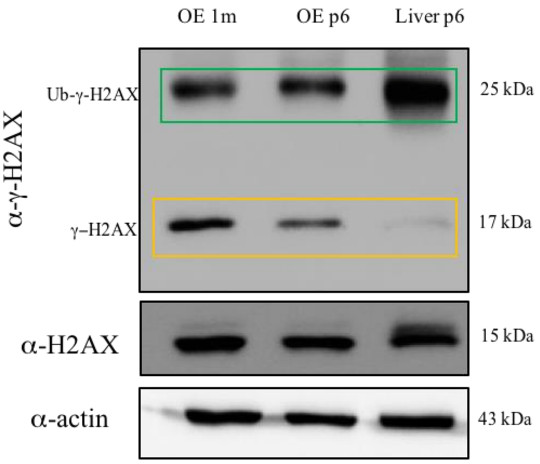


Figure-B 1. Western blot for endogenous γ-H2AX on OE and L tissues. Whole tissue protein lysates from C57BL/6J OE (p6 and 1m) and L (p6) were analyzed for endogenous γ-H2AX expression. Anti-γ-H2AX antibody recognized both phosphorylated H2AX (γ-H2AX) in the yellow square and phosphorylated plus ubiquitinated H2AX (Ub-γH2AX) in the green square, as expected from literature. Anti-H2AX was developed on the same gel and used as normalizer together with b-actin.

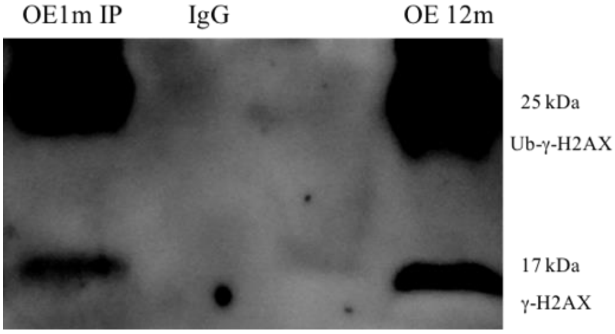


Figure-B 2. Western blot for endogenous γ-H2AX on IP samples. γ-H2AX-IP sample and IgG-IP sample protein lysates were incubated with anti-γ-H2AX antibody. Total protein lysate from OE tissues at 12m was used as positive control.

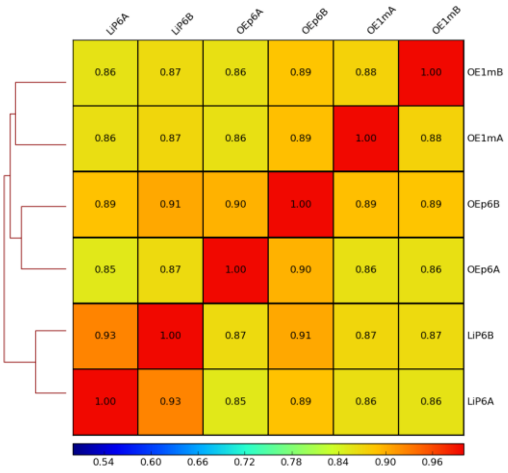


Figure-B 3. Correlation heatmap for ChIP-seq samples across all the conditions. The correlation analysis was performed by segmenting the dataset into bins of a defined length and calculating read abundance within the bins using as starting point directly the bam files used to call the peaks. The dendogram on the left grouped the biological replicates for each sample as expected.

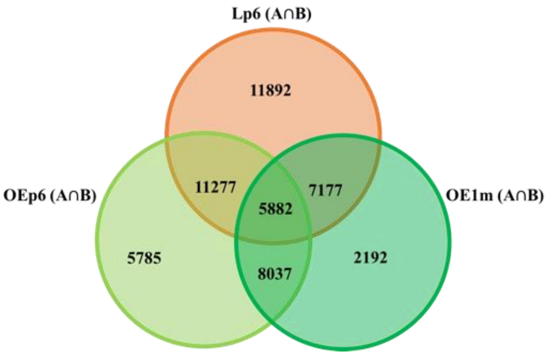


Figure-B 4. ChIP-seq sample peaks intersection. Venn diagram of peaks intersection among different ChIP-seq sample datasets.

Pol II	Liver	OlfactoryBulb	DNAseI	Brain	Liver	CTCF	Liver	OlfactoryBulb
Liver	11032 (100 %)	6948 (42 %)	Brain	369968 (100%)	89317 (53 %)	Liver	30374 (100%)	12308 (85 %)
OlfactoryBulb	6948 (63 %)	16591 (100%)	Liver	89317 (25 %)	167848 (100%)	OlfactoryBulb	12308 (41 %)	14510 (100%)

Table-B 1. ChIP-Seq Public datasets intersection. Pol II, DNAse and CTCF public datasets used for the analysis. We selected the best matching datasets available, Liver and OlfactoryBulb (OB). When OB was not available, we used dataset from whole brain. For each dataset the percentage of overlapping peaks is indicated.

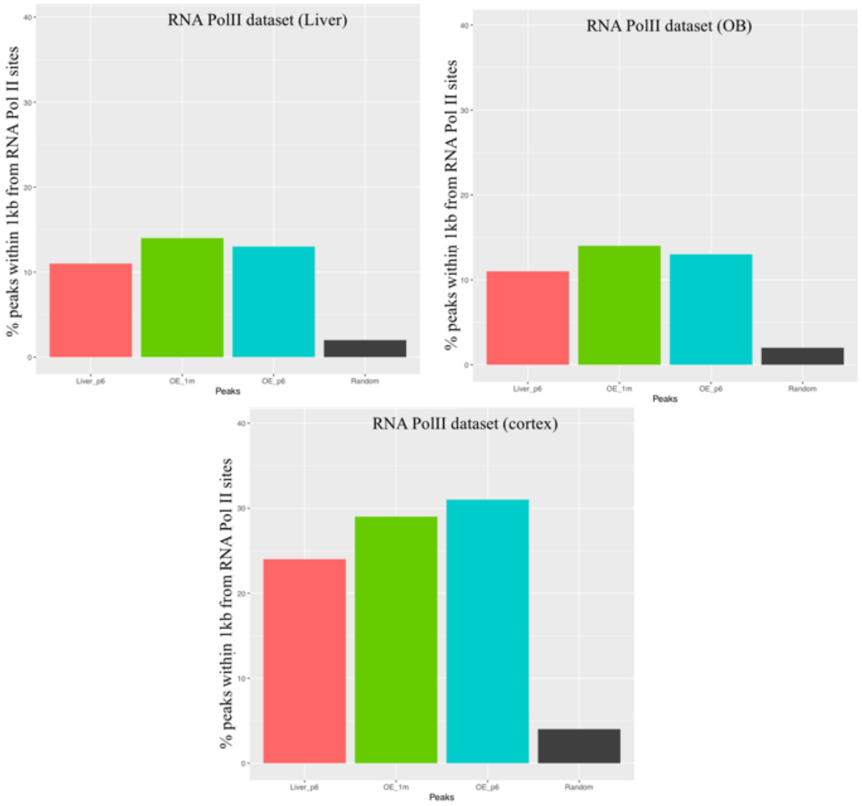


Figure-B 5. Percentage of peaks overalpping Pol II binding sites in different datasets. For each chart, the dataset used is indicated. Lp6 (red); OE1m (green); OEp6 (blue); random peaks (black).

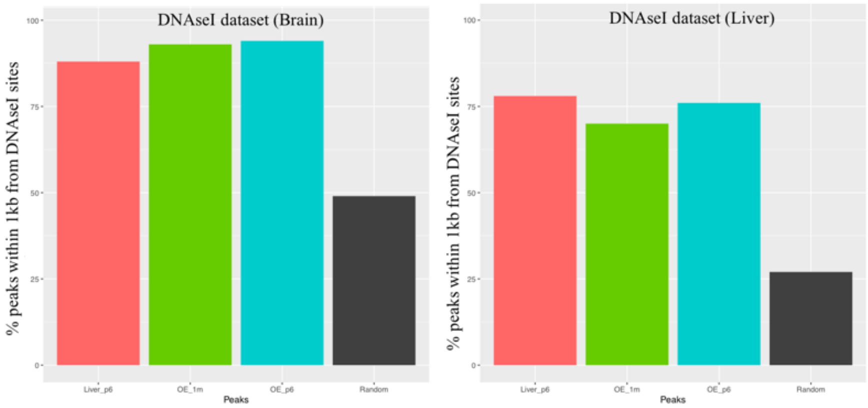


Figure-B 6. Percentage of peaks overalpping DNAseI sensible sites in different datasets. Lp6 (red); OE1m (green); OEp6 (blue); random peaks (black).

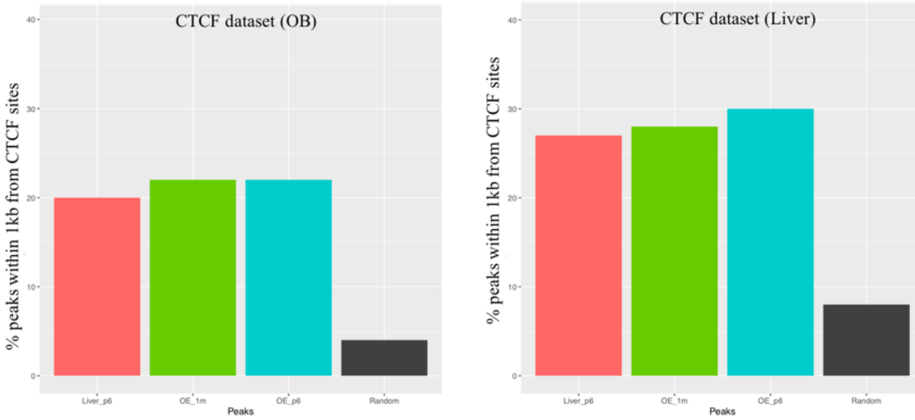


Figure-B 7. Percentage of peaks overalpping CTCF binding sites different datasets. Lp6 (red); OE1m (green); OEp6 (blue); random peaks (black)

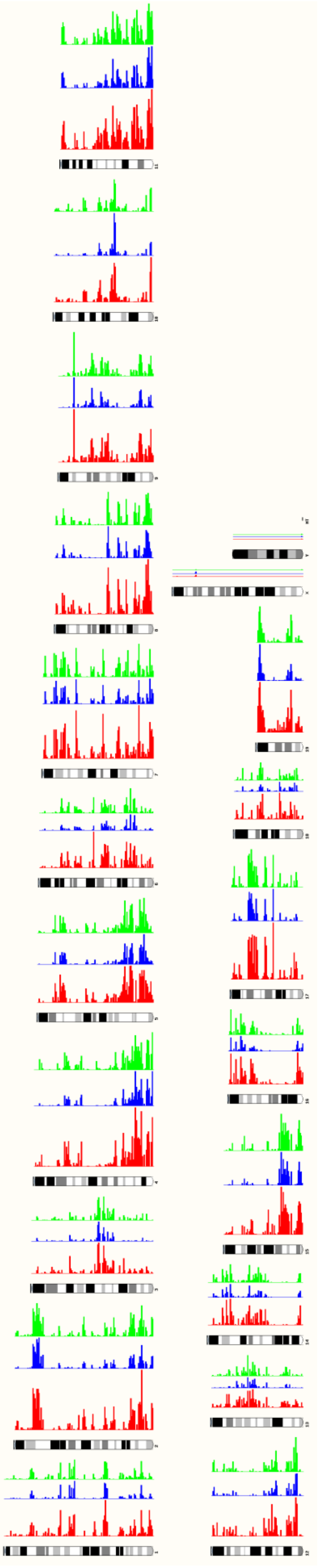


Figure-B 8. Peaks distribution with respect to the caryotype. Ensembl view of all mouse chromosomes. Chip-seq peaks are indicated with different color per sample: Liverp6 (red); OE1m (blue); OEp6 (green)

Siamo finalmente giunti alla fine di questa tesi di dottorato. Si conclude così non solo un lungo percorso lavorativo ma anche un capitolo importante della mia vita.
A Trieste ho trascorso i miei anni più intensi, spesso difficili quanto meravigliosi. Ho incontrato persone che mi lasceranno, nel bene e nel male, un ricordo indelebile.

Ringrazio il mio supervisor Stefano Gustincich, per la fiducia che ripone in me da tanti anni. Per avermi fatto conoscere alcuni degli argomenti più complessi e affascinanti della genomica e per avermi fatto capire che per fare scienza bisogna imparare a fare delle scelte. E io ancora non ho imparato a scegliere.
Ringrazio Remo per i sui consigli sinceri e per essere sempre gentile e ottimista, doti rare nel nostro campo.
Un enorme grazie è per Aurora, perché senza di lei questo lavoro non esisterebbe. Ti ringrazio Rori per la tua pazienza e per la tua dolcezza anche di fronte a situazioni estreme, come dimenticare quel ClustalW; per tutte le cicche in balcone con la tua giacca oversize "perché così non prendi freddo".
Un grazie di cuore a Margherita per la sua professionalità, coerenza e precisione. Grazie per aver letto e riletto la mia tesi, trasmettendomi grande tranquillità. Ma ti ringrazio soprattutto perché nel periodo più sconfortante del mio percorso, senza rendertene conto, mi hai trasmesso la voglia di lottare per non smettere di fare questo lavoro.
Un grazie va a tutte le colleghe e i colleghi dell'SG lab.
Ringrazio Marta per avermi dato sempre un grande esempio di integrità e correttezza. Per essere stata presente non solo come collega ma anche come amica, per tutte le pseudo discussioni scientifiche e per avere reso una trasferta di lavoro, una piacevole gita. Ringrazio Alice per essere stata la prima ad avermi pazientemente insegnato come si pensa e come si fa questo mestiere. Grazie a Carlotta per la sua sensibilità e incredibile gentilezza. Grazie davvero Carli, perchè sappiamo tutte che senza di te le nostre tesi non avrebbero un'impaginazione. Ringrazio Chiara per la sua forza e determinazione e per saper diffondere quell' allegria un po' pazza per tutto il lab.
Ancora grazie a Sara e Marta, per i loro consigli. A Gabriele e Abram per il loro buon umore. Ringrazio anche i miei futuri colleghi Genovesi perché già in in pochi giorni hanno saputo farmi sentire a casa.
Molto più di un grazie è per Lavi e Fra. Grazie Lavi per la tua dolcezza goffamente mascherata. Grazie Fra per la tua sensibilità e sincerità che mi hanno risollevata quando ne avevo più bisogno. Grazie a tutte e due per essere le migliori amiche e colleghe che

una persona possa desiderare. Vi ringrazio per aver condiviso con me ogni gioia e per aver diviso per tre ogni dolore di questo dottorato. Siete state per me la dimostrazione vivente di come l'unione faccia davvero la forza e di come la stima reciproca possa far dimenticare ogni fallimento. Vi ringrazio per essere "vacche" proprio come me con tutti i vostri pregi e con tutti i vostri difetti. Vi ringrazio per continuare a farmi credere di essere la vostra locomotiva senza rendervi conto che una locomotiva senza vagoni perderebbe presto il senso del suo viaggiare.

Ringrazio poi Gianluca Pietra perché senza di lui starei ancora cercando le famigerate cellule GFP. Ringrazio Lisa e Robi per avermi pazientemente insegnato ogni volta che nel panico non avevo idea di cosa fare. Ringrazio Andres per per avermi dedicato il suo tempo con tanta gentilezza. Grazie a Jess e Mica per essere sempre a nostra disposizione. Ringrazio la segreteria studenti per la loro efficienza gentile. Ringrazio le bariste e i cuochi per saperci coccolare come a scuola.

Ringrazio mia Mamma e mia Papà perché il desiderio di renderli orgogliosi è da sempre la mia più grande forza.

Grazie a mia sorella Alessia perchè in questi ultimi anni la sua determinazione mi è stata di grande esempio. Ti ringrazio Ale perché poter discutere di scienza in famiglia è un grande privilegio, soprattutto perché nessuno ci capisce. Grazie a Sandra e Piero per essere sempre presenti, attenti e affettuosi come una seconda famiglia.

E grazie anche a te Kellona che da quasi vent'anni scodinzoli quando mi vedi, pur non avendo davvero la coda. Un grazie alla mia Giandù per tutte le fusa che mi hanno tenuto compagnia durante la scrittura di questa tesi.

Un enorme grazie è per una persona speciale, Carlotta, anzi Carlice. Ti ringrazio per essere esattamente così come sei, per ogni tuo "affrontuleggiu", per ogni volta che abbiamo campeggiato nelle nostre molteplici stanze, per tutti gli esami preparati, sudati e rubati. Per quell'erasmus al circolo polare artico che solo noi avremmo potuto scegliere ma che senza di te sarebbe stato un vero incubo. Ti ringrazio per quel "Nescafè-cappuccino" che bevevamo in via bel poggio tutte le mattina anni fa e che mi sono ricomprata, solo per portarmelo a Genova. Per tutte le volte che ci siamo chieste se mai saremmo arrivate fino a qui. Ti ringrazio perché qui ci siamo arrivate insieme.

Un grazie sincero va alla mia amica Camilla che nonostante la distanza è come se vivessimo ancora in città-studi a cinque minuti l'una dall'altra. Ti ringrazio Cami per non farmi sentire mai sola, per avermi insegnato a non mollare e a rialzarmi sempre col sorriso. Grazie per esserci sempre e per essere da sempre il mio punto di riferimento.

Ringrazio la Fra, quella mia amica un po' matta e molto speciale. Ti ringrazio Chicca, per tutte quelle giornate in camera tua a studiare come mai più sono stata capace di fare. Grazie per le ore passate ad inventarci, tra una sigaretta e l'altra, strategie per evadere da un mondo che ci è sempre stato un po' troppo stretto. Grazie per avermi insegnato a lottare per quello che reputavo giusto anche se poi magari non lo era. Grazie per essere la mia fonte di ispirazione.

Ringrazio tutti i miei amici vicini e lontani. Grazie ad Andrea, Silvia, Toti, Lele, Cate, Fede, Matteo, Alessia, Carmen, Su, Isa, Fab, Moni, Tramo, Carlo, Ceci, Linda, Ale, Alejandro, Richard e tutti quelli che mi sto dimenticando perché come al solito sono sempre in ritardo. Vi ringrazio perché so che siete sempre lì quando ne ho bisogno.

Un grazie speciale a Giulia. Grazie per la tua dolcezza, per esserci prima ancora che io capisca di averne bisogno, per essere come una sorella. Grazie per aver condiviso con me l'amore per questa città che ci ha saputo tenere vicine e che, sono sicura, non saprà tenerci lontane per molto. Grazie per aver scelto Simone, perché così posso dirgli grazie per essere una persona meravigliosa, per avermi portato al mare col gesso, per aver coccolato il mio limone e per avermi preparato il polpettone di tonno. Ma grazie soprattutto perché "il 2 settembre 2018 è un sabato". Grazie a Lucia, perchè sei una delle rare persone che sa tirare fuori la mia parte più bella e sincera. Un grazie anche a Mat, Robo e Francesca per avermi fatto rubare il cuore da due cavalli. Grazie perché i due anni passati con voi mi hanno dato la forza di superare col sorriso anche le giornate lavorative più nere.

Per finire, sappiamo che Il mio Grazie più importante è per la persona senza la quale non avrei mai saputo ringraziarvi con la stessa sincerità. Ti ringrazio Gio per rendermi ogni giorno una persona migliore.