Scuola Internazionale Superiore di Studi Avanzati

PhD course in Functional and Structural Genomics

Detecting LINE-1 mediated structural variants from sequencing data: Computational characterization of genomic rearrangements occurring in human post-mortem brains in the pathologic context of Alzheimer's disease and in mouse olfactory epithelium at physiological conditions.

Thesis submitted for the degree of *"Philosophiae Doctor"*

Candidate:
Aurora Maurizio

Supervisor:
Prof. Stefano Gustincich

Co-supervisor:
Dr. Remo Sanges

Academic Year 2016/2017

Relax, there is nothing wrong with the transposition paper.
People aren't ready for this yet.

Barbara McClintock to Mel Green
1969

Abstract

One of the most intriguing discoveries in the recent decades is that "the genome is a work in progress", constantly gaining and loosing chunks of sequence, in order to provide new potentially favorable combinations for adaptation. The old genetic concept that the genome is static, has prevailed until the 1950s, when it was first suggested that there is a lot more to DNA than just genes. Indeed, genetic material is dynamic and the greatest part of most organisms genome is occupied by non-coding DNA, especially DNA fragments deriving from elements capable of moving to new locations: transposable elements (TEs). TEs are mobile DNA fragments, whose remnants occupy nearly half of mammalian genome and up to 90% of the genome of some plants (SanMiguel et al., 1996). Since almost the 1950, when Barbara McClintock discovered them in maize (McClintock, 1951), extensive efforts have been devoted to understand the function of these interspersed repeats. Unfortunately, due to their imperceptible activity, TEs have been largely underappreciated and dismissed as 'junk DNA'. When researchers identified long interspersed element-1 (**LINE-1** or L1) insertions to be responsible for haemophilia A, in 1988 (Kazazian et al., 1988), TEs gained new attention. LINE-1 elements are the only active, autonomous TEs present in the mammalian genome. These molecules, able to create polymorphisms among individuals and genomic mosaicism among populations of cells, are major sources of structural variations in humans and are responsible for 124 genetic diseases (Hancks and Kazazian, 2016). In particular, the discovery of LINE-1 mobilization in neurogenesis (Muotri et al., 2005, Coufal et al., 2009) urged the scientific community to investigate the potential involvement of mobile elements in neuropsychiatric disorders (Bundo et al., 2014 , Guffanti et al., 2016, Shpyleva et al., 2017) and neurodegenerative diseases (Li et al., 2012).

Nowadays, that LINE-1 activity has been proven in vitro (Moran et al., 1996) and in vivo (Ostertag et al., 2002), the establishment of the real rate of retrotransposition remains a challenge for scientists in this field. One of the main reasons is the lack of reliable methods to detect elements present in a small minority of cells, or unique to a single cell. This is exacerbated by the technical complexity of deconstructing non-reference, chimeric regions of the genomes through experimental or computational means.

Until very recently, assays using ligation-mediated PCR techniques have been considered the gold standard for proving and quantifying current retrotransposon activity. Unfortunately, both positive and negative changes in the number of repeats detected with these techniques can occur by a multitude of mechanisms not directly related to the retrotransposition molecular mechanism. Among the most common retrotransposition-independent rearrangements we can remember non-homologous recombination mediated deletions and duplications.

In this thesis, we focus on the effects of **LINE-1** elements on genome stability. To this purpose, we describe three different bioinformatics methods for the study of the hallmarks of LINE-1 mediated genome instability: direct **insertion**, double strand breaks (**DSB**s), and post-insertional **rearrangements**.

The increasing availability of large amounts of sequencing data produced by Next-generation sequencing technologies (NGS), calls for the development of genomics techniques targeted for retrotransposons study, to fully exploit the available resources. Therefore a scalable approach, such as the Splinkerette Analysis of Mobile Elements (SPAM) method proposed here, is of substantial interest to assist the current and future development in the study of transposable elements. Importantly, SPAM allowed us to target exclusively full-length LINE-1 elements (FL-L1) present in the frontal cortex (FC) and the kidney (K) of Alzheimer's disease affected patients (AD) and controls (CTRL) and to test if LINE-1 polymorphisms can be a relevant source of structural variants associated with AD risks. This is accomplished combining a PCR-based enrichment of FL-L1 elements with an ad hoc bioinformatic pipeline. The remarkable performance of our integrative method is achieved in part because of its ability to detect LINE-1 **insertion** sites with great precision and in part because of its scalability. Embedded in the methodology is the flexibility to perform the same technique in different organism and considering different classes of TEs. Using SPAM, we observed for the first time an unexpectedly high level of retrotransposition in the kidney. In association with the SPAM approach, we performed TaqMan based copy number variation (CNV) analysis to evaluate the content of potentially active L1s in the different tissues of AD and CTRL individuals. Finally, we employed high density arrays to compare the occurrence of FL-L1 elements in correspondence of genomic variations detected in AD and CTRL patients. Overall, we show that the content of FL-L1 sequences in AD is significantly lower than

in CTRL, that de-novo integrations are not associated to the disease but that FL-L1 polymorphisms can be a relevant source of structural variants associated to AD risk.

Then, we investigated which mechanism underlies the regulation of olfactory receptor choice in mouse olfactory epithelium, characterizing Olfr2 locus-specific genomic **rearrangements**. To perform this task, we combined PacBio single molecule sequencing with a complementary high-fidelity paired-end Illumina sequencing for accurate identification of breakpoints in a locus where a very high repeat concentration, especially LINE elements, provides more chances for recombination events to occur between retrotransposon fragments. Surprisingly, the analysis revealed hundreds of heterozygous structural variants in the vicinity of the locus, among which deletions are the most abundant. The presence and characteristics of particular genomic features associated with the observed deletions, suggest us that the same principal mechanism is operating in the formation of all the deletions: micro-homology mediated end joining (MMEJ) of double strand breaks (DSB). Further experiments will tell us if the observed SV are involved in the regulation of the receptor.

Intrigued by the idea that SV in Olfr2 region could be initiated by LINE-1 induced **DSB** lesions, for the first time, we profiled endogenous double strand breaks (DSB) distribution in mouse olfactory epithelium (at p6 and 1m) and liver (at p6). To this purpose, we performed a chromatin immunoprecipitation and sequencing (ChIP-Seq) analysis of γ-H2AX (an early response marker for DNA-DSBs). Little is known about the differential distribution of γ-H2AX throughout the genome at physiological conditions. In the light of our results, γ-H2AX signal is stronger in gene rich, transcribed regions where it co-localizes with regulatory sites. Thus, suggesting a possible involvement of DBSs in resolving topological stress and promoting interactions between regulatory regions.

The research described in this thesis is aimed at enhancing our understanding of the consequences of LINE-1 activity and their potential importance in health and disease.

Abbreviations

AD	Alzheimer's disease
AIS	Annotated integration sites
ARMDs	Alu-recombination-mediated-deletions
ATP	Adenosine triphosphate
BER	Base excision repair
CDS	Coding sequence
CFS	Common fragile sites
ChIP-Seq	Chromatin immune-precipitation sequencing
CNV	Copy number variation
DRS	Direct repeat site
DSB	Double strand break
EN	Endonuclease
EOD	Early Onset Alzheimer Disease
ERFS	Early replication fragile sites
FAD	Familial Early Onset Alzheimer Disease
FC	Frontal Cortex
FL-L1	Full length LINE-1
FOSTES	Replication Fork Stalling and Template Switching
G4 DNA	G-quadruplex DNA
GC	Global cell
gDNA	Genomic DNA
GFP	Green fluorescent protein
γH2AX	Phosphorylated-H2AX
GPCR	G-protein coupled receptor
H_2O_2	Hydrogen peroxide
HC	Horizontal cell
Hi	Hippocampus
HO•	Hydroxyl radical
HR	Homologous recombination
IHC	Immunihistochemestry
IP	Immunoprecipitation
IR	Inverted repeat
IR	Ionizing radiation
IRES	Internal-ribosomal entry site
IS	Integration sites
ITR	Inverted terminal repeat
K	Kidney

L	Liver
L1	Long interspersed nuclear element-1
LCM	Laser capture microdissecter
LCR	Locus control region
LCR	Low copy repeats
LINE-1	Long interspersed nuclear element-1
LSD1	Lysin demethylase 1
LTR	Long terminal repeat
MDA	Multiple displacement amplification
mf	MapFragment
MHC	Moajor histocompatibility complex
mm	Mismatch
MMEJ	Microhomology mediated end joining
MMR	Mismatch repair
MOE	Mouse olfactory epithelium
NAHR	Non-allelic homologous recombination
NanoCAGE	Cap analysis of gene expression
NHEJ	Non-homologous end joining
NIS	Non annotated integration sites
NPC	Neural precursor cell
O/N	Over night
O_2^-	Superoxide anion
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
PIS	Polymorphic integration sites
Pol II	RNA polymarase II
RNP	Ribonucleoprotein
ROS	Reactive oxigen species
RT	Retrotranscriptase
RT	Room temperature
RT-qPCR	Real time quantitative PCR
SC	Supporting cell
SINE	Short interseprsed nuclear element
SRS	Serial replication slippage

SSA	Single strand annealing
SV	Structural variation
TAF1	TATA-Box Binding Protein Associated Factor 1
TBP	TATA-binding protein
TE	Trasposable element
TNRs	Trinucleotide nucleotide repeat
TOP 2	Topoisomarase II
TPRT	Target primed reverse trancription
TSD	Target site duplication
TSS	Trascription start site
UPR	Unfolded protein response
UTR	Untranslated region
WB	Western blot
WGA	Whole genome amplification

List of Figures

List of tables

1 Introduction

1.1 Structural Variations (SVs)

Genomic mutations, affecting DNA sequence length and orientation, are major contributors to phenotypic diversity and disease (Alkan et al., 2011).

Differences between genomes can range from single nucleotide polymorphisms to large rearrangements called structural variants (**SV**s) such as insertions, deletions, inversions, translocations and copy-number-variations. These modifications can be beneficial, neutral or deleterious and act in concert to produce an enormous number of possible genomic configurations, driving evolution (Ewing and Jensen, 2014).

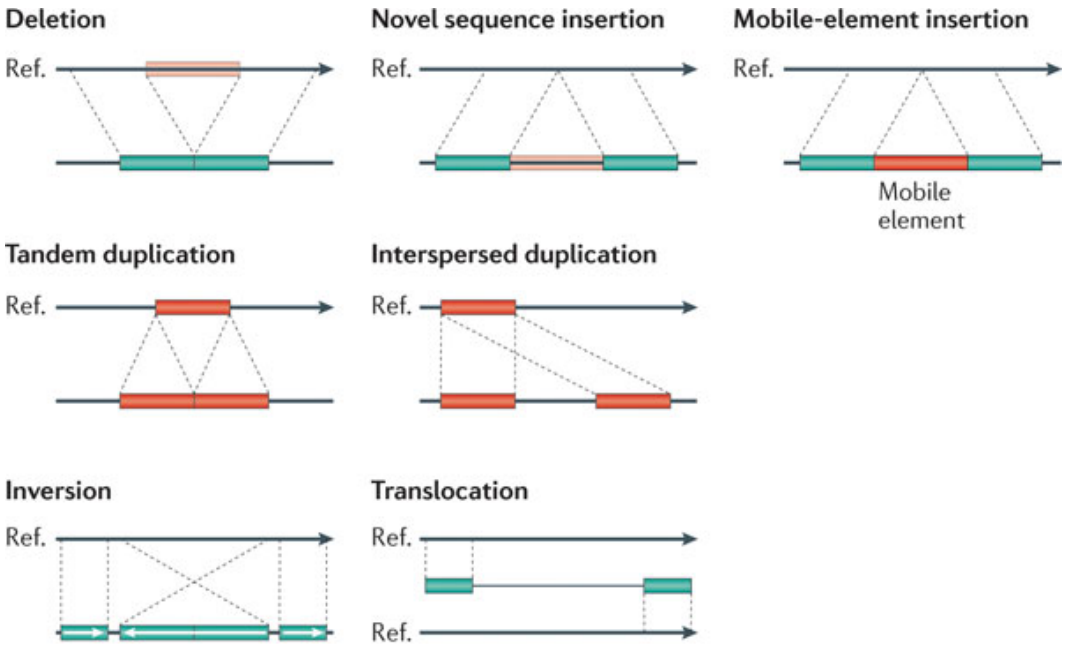


Figure 1.1 Classes of SV. The schematic depicts the most common classes of SV: deletions, novel sequence insertions, mobile-element insertions, tandem duplications, interspersed duplications, inversions and translocations (Alkan et al., 2011).

In the mouse genome for example, 29% of SVs result from DNA recombination-, replication- and repair-associated processes, but 54% is due to retrotransposons: LINEs (25%), LTRs (14%) and SINEs (15%), followed by satellite repeats (15%) and pseudogenes (2%) (Yalcin et al., 2011). In the human genome, retrotransposable elements occupy an impressive 40%, and are responsible for nearly the 10% of the reference specific indels larger than 100 bp (Xing et al., 2009). From these numbers is evident the key role of retrotransposons in mediating genome stability. Retrotransposable elements, dynamically contribute to genome reshaping by insertion mediated expansion and post-insertion mediated rearrangements.

Unfortunately, although increasing evidence highlights the important role of transposable elements mediated SVs in gene regulation and disease (including genetic disorders, psychiatric problems, and cancer), their global impact is still largely uncharacterized. This is mostly due to the technical complexity of deconstructing chimeric regions of the genomes through experimental or computational means.

Nowadays, high-throughput sequencing, is driving a quiet revolution in genetics and genomics. This technology, allowing the precise, quick and reasonably priced sequencing of multiple genomes, provides new opportunities for reliably detecting genomic polymorphisms in different individuals and cells. At the same time, the necessity to discover patterns in large, often noisy and overwhelming datasets, demands efficient bioinformatic tools for their analysis.

To this end, this thesis aims to carry out a detailed computational analysis on transposable elements mediated genomic rearrangements in a physiologic context such as olfactory receptor choice and a pathologic context such as Alzheimer's disease.

The important role of repetitive sequences in the generation of SVs is examined in this chapter.

1.2 Repetitive elements (REs)

Approximately 50% of the human genome and 40% of mouse genome consists of repeated sequences including **satellite-DNA (satDNA)**, segmental duplications (SDs) also called **low copy repeats (LCRs)**, **pseudogenes**, and **transposable elements (TEs)** (Muñoz-López and García-Pérez, 2010).

This great proportion of genome occupancy may be explained by the important role of repeated elements in genetic variation and regulation (Feschotte and Pritham, 2007). Repeated elements actively reshape the genome through a balanced give-and-take of sequence that creates diversity among individuals and variability among populations of cells, with important implications for health and disease (O'Donnell and Burns, 2010). Their regulation properties include chromatin preservation and nuclear organization, regulation of coding sequence expression of nearby genes and damage-repair through recombination mechanisms (Shapiro and von Sternberg, 2005).

In the following paragraphs, we briefly consider each class of repeats.

1.2.1 **Satellite-DNA**

Microsatellites or Simple Sequence Repeats (SSRs), are tandem repetitions of mono-di-tri or tetra nucleotides present in coding and non-coding regions of the genome (occupying the 3% of human and mouse genome) that may vary in length between individuals and generations. For this reason, when the repetitions are located in neutral regions of the genome, microsatellites can be employed for DNA fingerprinting and identification purposes. On the other hand, microsatellite expansions occurring within genes can lead to severe diseases such as fragile X syndrome (Richards et al., 1991) and Huntington's disease (Budworth and McMurray, 2013).

Minisatellites or variable number of tandem repeats (VNTRs) are repetitions of longer stretches of nucleotides (10-100 bp), generally GC rich and mainly associated with constitutive heterochromatin. Their function is reflected by their location in the genome. VNTRs are concentrated in delicate regions, such as centromeres (in mouse) and

telomeres (in humans) where they protect chromosome ends from losing coding sequence during cell divisions. Moreover, these repeats have also been implicated in centromere condensation, sister chromatid pairing, gene regulation and imprinting. Associated with chromosomal fragile sites, satellite-DNA, constitutes a very unstable part of the genome and is prone to rearrangements (Ramel, 1997).

1.2.2 **Low copy repeats (LCRs)**

Low copy repeats, also called segmental duplications, are long sequences (from 1 kbp to 400 kbp) characterized by a high degree of identity (commonly >95%) that occur mainly in centromeric, pericentromeric and telomeric regions, accounting for the ~5% of human DNA and of ~2% of mouse DNA (Sharp et al., 2005).

LCRs, are highly dynamic regions that mediate recombination events resulting in genomic instability and Non-Allelic-Homologous-Recombination (NAHR) mediated rearrangements (including deletions, duplications, inversions, translocations), that facilitate the formation of copy number variations (CNVs) (Bailey et al., 2004). LCR rearranged regions often contain genes, gene fragments, pseudogenes, ERV sequences and transposable elements such as LINEs (long interspersed repetitive elements) and SINEs (short interspersed nuclear elements). In particular, an enrichment of Alu repeat sequences has been reported at the boundaries of human segmental duplications, suggesting repetitive-element-homology-based-recombination as a possible source of LCR expansion. Interestingly, mouse segmental duplications are enriched in LTR and recent LINE-1 retrotransposons but (unlike humans) not in SINEs (Sheen et al., 2000) .

LCR-mediated CNVs can arise both meiotically and somatically as shown by the finding that identical twins can differ in CNVs and that different organs and tissues vary in copy number in the same individual (Hastings et al., 2009b). In humans CNVs appear to be implicated in evolution as well as in predisposition to dozens of neurological disorders such as autism (Sebat et al., 2007), schizophrenia (Walsh et al., 2008), intellectual disability (Cooper et al., 2011) and genomic disorders like Prader-Willi and Angelman syndromes (Ledbetter et al., 1981). Genomic disorders are a group of diseases caused by

rearrangements of the human genome due to inherent genomic instability that results in susceptibility to structural variation mutagenesis. The exceptional presence of elevated levels of aneuploidy and retrotransposition in neurons compared to other cell types encourages the suspect that somatic genome variation may contribute to functional diversity in the human brain.

Changes in copy number might change the levels of expression of genes included in the regions of variable copy number, provide the genes with new domains and therefore new functionalities and create gene sequence redundancy so that some copies become free to evolve new or modified functions or patterns of expression, while other copies maintain the original function (Goodier and Kazazian, 2008). Intrigued by these findings, in this thesis, we compare the occurrence of FL-L1 elements in correspondence of CNVs detected in the genome of Alzheimer's disease affected patients and controls, in order to characterize the repeat content in genomic variations potentially associated with the neurodegenerative disease.

1.2.3 **Pseudogenes**

Pseudogenes are not-fully-active gene copies that contain mutations in the coding sequence (Tutar, 2012). These elements can originate by spontaneous inactivating mutations, by transposable elements activity or by decay of genes that have been duplicated during evolution. According to the formation mechanisms they can be divided in two groups: processed pseudogenes and unprocessed pseudogenes. Processed pseudogenes are formed through LINE retrotransposition. These elements normally do not contain an upstream promoter sequence, nor introns, end in a poly(A) tail, and are flanked by short direct repeats. Unprocessed pseudogenes, originated from decay of duplicated genes, have introns and regulatory sequences but are usually inactive due to frameshift mutations and premature stop codons.

Among the potential functions of pseudogenes, we can list: gene expression regulation, formation of chimeric transcripts and creation of diversity reservoirs.

1.2.4 Transposable elements (TEs)

Transposable elements (TEs) occupy an impressive 45% of the human genome and 38% of mouse genome (Muñoz-López and García-Pérez, 2010). Interestingly, the genome of simpler organisms like fish, fly and worm is composed by a lower fraction of transposable elements (respectively 10%, 12% and 10%) suggesting a functional role of "junk" DNA in complex organism development and evolution.

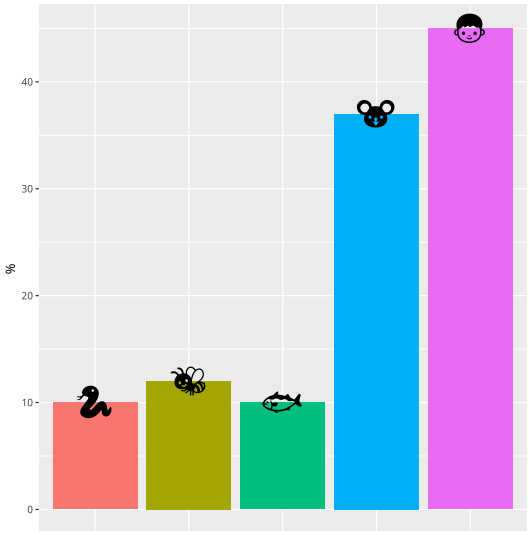


Figure 1.2.4 TE abundance among different organisms. Transposable elements (TEs) occupy the 45% of the human genome, the 38% of mouse genome, 12 % of fly genome and 10% of fish and worm genomes.

Transposable elements are indeed a rich source of evolutionary novelties and play an important role in adaptation providing new genomic material to the cell and creating new sequence combinations that may confer a fitness advantage in a population (Feschotte and Pritham, 2007). Moreover, due to their ability to respond quickly to environmental changes and stress conditions (e.g. chemicals, temperature, starvation, DNA-damage, osmotic shock and oxidative stress), TEs translate changes in the external environment

into changes at the genomic level (Casacuberta and González, 2013). Not surprisingly, TEs compose the largest part of most plant genomes. These are only few of the potential benefits of TEs which have also many negative effects including severe genetic disorders (Payer et al., 2017) and cancer (Burns, 2017). Given the potential deleterious effects of TEs the host has generated multiple mechanisms controlling their proliferation including methylation of repetitive element sequences (Bourc'his and Bestor, 2004) and chromatin condensation, RNA interference (RNAi)-based silencing (Aravin et al., 2008), APOBEC3 mediated nucleic acid editing (Refsland and Harris, 2013) and accumulation of repeat proteins like ORF1p in discrete cytoplasmic aggregates called stress granules (SG) in cell stress conditions (Goodier et al., 2007). Defects in these surveillance mechanisms were shown to increase transposable element activity with potentially dramatic consequences.

1.2.4.1 Classes of TE

There are two major groups of transposable elements, distinguishable by their transposition mechanism. Class II elements or DNA transposons comprise about 3% of the human genome and the 4% of the mouse genome and move by a so-called cut-and-paste mechanism(Lander et al., 2001; Mouse Genome Sequencing Consortium et al., 2002). Since no active DNA-transposons are present in mammals we will move directly to class I elements or retrotransposable elements (REs) (Campos-Sánchez et al., 2016). Retrotransposons move by a copy-and-paste mechanism involving reverse transcription of an RNA intermediate and insertion of its cDNA copy at a new position within the host genome. On the basis of the presence or absence of long terminal repeats (LTRs), all retrotransposons can be divided into two major groups. The first group consists of *LTR-containing elements* like LTR retrotransposons and tyrosine recombinase retrotransposons. The second group is called *non-LTR retrotranspsons*, and the main representatives of this group are long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and processed pseudogenes. Endogenous retroviruses (ERVs) and LINEs are referred to autonomous REs because they encode the proteins necessary for their proliferation and transposition. SINEs, SVA and processed

pseudogenes are non-autonomous elements and take advantage of LINE enzymatic machinery to retrotranspose.

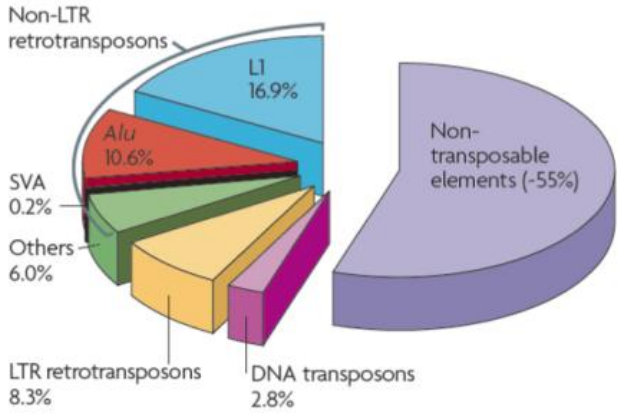


Figure 1.2.4.1 The transposable element content of the human genome. About 45% of the human genome can currently be recognized as being derived from transposable elements, the vast majority of which are non-LTR retrotransposons such as LINE-1, *Alu* and SVA elements. Figure readapted from (Casacuberta and González, 2013).

1.2.4.1.1 Long terminal repeat (LTR) containing elements

LTR-containing elements occupy about 8% of the human genome and the 10% of the mouse genome (Jern and Coffin, 2008). The consensus structure of LTR-retrotransposons contains long terminal repeats (LTRs) in direct orientation, a *gag* gene, encoding for a structural protein with nucleic acid binding activity, and *pol*, which encodes polyprotein with protease, reverse transcriptase, ribonuclease H, and integrase activities but lacks the *env* (envelope) gene present in the retroviruses (Novikova, 2009). Their activity is reportedly very limited in humans, but they still carry enormous potential to regulate gene expression and gene networks (Lamprecht et al., 2010). For example, the LTR of Endogenous retroviruses (ERVs) contain regulatory sequences such as promoters, enhancers and polyadenylation signals. ERVs constitute about 5% of the human DNA (Khodosevich et al., 2002) and ~10% of mouse DNA (Jern and Coffin, 2008) end are thought to be the inactive residues of ancient germ-cell retroviral infections. Most of the ERVs present in mammalian genomes result from homologous recombination between

two LTR and therefore lack the internal genes and contain *solo LTRs.* Even if most of the ERV present in the mammalian genome is unable to transcribe and transpose, ERVs have been proposed to be involved in cancer (Kassiotis, 2014), multiple sclerosis (Antony et al., 2011) and schizophrenia (Leboyer et al., 2013).

1.2.4.1.2 Long interspersed nuclear elements (LINEs)

Long interspersed nuclear elements (LINEs) are autonomous non-LTR retrotransposons that occupy the 17% of human genome (Lander et al., 2001) and the 19% of mouse genome (Mouse Genome Sequencing Consortium et al., 2002). LINEs-1 (or L1s) are believed to be the only currently active autonomous transposable-elements in humans. However, most of the LINE-1s present in the genome are inactive due to mutations, truncations and rearrangements. Only 100 elements are potentially able to retrotranspose in human and 3000 in mouse (Erwin et al., 2014).

The human FL-L1 is 6 kb long. It has a 900-nt-long 5′ untranslated region (UTR) that functions as an internal promoter for RNA polymerase II, two open reading frames (ORF1 and ORF2), a short 3′UTR, and a poly(A) tail. The mouse LINE-1 5′ UTR is distinguished from the human one by having tandem repeats. ORF1 encodes for an RNA binding protein, while ORF2 encodes for a protein with endonuclease and reverse transcriptase activity (Scott and Devine, 2017). LINE-1 integrations usually present typical hallmarks such as frequent 5' truncations, the presence of a 3' poly(A) tail and variable-length target site duplications (TSDs). These elements can potentially integrate at a very large number of sites in the genome since their endonuclease preferentially cleaves DNA at a short consensus sequence but, because of the specificity of the consensus motif, they are enriched in AT-rich genomic regions, that present a low-recombination frequency and are gene-poor (Jurka, 1997).

Endonuclease-independent retrotransposition occurs when LINE-1s integrate into already present DNA lesions, like the ones present in the telomeres, resulting in retrotransposon-mediated DNA repair. In this case, since retrotransposition is not involved, LINE-1s integrate at atypical target sequences, are mainly truncated at their 3' ends and lack TSDs (Morrish et al., 2002).

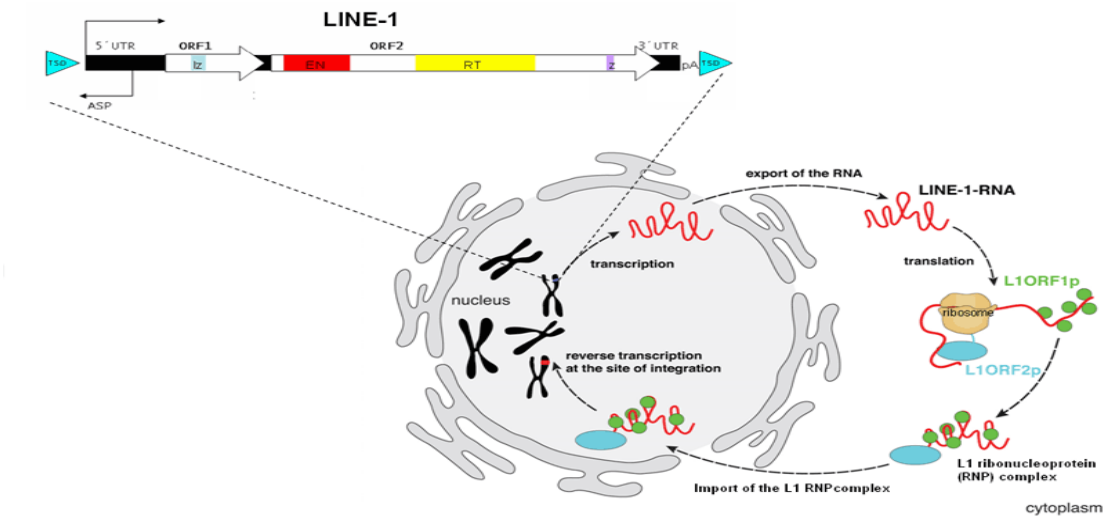


Figure 1.2.4.1.2 LINE-1 retrotransposition cycle. LINE-1 mRNA (red) is exported into the cytoplasm, translated, and L1 encoded proteins (LINE-1 ORF1p, LINE-1 ORF2p) bind to their own mRNA (cis preference) and form ribonucleoprotein complexes which are reimported into the nucleus. Subsequently, LINE-1 RNA is reverse transcribed and the cDNA is inserted into the genome by a mechanism named target primed reverse transcription (TPRT). Frequently, reverse transcription fails to procede to the 5' end, resulting in truncated non-functional LINE-1 de novo insertions. Readapted from www.pei.de.

The enzymatic machinery of a retrotransposition-competent LINE-1 principally transposes its own copies; this phenomenon is called "*cis*-preference" of LINE-1 transposition. However, LINE-1s are capable of transposing other non-autonomous sequences in "trans", like Alu retrotransposons, SVA and pseudogenes. When reverse-transcription fails to procede to the 5' end, the inserted copies result inactive. Otherwise, epigenetic (methylation of CpG dinucleotides, modifications of the histone tails) and post-transcriptional silencing methods promote the inactivation of potentially propagating, active LINE-1 elements.

Retrotransposition occurs mainly in the germ cells but it can also produce somatic alterations, leading to differences among individuals and populations of cells of the same

individual. According to the symbiotic theory (Reilly et al., 2013) "*it is advantageous for any transposable element to promote host mating, securing the propagation of the master elements to the next generation*". This speculation may explain why insertional events occur in the brain (neurons and glia) and in the testes: "*advantageous mutations in the brain can result in an increase of the cognitive faculties and therefore in better cultural and social performances which can promote host sexual reproduction"*. While LINE-1 activity in germline and early-embryonic cells, ensures LINE-1 transmission to the next generations.

LINE-1 retrotransposition has been detected in human embryonic stem cells (Coufal et al., 2009), in human fetal brain (Coufal et al., 2009) and during adult neurogenesis in the hippocampus (Muotri et al., 2009) suggesting that neural progenitor cells retain retrotransposition activity in adult stages. Moreover, adult human brain cells present a higher LINE-1 copy number than extra brain tissues. Muotri and colleagues in 2005 demonstrated that LINE-1 rerotransposons mobilize during neural development, furthermore the presence of megabase sized somatic CNVs in the brain hints at mobile-DNA role in neuronal mosaicism and plasticity. LINE-1s in particular, are known to produce large DNA rearrangements (mostly deletions and duplications) upon insertion and recombination (Hedges and Deininger, 2007), providing motifs that can be recruited by the host either for the regulation of its own genes or within its coding sequences. Indeed, LINE-1 mediated SVs altering the phenotype are very few. This is not surprising considering that protein coding genes occupy less than the 10% of the mammalian genome. Yet, the host places further controls on LINE-1 mobility (listed in chapter 1.1.4). Although most LINE-1 associated structural variations within the human populations with an allelic frequency higher than 1% appear to be neutral, specific disease phenotypes result from non-recurrent or private insertion and recombination between LINE-1 elements: Haemophilia A (Kazazian et al., 1988), glycogen storage disease (Burwinkel and Kilimann, 1998) and Alport syndrome (Segal et al., 1999) are just some examples. Moreover, LINE-1s activity has been shown altered in neuropsychiatric disorders such as autism (Shpyleva et al., 2017) and schizophrenia (Guffanti et al., 2016) but little is known about LINE-1 role in neurodegenerative diseases. In this thesis, we are going to investigate if FL-LINE-1 polymorphisms can be a relevant source of structural variants associated to Alzheimer's disease risks.

1.2.4.1.2.1 Effects of LINE-1 elements

In the previous section, we introduced LINE-1-mediated recombination, one of the many effects of LINE-1 elements peppering our genome with stretches of homologous sequences. But L1-retrotransposition affects the cells in many other ways:

- First of all, due to their conformation, LINE-1s can alter the expression of nearby genes (Elbarbary et al., 2016).

LINE-1 elements possess functional sense and antisense promoters so they can initiate both upstream and downstream transcription and even regulate tissue-specific expression of some genes (Lavie et al., 2004). Moreover, LINE-1s possess a polyA tail, that can lead to pausing in transcriptional elongation and formation of truncated transcripts when the retrotransposon integrates in an intron of a gene. Finally, splice sites within LINE-1s residing in introns can lead to new exons within genes (Yalcin et al., 2011).
Obviously, due to their mobilization effects, LINE-1 can affect genomic structure, disrupting the exons in which they are inserted and generating target site deletions and duplications. A LINE-1 insertion obliterated 46 kb of the gene encoding pyruvate dehydrogenase complex, component X (PDHX) and caused pyruvate dehydrogenase complex deficiency (Miné et al., 2007).

- LINE-1 can act as vectors for flanking sequences leading to their expansion in the genome (Pickeral et al., 2000).

When the weak LINE-1 poly(A) signal is ignored, and the LINE-1 transcription terminates at a downstream genomic signal, occurs the so called 3' transduction. This phenomenon accompanies from 10 to 20% of all L1 mobilizations. 5' transduction is most rare than 3' transduction and occurs when transcription initiates from a chance upstream promoter. The length of the additional sequences (that may include exons and regulatory sequences) varies from a small number of bases to over 1 kb. LINE-1s can also shuttle to new genomic locations other repeats like Alus and SVA that are unable to mobilize and take advantage of the LINE-1 machinery.

- LINE-1 can alter the chromatin state (Slotkin and Martienssen, 2007).

Since chromatin condensation suppresses the activity of LINE elements, after their integration these molecules become "boosters" that promote the spread of inactivation to the surrounding regions altering the expression of the nearby genes.

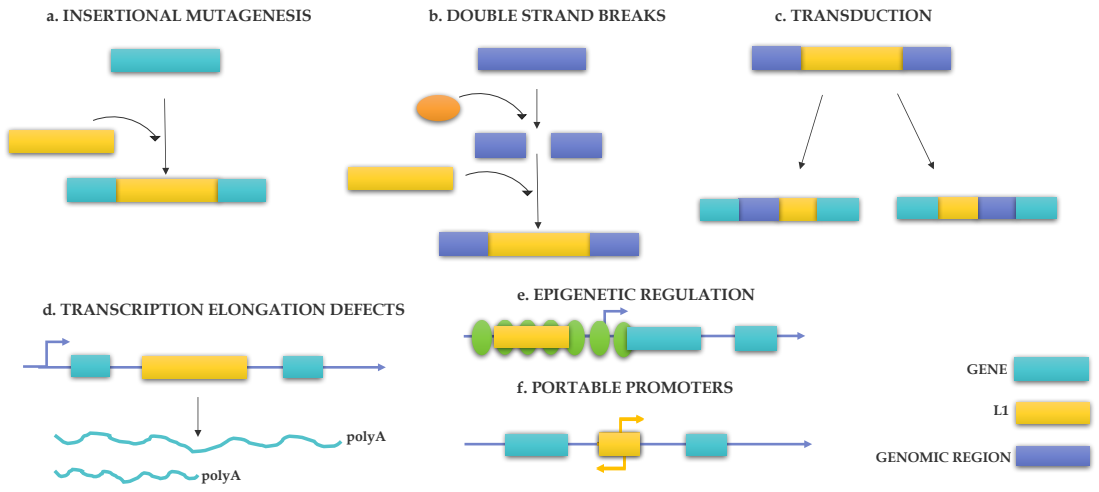


Figure 1.2.4.1.2.1 Effects of LINE-1 elements insertions. LINE-1 elements can affect genome structure and gene activity in many ways. a. Integrating a copy in an exon, LINE-1s can disrupt a gene, b. Repairing pre-existing DSB, c. Promoting 3' and 5' transduction. d. Integrating in an intron, the new LINE-1 copy can provide a premature polyadenylation signal to the host gene inducing the formation of truncated transcripts. e. Altering the chromatin state. f. Providing sense and antisense portable promoters.

1.2.4.1.2.2 Characterizing LINE-1 mediated SV

Because of TEs importance in shaping the genomic structure and their potentially dangerous retrotranspositional activity, the knowledge of the extent of LINE-1 mobilization is fundamental. To this purpose, a crucial step involves the ability to map integration sites at the genome-wide level. The repetitive nature of transposable elements makes this task really challenging, and several different approaches have been developed in the last years, leading to contrasting opinions about the real rate of somatic retrotransposition. Commonly used procedures including ligation-mediated PCR

techniques, or more recently array hybridization enrichment and high-throughput sequencing have suboptimal power. This is due to the struggle of determining the exact positions of elements that are present in highly homologous sequences in hundred thousand genomic locations and are often nested (inserted into pre-existing TEs). Despite ongoing progress in next-generation sequencing technologies (NGS), none of the actual approaches is capable of capturing the full spectrum of SV events with high sensitivity and specificity, especially in complex, repetitive regions. Among the physical constraints that impede an exhaustive LINE-1 detection, chromatin structure must be taken into account: is compact genome selecting against insertions into euchromatin or condensed regions are just difficult to study?

So far, our understanding of LINE-1 related SVs is limited by the low resolution of most recent surveys, such as those solely based on microarrays, which are not able to precisely indicate the breakpoint of the SVs. On the other hand, short-read dependent SVs detection tools are not optimized to detect long SVs (such as FL-L1 elements), especially when they exceed the paired-end insert size, and are prone to false positive calls due to alignment errors. Such errors may occur when the number of bases in the reads, matching the reference genome is too few and when the number of reads supporting a SV is small. The task becomes even more challenging when the goal is identifying individual somatic variations, present in a small minority of cells, or unique to a single cell. Increasing evidence is suggesting that somatic variations as well as polymorphic insertions (present in a restricted number of individuals) make up a consistent portion of each individuals LINE-1 profile (Streva et al., 2015). Since both somatic and polymorphic insertions arise from currently active, mobile LINE-1 elements, these SVs also represent the most interesting subset of repeats to study (Burns and Boeke, 2012). Unfortunately, they are also the less characterized. Intuitively, young events, private or population-specific, are underrepresented in the reference sequence of the genome, originally derived from a mixed pool of individuals. While, ancient, widespread, fixed mobile element relics, make the most abundant fraction of annotated repeats. Continuous efforts (1,000 Genomes Project (The 1000 Genomes Project Consortium, 2015), euL1db (Mir et al., 2015, p. 1), dbRIP (Wang et al., 2006) are made to fill the gap in the annotation of LINE-1 polymorphism. However, there is currently no comprehensive approach to confidently identify these variations. In order to increase the chances to target the rare events, extensive amplification and high depth of sequencing become fundamental. However, whole genome amplification introduces artifacts (Pugh et al., 2008) and higher coverage

of sequencing inevitably requires higher costs. Alternative, single molecule sequencing approaches and single-cell based approaches, provide solutions to some of the most vexing problems that face second generation sequencing. Single molecule sequencing approaches, like PacBio, allow reliable mapping of long SVs across repeat expansions, but suffer a high error rate (Rhoads and Au, 2015). Targeted, single-cell sequencing, offers the highest sensitivity in detecting sporadic SV events, but the percentage of the genome that is covered for each cell, and the total number of cells that can be tested with this approach, is limited. Meanwhile, as methods develop to detect insertions present in smaller proportions of cells, validation becomes progressively more difficult, imposing the use remarkably sensitive instruments, such as digital PCR to detect and quantify rare events.

These constraints result in inexact estimates of rare, large SVs, in the current research.

In this regard, in this thesis:

We develop a new technique for reliably and efficiently identifying annotated, polymorphic and somatic FL-L1 integration sites in the human genome with single-base resolution;

Then, we take advantage of dense arrays to correlate CNVs with genomic repeats such as the LINE-1 elements;

Finally, we combine different state of the art approaches to investigate the formation mechanisms of SVs in a complex region such as mouse Olfr2 locus.

1.2.4.1.3 Short interspersed nuclear elements (SINEs)

Short interspersed nuclear elements comprise about 12% of the human genome (Gogvadze and Buzdin, 2009) and the 5% of mouse genome (Walters et al., 2009). These elements are generally quite short (<700 bp) and do not code the proteins necessary to

their mobilization: they consist of a 5' region, a body and a poly(A) tail or, sometimes, another A-rich stretch on their 3′end. Therefore, they are not autonomous in their transposition. It is generally accepted that LINEs provide their reverse transcriptase to SINE, non-autonomous elements. In this way LINE elements allow the proliferation of a valuable factor of genetic variation (Beck et al., 2011). Among SINE elements, the members of Alu subfamily, represent the most abundant repeats in humans, occupying almost the 11% of the genome (Deininger, 2011). Multiple features predispose Alu elements to successful recombination, including their proximity in the genome (one insertion every 3 kb, on average), the high GC content of their sequence (~63%), and the remarkable sequence similarity (70%–100%) among Alu subfamilies. The recombinogenic-nature of these elements is reflected in the various forms of cancer and genetic disorders associated with Alu-mediated recombination events (Payer et al., 2017). Among the positive consequences of Alu-recombination-mediated-events we can list the evolution of the human glycophorin gene family. The event occurred through several duplication steps that involved recombination between Alu elements and Alu-recombination-mediated-deletions (ARMDs): one of the actors that played a role in shaping the unique traits of the human and chimpanzee lineages (Sen et al., 2006).

1.3 RE mediated genomic rearrangements

In addition to canonical insertion events, because of their high copy number and sequence similarity, retrotransposons can create genomic rearrangements by several additional recombination processes. Among the principal mechanisms we can find:

- nonallelic homologous recombination (NAHR) (Stankiewicz and Lupski, 2002) mediated insertion/deletion between two retrotransposons from the same family;

- nonhomologous-end-joining (NHEJ) mediated deletion (Han et al., 2008);

- and nonclassical-endonuclease-independent insertions of the retrotransposons (Morrish et al., 2002)

Proliferation of mobile elements significantly promotes genome instability (Kines et al., 2014) via insertional mutagenesis and generation of **double-strand-breaks (DSB)**, which have themselves been established as potent inducers of genetic instability. The consequences of genetic instability are particularly evident in humans, where a loss of repair capacity is associated with cancer predisposition, genetic syndromes and aging. Age dependent increase in transposition occurring in terminally differentiated neurons has been reported in *Drosphila* brain (Perrat et al., 2013). However, it is not known whether transposon expression is a cause or a consequence of aging and neurodegenerative diseases. Other repetitive genomic sequences, such as segmental duplications and microsatellites, also promote genetic instability resulting in disease phenotypes. Repetitive regions not only have the potential to form loops and non-B-DNA conformations (Zhao et al., 2010) but are potentially subjected to DNA breakage caused by active retrotransposition and persistent single strandedness due to extensive transcription, secondary structures or replication pausing which made preferential sites for double strand breaks (Hastings et al., 2009b).

1.3.1 Double strand breaks (DSBs)

DSBs are considered one of the most lethal forms of DNA damage as they can lead to dangerous mutagenic rearrangements or apoptosis (Lieber, 2010).

They form as the result of two single stranded nicks in opposing DNA strands occurring sufficiently close to one another (10-20 bp) (Khanna and Jackson, 2001). TEs can produce DSBs during the generation of a new genomic copies, a process that requires the disruption and repair of DNA. LINE-1 elements, as the only autonomous TEs present in the human and mouse genome, play a significant role in generating endogenous DSBs in host cells (Gasior et al., 2006). LINE-1 copies employ a self-encoded endonuclease to create a nick in the target DNA that produces a free DNA end that can be used as a cDNA primer to enter the genome at consensus (TT/AAAA) sequences resulting in the creation of hotspots for new EN cleavage and recombination-mediated rearrangements (Tremblay et al., 2000). Therefore, a high expression of LINE-1 protein may cause significant DNA damage when the integration of the LINE-1 copy does not occur, resulting in structural

variation formation. LINE-1 associated somatic CNV formation at LINE-1 loci harboring the endonuclease consensus sequence, is reportedly a consequence of extensive LINE-1 expression induced DSBs occurring during neural differentiation in the brain (McConnell et al., 2013, Cai et al., 2014).

Moreover, the protein kinase ATM, fundamental in many DNA repair signaling processes, was demonstrated to be involved in LINE-1 retrotransposition (Gasior et al. 2006, Coufal et al., 2011), thus suggesting an additional correlation between LINE-1 and DNA-DSBs repair systems. ATM is activated by double-strand DNA breaks and subsequently phosphorylates downstream substrates leading to the activation of a DNA damage checkpoint and cell cycle arrest.

Other endogenous agents producing DSBs are **transcription** (Kim and Jinks-Robertson, 2012), **replication stress** (Zeman and Cimprich, 2014) and **oxidative stress** (Woodbine et al., 2011). Among the exogenous agents producing DSBs we can list: ionizing radiation (IR), radiomimetic drugs, genotoxic stress, ultraviolet light, anti-cancer drugs such as DNA replication inhibitors, tobacco smoke and topoisomerase I and II inhibitors (Mehta and Haber, 2014).

In the next subsection, endogenous agents producing DSBs will be briefly examined in preparation for the study on endogenous DSBs in mouse OE, one of the topics of this thesis.

1.3.1.1 Transcription

Transcription is, one of the main endogenous processes increasing the occurrence of double-strand breaks in the genome that are repaired by recombination (Schwer et al., 2016).

Transcription stimulates recombination via collisions with the replication machinery, formation of R-loops, non-B DNA structures, engagement of topoisomerases, alteration in DNA base composition and promotion of DNA damage resulting in different types of rearrangements such as deletions, duplications, inversions and translocations (Kim and Jinks-Robertson, 2012).

Formation of transcription-associated DSBs caused by the binding of transcription factors to promoters can be related to torsional stress associated with transcription initiation. DSBs at gene promoters have been proposed to reduce topological stress and induce chromatin relaxation to facilitate transcription initiation and full expression of long genes (like genes involved in neuronal development and synaptic function).

Type II DNA topoisomerases (TOP2), RNA polymerase II (Pol II), CTCF and DNAse I are important players in transcription regulation.

Type II DNA topoisomerases enzymes regulate DNA topology by generating transient double stranded breaks during replication and transcription, recombination, DNA repair, chromatin remodeling, chromosome condensation, and segregation (Uusküla-Reimand et al., 2016).

CTCF and cohesin proteins are key architectural components of the genome that anchor long-range interactions which structure chromosomal domains, flanking the boundaries of topologically associating domains (TADs) (Ong and Corces, 2014). CTCF participates in transcription mediating the formation of loops aimed to promote interactions between various regulatory regions, such as promoters and enhancers. As reported by Madabushi and collegues in 2015, the transcription factor binding site motif is the most highly enriched at TOP2B binding sites and CTCF peaks are enriched in the surrounding of the TSS of genes that incur DSBs.

An additional cause of TSS associated DSBs is the formation of R-loops. R-loops are frequent in highly transcribed genes and result from a stable RNA:DNA hybrid generated when the nascent RNA reanneals to the transcribed strand leaving the non-transcribed single strand of DNA naked (Skourti-Stathaki and Proudfoot, 2014). The ssDNA filament can become the target of DNA-modifying enzymes such as activation-induced deaminase (AID) that can deaminate single-stranded DNA at cytidines leading to the generation of uracil lesions in DNA (Basu et al., 2009). The resulting U/G mismatches are then converted into DSBs through the co-opted activities of the base excision repair (BER) and mismatch repair (MMR) pathways.

During transcription, when transient separation of DNA complementary strands occurs, the naked single-stranded DNA may assume a particular non B-DNA conformation that stabilize R-loops (Zhao et al., 2010). Guanine rich (G-rich) sequences and trinucleotide repeats are particularly prone to form non B-DNA conformation (Zhao et al., 2010).

G-rich sequences form G-quadruplex or G4 DNA, which is comprised of a stacked array of G quartets that form loops which affect genome stability blocking transcription and

replication (Bochman et al., 2012). An important involvement of G4 DNA in genome instability is class-switch recombination is immune system, a process that allows activated B cells to modulate antibody production (Matthews et al., 2014).
Transcription thus has the potential to modify the genetic landscape by locally altering mutation rates, by stimulating loss of heterozygosity and by generating diverse types of rearrangements that include deletions, duplications, inversions and translocations (Kim and Jinks-Robertson, 2012).

1.3.1.2 Replication stress

Replication stress, defined as the slowing and stalling of the replication fork progression and/or DNA synthesis, is a serious problem for genome stability and cell survival (Zeman and Cimprich, 2014). It occurs when physical barriers impede the fork progression, when nucleotides are in limited number or misincorporated, when chromatin is condensed and when the replication origins are too sparse (Mirkin and Mirkin, 2007).
Among the physical barriers to replication fork progression we can list DNA lesions, ssDNA, unusual DNA conformations and conflicts between the transcription and replication machineries.
Repetitive elements that form secondary structures, for example, represent very difficult sequences to replicate leading to an increased chance DSBs and genomic instability. Not properly unfolded secondary structures and nucleotide repeats promote fork stalling and polymerase slippage with consequent expansions and contractions of the repeat sequence.
The genomic distribution of the replication-origin is another factor that significantly affects the performance of the replication fork. An exaggerate replication initiation, like the one produced by constitutively active oncogenes, can consume nucleotide reservoirs and slow down the replication fork speeds which in turn results in increased DNA damage (Wilhelm et al., 2016). On the other side, DNA sequences like the common fragile sites (CFS) that harbor too few replication origins can lead to under-replication and loss of genetic information (Franchitto, 2013). CFS are poorly accessible heterochromatic regions that replicate later than accessible, transcriptionally active euchromatic regions like Early Replicating Fragile Sites (ERFS). In ERFS sites, the concentration of genes is

high, as well as the DNA/RNA machinery collisions resulting in the formation of unstable ssDNA structures and R-loops (Mortusewicz et al., 2013). Nicks and gaps can be therefore both the result and the cause of replication stress. When encountered by the replicating machinery these lesions can trigger the activation of several stress response mediators that induce chromatin changes in the surroundings causing perturbations in the nearby gene expression (Downs et al., 2007).

1.3.1.3 Oxidative stress

Oxidative stress refers to the imbalance due to excess of reactive oxygen species (**ROS**) or oxidants over the capability of the cell to mount an effective antioxidant response (Ray et al., 2012). Oxidative stress results in macromolecular damage of nucleic acids, proteins, and lipids and is implicated in various disease states such as atherosclerosis (Mügge, 1998), diabetes (Ha et al., 2008), cancer (Liou and Storz, 2010), neurodegeneration (Andersen, 2004), and aging (Haigis and Yankner, 2010). ROS such as superoxide anion (O_2^-), hydrogen peroxide (H_2O_2), and hydroxyl radical (HO•), consist of radical and non-radical oxygen species formed by the partial reduction of oxygen. These molecules are generated through both endogenous and exogenous routes. Endogenous ROS are produced through leakage of these species from the mitochondrial electron transport chain (Zorov et al., 2014). Cytosolic enzyme systems, and by-products of peroxisomal metabolism are also endogenous sources of ROS. Generation of ROS also occurs through exposure to numerous exogenous agents and events including ionizing radiation (IR) (Yamamori et al., 2012), UV (Heck et al., 2003), cytokines (Yang et al., 2007), growth factors (Sattler et al., 1999), chemotherapeutic drugs (Arun et al., 2016), environmental toxins (Al-Gubory, 2014), and macrophages during the inflammatory response (Forman and Torres, 2001).

Apurinic/apyrimidinic sites (AP) are one of the most abundant signs of damage generated by ROS. Another major type of DNA damage is the single-strand-breaks SSB, followed by DSBs. DNA damage mainly targets gene promoter regions, as they contain GC-rich sequences that are highly sensitive to oxidative DNA damage and are not protected by transcription-coupled repair (Maynard et al., 2009).

The brain is especially susceptible to the assaults perpetrated by reactive oxygen species. This is because the organ is an important metabolizer of oxygen (one fifth of the body consumption), contains a large amount of fatty acids that can undergo peroxidation, is rich in pro-oxidant elements like iron and has relatively feeble protective antioxidant mechanisms. In brain tissue, ROS are generated by microglia and astrocytes and modulate synaptic and non-synaptic communication between neurons and glia (Popa-Wagner et al., 2013). ROS also interfere with increased neuronal activity by modifying the myelin basic protein and can induce synaptic long-term potentiation, a form of activity-dependent synaptic plasticity and memory consolidation (Massaad and Klann, 2011).

An increasing number of studies report the prevalence of oxidative stress and mitochondrial abnormalities in numerous neuropsychiatric disorders such as depression (Bakunina et al., 2015), Alzheimer's diseases (Manoharan et al., 2016), or schizophrenia (Bitanihirwe and Woo, 2011). Impaired mitochondrial function may contribute to the damage both by increasing reactive oxygen species and by reducing ATP required for DNA repair. A common denominator of all these pathologies is an increased inflammatory response of the brain. The oxidative burst occurring when neutrophils generate reactive oxygen species (ROS) during phagocytosis is in fact is a double weapon that contributes to host defense, but can also result in collateral damage of host tissues.

1.3.2 Age related DNA damage

DNA damage accumulation is regarded as one of the principal causes of aging, defined as progressive organic functional decline, with loss of homeostasis and increased probability of illness and death (Schumacher et al., 2008).

Among age related DNA-changes it is possible to list DSB, telomere erosion, mitochondrial damage induced by reactive oxygen species and extensive demethylation (Jung and Pfeifer, 2015).

The age specific global demethylation (Wood and Helfand, 2013) involves mainly repetitive regions of the genome and can result in the reactivation of retrotransposons and increase of genome instability. It can be observed in most cancer types (Szyf et al., 2004), inflammatory diseases (Strietholt et al., 2008) and age-related neurodegenerative diseases

like Alzheimer's disease (AD) (Chouliaras et al., 2013). Deterioration of central nervous system appears to be a critical part of the aging process (Mattson and Magnus, 2006). This may be due to the low DNA repair capacity of the post-mitotic brain tissue which can result in the accumulation of lesions generated by free radicals and reactive oxygen species (ROS). Thus, DNA damage may initiate a progressive cognitive decline in the elderly, affecting the expression of selectively vulnerable genes involved in synaptic plasticity, learning, memory and neuronal survival.

Senescence in the human frontal cortex (FC) is associated with a reduced expression of genes involved in signal transduction, long-term potentiation, memory storage, vesicle trafficking, and protein turnover. Among the transcripts known to increase their expression in the elderly FC we can list genes that mediate stress responses and repair like the ones involved in protein folding, antioxidant defense and metal-ion homeostasis and genes involved in in inflammatory and immune responses (Lu et al., 2004).

1.3.2.1 Alzheimer's disease (AD)

The ageing of the human brain is a cause of cognitive decline in the elderly and the major risk factor for Alzheimer's disease, especially late onset Alzheimer's disease (LOAD). Late onset Alzheimer's disease differs from the less frequent (5% of all AD cases) familial early onset Alzheimer's disease (FAD) for the tardive age of onset (well beyond 65) and for the diagnostic markers (Tanzi, 2012).

Familial forms of AD are caused by rare and usually highly penetrant mutations in three genes (APP*, PSEN1* and PSEN2)*,* all of which increase the production of the amyloid-β peptide (Aβ), the principal component of β-amyloid in extracellular senile plaques causing neuronal loss (particularly in hippocampus and cerebral neocortex) and brain atrophy. Other neuropathological changes occurring in AD are intra-neuronal neurofibrillary tangles (aggregates of hyper-phosphorylated and misfolded tau), neuropil threads (axonal and dendritic segments containing aggregated and hyper-phosphorylated tau), and dystrophic neurites (abnormal neuronal processes containing hyper-phosphorylated tau). AD is also associated with neurovascular dysfunction. High blood pressure and small vessel diseases are known to increase the risk of ischemic disease of

the brain which may trigger pro-inflammatory and endothelial reactions. Blood vessel diseases are estimated to contribute to approximately 40% of all dementias worldwide, including AD.

LOAD patients, compared with FAD patients, show an overrepresentation of inflammatory markers (creatinine), impaired renal function (high concentration of Blood Urea Nitrogen) and a higher frequency of the ε4-allele of the apolipoprotein E gene (*APOE*). Interestingly both Amiloyd Precursor Protein (APP) gene and Presenilin 1 (PS1) show gradual demethylation at the promoter in LOAD.

As reported in the previous paragraphs, extensive genomic demethylation is a typical sign of aging with consequences on gene expression and transposable elements mobilization. So, we started wondering about a possible involvement of LINE-1 elements, the only active retrotransposons in humans, in an LOAD. Recent studies already reported an involvement of LINE-1-elements in neuropsychiatric disorders like autism (Shpyleva et al., 2017) and schizophrenia (Bundo et al., 2014). Moreover, accumulating evidence indicates that the genomic DNA in the brain contains distinctive somatic genomic variations compared with non-brain tissue (Erwin et al., 2016a).

Considering these evidences, we decided to investigate, for the first time, the role of LINE-1 retrotranspsons in Alzheimer Disease, one of the most devastating neurodegenerative diseases.

1.3.3 γH2AX

The DNA-damage response (**DDR**) enables the cells to sense DNA damage, 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 DNA damage which can increase DNA accessibility and recruit the different repair proteins necessary to initiate the repair of DSBs (Turinetto and Giachino, 2015). Spontaneous γ-H2AX foci are detectable in both normal and cancer cells, likely as a result of endogenous DSBs. The basal level of foci varies with the cell type, but commonly 1–2 foci/cell have been observed in normal tissues while in proliferating cancer cell lines the number is larger

and more variable (Wang et al., 2014). A unique advantage of using γ-H2AX foci as a DSB biomarker is that these foci are formed regardless of the phase in the cell cycle (Mah et al., 2010). H2AX phosphorylation not only occurs at interphase but also during the mitotic phase where chromatin is more condensed. γH2AX also appears to be a great DSB marker due to its high sensitivity and almost immediate formation (within seconds) after DSB induction, while the maximal number of foci is reached within 1 min and the maximum in 9 to 30min after DNA damage. With time, also the size of the foci increases up to 30 Mbp. Importantly, the γ-H2AX foci level is linearly related to the number of DSBs. It must be said that, γ-H2AX foci may not exclusively reflect DSBs. DSB-independent background foci may be caused by ATR-mediated H2AX phosphorylation in growing cells with dis-regulated DNA metabolism and in response to heat (Wang et al., 2014).

Although, γ-H2AX foci formation is not an exclusive indicator of DSBs, it is still the best marker based on its cell phase-independent formation, tight correlation with repair kinetics and repair pathway independence. As γ-H2AX is formed de novo it is a more reliable DSB marker than other DNA repair proteins that are present in the cell even when there is no DNA damage. Moreover, γH2AX foci detection allows the distinction of the temporal and spatial distribution of DSB formation (Bonner et al., 2008). For this reason, we performed a chromatin immunoprecipitation and sequencing (ChIP-Seq) analysis of γ-H2AX to study endogenous double strand breaks (DSBs) distribution in mouse olfactory epithelium and liver.

1.4 SV formation

In the previous chapters were presented TEs, DSBs and SVs: the main actors responsible for genomic instability.

Due to the high density of repetitive elements in mammalian genomes, it is not surprising that we can find them as substrates for genomic SVs at a post-insertional stage. In fact, if we consider them as tracks of homologous sequences it is clear that they have the potential to alter those DNA repair processes which rely on homologous recombination (Burwinkel and Kilimann, 1998), thus resulting in genomic alterations.

1.4.1 Place of SV formation

Genomic architecture of a certain region is an important predictor of its stability.
Unstable genomic loci such as common fragile sites (CFS), recombination hotspots, AT rich palindrome sequences, core duplicons, and G-quadruplexes are more prone to recombination events. Among with repeat rich regions, heterocromatic regions (such as centromeres and telomeres), scaffold attachment sequences and replication origins and terminators are enriched in SV (Korbel et al., 2007, Huang et al., 2010). Noticeably, gene rich, accessible and early replicated genomic regions tend to be more elongated and exhibit more structural variations than gene poor, inaccessible and late replicated genomic regions (Hu et al., 2013).
In particular, very large genes (like neuronal genes) are more prone to DNA-breakage related events since:
1) the possibility of transcription and replication machineries collision increases when transcription requires more than a single cell cycle to complete (Helmrich et al., 2011);
2) TOP2B recruited to resolve positive supercoiling that arises during transcription induces DSBs that if not faithfully re-ligated, can potentially lead to genome rearrangements in flanking genomic regions (Uusküla-Reimand et al., 2016).

In this thesis, we provide further insights about how the physical properties of the DNA sequence underlying a certain locus, influence the propensity for a specific SV formation mechanism. In particular, we developed a computational pipeline for classifying SV occurring in the surroundings of a particular olfactory receptor gene, and inferring their formation mechanism.

1.4.1.1 Olfactory receptors (OR)

Rearrangement prone regions, like sub-telomeres and peri-centromeres, are characterized by a patchwork of repeats, frequent DSB, segmental duplications (which are known to induce NAHR and NHEJ) and wide-spread SV (Linardopoulou et al., 2005).

Interestingly, in this gene poor jungle, flourish most of the olfactory receptors genes (**ORs**) in both human and mouse genomes which appear as dense clusters distributed throughout the chromosomes. The curious locations of OR in these regions, and in repeat rich regions in general, raised a suggestive hypothesis: subtelomeres and pericentromeres may "function as nurseries for the generation of diversity in this multigene family" (Trask et al., 1998).

According to the above-mentioned hypothesis, in these dynamic regions OR can be duplicated or modified without affecting proximal, dosage sensitive genes.
Given the high frequency of CNVs (deletions, duplications and other complex rearrangements) affecting OR loci (Hasin et al., 2008), and in particular evolutionarily "young" OR genes and pseudogenes, some authors proposed that SV may be at the basis of OR organization. Moreover, the bias for CNV-enriched OR in close proximity to centromeres and telomeres as well between tandemly oriented segmental duplications, suggests that NAHR and NHEJ similar mechanism are likely to play a role in the diversification (new functions or regulation patterns) among OR.
Olfactory receptors are members of the seven-transmembrane-domain, large family of G-protein coupled receptors (GPCRs). With almost 1400 OR genes (including pseudogenes) in mouse genome and more than 750 in the human genome, ORs represent the largest mammalian gene superfamily (Niimura and Nei, 2005).
The coding regions of these genes are short, spanning only 1000 bp and intronless and are located in dense clusters throughout the chromosomes. These clusters are rarely interrupted by other genes and are located in gene poor, repeat-rich regions of the genome. In the nose ORs are expressed in the main olfactory epithelium (MOE) where they are believed to recognize odors (conscious odor perception) and in the vomeronasal organ (VMO) where they recognize pheromones (unconscious odor perception). Some OR are also transcribed in other tissues such as lung, kidney, colon, prostate, testis and germ cell tumors suggesting other non-olfaction related functions of ORs. In dendrites and axons of olfactory sensory neurons (OSN), OR genes are expressed in a monogenic and monoallelic fashion and the molecular mechanism regulating this activation/repression process is still unknown. What has been proven is that the process starts with a high number of silenced ORs and each neuron decides to de-repress only one allele of one gene of them. The produced OR protein elicits a feedback signal that prevents the activation of other OR genes. Speculations on this regard include chromatin

modifications, DNA editing and transposable-element-mediated regulation. So, chromosomal sequence context seems to play an important role in monoallelic gene expression.

The surroundings of OR genes share common feature with the flanking regions of other monoallelically expressed genes like vomeronasal, Igs and chaderins. The first common characteristic between monoallelically expressed genes is asynchronous replication (Donley et al., 2013). Asynchronous replication occurs when one allele replicates before the other but unlike monoallelic expression, asynchronous DNA replication is independent of whether a gene is expressed in a given cell type or not. Probably the genomic context in which the alleles reside plays the most important role. The regions surrounding monogenically expressed genes contain high densities of LINE-1 sequence (especially full-length), reduced density of SINE elements and a low GC content (Allen et al., 2003). LINE elements are known to be employed for repeat-induced gene silencing (the best-studied example is X inactivation in female cells where they promote heterochromatin to spread throughout the X chromosome (DISTECHE and BERLETCH, 2015)). SINE elements on the other hand are reportedly abundant in gene rich regions and depleted in proximity of imprinted genes. The GC poor content found in the surroundings of OR genes may be associated with the abundance of LINE-1 sequences in these regions: recent LINE-1 insertions prefer to integrate in AT rich regions due to the target specificity of ORF2 protein (Tremblay et al., 2000).

1.4.2 Moment of SV formation

SVs such as insertions and deletions can arise both meiotically (the rearrangement occurred in germ cells, is present in every tissue of the individual and can be inherited) and mitotically (different organs and tissues vary in copy number in the same individual). Genomic disorders such as α-thalassemia, which is caused by α-globin gene deletions, derive from a rearrangement occurred in the germ cells (Horst et al., 1984) while disorders such as cancer depend mostly on a somatically occurred mutation (Piccolo and Frey, 2008).

Recurrent rearrangements that share the same size and genomic content in unrelated individuals are more frequently observed in early replicating regions of the genome whereas non-recurrent rearrangements that have a unique size and genomic content at a given locus in unrelated individuals are more frequently observed in late replicating regions (Gu et al., 2008). Early replicating regions are known to be hypomethylated gene rich regions, with a high GC content, that contain actively transcribed genes. As already discussed, transcription itself is an important agent of genome instability (Kim and Jinks-Robertson, 2012). At the same time, regions of reduced rates of replication, prone to polymerase pausing, are also liable to passive breakage under prolonged stalling conditions (Mirkin and Mirkin, 2007).

1.4.3 Mechanism of SV formation

While point mutations usually reflect errors of DNA replication and repair, gross genomic rearrangements, such as the ones examined in this work, are often the result of other mechanisms mediated by genomic structural features.

When a rearrangement occurs in the genome, the SV breakpoint regions can give important clues about the mechanism mediating its formation. The breakpoints are the novel sequence junctions (start-end coordinates) identified by comparing the structure of a rearranged genome to that of the reference genome, so their annotated position is based on the coordinate system of the reference genome (Quinlan and Hall, 2012). This concept can cause some confusion since it entirely depends on the accuracy of reference genome annotation. For example, an insertion in the experimental genome may reflect a deletion in the reference genome.

Interestingly, breakpoints are often clustered in a highly nonrandom manner. Unstable loci, more prone to recombination events, are often genomic regions enriched in repeats such as centromeres, telomeres and sub-telomeres. Therefore, repeated sequences and repetitive elements in particular, providing large regions of sequence similarity, appear to be ideal substrates for recombination events.

Indeed, all the general mechanisms that give rise to SVs listed above involve repeated sequences to a certain extent.

Insertional mechanisms:

- transposition of mobile elements.

Recombination based mechanisms:

- non-allelic homologous recombination (NAHR);
- non-homologous end-joining (NHEJ);
- microhomology-mediated end joining (MMEJ).

Replication-based-mechanisms:

- Serial Replication Slippage (SRS);
- Fork Stalling and Template Switching (FoSTeS);
- Microhomology-Mediated Break-Induced Replication (MMBIR).

1.4.3.1 Insertional mechanisms

Mobile elements, extensively described in the dedicated section, are DNA sequences that are capable of integrating themselves or a copy into the genome at a new site within the cell of its origin. DNA transposons, mobilize through a cut-and-paste mechanism and are inactive in human and mouse. DNA retrotransposons, such as active LINE elements, use a copy-and-paste mechanism to insert extra copies of themselves into new genomic locations making up for the 0.3% of all mutations in the human genome (Ayarpadikannan and Kim, 2014).
Consequences of these mutations might be:

- DSB formation;
- gene disruption if a transposable element integrates into an exon;
- alternative splicing if the transposable element integrates into an intron;

- premature polyadenylation and consequent formation of truncated transcripts;
- chromatin state alterations;
- regulation of gene expression;
- repair of already present DSB like the ones present in the telomeres;
- increased recombination rate;
- unequal crossing-over.

The genetic instability produced by mobile-elements inserting into new genomic locations may have severe consequences for the cell or the organism. The effects of harmful unrepaired insertions may be evident (such as Apert syndrome (Bochukova et al., 2009) and Duchenne-muscular-dystrophy (Smith et al., 2011)) or become apparent only later in life (such as mental disorders and cancer (Piccolo and Frey, 2008)).

Non-homologous end joining (NHEJ) repair (described in a few paragraphs) resolves most of DSBs left by the endonuclease during unsuccessfull retrotransposition events, which occasionally results in deletions and rearrangements in human cells. Viceversa, in cells deficient in non-homologous end joining (NHEJ), mobile elements can insert into pre-existing DNA breaks, repairing them (Sen et al., 2007).

1.4.3.2 Recombination based mechanisms

1.4.3.2.1 Non-allelic homologous recombination (NAHR)

Non-allelic (or ectopic) homologous recombination occurs during mitosis and meiois and involves genomic regions comprised between long stretches of directly-oriented or inverted repeats (LCR, Alu, LINE) that share almost perfect homology (> 95%) to repair DNA breaks and gaps (Gu et al., 2008). Interspersed TEs therefore play an important role in the DSB repair pathway offering alternative non-allelic tracts of homology with which the invading strand can anneal. NAHR can result from crossover between interacting homologies in non-allelic position on the same chromosome (produces mostly deletions)

or homologous chromosomes (produces deletions and duplications), rarely between non-homologous chromosomes. If the same chromosomal position in the sister chromosome or homologue is employed for the repair, no change in structure occurs and the HR process leaves no trace. Homologous recombination is unavailable in non-cycling cells such as post mitotic neurons where the sole pathway available is NHEJ (Carvalho and Lupski, 2016).

1.4.3.2.2 Non-homologous end joining (NHEJ)

Non-homologous end joining is the number one DSB repair mechanism. NHEJ does not require a homologous template (sister chromatid or homologue) to join the break ends but employs short homologous sequences (1–4 bp microhomologies) exposed in single-stranded overhangs on the DSB ends for base pairing. It is capable to re-store the pre-break situation by direct ligation of compatible ends, but frequently leaves small insertions (often from retrotransposons and mitochondrial DNA) or deletions (1-10 bp) at the breakpoint. Consequences of an inefficient repair are gross chromosomal rearrangements such SVs. NHEJ can function in both dividing and non-dividing cells, available during all phases of the cell cycle, it is most active during G1. In the absence of NHEJ damaged cells may activate MMEJ, in extreme cases apoptosis. An unrepaired DSB can become the substrates for frequent translocations (Lieber, 2010).

1.4.3.2.3 Microhomology-mediated end joining (MMEJ)

Microhomology-mediated end joining is an alternative non-homologous DSB repair pathway that relies on microhomologies (5-25 bp) on either side of the break to join and stabilize the broken DNA (McVey and Lee, 2008). Repair by MMEJ therefore leads to deletion of the DNA sequence between the microhomologies. Recurrent rearrangements, occurring in recombination hotspots and presenting clustered breakpoints, are mostly the

result of non-allelic homologous recombination (NAHR) mechanisms while non-recurrent rearrangements that vary in size and have scattered breakpoints are probably reconducible to replicative and non-replicative microhomology-mediated mechanisms such as NHEJ, MMEJ and MMBIR (Verdin et al., 2013). Among the unique features which are characteristic of MMEJ over NHEJ we can list the longer microhomology stretches (5-25 be in MMEJ vs 1-4 bp in NHEJ), and the highly mutagenic nature of the first. MMEJ probably assumes a more important role in break repair if DNA ends are not readily compatible. For this reason it always results in deletions and is frequently associated with translocations (McVey and Lee, 2008).

Recently it was demonstrated how MMEJ resolves somatic deletions generated by LINE-1 endonuclease cutting activity in the brain (Erwin et al., 2016a). In this thesis we provide further evidence supporting this very interesting result, characterizing the breakpoints of hundreds of deletions detected in olfactory epithelium.

1.4.3.3 Replication based mechanisms

Low fidelity and reduced processivity of the error prone replication process often result in aberrant replication and SV formation (Liu et al., 2012).

1.4.3.3.1 Replication slippage

Replication slippage (Streisinger et al., 1966) occurs during DNA replication in repetitive regions such as microsatellites when the polymerase enzyme encounters a hairpin or another non-linear DNA conformation. At this point the primer and the template strands can dissociate and reanneal in correspondence of another repeat beyond the skipped barrier, where polymerase reloads and replication resumes. Replication-slippage results

in deletions, repeat expansions and frameshift mutations. Serial replication slippage (**SRS**) occurs when two or more consecutive forward slippages occur in cis or trans orientation resulting in complex rearrangements (Chen et al., 2005).

1.4.3.3.2 Replication fork stalling and template switching (FoSTES)

According to the **FoSTES** model genomic regions containing symmetrical features like low copy repeats (LCRs) may confuse the DNA replication machinery, causing single or multiple replication fork stalling and switching events before resuming replication on the original DNA template, resulting in rearrangements such as deletions and duplications. The presence of complementary template microhomology (2–5 bp) allows annealing and priming during mitosis with the consequent formation of a 'join point' between two distant segments of the genome resulting in close proximity in three-dimensional space (Lee et al., 2007).

1.4.3.3.3 Microhomology-mediated break-induced replication (MMBIR)

Microhomology-mediated break-induced replication is invoked when the necessity is to repair a single-stranded DNA damage event generated during replication. Single-stranded DNA stretches occur in replication forks, from stalled transcription complexes, at excision repair tracts, or at secondary structures in DNA such as cruciforms or hairpins caused by inverted repeats and possibly in other situations such as in promoter regions and replication origins.

A typical characteristic of SVs (deletions, duplications, translocations, and inversions) resulting from MMBIR is that microhomology junctions are followed by stretches of DNA sequence derived from elsewhere. Interestingly, during the repair process, MMBIR could increase genome subsceptibility to future MMBIR events creating LCRs that are

going to provide the homology necessary for NAHR and the formation secondary structures (Hastings et al., 2009a).

Thus, such “errors of replication” may provide a mechanism for the maintenance of genome plasticity and, conceivably over longer periods of time, genome evolution.

2 Materials and Methods

Due to the complex nature of structural variants, it is not straightforward to find a simple way to characterize all types of them. Therefore, since the complete range of structural DNA variation cannot be investigated with a single procedure, we decided to focus on the effects on genome stability of one class of mobile elements that are LINE in different contexts, adopting specific strategies for:

- identify novel FL-L1 insertions
- quantify FL-L1 insertions in different tissues
- characterize LINE-1 content in genomic variations
- explore LINE role in the generation of structural variants such as deletions
- profile double strand breaks: cause and effect of structural variants

In this chapter, we are going to describe all the techniques that have been used to address each of the specific tasks described in this thesis.
In each section, bioinformatics analyses are complemented with essential wet lab experiments and validations. Accordingly, even if this thesis is focused on the computational methods to study transposable elements mediated SVs, for the sake of knowledge, in this chapter, we are going to describe also the experimental techniques that have been used to address each of the three specific tasks described in this thesis.

In the first part of the chapter, we illustrate a novel experimental technique developed in our laboratory called SPAM, SPlinkerette Analysis of Mobile Elements. This procedure allows us to target exclusively FL-L1 elements present in the frontal cortex (FC) and the kidney (K) of Alzheimer's disease affected patients (AD) and controls (CTRL), combining a PCR-based enrichment of LINE-1 5'end and their flanking genomic portions with an ad hoc bioinformatic pipeline. Then, we describe the TaqMan based copy number variation (CNV) analysis, carried out to evaluate the content of potentially active LINE-1s in the different areas of the brain and kidney of AD and CTRL individuals. Finally, we

show how we employed high density arrays to compare the occurrence of FL-L1 elements in correspondence of genomic variations detected in AD and CTRL patients.

In the second part, we present the steps of the study performed to explore the possible function of LINE-1 induced somatic genomic variation in the regulation of olfactory receptor choice in mouse olfactory epithelium. To perform this study, we combine the benefits of short Illumina reads and long PacBio reads to describe the SV profile of the surroundings of an active olfactory receptor gene.

In the third part we describe the approach adopted to study endogenous double strand breaks (DSBs) distribution in mouse olfactory epithelium and liver. To this purpose, we performed a chromatin immunoprecipitation and sequencing (ChIP-Seq) analysis of γ-H2AX.

2.1 Analysis of FL-L1 elements in the genomes of AD post-mortem brains

2.1.1 Identification of novel FL-L1 insertions: The SPAM technique

Several strategies have been adopted to map the exact insertion sites of repetitive elements, most of them based on ligation-mediated PCR techniques (Arnold and Hodgson, 1991, Eggert et al., 1998), or more recently on array hybridization enrichment (Shukla et al., 2013) and high-throughput sequencing (Ewing and Kazazian, 2011, Lee et al., 2012). Results have been overall mixed.

This is due to the challenges of determining the exact positions of elements that are present in highly homologous sequences in hundred thousand genomic locations and are often nested. Moreover, the available short-read dependent SV detection tools are not optimized to detect long insertions, especially when they exceed the paired-end insert size.

Most studies have focused on the 3'end LINE-1 region (Ewing and Kazazian, 2010, Erwin et al., 2016), aiming at the identification of both the integer (1%) and the 5'truncated forms (99%) of LINE-1 elements present in the human genome. Streva and colleagues, in 2015 proposed a method to investigate specifically polymorphic LINE-1 in the human genome, to this aim, they focused on the 5'end LINE-1 region. In the meanwhile, we were developing our technique. Importantly, in this work, we analyze only the small fraction of LINE-1s that retain their potential to impact genomic structure and gene expression: FL-L1 elements. Intact LINE-1-elements are 6kb long molecules that harbor the complete machinery necessary for their retrotransposition and therefore are still able to mobilize and give rise to novel LINE-1 integration sites.

This technique consists of a specific series of steps: the **Splinkerette enrichment PCR**, Illumina **sequencing** of the amplicons, **bioinformatic pipeline** and **validation PCR**.

2.1.1.1 Samples

SPAM was performed on gDNA samples extracted using a standard phenol-chloroform extraction method from frontal cortex and kidney of 4 AD patients and 4 not affected individuals of a Brazilian cohort, provided by the Brain Bank of Sao Paulo.

CODE	GENDER	AGE	CLINICAL	BRAAK	CERAD	CONDITION
7660	M	72	2	6	C	AD
2682	M	82	3	6	C	AD
7466	M	87	0.5	4	B	AD
9345	F	90	1	4	A	AD
2149	F	76	0	1	0	CTRL
9269	F	88	0	3	0	CTRL
6868	F	89	0	2	0	CTRL
929	F	61	0	0	0	CTRL

Table 2.1.1.1 SPAM samples. Spam was performed on the FC and the K of four AD affected patients and 4 CTRLs.

2.1.1.2 Splinkerette enrichment PCR

In order to effectively target the boundary between a FL-L1 and its flanking genomic regions, we set up the Splinkerette Analysis of Mobile Elements (SPAM) technique, inspired by the Splinkerette PCR (spPCR) protocol. This technique, developed to amplify the genomic DNA flanking a known sequence, allows efficient and specific mapping of transposable elements. Designing the primers at the very 5' of the LINE sequences (Lavie et al., 2004) allowed us to target only the FL-L1 elements.

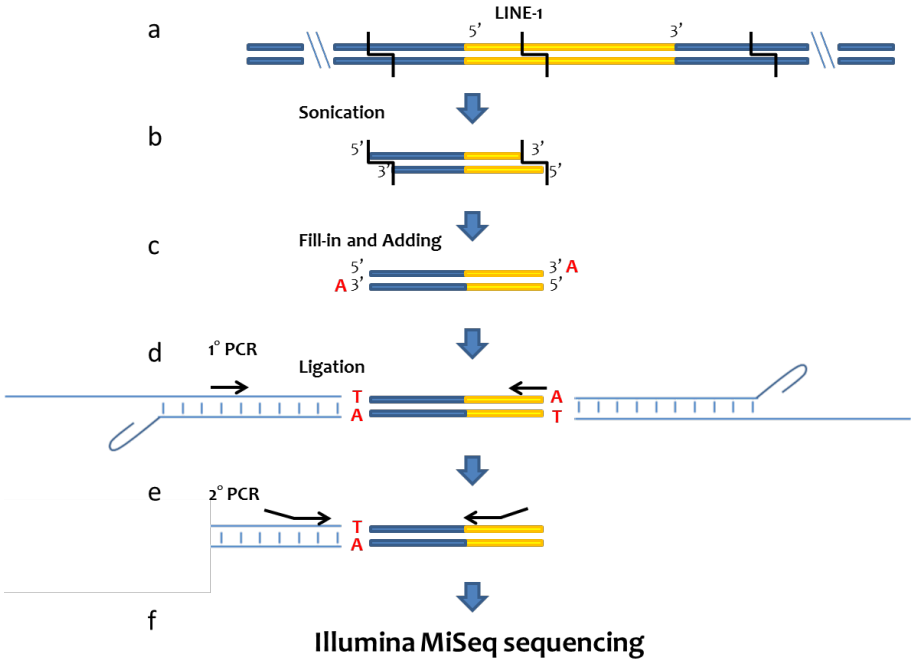


Figure 2.1.1.2 The SPAM PCR schematic protocol. (a) Representation of a LINE-1 sequence (in yellow) inserted in the genome. (b) The SPAM protocol starts with genomic DNA (gDNA) fragmentation by sonication, that produces sticky-ends fragments. (c) gDNA fragments are filled in in order to obtain blunt ends and an Adenine is added to the 3'ends. (d) Fragments ligation to synthetic double strand adapters, followed by a first round of PCR with primers complementary to the adapter and the LINE-1's 5'UTR. (e) Nested PCR with primers complementary to the adapter and the LINE-1's 5'UTR, harboring the Illumina barcodes and adapters. (f) 2x300 paired-end MiSeq Illumina sequencing.

SPAM oligonucleotides	Sequence 5' -> 3'
Long strand adapter	CGAAGAGTAACCGTTGCTAGGAGAGACCGTGGCTGAATGAGACTGGTGTCGACACTAGTGGT
Short strand adapter	Phosph - CCACTAGTGTCGACACCAGTCTCTAATTTTTTTTTTCAAAAAAAG
Fw1	CGAAGAGTAACCGTTGCTAGGAGAGACC
Fw2	GTGGCTGAATGAGACTGGTGTCGAC
Rev1	CGTCCGTCACCCCTTTCTTTGACTCG
Rev2	CTTGCGCTTCCCAGGTGAGG

Table 2.1.1.2 Spam primers and adapters.

Primer design performed by Paolo Vatta and experiment performed by Marta Maurutto

2.1.1.3 Library preparation

Two µg of genomic DNA was sheared in 100 µL nuclease-free water by sonication with a Bioruptor NGS, to obtain an average fragments size of 300 bp (7 cycles: 30 seconds of sonication at max power followed by 90 seconds of stop). Since sonicated gDNA fragments present irregular sticky ends, samples were processed with an end-repair reaction (inspired by Illumina protocols for sequencing libraries preparation) in order to get blunt-ended fragments. Samples were end-repaired for 30 minutes at 20°C in a final volume of 50 µL, using 5 µL 10X T4 Ligase buffer with 10 mM ATP (NEB), 2 µL 10 mM dNTPs, 1 µL T4 DNA Polymerase (NEB), 1 µL Klenow large fragment (NEB) and 1 µL Polynucleotide Kinase (NEB). Reactions were purified using the Qiaquick PCR purification kit (Qiagen) following manufacturer's instructions. A 3'-end Adenine was added to the blunt end purified fragments in a final volume of 50 µL, using 5 µL NEB buffer 2, 1 µL Klenow fragment (3'→5' exo-) and 10 µL 1 mM dATPs, for 30 minutes at 37°C. Fragments were purified using the Qiaquick PCR purification kit (Qiagen) and eluted in 30 µL of milliQ sterile water.

SPAM adapters were designed in order to have blunt ends and a 3' protruding T at the longest strand. The adapters were assembled as follows: long strand and short strand oligonucleotides (Sigma) were resuspended in TE buffer and mixed to a final concentration of 50 µM, denatured at 95°C for 5 minutes and annealed by slow cooling to RT. In order to increase variability, adapters were ligated to sheared genomic DNA by three independent ligations per sample, using 8 µL of gDNA, 2 µL of adapters 50 µM, 5 µL of 10X T4 Ligase buffer with 10 mM ATP (NEB) and 2.5 µL of T4 DNA Ligase (NEB) in a final volume of 50 µL. The reactions were performed at 16°C ON, plus 1 hour at 37°C after the addition of extra 0.5 µL T4 DNA Ligase. The ligations of each sample were pooled together, purified using the Qiaquick PCR purification kit (Qiagen), and eluted with 40 µL of 10 mM Tris-HCl pH 7.4.

Nested PCRs were performed to enrich LINE-1 insertion sites using forward primers specific to the adapter sequence, and reverse primers specific to the LINE-1's 5'UTR sequence (Uren et al., 2009). Tags and barcodes necessary for the Illumina MiSeq sequencing were added to nested forward and reverse primers. Three independent primary PCRs per sample were performed in a final volume of 50 µL in a thermocycler (ABI) using 5 µL of purified ligated gDNA fragments, 240 nM S forward primer specific

to the adapter, 240 nM 5UTR 1° reverse primer specific to LINE-1's 5'UTR, 250 nM dNTPs, 1X High Fidelity PCR buffer (Invitrogen), 2 mM MgSO4 (Invitrogen) and 1.25 U of Platinum® Taq High Fidelity (Invitrogen). PCR protocol was as follows: 2min at 94°C, 30 cycles of 15s at 94°C, 30s at 68°C and 3min at 68°C, followed by a final elongation of 5min at 68°C.

Three nested PCRs were performed in a final volume of 50 µL using 1 µL of primary PCR product, 240 nM 2° SPLINK2, 240 nM 2° reverse primer specific to LINE-1's 5'UTR, 250 nM dNTPs, 1X High Fidelity PCR buffer (Invitrogen), 2mM MgSO4 (Invitrogen) and 1.25 U of Platinum® Taq High Fidelity (Invitrogen) with the following protocol: 2min at 94°C, 25 cycles of 15s at 94°C, 30s at 60°C and 5min at 68°C, and final elongation of 5min at 68°C.

The three nested PCR reactions were mixed, denatured at 95°C for 10 minutes and rapidly cooled down by the addition of milliQ water at 4˚C up to 500 µL. Reactions were purified using Microcon® DNA Fast Flow Centrifugal Filters (Millipore) following manufacturer instructions. A small amount of purified amplicons were stored for future PCR validations, while the rest (~70 µL) were precisely quantified by Bioanalyzer, and loaded on a 2% TAE electrophoretic gel. Smeared samples were cut from size 200 bp to 1000 bp and gel extracted with QIAquick Gel Extraction Kit following manufacturer instructions, in order to remove any large concatamer and amplicons too short to contain a sufficient amount of genomic and LINE-1 sequence to be successfully mapped. Samples were quantified again with Bioanalyzer, pooled and sequenced.

Experiment performed by Marta Maurutto

2.1.1.4 Sequencing

SPAM samples were sequenced using the Illumina MiSeq technology (300 bp paired-end set-up). Different reverse nested primers with different barcodes were used to perform the secondary PCR, allowing samples to be sequenced in multiplex. Sequencing was performed by IGA Technologies (Udine, Italy). For each sample, we obtained on average 8.6 million reads.

2.1.1.5 Nomenclature

The following nomenclature, assigned to the outcomes of the bioinformatics analyisis, will be used throughout the thesis: reads in a pair will be referred with **R1** (forward) and **R2** (reverse); **fragment** is used to indicate an assembled reads pair (R1 + R2); **MapFragment** indicates a fragment containing the expected LINE 5' portion and a mappable unique genomic sequence; **MapCluster** indicates the integration site after clustering of two or more non-identical overlapping MapFragments; Annotated Integration Site (**AIS**) is used to indicate an integration site already known; Non-annotated Integration site (**NIS**) is used to indicate an integration site not present in RepeatMasker and not reported in literature; Polymorphic Integration site (**PIS**) is used to indicate an IS not present in the reference genome, but annotated as retrotransposon insertion polymorphism (MRIP) in the euL1db. This database collects results obtained in 32 studies containing >900 samples, >140,000 sample-wise insertions and almost 9000 distinct merged insertions.

Paired-end reads	Sequences (forward and reverse) coming from both ends of an amplicon
Fragment	Assembly of forward and reverse reads according to their overlap region
Mapfragment	Uniquely mapping fragment containing the specific portion of the L1 sequence ampified and a mappable unique genomic sequence
Mapcluster	IS
IS	INTEGRATION SITE, genomic reagion where a cluster of at least 2 overlapping mapfragments have been mapped
AIS	ANNOTATED INTEGRATION SITE, IS present in the reference genome
NIS	NON ANNOTATED INTEGRATION SITE, IS not present in the reference genome
PIS	POLYMORPHIC INTEGRATION SITE, IS not present in the reference genome but annotated in the euL1db (MRIP)
Germinal IS	IS present in both the frontal cortex and the kidney of the same individual
Single Tissue IS	IS present in only one tissue of the individual
Private IS	IS present in only one individual
Public IS	IS present in more than one individual

Table 2.1.1.5 SPAM nomenclature In the table is schematically described the nomenclature used throughout the thesis.

2.1.1.6 Bioinformatic pipeline

Before proceeding with the paired-end reads assembly, a quality check of raw reads was performed with FastQC (version v0.10.1) (Andrews, 2010). ADEPT error-detection program (version 1.1) (Feng et al., 2016) was employed to assess and correct PCR and sequencing errors within paired-end reads (-p parameter) before removing read duplicates

with FastUniq (parameters –t q) (version 1.1) (Xu et al., 2012). Trimmomatic (version 0.32) (Bolger et al., 2014) was employed to trim the adapters and clip and filter low quality bases, using the following parameters: leading = 0, trailing = 0, sliding = 3:10, minlen = 25, clip = '2:30:12:10:true'.

Every read-pair was analyzed and classified accordingly to several controls to ensure the selection exclusively of those pairs supporting a unique genomic LINE-1 integration site. The first part of the pipeline is made of a Perl script making extensive usage of BioPerl libraries. We assembled every read pairs using Cap3 (Huang and Madan, 1999) with R2 in reverse complement without using quality values. The read pairs that did not align for at least 30 bp at 80% identity were classified as **NORALIGN** and discarded. The assembled fragments were checked for the presence of the nested forward primer plus the final part of the adapter's long strand (GTGGCTGAATGAGACTGGTGTCGACACTAGTGGT). We noticed the formation of concatamers in our initial analysis and all of them contained the hairpin sequence (TTTTTTTGAAAAAAA) in reverse complement inside the fragment sequence. Therefore, every fragment has been checked against the hairpin sequence allowing 2 mismatches (mm) at maximum. Fragments containing the reverse complement of the hairpin sequence were classified as **HAIRPIN** and discarded. Then we checked the remaining fragments for the presence of human LINE sequence using BLASTN with default parameters against a selection of human LINE sequences extracted from the REPBASE database. The fragments not containing the expected LINE-1 sequence in the expected position were classified as **NOREP** and discarded. We trimmed the portions of the remaining fragments overlapping the adapter's long strand and the LINE-1, discarding and classifying as **NOSPACE** all the trimmed fragments resulting shorter than 30 bp. The remaining fragments were mapped on the human reference genome Hg19 (GRCh37 Feb. 2009, downloaded from UCSC) using BLASTN with the following parameters: word_size 20, evalue 1e-10, perc_identity 90. If the length, evalue and percentage identity of the first two BLAST hits were identical, the fragment was classified as **BLNOUNIQUE** and discarded. If the BLAST result was shorter than 30 bp or less than 50% of the query fragment, the fragment was classified as **BADBLRES** and discarded. Similarly, if no BLAST results were obtained, the fragment was classified as **NOBLAST** and discarded. All the remaining fragments were classified as **BLASTRES** and the match collected in the result table. Among all the putative BLASTRES we retained only the ones containing the entire sequence (0 mismatches) of the synthetic

adapter flanking the genomic portion. Moreover, in order to avoid technical redundancy, we retained only one **BLASTRES** among identical BLASTRES of the same individual. These were classified as MapFragments. The way in which MapFragments were mapped allowed us to precisely predict the IS at the boundary between the 5' of the FL-L1 and the flanking genomic region. However, we allowed a range of 10 bp around the predicted IS nucleotide in order to include possible mismatches or sequencing errors that could create a slightly shifted mapping of the IS. All the IS in the range of 10 bp and on the same strand were clustered together to form what we named MapCluster. MapFragments and MapClusters were associated to overlapping/flanking coding gene structures (gene, exon, intron, 5'UTR, 3'UTR, promoter, intergenic region) based on Ensembl 75 (Feb 2014). Every intergenic element was associated to the closest gene up to a distance of 10Kb. MapFragments and MapClusters were annotated as AIS, NIS or PIS according to the overlap/distance of the IS and the 5' end of the closest LINE-1 sequence using the annotation from RepeatMasker, downloaded from the UCSC genome browser. Elements whose IS had a distance bigger than 30 bp from the closest LINE-1 5' end and/or displayed opposite orientation were considered as NIS or PIS. IS that were detected in more than one tissue of the same individual were classified as "Germinal", while IS detected in only one tissue were defined as "Single tissue". IS detected in only one individual were classified as "Private", IS detected in more than one individual as "Public".

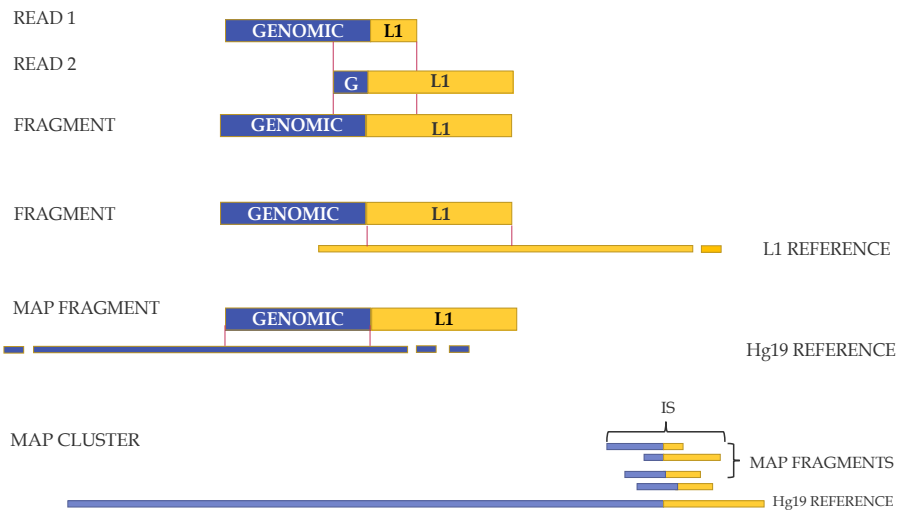


Figure 2.1.2.6 Schematic representation of the principal steps of the SPAM bioinformatics pipeline. Every read pair (R1 and R2) is assembled using Cap3. If the resulting Fragment contains the expected LINE-1 portion and its genomic portion maps on a unique position over the genome it is called a MapFragment. Overlapping MapFragments create a MapCluster.

2.1.1.6.1 FL-L1 coverage

To assess the ability of SPAM to detect FL-L1 elements, we first sought to determine if the technique was efficient for retrieving reference, full-length elements.
The two LINE-1-specific primers used in the SPAM reaction were both designed at the very 5' of a canonical LINE-1 (the sequence comprised between the two primers goes from base 191 to base 256) in order to match only the full-length L1Hs.
To identify the reference LINE-1 sequences detectable with the two primers and, among them, the LINE-1 elements effectively targeted by SPAM (AIS), we mapped the two primers on the Hg19 genome admitting an increasing number of mismatches with Bowtie (version 1.0.0, -v 0,1,2,3 -y -a -c). The resulting fragments comprised between the two primers were then associated with the closest LINE-1. Overlapping LINE-1s represented the set of reference LINE-1s theoretically targetable with our primers. Intersecting the coordinates of the AIS with the reference LINE-1s targetable with our primers admitting 0,1,2,3 mismatches, we could infer the number of likely FL-L1s effectively targeted with SPAM.

An important information about the position where MapFragments match the reference FL-L1 sequence can be retrieved in the *blastres.tab tables, at the column hrepend (end of the match in the LINE-1 repeat). This is an important parameter that allows us to evaluate if we got the right match based on the used repeat specific primers.

2.1.1.6.2 SPAM Efficiency

In order to measure SPAM efficiency in detecting AIS, PIS and NIS we tried to put the minimum threshold of MapFragments necessary to define an IS at 3 or 5 instead of 2 and we plotted the total number of IS detected in the three different categories according to each threshold. The same analysis was performed taking into account the tissue and the condition.

Then, a rarefaction-like analysis was performed by plotting the average number of different IS identified sampling an increasing number of reads, in order to assess whether the sequencing depth was sufficient to retrieve all the IS (AIS, PIS, and NIS) present in our samples. Looking at the curve slope in the rarefaction plot it is possible to infer whether IS detection is close to saturation or not. Indeed, when the curve becomes flat it means that a higher sequencing depth would result in very few additional IS. The analysis was performed for each sample using R statistical software v. 3.3.2, starting from the total number reads of each sample and performing 100 simulations with increasing numbers of randomly sampled reads (step 100000) in order to count the average number of unique MapClusters (average number of different IS per sample size in 100 simulations) obtained by increasing the sampling size of reads.

2.1.1.6.3 Chromatin accessibility

In order to explore LINE-1 integration sites distribution in the genome we compared the average distance in bp (bedtools closest version 2.16.2) of AIS, AIS that we targeted admitting 2 mm to the primers, AIS that we should have targeted with 2 mm but we did not, NIS and PIS from DNAse I cluster regions present in the human genome (coordinates of DNAse1 hypersensitivity regions were downloaded from UCSC genome browser) with that of a random sample of IS (generated through bedtools shuffle).

2.1.1.6.4 Differential integration analysis

We performed a differential integration analysis in order to find genomic locations differentially targeted by LINE-1 insertions in frontal cortex and kidney of both conditions (AD and CTRL). The analysis generated a list of genes showing differences in the number of associated MapFragments and a list of IS with a differential coverage of MapFragments in the two conditions.

The statistical analyses were performed using the edgeR statistical software version 3.2.0 and corresponding statistical tests. The analyses were performed at the MapCluster level, considering the number of MapFragments associated to each MapClusters, and at the gene level considering the number of MapFragments associated to every gene (max distance 10Kb) and the number of MapClusters associated to every gene (max distance 10Kb). In the analysis, the number of MapFragments are used as a quantitative measure in a similar way as the counts of mapped reads is used in an RNAseq experiment. The MapFragments counts were normalized using the calcNormFactors, estimateCommonDisp and estimateTagwiseDisp functions and the differential analysis using the exactTest function of the edgeR package. Resulting p-values were corrected using the FDR method. We considered as significant the results showing a p-value ≤ 0.1.

2.1.1.6.5 Gene ontology enrichment analysis

To assess the implications of IS associations on gene function, we examined NIS and PIS closest genes (≤ 10000 bp) with respect to Gene Ontology (GO) functional category classification using GOseq (Young et al., 2010) R Bioconductor package (version 1.22.0). This package provides methods for performing Gene ontology analysis, taking into account the length bias. This normalization is very important to avoid biased enrichments in long genomic loci (typical of neuronal genes). Only results showing the GOseq parameter numDEInCat (in our study corresponding to the number of GO term associated test genes) higher than 10 and over-represented pvalue < than under-represented pvalue were FDR adjusted and reported for NIS. Such filters were not applied for PIS, due to the small number of IS associated genes in this category.

We considered the three GO divisions: biological process, molecular function, and cellular component. The proportion of genes associated with each GO term was compared between: AD vs AD + CTRL, CTRL vs AD + CTRL, K versus FC + K, FC versus FC + K, FC AD vs FC, FC CTRL vs FC, K AD vs K, K CTRL vs K, random IS (of the same number and width of the real NIS+PIS) versus FC + K.

Gene ontology analysis considering the same comparisons was performed also with GREAT (version 3.0.0), using the following settings: single nearest gene within 1000 Kb.

2.1.1.7 Validation PCR

The validation was performed by a unique round of PCR using three primers: one forward primer designed on the genomic DNA at the insertion site (upstream the LINE-1 element), one reverse primers designed against the very beginning of the LINE-1 5'UTR, and one reverse primer designed on the genomic sequence downstream the LINE-1 insertion in order to detect in the same PCR reaction the presence or absence of the LINE-1 insertion. In particular, the primers used for the HLA-genotyping assay were designed on the sequences of the human reference genome Hg19 (human assembly GRCh37 Feb. 2009, also known as hg19, downloaded from UCSC) with (HLA haplotypes DBB/MANN) or without the HLA insertion. The PCR was performed on 200 ng of genomic DNA in a final volume of 20 µL (except for the IS-HLA assay that was performed in 50 µL) using 240 nM of each primer, 200 µM dNTPs, 1X ExTaq PCR buffer (Takara) and 1.25 U of ExTaq (Takara) with the following protocol: 10 min at 94°C, 40 cycles of 30s at 94°C, 30s at 60°C, 1 min at 72°C, and final elongation of 5min at 72°C. PCR products were run on a Midori Green stained 2% agarose gel, and, for the first test, bands were extracted using the Qiagen Gel extraction kit following manufacturer's instructions and Sanger sequenced to confirm the specificity of the PCR products.
For analysis of Hardy Weinberg Equilibrium (HWE), allele frequencies and Odds ratio Chi square test was used. Statistical significance was defined as $p<0.05$.
The ddPCR experiment was performed with the QX200 Droplet Digital PCR System by Bio-Rad. The ddPCR assay designed on the housekeeping gene RPP30 was previously published by White and colleagues (White et al., 2014), as well as the probe and the reverse primer designed on the 5'UTR of the LINE-1 element. We designed the forward primers specific for each different genomic location where the IS were inserted and we adapted the protocol with some minor modifications. Each reaction was performed in 20 µL using the ddPCR Supermix for Probes (No dUTP) by Bio-Rad with 900 nM of each primer, 250 nM of each probe and 50 ng of gDNA. The cycling conditions were: 10 min at 95°C, 40 cycles of 30 s at 94°C and 1 min at 64°C, with a final incubation of 10 min at 98°C. A 2°C/s ramp rate was performed at each step of the PCR. Data analysis was performed using the QuantaSoft Software by Bio-Rad.

Validation performed by Marta Maurutto

2.1.1.8 Gene expression analysis of the genes located near AIS and PIS

Genotype data of structural variants (SVs) of 445 individuals from 5 different populations (Great Britain (GRB), Finland (FIN), Yoruba (YRI), Northern Europeans from Utah (CEU) and Tuscany (TSI)) were downloaded from the 1000 Genomes Project (phase 3) website, while gene expression data from a lymphoblastoid cell line (LCL) of the same individuals were obtained from the 'RNA sequencing project' section of the GEUVADIS website on the same date.

AIS and PIS characterized by an FDR corrected p-value <0.1 in the differential integration analysis were compared with all SVs from the previously downloaded 1000 GP data through a bedtools analysis, in order to find matching SVs. The 'closest' function of the toolkit allows comparing two sets of genomic coordinates reporting the closest hit for every set of coordinates in the query. AIS and PIS genomic coordinates were used as the query, while SVs genomic coordinates from the 1000 GP represented the database. Database SVs found to be closer than 10 bps to significant AIS and PIS were considered to be a match.

The BBduk functionality of the toolkit BBTools has been used to assess the presence of a query nucleotide sequence specific for PIS which did not match with SVs from the 1000 genome project in fastq data relative to individuals from the 1000 genome project.

In order to assess the impact on gene expression of matched SVs, the expression levels of each matched SV's closest coding gene between individuals carrying the SV ("1/1" or "1/0") and individuals not carrying the SV ("0/0") were compared. Furthermore, the same analysis was performed for all genes found in a range of 500 kbp around each SV.

A custom python script was used in order to associate each genotype with the correct gene expression data for each SV and each individual. P-value was calculated through R's Student's t-Test function. The same analysis was performed for assessing the expression of each exon of the genes of interest.

Analysis performed by Giovanni Spirito

2.1.2 Quantification of FL-L1 insertions in different tissues

In order to study the copy number variation of potentially FL-L1 elements in the human genome we decided to adopt the qPCR technique with Taqman probes, as previously performed by Coufal and colleagues for the total LINE-1 elements (full-length and truncated). In particular, we designed a new Taqman assay on the 5'UTR sequence of a canonical L1Hs element (Lavie et al., 2004) to be sure to amplify only the complete full-length sequences.

Experiment performed by Marta Maurutto

2.1.2.1 Samples

The genomic DNA samples used in the copy number variation (CNV) analysis were extracted from human autopsy specimens of two different cohorts of patients.
The Spanish cohort, received from the Bellvitge Neuropathology Institute in Barcelona comprised samples of frontal cortex from 10 AD patients at the final Braak stages V-VI (severe AD), 10 patients at Braak stages I-II (mild AD), and 7 healthy controls.
The Brazilian cohort, provided by the Brain Bank of Sao Paulo, comprised samples of frontal cortex, temporal cortex, hippocampus, cerebellum and kidney from 10 AD patients at Braak stages IV-VI and 10 not affected individuals at Braak stages 0-II.
Genomic DNA was extracted using a standard phenol-chloroform extraction method.

Cohort	**Condition**	**Patients**	**Braak NFT**	**Age**	**Gender (F:M)**	**PMD**
Spanish cohort	CTRL	7	0	70 ± 8	2:5	3 ± 1
	early AD	14	I-II	73 ± 12	4:9	8 ± 6
	late AD	10	V-VI	80 ± 4	5:5	10 ± 5
Brazilian cohort	CTRL	9	0-III	80 ± 15	6:3	N.A.
	AD	8	III-VI	83 ± 10	4:4	N.A.

Table 2.1.2.1 CNV Samples. Two cohorts of AD affected patients and CTRLs were screened for FL-L1 CNV.

2.1.2.2 FL-L1 CNV analysis

The quantitative real-time PCR with Taqman probes was performed according to Coufal et al. with minor technical modifications. The new assay employed in the copy number variation analysis of full-length elements was designed on the 5'UTR sequence of an L1Hs (L1-Ta) element. According to the sequences present in the LINE-1 database (http://l1base.molgen.mpg.de/), 114 of the 146 elements in the Human FL-L1 (Ens84.38) database are detected by the assay. Coordinates in the L1Hs sequence for the assay are from base 98 to base 254. The assay for the invariant inner control was designed on the glyceraldeyde 3-phosphate dehydrogenase (GAPDH) sequence. A target area of 204 bases between exon 3 and exon 4 was selected. Taqman probes and primers were designed using the online tool Primer3 (http://primer3.ut.ee/). The experiment was performed in triplicate, and data were analyzed with the 2^{-Ct} method.
The three replicas were analyzed individually and considering the mean values for each sample with a two-tailed Mann Whitney statistical test. $P < 0.05$ was considered significant.

Taqman Assay	Forward primer (5'→3')	Reverse primer (5'→3')	Probe (5'→3')
L1 5'UTR	GAGGTACCGGGTTCATCTCA	TCACCCCTTTCTTTGACTCG	TAGGGAGTGCCAGACAGTGG **(FAM)**
GAPDH	CCCTTCATTGACCTCAACTACATG	TGGGATTTCCATTGATGACAAGC	CGTTCTCAGCCTTGACGGTGCCAT **(VIC)**

Table 2.1.2.2 CNV Primers

Experiment performed by Marta Maurutto

2.1.3 Characterization of LINE-1 content in genomic variations

LINE-1 content in CNVs was evaluated performing an analysis of genomic CNVs using the Illumina Infinium high-density chip. We considered both the number of retrotransposons that were in overlap with the CNVs and their coverage values in order to increase the resolution of the study of LINE-1 copy number performed with the qPCR technique.

2.1.3.1 Samples

Sample ID	Sex	Age	Cohort	Type	Braak (NFT)	CERAD	tissue	DRB1*	DRB1*	DQA1*	DQA1*	DQB1*	DQB1*	
C02_S001_C	F	65	Spanish	CTRL	0	0	FC	03:01:01:01	03:01:01:01	05:01:01:01	05:01:01:01	02:01:01	02:01:01	
C02_S002_C	M	67	Spanish	CTRL	0	0	FC	03:01:01:01	13:01:01	01:03:01:01	05:01:01:01	02:01:01	06:03:01	
C02_S003_C	F	69	Spanish	CTRL	0	0	FC	04:02:01	07:01:01:01	02:01	03:01:01	02:02:01	03:02:01	
C02_S004_C	M	85	Spanish	CTRL	0	0	FC	03:01:01:01	13:01:01	01:03:01:01	05:01:01:01	02:01:01	06:03:01	
C02_S005_C	M	78	Spanish	CTRL	0	0	FC	03:01:01:01	04:04:01	03:01:01	05:01:01:01	02:01:01	03:02:01	
C02_S006_C	M	66	Spanish	CTRL	0	0	FC	04:03:01	15:01:01:01	01:02:01:01	03:01:01	03:04:01	06:02:01	
C02_S007_C	M	61	Spanish	CTRL	0	0	FC	11:04:01	15:01:01:01	01:01:03	05:01:01:01	03:01:01:01	06:02:01	
C09_S020_C	F	66	Spanish	CTRL	0	0	FC	04:07:01	07:01:01:01	02:01	03:01:01	02:02:01	03:01:01:01	
C09_S036_C	M	75	Spanish	CTRL	0	0	FC	04:03:01	15:01:01:01	01:02:01:01	03:01:01	03:02:01	06:02:01	
C04_S004_A	M	79	Spanish	AD	5	NA	FC							
C04_S006_A	M	78	Spanish	AD	5	NA	FC							
C04_S009_A	F	85	Spanish	AD	5	NA	FC							
A08/00153	F	74	Spanish	AD	5	NA	FC							
C04_S012_A	F	81	Spanish	AD	5	NA	FC							
C04_S013_A	M	87	Spanish	AD	5	NA	FC							
C04_S014_A	M	84	Spanish	AD	5	NA	FC							
C04_S018_A	F	77	Spanish	AD	6	NA	FC							
C09_S001_A	M	84	Spanish	AD	4	C	FC	01:01:01	07:01:01:01	01:01:01	02:01	02:02:01	05:01:01:01	
C09_S002_A	M	75	Spanish	AD	6	NA	FC	07:01:01:01	10:01:01	01:01:01	02:01	02:02:01	05:01:01:01	
C09_S006_A	M	84	Spanish	AD	4	B	FC	07:01:01:01	07:01:01:01	02:01	02:01	02:02:01	02:02:01	
C09_S007_A	M	75	Spanish	AD	5	B	FC	07:01:01:01	15:01:01:01	01:02:01:01	02:01	02:01:01	06:02:01	
C09_S011_A	M	77	Spanish	AD	5	C	FC	15:01:01:01	15:01:01:01	01:01:01	01:02:01:01	06:02:01	06:02:01	
C09_S019_A	M	89	Spanish	AD	4	B	FC	03:01:01:01	13:01:01	01:01:03	05:01:01:01	02:01:01	06:09:01	
C09_S018_A	M	79	Spanish	AD	4	B	FC	07:01:01:01	14:01:01	01:01:01	02:01	02:02:01	05:03:01:01	**Excluded**
C09_S020_A	M	86	Spanish	AD	5	B	FC	04:02:01	04:05:01	03:01:01	03:01:01	02:02:01	03:02:01	
C09_S003_A	F	75	Spanish	AD	4	B	FC	07:01:01:01	11:04:01	01:02:01:01	02:01	02:02:01	05:01:01:01	
C09_S004_A	F	79	Spanish	AD	4	B	FC	01:01:01	11:04:01	01:01:01	05:01:01:01	03:01:01:01	05:01:01:01	
C09_S005_A	F	96	Spanish	AD	5	C	FC	03:01:01:01	11:01:01	05:01:01:01	05:01:01:01	02:01:01	03:01:01:01	
C09_S008_A	F	83	Spanish	AD	4	C	FC	01:01:01	14:01:01	01:01:01	01:01:01	05:01:01:01	05:03:01:01	
C09_S009_A	F	81	Spanish	AD	4	C	FC	11:04:01	13:01:01	01:01:03	05:05:01:01	03:01:01:01	06:09:01	
C09_S013_A	F	86	Spanish	AD	6	C	FC	04:02:01	07:01:01:01	02:01	03:01:01	02:02:01	03:02:01	
C09_S015_A	F	81	Spanish	AD	5	C	FC	07:01:01:01	11:01:01	02:01:01	05:01:01:01	02:01:01	03:01:01:01	
C09_S017_A	F	81	Spanish	AD	4	C	FC	11:01:01	13:02:01	01:01:03	05:05:01:01	03:01:01:01	06:04:01	
C01_S001_C	M	75	Brazilian	CTRL	2	A	FC, CER, K	07:01:01:01	13:01:01	01:02:01:01	02:01	02:02:01	05:01:01:01	
C01_S002_C	M	79	Brazilian	CTRL	0	A	FC, CER, K	03:02:01	11:01:01	04:01:01	05:03	03:01:01:01	04:02:01	
C01_S004_C	M	85	Brazilian	CTRL	2	0	FC, CER, K	01:01:01	08:04:01	01:01:01	05:01:01:01	03:01:01:01	05:01:01:01	
C01_S005_C	F	61	Brazilian	CTRL	0	0	FC, CER, K	01:02:01	11:01:01	01:01:01	05:01:01:01	03:01:01:01	05:01:01:01	
C01_S006_C	F	75	Brazilian	CTRL	2	0	FC, CER, K	03:01:01:01	04:06:01	03:02	05:01:01:01	02:01:01	04:02:01	
C01_S007_C	F	76	Brazilian	CTRL	1	0	FC, CER, K	04:02:01	13:01:01	01:03:01:01	03:01:01:01	03:02:01	06:03:01	
C01_S008_C	F	88	Brazilian	CTRL	3	0	FC, CER, K	04:03:01	11:03:01	03:01:01	05:01:01:01	03:01:01:01	03:04:01	
C01_S009_C	F	89	Brazilian	CTRL	2	0	FC, CER, K	01:01:01	11:01:01	01:01:01	05:01:01:01	03:01:01:01	05:01:01:01	
C01_S010_C	F	92	Brazilian	CTRL	3	A	FC, CER, K	01:01:01	13:01:01	01:01:01	01:03:01:01	05:01:01:01	06:03:01	
C01_S001_A	M	72	Brazilian	AD	6	C	FC, CER, K	01:01:01	04:02:01	01:01:01	03:01:01	03:02:01	05:01:01:01	
C01_S002_A	M	80	Brazilian	AD	3	C	FC, CER, K	03:01:01:01	14:01:01	01:01:01	05:01:01:01	02:01:01	05:03:01:01	
C01_S003_A	M	82	Brazilian	AD	6	C	FC, CER, K	03:01:01:01	07:01:01:01	02:01	05:01:01:01	02:01:01	02:02:01	
C01_S004_A	M	87	Brazilian	AD	4	B	FC, CER, K	07:01:01:01	07:01:01:01	02:01	02:01	02:02:01	02:02:01	
C01_S005_A	F	80	Brazilian	AD	4	B	FC, CER, K	01:01:01	07:01:01:01	01:01:01	02:01:01	02:02:01	05:01:01:01	**FC excluded**
C01_S006_A	F	83	Brazilian	AD	6	C	FC, CER, K	04:05:01	13:01:01	01:03:01:01	03:01:01	03:02:01	06:03:01	**K excluded**
C01_S007_A	F	90	Brazilian	AD	4	A	FC, CER, K	01:01:01	04:02:01	01:01:01	03:01:01	03:02:01	05:01:01:01	
C01_S008_A	F	92	Brazilian	AD	4	B	FC, CER, K	07:01:01:01	07:01:01:01	02:01	02:01	02:02:01	02:02:01	

Table 2.1.3.1 Illumina Infinium high-density chip samples

2.1.3.2 The Illumina Infinium high-density chip assay

Illumina® Infinium OMNI 5 arrays were selected due to their exceptional SNP probes density and ultra-high coverage of the whole Human genome (>4.3e+9 probes). 96 samples were genotyped according to manufacturer's protocols (Illumina Infinium LCG Quad Assay) and analyzed using the PennCNV tool (v. 2014 May 07, parameters: '-confidence') (Wang et al., 2007). The required PFB file (Population Frequency of B allele) was downloaded from the PennCNV site (downloaded from: YALE_Merged_PFB_hg19.pfb, v. 2014 Aug 18). We strictly followed the recommended steps, during which, 91193 records were discarded due to a lack of PFB information for the markers. In addition, samples C01_S005_A_FC, C09_S018_A and C01_S006_A_K showed quality issues and metadata inconsistency and were not considered for further analyses. Identified CNVs were filtered by confidence scores, keeping only CNVs with a score higher than 30 which led to the identification of 6040 total CNVs. Finally, CNVs that were probably split were joined using the PennCNV's clean.pl script with default parameters leading to a final number of 5675 total CNVs.

Analysis performed by Gabriele Leoni

2.1.3.3 CNV annotation and bioinformatics analysis

Filtered and joined CNVs were grouped by sample, tissue, and type of CNV. Dissimilarities in the distributions of lengths and counts were evaluated for each group of CNV using the R statistical software (version 3.3.2.). We generated several LINE-1 collections from different sources. LINE-1 sequences were first retrieved from three L1base dbs in genome build hg38 (FL-L1, LINE-1 with intact ORF2 and LINE-1 longer than 4500 bp) (Penzkofer et al., 2017). We converted their coordinates by aligning the sequences to genome build hg19 using a local megablast (version 2.4.0, parameters: '-perc_identity 100'). From the results, we kept only coordinates that presented full coverage and 100% of sequence identity with the reference genome. From UCSC table,

RepeatMasker annotation was retrieved in genome build hg19 and in addition, also coordinates of the elements targeted by the qPCR experiments, and TaqMan probes coordinates were taken into account.
Additionally, from the RepeatMasker annotations, we generated a subset composed of only-LINE-1 sequences longer than 4500 bp.
Bedtools suit (v2.25) was used to calculate the coverage and the number of overlaps between each CNV and LINE-1 from the different collections (bedtools coverage: default parameters, bedtools intersect: "-wa" and "-loj"). Coverage was intended as the percentage of nucleotides in a CNV covered by an LINE-1 element over its total length. For each sample, we also calculated a total coverage by using the sum of the CNVs lengths and the sum of the LINE-1 fragments in overlap with them, and we used these values to draw the boxplots and perform the statistical analyses.

Analysis performed by Gabriele Leoni

2.2 Analysis of LINE mediated SV in Olfr2 locus

2.2.1 Exploring LINE role in the generation of structural variants such as deletions

Despite ongoing progress in Next-generation sequencing technologies (NGS), none of the actual approaches is capable of capturing the full spectrum of SV events with high sensitivity and specificity, especially in complex, repetitive regions.
The highly repetitive structure of Olfr2 locus and the somatic nature of the regulatory variation that we are looking for, make our task really challenging.
Combining PacBio single molecule sequencing for reliable mapping across repeat expansions and low complexity regions with a complementary high-fidelity paired-end Illumina sequencing for accurate identification of breakpoints we try to overcome the limits of the aforementioned techniques. While with Multiple Displacement Amplification we tried to increase the possibility to target the rare event occurred in the GFP-positive cells expressing the receptor collected by Laser Capture Microdissection.

2.2.1.1 Animals

B6;129P2Olfr2tm1Mom/MomJ (Jackson) mice 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.
The hybrid B6;129P2Olfr2tm1Mom/MomJ mice (C57BL/6 x129 genetic background) were chosen because they contain a GFP gene inserted at the 3' of the Olfr2 locus (Bozza et al. 2002), an OR expressed in a small amount of OSN. Since this conformation gives

rise to a bicistronic mRNA, cells that naturally activate the transcription of Olfr2, become fluorescent, therefore amenable to be identified.

2.2.1.2 Sample preparation for Laser Capture Microdissection

OE samples were prepared from 6 days old B6;129P2-Olfr2tm1Mom/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% sucrose1× 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 the 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.

Experiments performed by Alice Urzi

2.2.1.3 Whole genome amplification (WGA)

To evaluate the most efficient WGA protocol we amplified genomic DNA (from 10 olfactory neurons and from bulk genomic DNA from OE) using multiple displacement amplification (MDA).

Experiments performed by Alice Urzi

2.2.1.4 Multiple Displacement Amplification (MDA)

Different pools of 10 GFP positive cells were collected by LCM from B6;129-Olfr2-GFP mice at the age of p6and immediately amplified with Multiple Displacement amplification.

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 manufactures 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 in the experiment.

As a control of successful Olfr2 locus amplification we checked the presence of Olfr2 coding sequence (CDS) in the amplified product performing control PCR on each MDA replicate, before proceeding with subsequent long-range amplifications.

Experiments performed by Alice Urzi

2.2.1.5 Long-Range PCR amplification of 50 kb Olfr2 locus

Purified MDA products and bulk genomic DNA coming from the very same mouse line (without any MDA amplification) were used as template for long-range 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.
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 also from bulk genomic DNA, extracted from OE (gDNA-OE) and not amplified by MDA. Bulk OE DNA sample is a "negative biological control" because it consists of whole OE cell population among which Olfr2-expressing cells (GFP positive) represented around the 0.1% of the total olfactory sensory neurons.
The best PCR products from each of 11 MDA biological replicas were purified and pooled together for Pac Bio sequencing in parallel with PCR products from bulk OE genomic DNA

For the subsequent Illumina sequencing, the 50 kbp sequence locus was divided into 13 amplicons with a size ranging from about 400 bp to about 5k bp (amplicon 2 was divided into three sub-amplicons for Illumina sequencing:2.1, 2.2 and 2.7). We performed the PCR amplification using a long-range 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).

For Illumina sequencing we performed long-range PCR of Olfr2 locus on two different MDA biological replicas out of 11 (MDA-V and MDA-XI) and on bulk genomic DNA from OE (OE sample). Finally, 13 PCR amplicons for each sample were sequenced for a total of 39 Illumina libraries.

Experiments performed by Alice Urzi

2.2.1.6 Illumina sequencing, read quality check and mapping

A 300 bp Illumina MiSeq paired-end sequencing was performed at IGA technology Services (Udine, Italy). Quality check and trimming were performed with FastQC (version v0.10.1) (Andrews, 2010) and Trimmomatic (version 0.32) (Bolger et al., 2014) on Illumina reads.

In preparation for variant-calling, the alignment of high-fidelity Illumina paired-end reads over mouse reference genome (mm10, NCBI build GRCm38) was performed using Burrows-Wheeler mapping software (version 0.7.10) BWA-MEM (default parameters for paired-end mapping and –M option). The gapped aligner was chosen in anticipation of the subsequent SV detection (Li and Durbin, 2009).

2.2.1.7 Pac Bio sequencing

About 5kb long-range PCR products of each amplicon of the two samples (OE and MDA) were pooled together in equimolar ratio to reach 2micrograms of total DNA and sequenced on the PacBio RSII platform at GATC Biotech in Germany, to obtain long PacBio reads.

2.2.1.8 Variation discovery

Sorted and indexed BAM files (samtools, version 0.1.18) were scanned for the presence of variations (small insertions, deletions, tandem duplications and inversions) with Pindel (default parameters, insert size 500) (Ye et al., 2009).
Pindel was chosen as this tool is known to reliably identify medium sized SVs, especially large deletions, starting from paired-end reads. Pindel employs a combined Split Read/read-Pair approach, searching for clusters of split reads using balanced splits as seeds and evaluation of the span and orientation of paired-end reads (Karakoc et al., 2012).
The analysis was limited to the 50 Kb Olfr2 locus extended by 10 Kb at the 3' and 10 Kb at the 5'. Only variations in the 70 Kb regions were considered (parameter –c chr7:106967606-107037605).
Variations not covered by at least 5 reads in at least one sample (MDAV, XI or OE) were discarded. Variation coverage was calculated by dividing the number of alternative allele supporting reads by the coverage of reference reads.

2.2.1.9 Deletion validation with single molecule PB reads

Variant callers are known to be prone to false positive calls due to alignment errors. Such errors may occur when the number of bases in the reads, matching the reference genome, is too few and when the number of reads supporting a SV is small. This problem exacerbates in highly repetitive regions.
For this reason, we decided to use a PacBio read data set, complementary to the Illumina one, to increase the accuracy of variation prediction in Olfr2 locus.
Illumina split reads supporting the deletions were aligned over PB long reads using blastn (version 2.2.29+, parameters word_size 20, perc_identity 80, evalue 1e-10). Each Illumina read supporting a deletion present in the OE sample was aligned over all PB reads coming from the same sample and over all reads coming from the MDA sample and vice-versa).

In order to consider validated with PB a deletion detected with Pindel, at least one query Illumina fasta sequence supporting the deletion should align for its entire length (± 10 nucleotides) over at least one corresponding subject PB read, regardless the sample.

2.2.1.10 Repeat coverage

Repeat description in correspondence of the deletions longer than 50 bp was performed with bedtools intersect (version 2.16.2) (Quinlan and Hall, 2010) according to UCSC's repeatmasker (http://www.repeatmasker.org) annotation (version open-4.0) .

2.2.1.11 Deletion Clustering

This analysis was performed in order to reduce the complexity of the detected deletions. Deletions supported by 5 or more Illumina reads in at least one sample, supported by at least 1 PB read were clustered together if they overlapped a minimum of two LINE elements reported in the reference genome.

2.2.1.12 PCR validation assay

Selected deletions were validated by PCR assays performed with ExTaq DNA-Polymerase (Takara) following the manufacture's protocol. A list of validation primers used is shown in table 2.2.1.12.
According to the putative band length expected by the primers position and observed in the gel, it was possible to precisely retrieve bioinformatically the exact Pindel ID of the validated deletion, among many overlapping possible ones. Sanger sequences, then, were

manually checked using BLAT (Kent, 2002) and BLAST (Altschul et al., 1990) to verify if they reflected the deletions detected in silico accordingly with the Illumina and PB split reads.

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

Table 2.2.1.12 List of primers used for Pindel deletion validation PCR assays. Primers were designed on the forward strand.

2.2.1.13 DRS discovery

This analysis was limited to search microhomology regions in a range of 60 bp (30 bp upstream and 30 bp downstream) around the deletions breakpoints of the 125 clustered deletions.

60 bp 3' and 5' reference fasta sequences (bedtools getfasta version 2.16.2) were scanned with a custom python script to look for the longest direct repetitive sequences (DRS) among them.

Annotation of the LINE-1s present in the reference genome (repeat masker version open-4.0) in those regions was performed with bedtools intersect (version 2.16.2).

Clustal omega (Larkin et al., 2007) multiple sequence alignment was employed to look for a relationship between the 125 microhomology regions.

Blast2 and ClustalO were employed to determine the percentage of sequence conservation between the repetitive elements at the deletion breakpoints.

The same analysis was performed creating a random set of deletions in the 70kb locus (bedtools shuffle version 2.16.2).

To assess the significance of the difference in the proportions of the real and random microhomology regions a statistical test of Equal or Given Proportion (prop.test) was performed with R software. We evaluated statistically whether the observed length

microhomology regions was significantly different from what was expected by chance performing a Student's t-Test (t.test) to compare the average length of real and random DRSs.

GC content of real and random motifs was measured with the online tool Genomics % GC Content calculator (Science Buddies).

2.3 Chip Seq analysis of endogenous γ-H2AX in mouse olfactory epithelium and liver

2.3.1 Profiling double strand breaks: cause and effect of structural variants

Endogenous DNA double-strand breaks are thought to be the principal cause of genomic instability in all cells since their misrepair may lead to mutations, deletions, and rearrangements.

γH2AX is known to be an important, early player of the repair cascade on the chromatin flanking the DSBs. For this reason, it is often employed as a marker of DSB. However, the determinants controlling the distribution of γH2AX are still unknown.

Most research relying on chromatin immunoprecipitation (ChIP) methods to understand how γH2AX contributes to double-strand break repair in mammalian cells starts from artificially induced, often target specific DNA breaks (Madabhushi et al., 2015, Katsube et al., 2014, Redon et al., 2009).

In our exploratory study, we try to profile spontaneous γH2AX signal at physiological conditions.

2.3.1.1 Samples

In order to investigate how γ-H2AX distributes in the mouse genome we performed a chromatin immune-precipitation (IP) and sequencing experiment in C57BL/6J mice, analyzing OE at 6 days (p6) and 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.

2.3.1.2 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 μg 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/chloroform protocol. 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.

Experiments performed by Alice Urzi

2.3.1.3 ChIP samples sequencing and peak calling

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 adapter sequences. First, raw reads were trimmed against adapters 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 (Li and Durbin, 2009). Duplicates were marked using picard tools and reads with mapping quality below 60 were filtered out together with duplicates and improper pairs. The subsequent filtering, sorting, and mate fixing steps were performed using samtools (version 0.1.19) (Li et al., 2009). Peak calling was performed on the filtered data using epic (version 0.1.18), a peak caller based on SICER (Xu et al., 2014) suitable to identify diffused domains of enrichment, which is the pattern expected for the γ-H2AX signal. Peaks overlapping blacklisted

genomic regions were removed using intersectbed from bedtools suite. Blacklisted regions are artifact regions that tend to show artificially high signal (excessive unstructured anomalous reads mapping), often corresponding to repetitive regions such as centromeres and telomeres. 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. 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.

Analysis performed by Fei Sang and Margherita Francescatto.

2.3.1.4 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.

Analysis performed by Margherita Francescatto.

2.3.1.5 Gene ontology enrichment analysis

Gene ontology (GO) enrichment analysis was performed using GREAT online tool (McLean et al., 2010). 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.

Analysis performed by Alice Urzi.

2.3.1.6 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).

Analysis performed by Margherita Francescatto.

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

Liver CAGE expression data

Liver expression data was derived from the mouse tissue catalog 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_phase1and2combined_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 filtered in order to retain only CAGE peaks with at least 1tpm (tpm=tags per million) in all samples.

OE expression data

Expression data from the work of Ibarra-Soria and colleagues (Ibarra-Soria et al., 2014) was 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.

Analyses performed by Margherita Francescatto.

2.3.1.8 Comparison of ChIP-seq peaks with chromatin segmentation of the L mouse genome

We downloaded 11 ChIP-seq datasets:

- 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.

The data was 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.

Analyses performed by Margherita Francescatto.

2.3.1.9 Comparison of ChIP-seq peaks distribution with respect to gene clusters.

Top 10 mouse gene cluster (ranking based on the number of genes associated with each cluster) features and coordinates were kindly provided us by Massimiliano Volpe (Stazione Zoologica Anton Dohrn, Naples). Massimiliano developed a tool to identify potential genic clusters in the genome. This tool combines in a unique cluster each genomic interval in which at least two genes, residing in the same chromosome are closer than 500kb and share the same domain (Biomart Pfam annotation). OE, L and random peaks distribution with respect to mouse gene clusters (all 10 clusters, olfactory receptor clusters, Vomeronasal Clusters, Immunoglobulin Clusters, Zinc Finger clusters and Homeobox clusters) was performed with bedtools closest (version 2.16.2). Random peaks were created with bedtools shuffle (version 2.16.2), imposing no overlap between random peaks and real peaks. We used again bedtools shuffle to randomize the distribution of gene clusters. This was performed in order to compare the distribution of real peaks and random peaks with respect to real clusters, with the distribution of real peaks and random peaks with respect to random clusters.

2.3.1.10 Comparison of ChIP-seq peaks distribution with CTCF, Pol II and DNAse data.

CTCF and Pol II datasets (olfactory bulb 8weeks, liver 8weeks) were downloaded from UCSC genome browser, ENCODE Chip-seq (Robertson *et. al* 2017): https://www.genome.ucsc.edu/cgi-bin/hgFileUi?db=mm9&g=wgEncodeLicrTfbs. Liver (8 weeks, C57BL/6 DNAse I DGF) and Whole Brain (8 weeks, C57BL/6 DNAse I DGF) DNAse I datasets (no olfactory epithelium DNAse I dataset was available) were downloaded from http://ucscbrowser.genap.ca/cgi-bin/hgTrackUi?db=mm9&g=wgEncodeUwDgf.
Peak distribution and gene cluster distribution with respect to CTCF, Pol II and DNAse data was assessed with bedtools closest (parameters – d, t first version 2.16.2).

Because the peak calling is not an exact process, we accepted two features to overlap if they were located within 1kb of each other.

2.3.1.11 ChIP-seq peaks with respect to different class of repeats

OR gene clusters, L, OE and random Peaks were associated with the proximal repeats (LINE, SINE, LTR and satellite from rmsk, version) with bedtools closest (parameters –d, t first, version 2.16.2).

3 Results

3.1 Analysis of FL-L1 elements in the genomes of AD post-mortem brains

In this first section of results, we focus on FL-L1 elements induced SVs. FL-L1 are the only autonomous transposable elements present in the human genome able to influence chromosome integrity and gene expression upon reinsertion (Belancio et al., 2009).

LINE-1s mobilize during neuronal differentiation and generate "genomic plasticity" in neurons by causing variation in genomic DNA sequences and by altering the transcriptome of cells (Singer et al., 2010). Neuronal genetic diversity resulting from LINE-1 mediated copy number variations (CNVs), and LINE-1 integration, therefore, could result in individual differences in behavior and disease.

Given that:

- LINE-1s activity has been shown altered in neuropsychiatric disorders (Bundo et al., 2014, Shpyleva et al., 2017),
- age specific global demethylation involves mainly repetitive regions of the genome and can result in the reactivation of retrotransposons (Bollati et al., 2011),
- late onset Alzheimer's disease (LOAD) is an age related neurodegenerative disease (Isik, 2010),

we decided to compare the activity and distribution of FL-L1 in the genome of LOAD affected patients and CTRLs.

This study, performed with different methodologies, on a brain and an extra brain tissue, aims at investigating if FL-L1s polymorphisms can be a relevant source of structural variants associated with AD risks.

3.1.1 Identifying novel FL-L1 insertions

3.1.1.1 The SPAM technique

To understand the impact of retrotransposition in AD, it is essential to delineate the position of LINE-1 insertions in the genome. In order to unambiguously map only active LINE-1 elements present in the human genome, still able to mobilize and give rise to novel LINE-1 integration sites, we established a novel method called SPlinkerette Analysis of Mobile elements (SPAM). This technique, inspired by the splinkerette PCR (spPCR) protocol (see methods), is based on a specific targeting of FL-L1 elements, followed by an accurate bioinformatics estimation of their distribution.

SPAM analysis was performed on frontal cortex and kidney samples of 4 AD patients and 4 CTRLs.
From the neuropathological point of view, the controls were clinically healthy individuals, at Braak stages 0-III, while AD patients were demented individuals at Braak stages IV-VI (view table in materials and methods). Braak staging is a classification system of AD progression based on the spreading of neurofibrillary tangles in symptomatic and non-symptomatic individuals (Braak and Braak, 1995). During the first two stages, clinically silent, neurofibrillary tangles (**NFT**s) are confined mainly to the transentorhinal region of the brain; at stages III and IV, incipient Alzheimer's disease, there is also an involvement of limbic regions, at the last two stages (V-VI), called the neocortical stages, NFTs are present in all the subdivisions of the cerebral cortex and the disease is fully developed. Progression through Braak Stages benchmarks regressions in cognitive function.

To perform the experiment, gDNA was extracted from frozen tissues using a standard phenol-chloroform extraction method and sheared by sonication. This was done in order to avoid amplification biases linked to the irregular genomic distribution of restriction enzyme recognition sites. gDNA fragments were then end-repaired, an adenine was added at the 3'end of the blunted gDNA fragments and finally ligated to a synthetic double stranded adapter, harboring an extra thymine protruding at the 3'end.

After the first round of PCR, a nested round of PCR was performed with primers specific for the adapter and the 5'UTR LINE-1 sequence, in order to amplify the genomic region upstream to the LINE-1 element, comprised between the adapter and the LINE-1 5'UTR. The tags and barcodes inserted in the nested primers allowed the following Illumina MiSeq sequencing in multiplex with 300 bp paired-end set-up. Considering the total LINE-1 elements reported in RepeatMasker, the primers used in the SPAM protocol can detect the majority of FL-L1 elements of the Hs family (the most recent and still active LINE-1 family in the human genome), and part of the LINE-1 elements from the older L1Pa family.

3.1.1.2 The SPAM Bioinformatic pipeline

A series of computational analyses have been used to unambiguously map amplified LINE-1-containing genomic fragments. After quality check performed using FastQC (Andrews, 2010) we performed a paired-end error correction with ADEPT (Feng et al., 2016) and a *De Novo* read duplicate removal with FastUniq (Xu et al., 2012).

Paired end error correction was performed in preparation for the important *De Novo* read duplicate removal step. Error detection (and correction) at the single nucleotide level becomes very important in a technique like ours, where the goal is to unambiguously detect somatic integration events which can be characterized by low read coverage. *De Novo* read duplicate removal then, took care of duplicates introduced by PCR amplification which otherwise could have led to subsequent misleading coverage of the detected FL-L1 integration sites. Trimming was performed with Trimmomatic (Bolger et al., 2014) to cut off bases below the threshold quality. At this point, MiSeq paired-end reads were assembled to create sequencing data for the entire PCR amplicon, overcoming the limit of Illumina read size. Indeed, only the assembled fragments containing the correct synthetic adapter sequence and aligning on an LINE-1 consensus sequence were aligned on the human reference genome Hg19. To rigorously define data produced at each experimental and bioinformatics step (see materials and methods), we introduced an *ad hoc* nomenclature reported in table 3.1.1.2.

Paired-end reads	Sequences (forward and reverse) coming from both ends of an amplicon
Fragment	Assembly of forward and reverse reads according to their overlap region
Mapfragment	Uniquely mapping fragment containing the specific portion of the L1 sequence ampified and a mappable unique genomic sequence
Mapcluster	IS
IS	INTEGRATION SITE, genomic reagion where a cluster of at least 2 overlapping mapfragments have been mapped
AIS	ANNOTATED INTEGRATION SITE, IS present in the reference genome
NIS	NON ANNOTATED INTEGRATION SITE, IS not present in the reference genome
PIS	POLYMORPHIC INTEGRATION SITE, IS not present in the reference genome but annotated in the euL1db (MRIP)
Germinal IS	IS present in both the frontal cortex and the kidney of the same individual
Single Tissue IS	IS present in only one tissue of the individual
Private IS	IS present in only one individual
Public IS	IS present in more than one individual

Table 3.1.1.2 Output of the bioinformatics pipeline. For each sample and tissue are reported the total number of raw reads, cleaned reads resulting from ADEPT FastUniq and Trimmomatic filtering steps, total MapFragments, and specific AIS, PIS and NIS MapFragments and MapClusters.

Uniquely mapped fragment containing the full sequence of the synthetic adapter, the amplified L1-5'UTR sequence, and a flanking unique genomic sequence were called MapFragments. Overlapping MapFragments were then classified in MapClusters which in turn were divided in Annotated Integration Sites (AIS) or Non annotated Integration Sites (NIS) according to the presence of the LINE-1 element in the reference genome and in Polymorphic Integration Sites (PIS) if not reported in the reference genome, but annotated in the euL1db. If polymorphic IS are expected to be present in more than one individual of the species, the fraction of NIS present in only one individual (that we called private) represents potential *De Novo* events that occurred either in the parental genome of the individual carrying the insertion (present in every cell of the individual), or in some cell of the individual (somatic events).

The relaxation of the epigenetic repression that occurs during early developing germline is a dangerous window in which LINE-1 elements can mobilize and integrate into new genomic locations (Zamudio and Bourc'his, 2010). A new integration occurring in the germline is likely to be vertically inherited. Moreover, LINE-1 transcripts appear to be competent for mobilization also in the early embryos after being carried over by the gametes through fertilization (Gerdes et al., 2016). Inherited insertions and insertions occurred in the early embryo development, are going to be present in all the cells of the individual and can be transmitted to the next generation. On the other hand, somatic insertions present only in one or a subset of cells, are not inherited from a previous generation. These insertions are going to be very hard to detect without a single-cell, high coverage, sequencing approach, unless they do not undergo clonal expansion (Doucet-O'Hare et al., 2015).

Therefore, we further categorized the IS as “germinal” or “Single tissue” according to the presence of the IS in both tissues (these events are indicative of a germinal insertion event) or only one tissue (these events are suggestive of a somatic insertion or a lack of saturation). IS detected in only one individual were classified as “Private”, IS detected in more than one individual as “Public”.

3.1.1.3 FL-L1 IS characterization

For each sample processed with SPAM, we obtained on average 8.6 million reads which were analyzed with the described bioinformatics pipeline. This analysis allowed the identification of 4634 total IS: 1197 AIS, 3312 NIS and 125 PIS (Figure 3.1.1.3a).

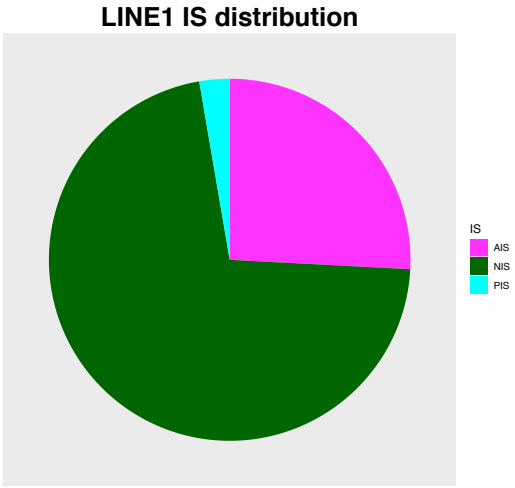


Figure 3.1.1.3a IS characterization. NIS constitute the most abundant fraction of the detected IS (72%), followed by AIS (26%) and PIS (3%).

In particular: 1099 AIS, 282 NIS, and 93 PIS were observed in the AD frontal cortex, 1100 AIS, 371 NIS and 92 PIS were observed in the CTRL frontal cortex, while 1075 AIS, 512 NIS and 91 PIS were observed in the AD kidney, and 1107 AIS, 2290 NIS and 92 PIS were observed in the CTRL kidney, with an always surprisingly higher number of NIS detected at the level of kidney as compared to frontal cortex.

sample	Raw reads	Cleaned reads	Total MF	MF AIS	MF PIS	MF NIS	AIS	PIS	NIS
AD_2682_FC	10488367	7093742	65597	60398	3727	1472	1045	48	105
AD_2682_K	8080106	5340255	43106	39357	2555	1194	955	47	182
AD_7466_FC	4413193	3066722	26412	24387	1385	640	906	40	83
AD_7466_K	11818784	8177208	35038	32250	1688	1100	962	40	237
AD_7660_FC	9476392	6080119	51718	47366	2950	1402	1006	51	120
AD_7660_K	7009944	4716034	34132	31076	2008	1048	920	48	138
AD_9345_FC	9185377	6104185	18013	16697	925	391	865	51	49
AD_9345_K	8258682	5723887	35596	32020	1810	685	933	50	29
CTRL_2149_FC	7411790	4692759	52648	48554	2945	1149	981	47	84
CTRL_2149_K	7904994	5344456	21961	18869	1098	1994	838	45	776
CTRL_6868_FC	7575055	4904224	48746	45147	2478	1121	1011	42	99
CTRL_6868_K	8470822	5481726	52450	48632	2611	1207	1035	41	120
CTRL_9269_FC	8296109	5717890	47049	43151	2619	1279	967	50	140
CTRL_9269_K	14005434	8719913	55092	48531	2982	3579	1010	53	1222
CTRL_929_FC	8260141	5385605	44890	41360	2286	1244	989	43	132
CTRL_929_K	7598435	4904250	20841	18905	1031	905	883	44	232

Table 3.1.1.3a Output of the bioinformatics pipeline. For each sample and tissue are reported the total number of raw reads, cleaned reads resulting from ADEPT FastUniq and Trimmomatic filtering steps, total MapFragments, and specific AIS, PIS and NIS MapFragments and MapClusters.

Almost the 92% (1063) of total AIS and 93% (117) of total PIS were detected in both tissues of the same individual, while only the 1.8% (60) of NIS was classified as germinal, suggesting the presence of a high frequency of somatic retrotranspositional events occurring in small subgroups of cells (Fig 3.1.1.3b and c).

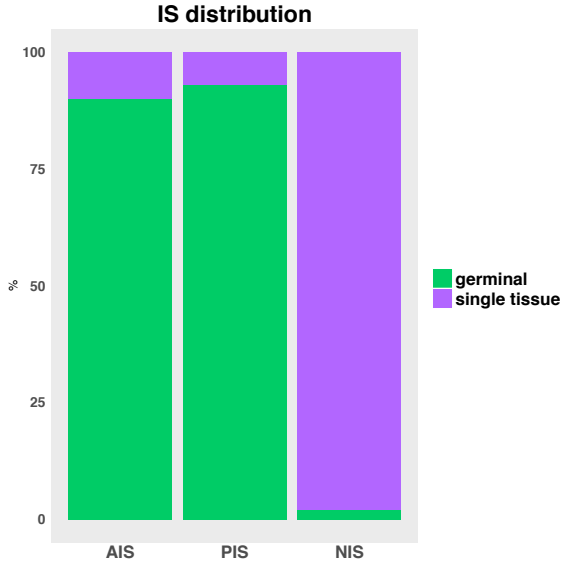


Figure 3.1.1.3b Germinal and single tissue IS. Almost the 92% (1063) of total AIS and 93% (117) of total PIS were detected in both tissues of the same individual, while only the 1.8% (60) of NIS was classified as germinal, suggesting the presence of a high frequency of somatic retrotranspositional events occurring in small subgroups of cells.

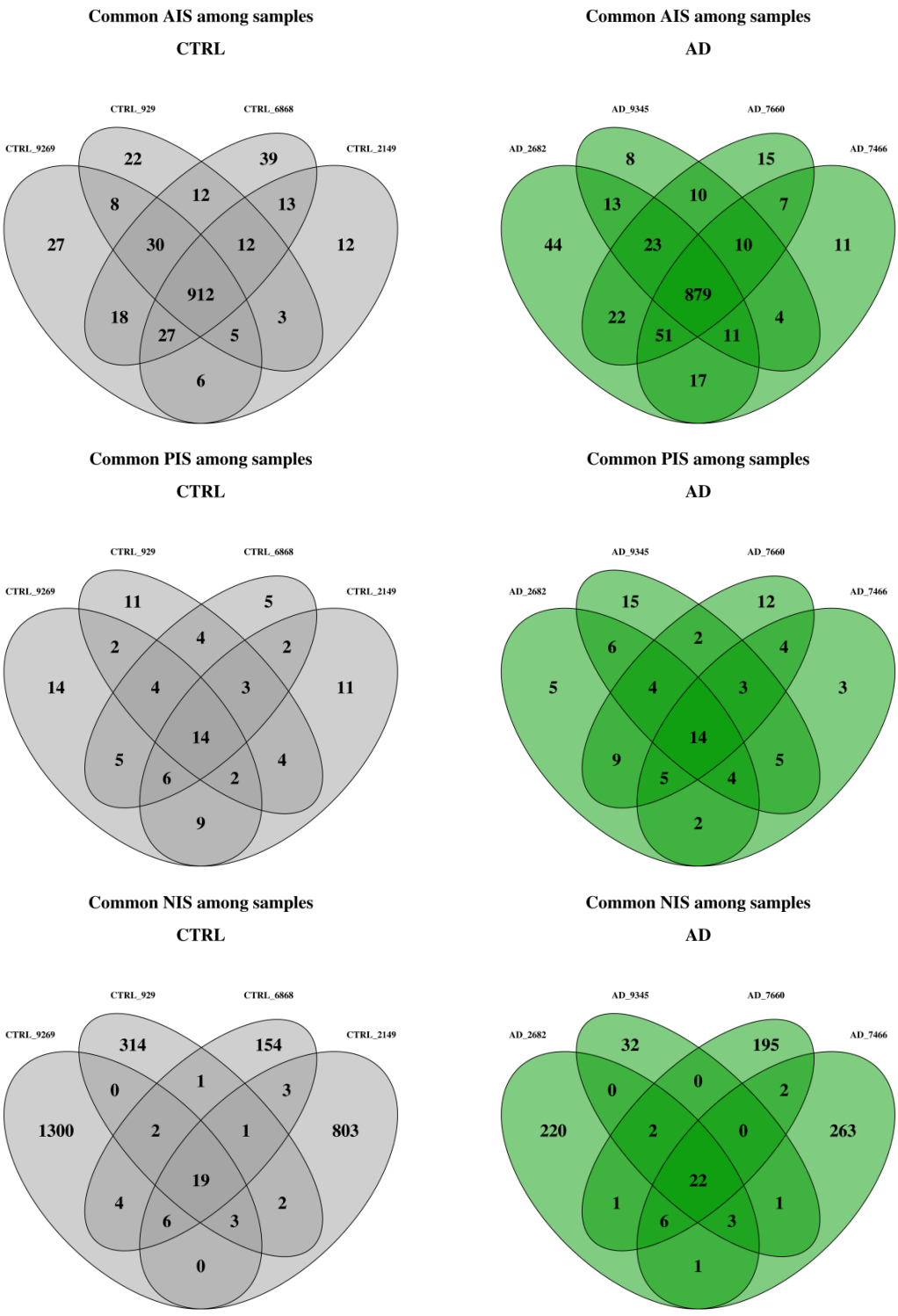


Figure 3.1.1.3c Germinal and somatic IS per condition. The Venn diagrams represent the number of Public and Private AIS, PIS and NIS between the individuals. The number of shared AIS and PIS between individuals (Public) is proportionally much higher than the number of shared NIS. As proposed by Ewing and Kazazian in 2010, non-reference insertions are also plausibly more recent insertions that are more likely to be absent from the reference due to lower allele frequency.

As shown in Figure 3.1.1.3d, we noticed that AIS tended to be defined by a higher number of MapFragments compared to NIS (on average ~38 mf/AIS depending on the sample, versus ~47 mf/PIS, ~9 mf/NIS), and this is likely due to the fact that AIS and PIS which are fixed in the human genome or deriving from early germinal retrotranspositional events, are more easily detectable by the technique. On the other hand, NIS, are probably present in one or fewer cells of the sample.

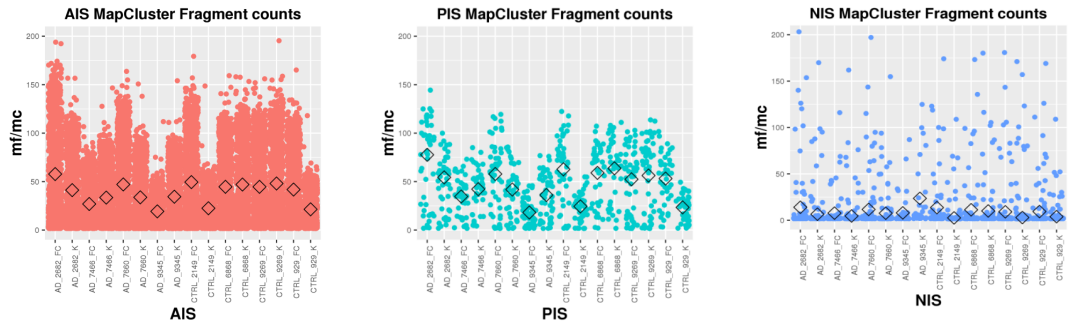


Figure 3.1.1.3d MapCluster fragment counts. The graphs report the distribution of the number of MapFragments per MapCluster for each individual. AIS and PIS are on average defined by a higher number of MapFragments compared to NIS.

It is not surprising that the majority of both AIS (~92%) and PIS (~65%) were classified as public (present in more than one individual), while public NIS represented only the 1.5% of total NIS (Fig. 3.1.1.3e).

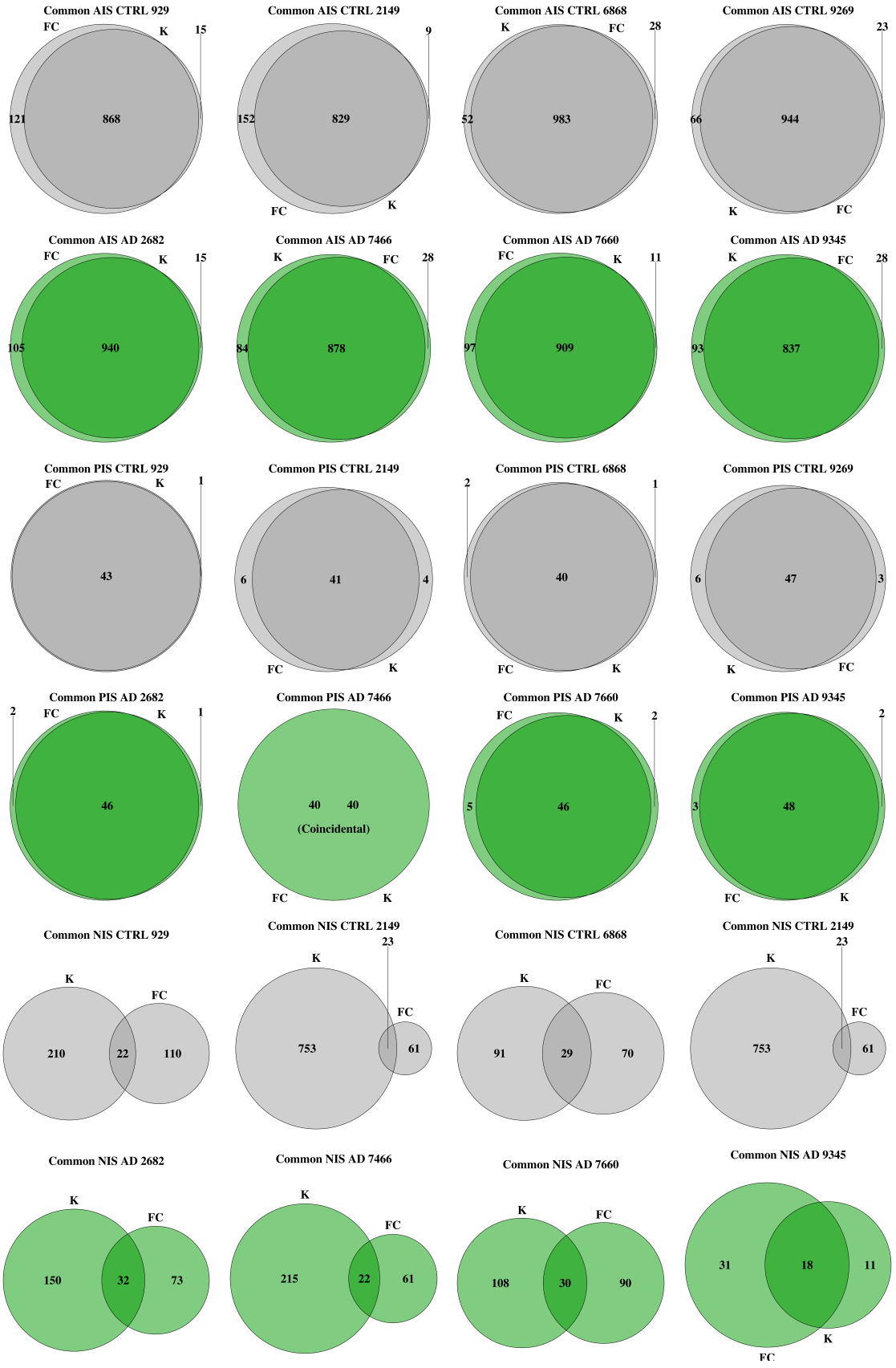


Figure 3.1.1.3e Common and private IS. The Venn diagrams represent the number of germinal and single tissue IS in frontal cortex and kidney of each individual. The number of shared AIS and PIS between the two tissues is proportionally much higher than the number of shared NIS.

3.1.1.4 SPAM reveals extensive somatic retrotransposition in the kidney

One of the most unexpected outcomes of this study is the incredibly high number of NIS detected in the kidney of AD and CTRL patients.

The total number of AIS and PIS that we detected in the kidney was comparable to the one detected in the frontal cortex, with the majority of them common between the two tissues. On the other hand, looking at the total number of NIS, it is apparent the difference between the two tissues. 613 NIS were detected in the frontal cortex and 2764 total NIS were detected in the kidney. Only a small number is common between the two tissues.

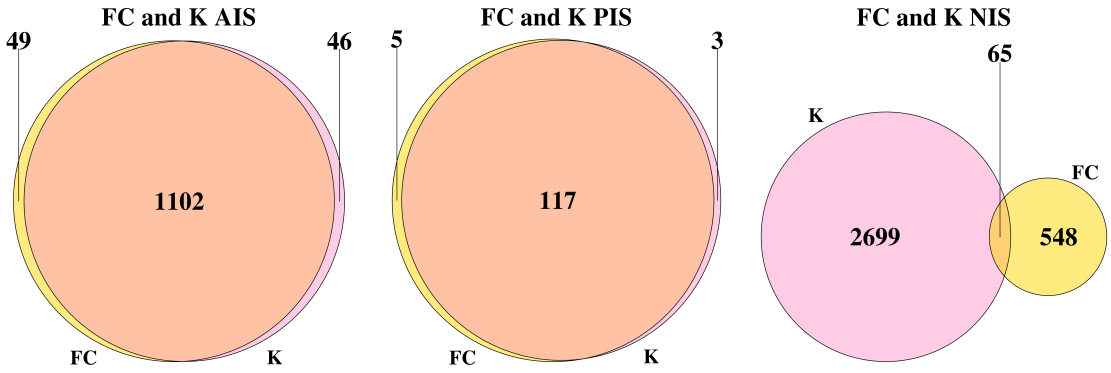


Figure 3.1.1.4 Somatic retrotransposition in the kidney. The 81.5% of non-annotated integrations detected with our technique occurred only in the kidney. The majority of AIS and PIS is common between the two tissues: kidney specific AIS and PIS represent the 4% and 2.4% respectively.

SPAM analysis, therefore, revealed the presence of an unexpectedly high level of somatic retrotransposition in the kidney, commonly considered a static organ, with very low cellular turnover capability. Only recently, Rinkevich and colleagues demonstrated that cellular precursors that work as a staminal niche are present in the mouse kidney, constantly maintaining and preserving the renal tissue throughout life. The presence of proliferating cells in the kidney might be the cause of such a high number of detected FL-LINE-1s. We remark also that both patients and CTRL individuals are elderly people, and kidney disfunction is a consequence of age. Impaired renal function, which might trigger hypertension, proinflammatory and endothelial reactions in the elderly, consequently, might promote cerebrovascular pathology and AD (Panegyres and Chen, 2013). Concerning our samples, we observed a higher level of retrotranposition in the kidney of CTRL individuals. Therefore, other age related pathologies, not directly related

to AD (e.g. diabetes or even cancer), might explain the observed LINE-1 burst in the kidney.

Clinical data or medical records of patients and controls included in the analysis might be useful to understand whether these individuals were affected by other pathologies.

3.1.1.5 FL-L1 IS genomic distribution

Retrotransposition events can potentially introduce a new LINE-1 element in every genomic position: a coding gene, a regulatory feature or a neutral region. So, we also considered the location of the IS with respect to gene annotations, and we observed that 1580 (~47.7%) NIS were in overlap with an annotated gene. Of these, 135 overlapped an exon, 1405 an intron and 40 a promoter region (Fig 3.1.1.5). Concerning PIS, only 43 (~34.4% of the total) were in overlap with an annotated gene, in particular at the level of an intron. Concerning AIS, ~33% (397) of the total AIS was gene-associated, of which 20 AIS were exonic, 367 intronic and 10 associated to a promoter.

Moreover, 785 NIS and 24 PIS were in antisense orientation in respect to the gene and 795 NIS and 19 PIS shared the same strand of the gene, while 266 AIS were in antisense orientation and 131 AIS were in sense position.

The reason why a high percentage of the AIS and PIS are in antisense orientation in respect to genes, while the percentage of sense and antisense NIS is equivalent, might be linked to a mechanism of negative selection: AIS, that are fixed in the genome have been likely subjected to an evolutionary pressure against dangerous integrations in sense orientation (Erwin et al., 2014). On the contrary, PIS, which derive from more recent integrations, and NIS that did not undergo the same degree of negative selection, display a higher amount of sense integrations.

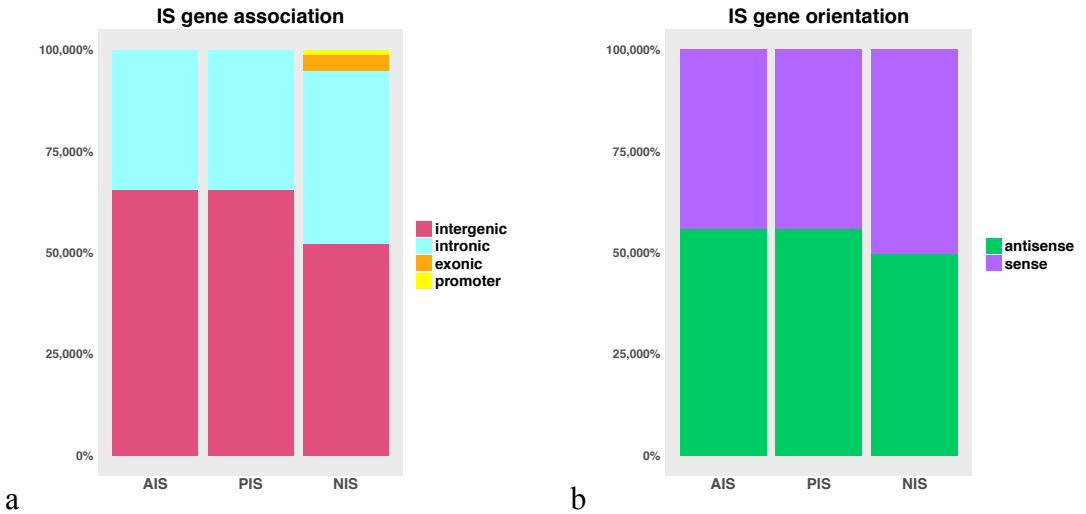


Figure 3.1.1.5 IS genomic distribution. (a) AIS, PIS and NIS distribution in the genome (intergenic, exonic, intronic, promoter) and (b) proportion of sense and antisense genic IS.

3.1.1.6 Mitochondrial IS

Besides the FL-L1 insertions falling in the genomic DNA that where included in the main analysis of this study, SPAM revealed also the presence of MapFragments falling in the mitochondrial DNA.

These mf corresponded to 45 IS: 20 IS (for a total of 61 mf) in AD affected patients and 42 IS (for a total of 144 mf) in CTRLs. 4 of these IS were exonic, 31 were located in the promoter and 10 were intergenic. Giving the high concentration of genes in the small mitochondrial (mt) genome it is not surprising to find the majority of the IS associated to a gene. Differently from the nuclear ISs, the genomic portion of different MapFragments supporting the same mt Mapcluster showed a staggered pattern at the breakpoint junction. This may suggest the presence of multiple integrations occurring in the same place in multiple mitochondria.

Surprisingly, one IS, detected in one single individual, appeared to be germinal (present in both the FC and the K of the same individual). The germinal IS was validated with two rounds of nested PCR on the gDNA. Sanger sequencing of the validated band confirmed the result. However, it remains to be assessed if we are dealing with Nuclear Mitochondrial DNA sequences (NUMTs) or LINE-1 integrations in the mitochondrial genome. Indeed, NUMT are known to be associated with transposable elements that are thought to mediate mtDNA integration in the nuclear genome (Ju et al., 2015).

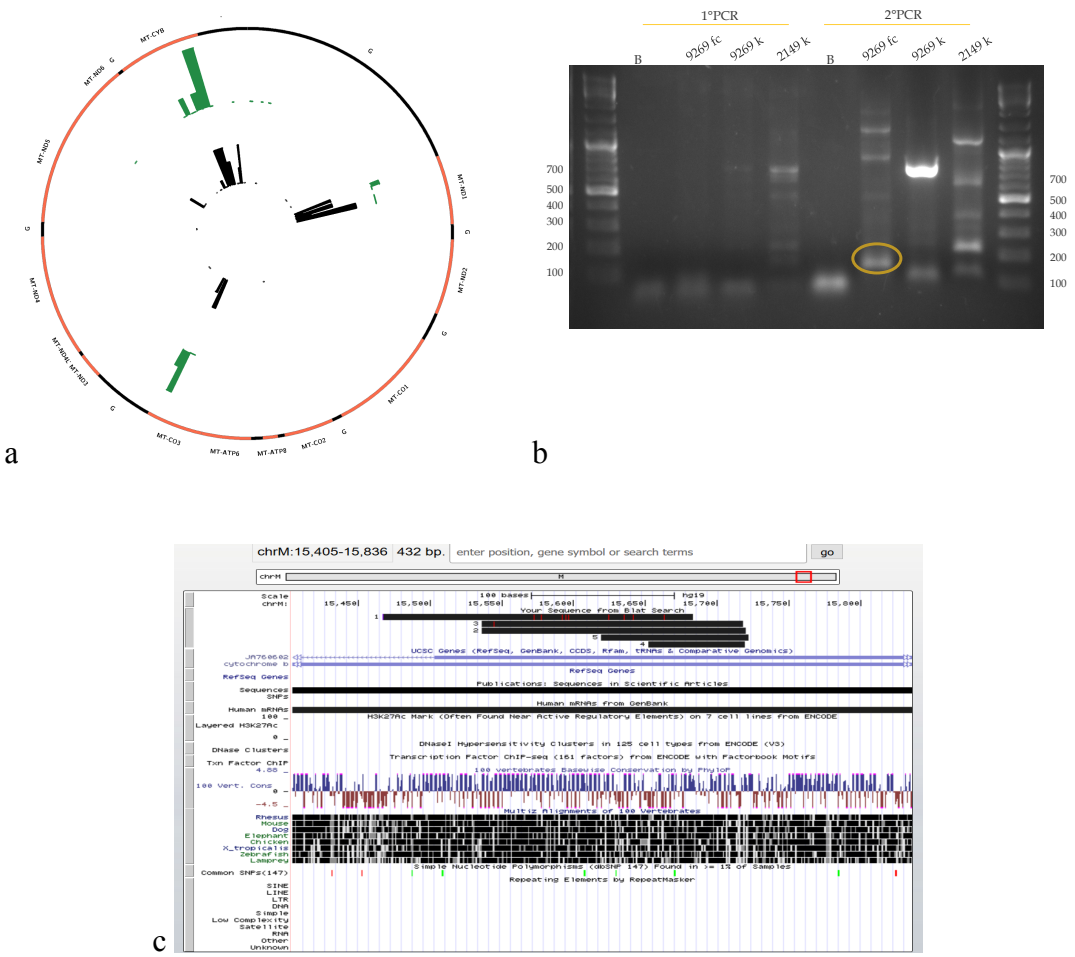


Figure 3.1.1.6 Mitochondrial IS. (a) In the Circos is represented the mitochondrial distribution of the detected IS. IS detected in the AD sample are represented in green, IS detected in the CTRL sample are represented in black. Genes are colored in orange. (b) The germinal mitochondrial IS was validated with two rounds of nested PCR on the gDNA. (c) Genome browser screenshot of MapFragments supporting the same mt MapCluster. Curiously, the MapFragments show a staggered pattern at the breakpoint junction.

3.1.1.7 FL-L1 coverage

SPAM technique was developed to target exclusively FL-L1 elements. To this aim, the primers of the enrichment PCR were designed at the very 5' of the LINE sequences (Lavie et al., 2004). The positive outcome of the enrichment is appreciable in figure 3.1.1.7. As expected, fragments that passed the filtering steps, align at 211 bp from the beginning of the putative 6000 bp FL-L1 sequence, exactly where the nested primers were designed. In particular, considering the AIS that we effectively targeted among all the targetable ones, we observed that SPAM is able to identify the 93% of the LINE-1 integration sites

detectable with our primers with 0 mismatches. Admitting an increasing number of mismatches, we observed that the percentage of the detected L1Hs decreased, while the number of detected L1PA, increased.

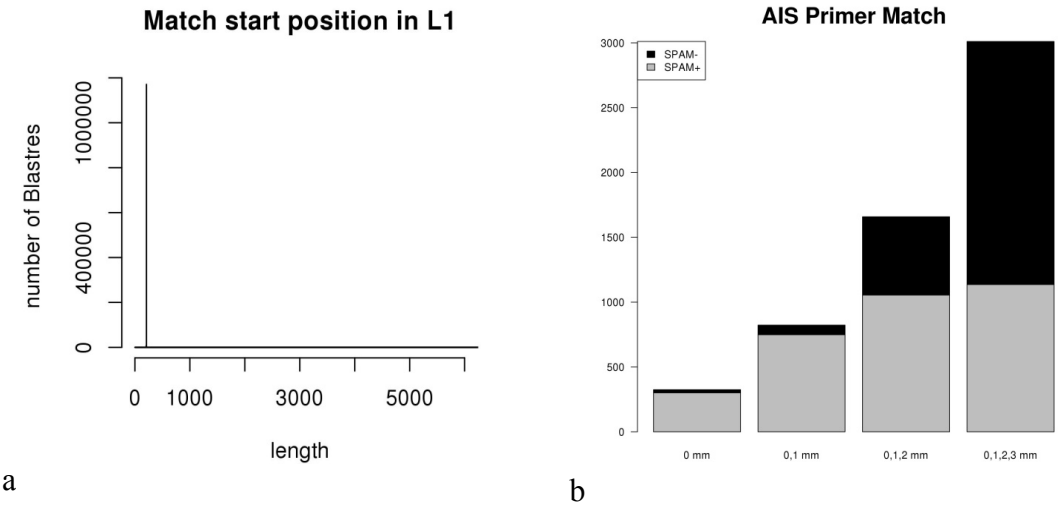


Figure 3.1.1.7 SPAM specificity. (a) The first graph reports the total number of Blastres (y axis) that map at 211 bp from the beginning of the FL-L1's 5'UTR (x axis), that is exactly the position where the SPAM nested primer was designed. (b) The second graph represents the AIS primer match, demonstrating that admitting an increasing number of SPAM primer mismatches the percentage of the SPAM detected AIS decreases. In grey (SPAM +) is reported the number of AIS effectively detected with the SPAM bioinformatics pipeline, in black (SPAM -) the number of AIS that we should target according to our primers match on the reference genome but we do not.

3.1.1.8 SPAM Efficiency

In order to make the most of our data and, at the same time, exclude from the analysis MapFragments deriving from PCR duplicates (see materials and methods) we put the minimum threshold of mf necessary to define an IS at 2 non 100% overlapping MapFragments. If we put this threshold at 3 or 5 we notice that the total number of AIS remains almost the same while we notice a very pronounced decline in the number of NIS. This suggests us that AIS and PIS are germinal events, covered by many MapFragments while NIS are likely to be somatic events, covered by a few MapFragments. This trend remains the same regardless the tissue or the condition considered.

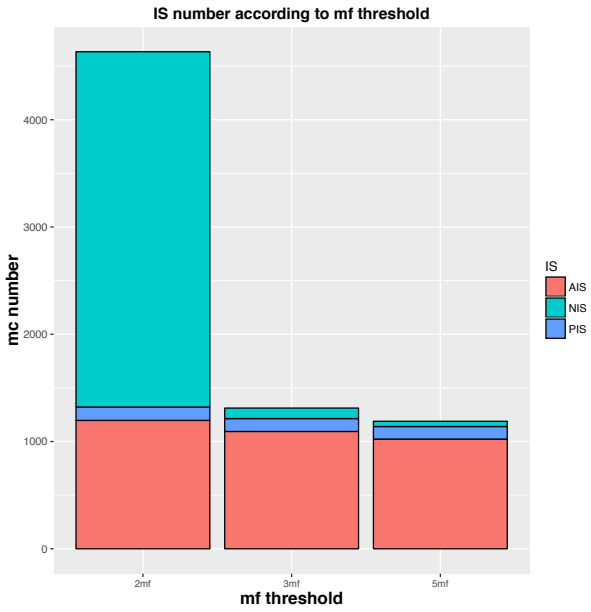


Figure 3.1.1.8a AIS and NIS number according to the minimum MapFragment threshold defining an IS. In the barplot is represented the total number of AIS, PIS, and NIS detectable admitting an increasing number of MapFragments to define a MapCluster. The number of AIS and PIS remains almost the same regardless the MapFragment threshold, while the number of NIS dramatically decreases.

Then, a rarefaction analysis was performed in order to investigate SPAM ability in detecting new LINE-1 insertions with these experimental settings, in particular with this sequencing depth. By plotting the average number of different new IS computed sampling an increasing number of reads, the detection of AIS and PIS almost reached the saturation, while the detection of NIS did not, suggesting a high number of different somatic insertions present in the cells analyzed and a lack of saturation at our sequencing depth (Fig. 3.1.1.9b), most of them undetected at the depth used in our sequencing. This is particularly true for two samples, both K CTRL (samples 9269 and 2149). We cannot exclude that the high LINE-1 content in the kidney of these 2 CTRLs may be caused by another pathology (e.g. cancer) affecting the tissue.

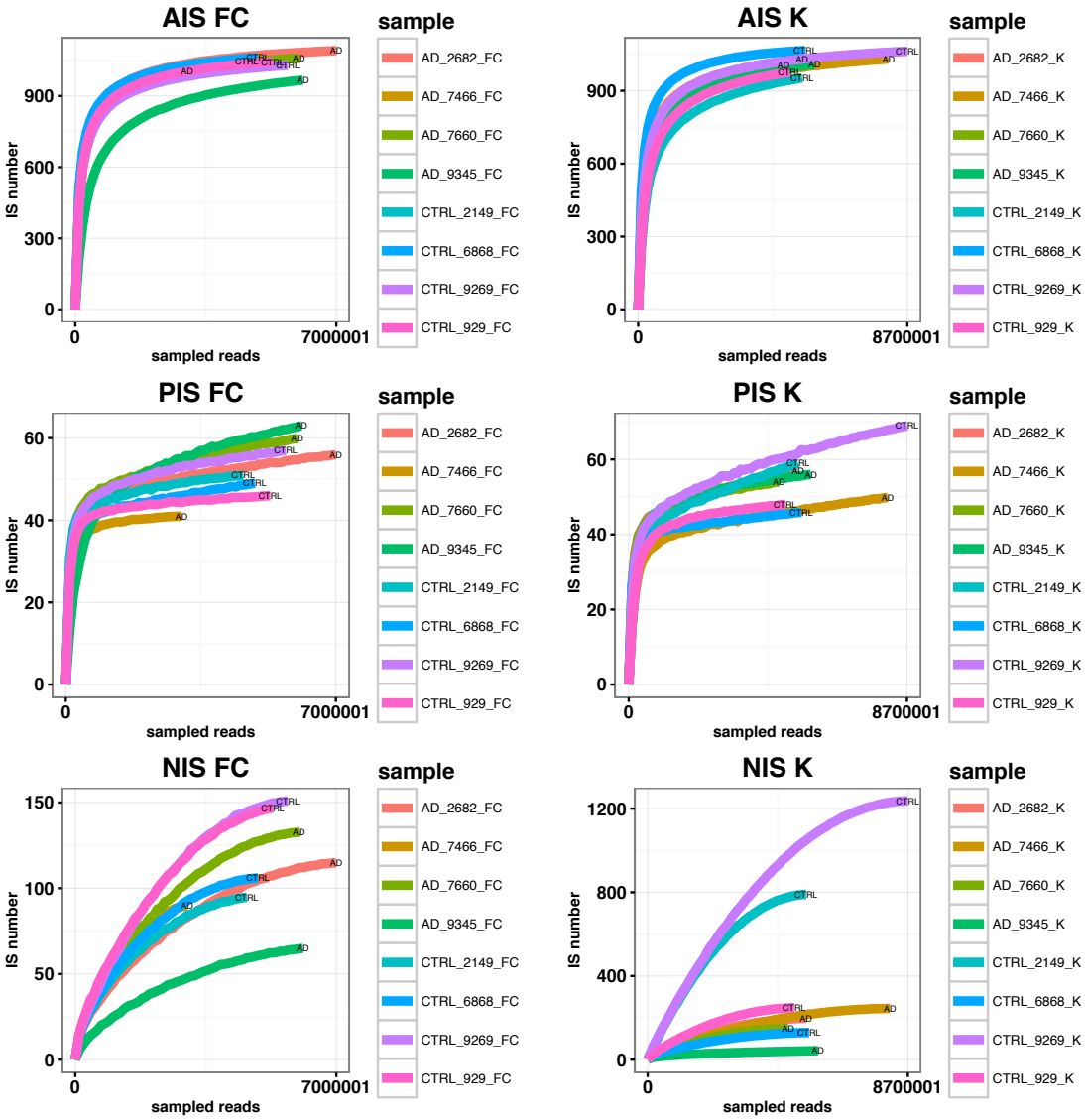


Figure 3.1.1.8b Rarefaction plots. By plotting the average number of different new IS computed sampling an increasing number of MapFragments, the detection of AIS almost reached the saturation, while the detection of NIS did not, suggesting the presence of a high number of different somatic insertions present in the DNA samples.

3.1.1.9 Technical Validation by PCR of IS detected by SPAM

In order to estimate SPAM capacity in detecting FL-L1 elements we tried to validate AIS, PIS, and NIS. For each of these IS we performed a first round of PCR and a nested PCR using two couples of primers with the forward primers designed on the genomic DNA at the insertion site, and the reverse primers designed against the very beginning of the

LINE-1 5'UTR. Primary PCRs were performed on genomic DNA and/or SPAM product. Digital Droplet PCR (ddPCR) was also employed in order to increase the chances to detect very rare events (e.g. somatic NIS covered by 2 or a few MapFragments). For the droplet PCR we selected ISs covered by MapFragments containing the full probe and reverse primer designed on the 5'UTR of the LINE-1 element (without mismatches) by White and colleagues in 2014. Overall, we were able to validate only germinal IS covered by two or more MapFragments.

IS SPAM	Tissue SPAM	MapFragments min coverage per sample	Closest gene	Genomic location	gDNA PCR	SPAM product PCR	ddPCR	Validated in	Zygosity
AIS	FC, K	>=41	RB1	Intronic	Y	Y	N	FC, K	
AIS	FC, K	>=61	PRKG1	Intronic	Y	Y	N	FC, K	
AIS	FC, K	>=2	ROBO2	Intergenic	Y	N	Y	FC, K	Homozygous
AIS	FC, K	>=3	KIRREL3	Intergenic	Y	Y	N	FC, K	
AIS	FC, K	>=2	COL11A1	Intronic	Y	Y	N	FC, K	
PIS	FC, K	>=33	HYDIN	Intronic	Y	N	N	FC, K	
PIS	FC, K	>=30	ERC2	Intronic	Y	N	N	FC, K	
PIS	FC, K	>=35	MED12L	Intronic	Y	N	N	FC, K	
PIS	FC, K	>=15	COMMD10	Intronic	Y	N	Y	FC, K	
NIS	FC, K	>=2	KCND3	Intergenic	Y	Y	N	FC, K	Homozygous
NIS	FC, K	>=57	CABS1	Intergenic	Y	N	Y	FC, K	Heterozygous
NIS	FC, K	>=58	FAM98A	Intergenic	Y	N	Y	FC, K	Heterozygous
NIS	K	>=2	RARLYL	Intronic	Y	N	Y	FC, K	Heterozygous
NIS	FC, K	>=23	CCNA1	Intergenic	Y	N	Y	FC, K	Heterozygous
NIS	K	>=2	RYR3	Intronic	Y	N	Y	FC, K	Heterozygous
NIS	K	>=2	DIS3L2	Intronic	Y	N	Y	FC,K	Homozygous
NIS	FC	>=2	FRG2C	Intergenic	Y	N	Y	FC,K	Homozygous

Table 3.1.1.9 Technical Validation by PCR of IS detected by SPAM: In this table are reported the main features of the IS validated by PCR. In column *IS SPAM* we define the type of IS; in *Tissue SPAM* we report the tissue in which the IS was detected; in *MapFragments min coverage per sample* we report the minimum coverage in terms of MapFragments of that IS among different samples; in *Closest gene* we report the name of the first IS closest gene; in *Genomic location* we report the position of the IS with respect to its closest gene; in *gDNA PCR, SPAM product PCR* and *ddPCR* we report the validation strategy; in *Validated in* we report the tissues in which the IS was validated regardless the SPAM prediction; in *Zygosity* we report the condition of the ISs validated with ddPCR.

3.1.1.10 Chromatin accessibility

Chromatin state of the master LINE-1 elements and chromatin states of the target regions are important factors influencing LINE-1 mobilization.

Chromatin state varies across the genome creating fluctuations in DNA fragility. From a biological point of view, active, fragile chromatin coincides with nuclease hypersensitivity, and facilitates LINE-1 integration, whereas condensed heterochromatin hinders it (Singer et al., 2010). From a technical point of view, the accessibility of chromatin could be an important determinant of experimental bias. Heterochromatic regions may be more resistant to shearing by sonication than euchromatic regions

(Teytelman et al., 2009), resulting in a loss of information and an underestimate of LINE-1 elements located in these inaccessible regions.

In order to ascertain the role of chromatin state during integration and evaluate if PCR amplification was favored within open chromatin regions we compared the average distance of AIS, AIS targeted admitting 2 mm in the primers, AIS that we should have targeted with 2 mm but we did not, NIS and PIS from open chromatin regions of the human genome with that of random IS. Purifying selection may also explain IS distribution with respect to open chromatin regions (DNAse I accessible sites). Interestingly, all classes of IS, excluding PIS, appeared to be closer to DNAse I hypersensitive regions (Student's t-Test pvalue 2e-16) with respect to random IS (Fig 3.1.1.10). Since the effectively detected AIS were not closer to open chromatin regions than the AIS that we did not target, we can exclude that this was dependent on a technical issue. Interestingly, PIS appeared to be more distant from DNAse I hypersensitive regions than AIS and NIS, and therefore probably less prone to expression and in case retrotransposition. This may suggest that a closed chromatin material might be ideal for the host genome to coopt new genetic material.

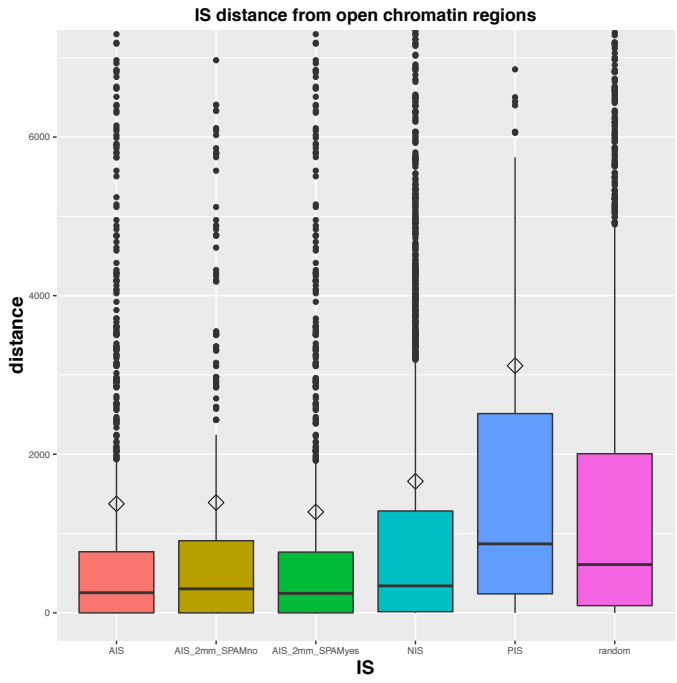


Figure 3.1.1.10 Chromatin accessibility. By comparing the average distance of AIS (red), AIS that we should have targeted with 2 mm but we did not (ocra), AIS targeted admitting 2 mm in the primers (green), NIS (light blue) and PIS (blue) from open chromatin regions of the human genome with that of random IS (pink), we observed that all classes of IS appeared to be significantly closer to open chromatin regions than random IS. PIS appeared to be more distant from DNAse I hypersensitive regions than AIS and NIS, meaning that probably they are less prone to expression and in case retrotransposition.

3.1.1.11 LINE-1 differential integration in AD and CTRL samples

In order to understand whether IS can present a different pattern of distribution in AD samples and controls, we tried to find genomic locations differentially targeted by LINE-1 insertions (AIS, NIS and PIS) in frontal cortex, kidney, and in the two conditions, AD and CTRLs, by performing a differential integration analysis. Statistically significant differences were detected in differential coverage of MapFragments per specific MapClusters and MapFragments per Gene.

Considering an FDR-adjusted p-value <0.1, 18 AIS, 42 PIS could be considered significant in the comparison between AD and CTRLs (frontal cortex and kidney together). No significant NIS were observed but we cannot exclude that this result could have been determined by the lack of saturation.

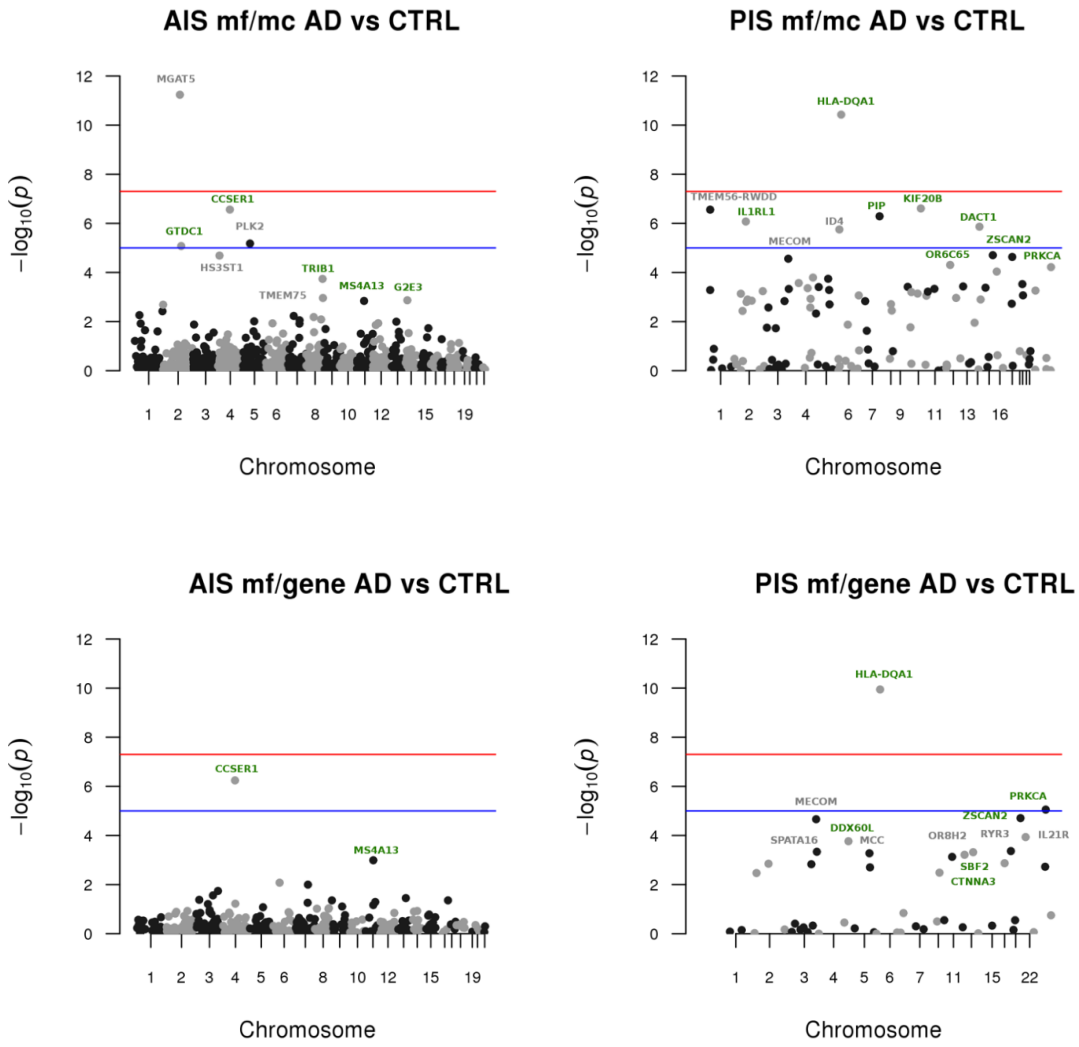


Figure 3.1.1.11 LINE-1 differential integration in the genome. The Manhattan plots report IS and genes showing differences in the number of associated MapFragments in the comparison AD vs CTRL. The AIS and PIS having a significant differential MapFragments association (FDR-adjusted p-value <0.1) were chosen for validation by PCR. On the x axis are reported the chromosomes while on the y is reported the –log10 of the pvalue (not adjusted). The blue threshold indicates a pvalue of –log10(1e-5) while the red threshold indicates a pvalue of –log10(5e-8).

For both AIS and PIS we decided to validate by PCR the IS with the lowest FDR-adjusted p-value associated to the following genes: MGAT5, PLK2/SNK, CCSER1 and HS3ST1 for the AIS category, and TMEM56-RWDD, ID4, KIF20B, PIP, IL1RL1, DACT1 and HLA-DQA1 for PIS to confirm the confidence of the computational predictions. For all these IS, an assay comprising a FW primer complementary to the upstream genomic sequence, a REV primer on the 5'UTR LINE-1 sequence and a REV primer on the downstream genomic sequence were designed, in order to be able to amplify from gDNA, in the same PCR reaction, part of the allele containing the LINE-1 5'UTR and/or the

allele without the insertion. All the gDNA samples used in these analyses derived from individuals with an age ≥ 70 years. We report our results for each gene below.

Concerning AIS:

- the IS associated to **MGAT5** (*N*-acetylglucosaminyltransferase V), a glycosyltransferase involved in cellular survival and migration (de Freitas Junior and Morgado-Díaz, 2016) was amplified in all samples analyzed with SPAM (AD and CTRLs) apparently in homozygosis, meaning that the differential integration previously observed, was determined by a failure of SPAM in detecting the AIS in all samples;
- the IS associated to **PLK2/SNK** (Polo-like kinase 2/serum-inducible kinase), a regulator of synaptic plasticity, recently demonstrated to be involved in Aβ production (Lee et al., 2017) was detected by SPAM in two out of four SPAM CTRLs and none AD samples. Since the bionformatic result was confirmed by PCR, we decided to extend the PCR validation to a larger sample. We processed a total of 187 AD and 25 CTRL samples: the allelic frequencies that we obtained were 0.08 and 0.06 respectively, meaning that no significant difference between AD and CTRL could be detected.
- The AIS associated to **CCSER1** (Coiled-Coil Serine Rich Protein 1), a gene involved in alpha-synuclein gene triplications associated with Parkinson's Disease (Olgiati et al., 2015), was detected by SPAM in two out of four AD samples and none CTRL. In this case, we processed a total of 100 AD and 65 CTRL samples: the allelic frequencies that we obtained were 0.12 for both groups.

Concerning PIS with significantly higher coverage in CTRLs as compared to AD samples:

- The IS associated to **TMEM-RWDD** gene (transmembrane protein 56- RWD domain containing), was searched on a total of 100 AD and 25 CTRL samples: the allelic frequencies that we obtained were 0.135 and 0.180 respectively.
- A total of 28 AD and 25 CTRL samples were analyzed for the IS associated to **ID4** (Inhibitor of DNA Binding 4), a transcription factor expressed during neurogenesis in

CNS and peripheral nervous system. The allelic frequencies that we obtained were 0.036 and 0.040 respectively.

Concerning PIS with significantly higher coverage in AD samples as compared to CTRLs:

- **KIF20B** (Kinesin Family Member 20B) is a kinesin involved in the morphogenesis of the cortical pyramidal neurons (McNeely et al., 2017). The IS associated to the gene was screened in 99 AD and 25 CTRL samples and the allelic frequencies that we obtained were 0.076 and 0.060 respectively.
- **PIP** (Prolactin Induced Protein) is a protein involved in the reproductive and immunological systems (Hassan et al., 2009). The IS associated to the gene was screened in 100 AD samples and 25 CTRLs and the allelic frequency that we obtained for both the groups was 0.1.
- **IL1RL1** (Interleukin 1 Receptor Like 1) gene encodes for a receptor (ST2) that interacts with IL-33, contributing to the regulation of immune responses and tissue homeostasis (Balato et al., 2016). The IS associated to this gene was screened on 96 AD samples and 61 CTRLs: the allelic frequencies that we obtained were 0.078 and 0.041 respectively.
- The gene **DACT1** (Dishevelled Binding Antagonist Of Beta Catenin 1) encodes for a protein with an important role as intracellular signaling regulator, whose mutations have been demonstrated to be risk factors for human neural tube defects (Shi et al., 2012). We validated the presence of a DACT1-associated PIS on 28 AD samples and 25 CTRLs and the allelic frequencies were 0.161 and 0.100 respectively.
- The PIS showing the most significant result (since present in 3 out of 4 AD samples and none of the CTRLs) was located in the intergenic region between the **HLA-DRB1** and the HLA-DQA1 genes, inside the MHC class II locus, the most variable region in the human genome. This PIS (IS-HLA) corresponds to a known FL-L1 polymorphism of the human population and reported to be present in the so-called MANN and DBB MHC haplotypes (Horton et al., 2008). The analysis, performed on 410 AD samples and 239 CTRLs, actually did not show a different incidence of the IS-HLA, whose allelic frequency in AD samples was 0.138 and in CTRL samples 0.126.

3.1.1.12 Gene ontology enrichment analysis

To assess the implications of FL-L1 IS on gene function we examined the set of NIS and PIS associated genes with respect to gene ontology (GO) functional category classifications. To avoid biased enrichments associated to longer genomic loci (typical of neuronal genes) we used the GOseq package (see materials and methods for details) which specifically avoids this bias. To analyze enrichments at the tissue level, the set of genes associated to the NIS from FC and K were compared to the set of genes associated to all the identified NIS from the whole experiment (FC + K). In addition, to analyze enrichments at the disease level, the set of genes associated to the NIS from FC or K of AD patients were compared to the set of genes associated to all the NIS identified in the specific tissue from the whole experiment. The strategy of using the total sets of identified NIS (from whole experiment or from each specific tissue) as a reference in the comparison further ensures us that our results will not be affected by any length bias.
Of the different comparisons performed only two gave significant results suggesting intriguing functional outcomes of retrotransposition events and their involvement in AD.

At the tissue level, the genes associated to NIS in FC with respect to the genes associated to all the identified NIS (FC + K) are significantly enriched for biological processes such as cell junction organization, neuron differentiation; molecular functions associated to protein phosphatase activity; cellular components such as cell projection and plasma membrane. This result suggests that retrotransposition effectively tags genes associated to neural differentiation, learning and memory (Temtamy et al., 2008).

Category	over_represented_pvalue	under_represented_pvalue	numDEInCat	numInCat	Term	Ontology	fdr_adj_pval
GO:0004721	0.00013446827445671	0.999979680582298	12	20	phosphoprotein phosphatase activity	MF	0.00726128682066236
GO:0034332	9.77E+09	0.999987988825411	11	17	adherens junction organization	BP	0.023777240519704
GO:0045216	0.00011268834369528	0.999977652275301	15	28	cell-cell junction organization	BP	0.023777240519704
GO:0045665	0.000324087877142602	0.999956409156197	10	16	negative regulation of neuron differentiation	BP	0.045588361384726
GO:0071944	0.00172125316722288	0.99889171428485	124	490	cell periphery	CC	0.0505682488727613
GO:0005886	0.00192470890443242	0.998757147823343	122	482	plasma membrane	CC	0.0505682488727613
GO:0042995	0.00223095215615123	0.998692401295401	66	235	cell projection	CC	0.0505682488727613
GO:0016791	0.00280254541247162	0.999221940658426	13	29	phosphatase activity	MF	0.0756687261367336
GO:0034330	0.000749440714955934	0.999806231687736	15	32	cell junction organization	BP	0.0790659954278511
GO:0042578	0.00454860500920058	0.998350887304303	17	44	phosphoric ester hydrolase activity	MF	0.0818748901656104

Table 3.1.1.12a GO analysis of NIS associated genes in all FC. In the table are reported the GO terms significantly enriched (FDR adj pvalue < 0.1) in the set of genes associated to FC (distance <= 10000 bp) with respect to FC + K NIS associated genes (distance <= 10000 bp). The analysis was performed with GOseq package (Young et.al 2010), which returns a data frame with several columns. The first column gives the name of the GO category, the second gives the p-value for the associated category being over represented amongst test genes. The third column gives the p-value for the associated category being under represented amongst test genes. The p-values have not been corrected for multiple hypothesis testing. The fourth and fifth columns give the number of test genes in the category and total genes in the category respectively. Then are reported the GO term, its ontology (MF, BP or CC) and the FDR adjusted over represented pvalue.

At the disease level, the genes associated to NIS in FC from AD patients with respect to the genes associated to NIS in FC from all samples (AD + CTRL) result significantly enriched for functions related to signal transduction and receptor activity. This result suggests that in the FC of AD patients, loci associated to signal transduction pathways are more targeted by retrotransposition events. Indeed, it was demonstrated that these genes present reduced expression in FC of elderly people (Lu et al., 2004). Finding these pathways enriched in the FC of the patients may be suggestive of a potential correlation between aging, LINE-1 mobilization and progression of the disease.
The same analysis executed on the PIS did not result in any significant enrichment.

Category	over_represented_pvalue	under_represented_pvalue	numDEInCat	numInCat	Term	Ontology	fdr_adj_pval
GO:0038023	0.000291049751968502	0.999950504227471	21	26	signaling receptor activity	MF	0.00582099503937004
GO:0004871	0.000763759977069175	0.999798027819755	27	37	signal transducer activity	MF	0.00654249747810952
GO:0004888	0.000981374621716427	0.999814490086967	19	24	transmembrane signaling receptor activity	MF	0.00654249747810952
GO:0001071	0.0171114882315186	0.99665451350175	11	14	nucleic acid binding transcription factor activity	MF	0.0623013717457329
GO:0003700	0.0171114882315186	0.99665451350175	11	14	transcription factor activity, sequence-specific DNA binding	MF	0.0623013717457329
GO:0004872	0.0217314965995516	0.991810135519482	21	32	receptor activity	MF	0.0623013717457329
GO:0060089	0.0218054801110065	0.990644924526187	27	43	molecular transducer activity	MF	0.0623013717457329

Table 3.1.1.12b GO analysis of NIS associated genes in AD FC samples. In the table are reported the GO terms significantly enriched (FDR adj pvalue < 0.1) in the set of genes associated to AD FC NIS (distance <= 10000 bp) with respect to FC NIS associated genes (distance <= 10000 bp).

The GO analysis performed with the tool GREAT on the same comparisons showed a unique but interesting significant result. Panther pathway term: Alzheimer's disease presenilin pathway emerged in the comparison NIS AD K against AD + CTRL K.

3.1.1.13 AIS and PIS influence on gene expression

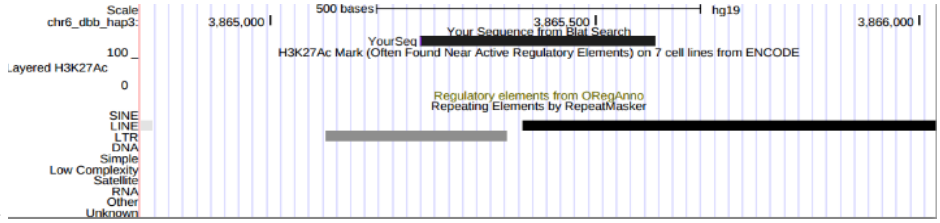
In order to evaluate whether the differentially integrated FL-L1 elements could have an impact on the expression of nearby genes, we compared the genomic coordinates of the 18 AIS and 42 PIS with an FDR corrected p-value <0.1, with those of known genomic structural variations (SVs), and screened the expression level of the closest coding genes in 445 individuals harboring or not the SV.

As expected, no AIS matched with SVs, while one AIS was found in overlap with a 9.5 Kb deletion (hg19 coordinates: chr2:4,777,625-4,787,178), corresponding to CTCF binding site. Since only one individual carrying this SV was found, no further analyses on the impact of the LINE-1 insertion on gene expression could be performed.

On the other hand, 28 PIS matched with known SVs. For most of them no estimations on the functional impact could be made, because of the unavailability of complete expression data, or rare incidence of the SV, or because the analysis did not reach the statistical significance. The latter was mostly due to the high variance in expression levels of the same genes in different individuals that were randomly selected from different populations.

However, we identified a PIS able to negatively influence the expression of its 4 closest genes and of a non-coding expressed anti-sense RNA (Fig 3.1.1.13d). This FL-L1 element is located in the chromosome 6 HLA locus (hg19 coordinates: chr6:32589880-32589881), 315 bp downstream the IS-HLA, and results to be present in 5.6% of the overall analyzed population (25 on 445 individuals) and 4.8% of all individuals included in the 1000 GP (122 out of 2504 total individuals). Since the expression of each exon of the four nearby genes is significantly downregulated, we can infer that the IS has an effect on the overall gene expression. The same conclusion was taken for the HLA-IS (detected in 70 out of 445 individuals) described in section 3.1.1.11. Interestingly, the insertion in the HLA locus disrupts an LTR sequence which was previously demonstrated to be a transcribed enhancer region (Thurman et al., 2012, Andersson et al., 2014) with a possible role as gene expression modulator, at least in immune cells. Transcription factor binding sites for TBP, TAF1 and Pol II are present in correspondence of the IS, suggesting a possible regulatory role of the element (Figure 3.1.1.13d).

a



b

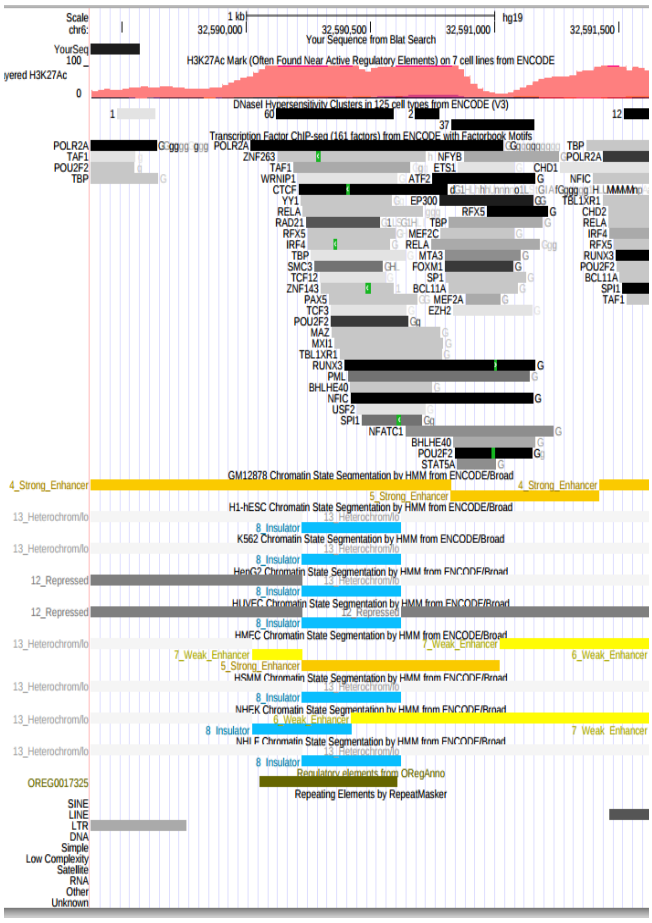
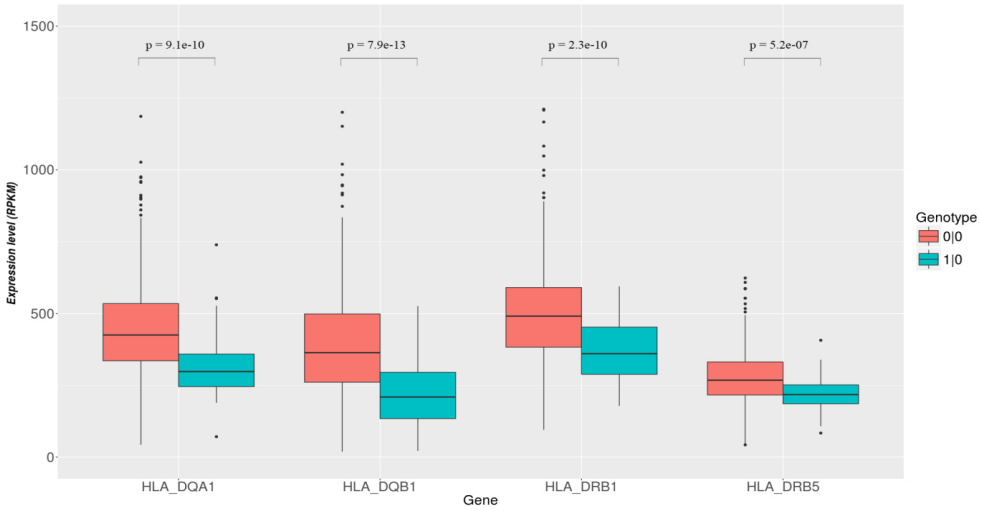


Table 10 Major indels in the form of retrotransposible elements

Chr6 pos'n	Flanking loci	Presence in haplotype								Details
		PGF	COX	QBL	SSTO	APD	DBB	MANN	MCF	
29002370	TRIM27:C6orf100	C	C	C	C	?	?	C	C	**Complex region (A)**
29440424	OR5V1:OR12D3	✓	✓	?	✓	?	?	X	X	AluYa5
29784097	C6orf40:HCP5P15	✓	X	✓	✓	?	X	X	X	AluYa5/8 175..304
29788451	Within HCP5P15	X	X	✓	X	?	✓	✓	X	AluYa5/8 176..310
29794763	HCP5P15:HLA-F	✓	X	X	✓	?	X	X	X	SVA_E plus simple rpt.s
29922942	HLA-G:MICF	✓	X	✓	✓	✓	✓	✓	✓	L1ME3B 5940..6165
29954495	MICF:HLA-H	✓	X	X	X	X	X	X	X	HERVK9 inserted in MEF
30008633	HLA-K:HLA-21	✓	X	X	✓	X	X	✓	?	SVA E/F plus simple rpt.
30106475	HCG8:ETF1P1	X	✓	X	X	✓	✓	X	X	AluYb8
30547387	SUCLA2P:RANP1	X	X	X	✓	?	X	X	?	AluJb 1..283 and parts of MLT1D/L1PBa
31079582	C6orf205:HCG22	X	X	✓	X	X	X	?	X	AluYb8 37..297
31117638	C6orf205:HCG22	✓	X	X	✓	✓	X	✓	✓	AluY (whole & part) and MER63 1017..1062
31301931	HCG27:HLA-C	✓	✓	X	✓	?	✓	✓	✓	HERV3 part (6489...7339)
31320352	HCG27:HLA-C	✓	X	X	X	?	X	X	X	SVA_F 349..850 plus GC rich rpt.
31358220	RPL3P2:WASF5P	X	X	✓	X	?	X	X	X	AluY 35..306
31400900	WASF5P:HLA-B	✓	✓	✓	✓	?	X	X	X	AluSp plus L1PREC2 part (3205...4617)
31405648	WASF5P:HLA-B	✓	X	✓	✓	?	X	x	x	HERVIP10F (part) and AluSg (only cf CX DB)
31418854	WASF5P:HLA-B	✓	✓	✓	✓	?	✓	✓	X	L1PA5 part (5503..5876)
31530995	MICA:HCP5	✓	X	?	✓	?	?	X	?	SVA B/F plus simple rpt.s
32421915	within C6orf10	✓	X	X	✓	X	X	✓	X	AluYb8
32486228	BTNL2:HLA-DRA	✓	✓	✓	✓	✓	X	X	X	L1P1/L1HS parts
32655545	HLA-DRB1 intron 5	✓	x	x	X	?	✓	✓	?	AluYa5 within more or less partial LTR12
32660731	HLA-DRB1 intron 1	X/X	✓/X	X/X	✓/✓	?	✓/✓	✓/✓	?	Tigger4/AluSx
32661119	HLA-DRB1 intron 1	C	C	C	C	?	C	C	?	**Complex region (B)**
32663167	HLA-DRB1 intron 1	X/✓	✓/✓	✓/✓	✓/X	?	✓/X	✓/X	?	AluSq/AluY
32669534	HLA-DRB1:HLA-DQA1	C	C	C	C	?	C	C	?	**Complex region (C)**
32679461	HLA-DRB1:HLA-DQA1	✓	X	X	X	?	X	X	?	AluY
32693271	HLA-DRB1:HLA-DQA1	✓	✓	✓	✓	?	X	✓	?	L1PA4 (parts)
32697545	HLA-DRB1:HLA-DQA1	X	X	X	X	?	✓	✓	?	L1HS 7..6032
32701428	HLA-DRB1:HLA-DQA1	✓	X	✓	✓	?	x	X	x	L1PA2 part and from CX: MER2B and AluY
32728179	HLA-DQA1: HLA-DQB1	C	C	C	C	?	C	C	C	**Complex region (D)**
32739664	within HLA-DQB1	X	X	✓	X	?	X	✓	X	AluY
32743646	HLA-DQB1: MTCO3P1	X	X	X	X	?	✓	X	X	LTR13
32746780	HLA-DQB1: MTCO3P1	X	X	X	X	?	✓	X	✓	L1PA4 (parts)

c



d

Figure 3.1.1.13 HLA locus. (a,b) Genome browser screenshot of the PIS located in the intergenic region between the HLA-DRB1 and the HLA-DQA1 genes. The L1HS insertion in the HLA locus disrupts an LTR sequence which was previously demonstrated to be a transcribed enhancer region. **(c)** This PIS (IS-HLA) corresponds to a known FL-L1 polymorphism reported to be present in the so-called MANN and DBB MHC haplotypes (Horton et al., 2008). **(d)** The PIS in the HLA locus significantly downregulates the expression of its nearby genes.

3.1.2 Quantification of FL-L1 insertions in different tissues

3.1.2.1 FL-L1 CNV analysis in AD post-mortem brains.

In order to study the copy number variation of potentially FL-L1 elements in the human genome we decided to adopt the qPCR technique with Taqman probes, as previously performed by Coufal and colleagues for the total LINE-1 elements (full-length and truncated).

In particular, we designed a new Taqman assay on the 5'UTR sequence of a canonical L1Hs element in order to be sure to amplify only the complete full-length sequences.

The primers and probe were blasted against the Human full-length, intact LINE-1 Element (Ens84.38) to identify which elements were detected by the assay designed. Of the 146 elements contained in FL-L1, 114 are recognized by the assay (Penzkofer et al., 2017). The analysis was performed on the gDNA extracted from frontal cortex, temporal cortex, cerebellum, hippocampus, and kidney of 9 healthy individuals and 8 AD patients, belonging to the same Brazilian cohort. From the neuropathological point of view, the controls were clinically healthy individuals, at Braak stages 0-III, while AD patients were demented individuals at Braak stages III-VI.

By performing the qPCR for the FL-L1 element we observed the presence of a lower amount of FL-L1s in the AD frontal cortex tissue as compared to controls, while no differences could be observed in the temporal cortex, nor in the hippocampus, which are both heavily involved and damaged by the disease (Fig 3.1.2.1). An unexpected finding was the lower copy number in the cerebellum and kidney of AD patients. Indeed, cerebellum is a brain structure that is not primarily involved in AD as well as the kidney, but apparently not stable from the retrotransposon-mobilization point of view. The same analysis was performed on a Spanish cohort composed of healthy controls, patients at Braak stages I-II (early AD), and patients at the final AD Braak stages V-VI (late AD). In this case, we observed a progressive and always significant decrease in the content of LINE-1 full-length sequences starting from the healthy controls group to the group of patients affected by late AD. In particular we observed a decrease of ~14% between controls and early AD patients, ~20% between controls and late AD patients, and finally

~7% between early and late AD patients, confirming the trend previously observed in the Brazilian cohort (Fig. 3.1.2.1b).

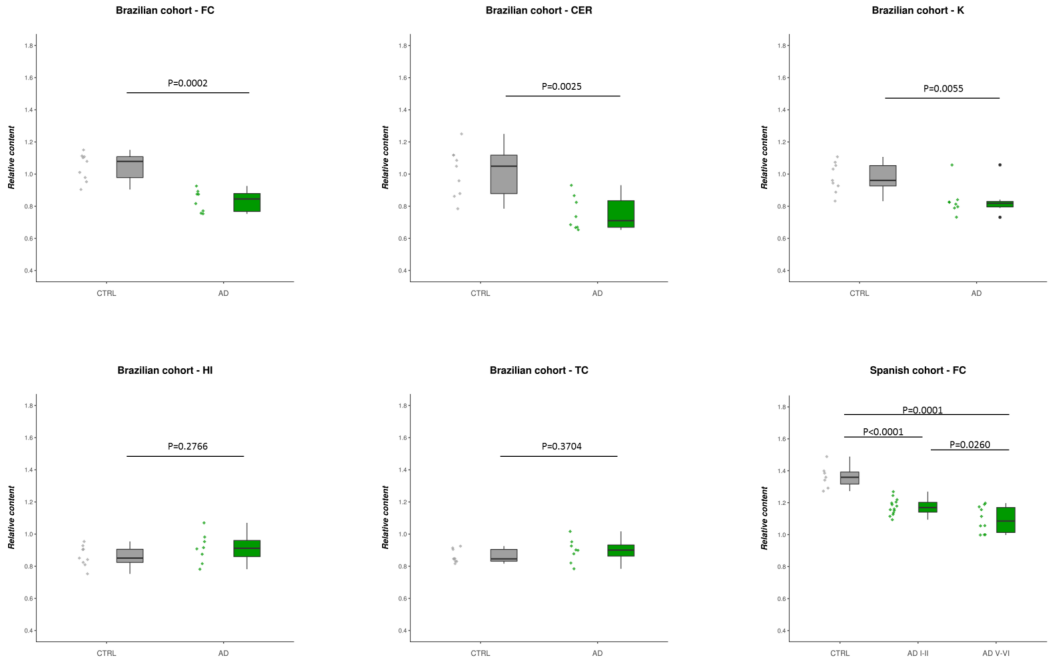


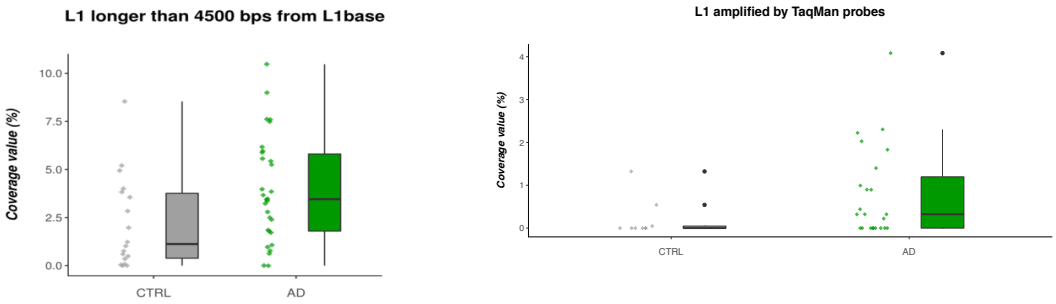
Figure 3.1.2.1 The LINE-1 copy number variation analysis. We performed a qPCR analysis with Taqman probes of FL-L1 copy number in different brain tissues and kidney of different cohorts of patients. For the Brazilian cohort (8 AD patients and 9 controls) we observed a lower amount of FL-L1 elements in AD frontal cortex, cerebellum and kidney compared to controls. No differences were observed in the temporal cortex and hippocampus. In the case of the Spanish cohort (7 healthy controls, 14 patients at Braak stages I-II, and 10 patients at the final AD Braak stages V-VI) only frontal cortex samples were analyzed, and a progressive and always significant decrease in the content of FL-L1 sequences starting from the healthy controls group to the group of patients at the final stages of the disease were detected.

3.1.3 Characterization of LINE-1 content in genomic variations

3.1.3.1 Genomic CNV analysis with Illumina Infinium high-density chip

In order to increase the resolution in the study of LINE-1 copy number and to understand whether the lower amount of LINE-1 elements observed in AD patients with the qPCR experiment could be due to a loss of larger genomic fragments resulting from genomic rearrangements, we performed an analysis of genomic CNV using the Illumina Infinium high-density chip on the same cohort of Brazilian individuals examined with the Taqman assay, 7 Spanish CTRLs, and 8 Spanish Late AD affected patients examined with the Taqman and 2 more Spanish CTRLs and 16 late AD patients that were not examined before.

CNVs were investigated for LINE-1 content taking advantage of various LINE-1 collections. We evaluated both the number of retrotransposons that were in overlap with the CNVs and their coverage values (see Materials and Methods for further information). Interestingly, significant differences in agreement with the results obtained with the qPCR experiments could be observed. In particular, in the Spanish FC Hetero-dels, the average coverage values of FL-L1 elements in AD samples (23 individuals) was higher when compared to CTRL FC (9 individuals) (Figure 3.1.3.1a). This difference remained significant also after the adding of the Brazilian FC samples, although Brazilian samples alone showed the same but not significant trend, probably because of the small sample size (7 AD and 9 CTRLs). The Spanish cohort, presented statistically significant differences also in the coverage of the LINE-1 elements targeted by the Taqman assay, which are mostly non-integer L1s longer than 4500 bp (~90% of the total) (Figure 3.1.3.1b). These evidences are in agreement with our Taqman results, proving that putatively FL-L1 elements are significantly enriched in heterozygous deletions in AD patients, as compared to CTRL samples.



a b

Figure 3.1.3.1 Illumina Infinium high-density chip assay. Spanish cohort. (a) Performing the analysis considering only the LINE-1 elements longer than 4500 bps reported in the L1base, a higher coverage of these elements could be detected in AD frontal cortex samples at the level of deletions in heterozygosis, compared to control samples. This data confirm the presence of a lower amount of FL-L1 elements in AD samples than in CTRLs as already observed with the qPCR experiment (b) By performing the same kind of analysis, but considering only the LINE-1 elements detectable with the primers employed in the Taqman assay, we observe the same, significant trend.

In this section of the thesis we profiled the different FL-L1 makeup of AD affected patients and CTRLs. FL-L1, due to their potential for instability, can be a relevant source of structural variants and allelic heterogeneity in the human genome. It is not coincidence that one of the most interesting polymorphisms found in this study is located in the MHC locus, the most variable region of the genome. Indeed, regions containing active genes are known to be more mutable than regions containing non-active genes as well as regions containing a high density of genes such as gene clusters.

The high gene density, extreme polymorphism, and clustering of genes are all characteristics shared by one of the largest mammalian multigene families: olfactory receptor genes (OR). So, as HLA genes are characterized by extreme genetic polymorphism, OR genes may be expected to exhibit pronounced variability as well, leading to a multitude of different and specialized OR loci in individual genomes in order to accomplish the complex task of reliably distinguishing thousands of odors (Andersson et al., 2014).

In the next section, the genomic landscape of cells expressing an OR gene and cells not expressing the same OR gene, will be compared in order to look for alterations from the reference and differences in DNA sequence between the active and an inactive allele.

3.2 Analysis of LINE mediated SVs in Olfr2 locus

3.2.1 Exploring LINE role in the generation of structural variants such as deletions

Mobile elements affect genome stability creating structural variations through their de novo insertions and post-insertional genomic rearrangements (O'Donnell and Burns, 2010).

In the previous chapter, we discussed about how active LINE-1 mobilize, integrating extra copies of themselves into new genomic locations. However, LINE elements capable of autonomous retrotransposition represent only a small fraction of the total forms present in human and mouse genome. In fact, most of the retrotransposons are no more able to mobilize due to 5' truncations, mutations and rearrangements (Brouha et al., 2003).

In this chapter, we will focus on another significant source of genetic variation which can be related with LINE elements introduced in the previous chapter: genomic rearrangements.

Using a high-throughput sequencing approach, we explore LINE induced genomic variation in a locus where a very high repeat concentration provides more chances for recombination events to occur between retrotransposons, trying to face the technical difficulty of mapping breakpoints within repeated elements.

With this analysis, our aim is to investigate mechanisms underlying the regulation of olfactory receptor choice in mouse olfactory epithelium, characterizing locus-specific genomic rearrangements.

3.2.1 Olfr2 Locus

The olfactory receptor (OR) genes are the largest G-protein coupled receptor mammalian gene family and are expressed in a monogenic and monoallelic fashion in olfactory sensory neurons (OSNs). Each mouse olfactory sensory neuron monoallelically expresses one out of approximately 1400 OR genes (Niimura and Nei, 2005).

Among all the OR genes we decided to focus on Olfr2 due to the availability of a knock-in mouse model (B6;129-Olfr2-GFP mice). In this mouse a GFP sequence was inserted at the 3' of *Olfr2* coding sequence, an olfactory receptor (OR) expressed in a small number of olfactory sensory neurons (OSNs). Since this configuration gives rise to a bicistronic mRNA, OSNs naturally expressing Olfr2 co-express also GFP and cells that naturally activate the transcription of Olfr2 become fluorescent and therefore easy to detect and isolate.

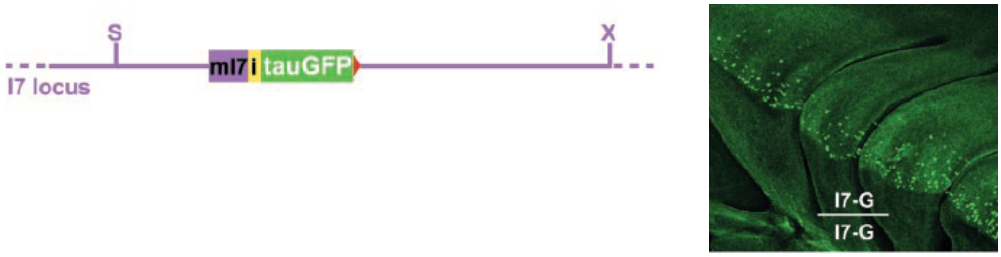


Figure 3.2.1 GFP construct inserted at the 3' of Olfr2. In this experiment, we took advantage of a transgenic mouse line where a GFP gene was inserted at the 3' of Olfr2 in order to easily identify fluorescent cells expressing the receptor. Both pictures come from Bozza et al.,2002.

3.2.2 Characterizing genomic rearrangements in the expressed locus

Eleven overlapping PCR products, each nearly 5000 bp long for a total of nearly 50 kb of sequence around Olfr2 transcription start site (TSS) were amplified from GFP+ cells (collected by LCM and amplified by MDA) expressing the receptor, and from bulk genomic DNA (extracted from OE of an age-matched mouse from the same litter and not amplified by MDA). We consider the GFP+ cells (MDA sample) the test where we expect to find the regulatory element, and the bulk genomic DNA (OE sample) the control. OE sample is both a "technical positive control" not subjected to MDA amplification and therefore "artifacts free" and a "negative biological control" where we do not expect to find the SV which might potentially regulate the expression of the receptor, if any. Bulk genomic DNA consists of a whole OE cell population in which Olfr2-expressing cells (GFP positive) represented only the 0.1% of the total olfactory sensory neurons.

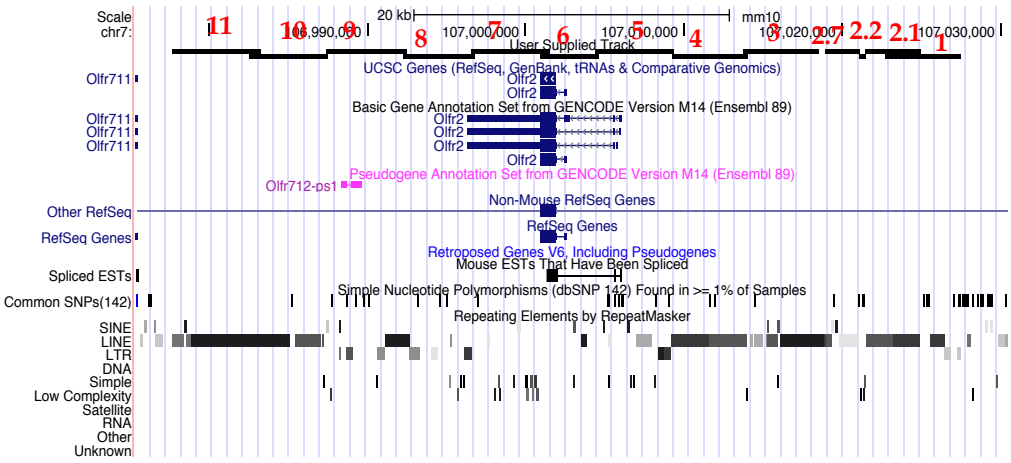


Figure 3.2.2 Amplicons distribution over Olfr2 locus. Genome browser screenshot of the 50kb Olfr2 locus. The amplicons are numbered from 1 to 11, starting from the 5' to the 3'. GFP construct is located in correspondence of amplicon 7.

3.2.4 Repetitive elements in OR clusters

Olfr2 locus genomic environment is very complex, with a high density of annotated TEs (Fig 3.2.2), especially LINEs.

This is not surprising since monoallelically expressed genes are known to be flanked by high densities of LINE-1 elements, especially FL-L1s, and only a few short interspersed nuclear elements (SINEs), in sharp contrast with biallelically expressed genes (Allen et al., 2003).

80 different repetitive elements are annotated by Repeat Masker in the 50 kb regions around the Olfr2 locus, covering almost the 70% of the locus. The remaining 30% consist of non-repetitive genomic DNA. LINE-1s appeared to be the most abundant TEs present in the locus (29 elements, covering 27341bp), followed by LTRs (12 elements covering 3944 bp), SINEs (11 elements occupying 1594 bp), simple repeats (18 elements, 1227 bp) and low complexity repeats (10 LCR, 703 bp).

Interestingly, LINE-1 enrichment is a characteristic not only of our locus but of all monoallelically expressed genes and specifically of OR-clusters, which show a peculiar enrichment for this class of retrotransposons compared with other classes of repeats. A high LINE-1 concentration in a locus is known to function as substrate for homologous pairing and in general increase chances for recombination events to occur between retrotransposons (Han et al., 2008), which can lead to chromosomal breaks and rearrangements. Therefore, in order to handle the problem of SV reconstruction in this repeat-rich region, we combined PacBio single molecule sequencing for reliable mapping across repeat expansions with a complementary high-fidelity paired-end Illumina sequencing for accurate identification of breakpoints.

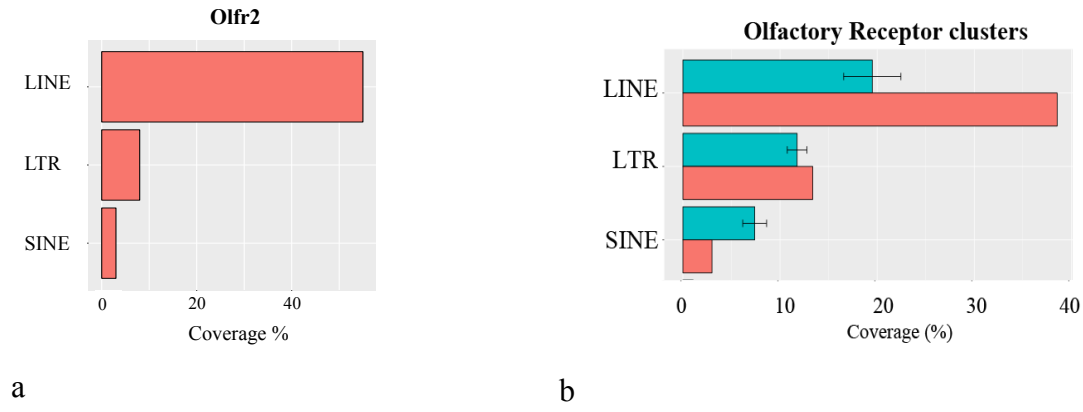


Figure 3.2.4 Repeat occupancy of OR. (a) LINE elements, covering the 55% of Olfr2 locus, become the most abundant class of repeats in this region; (b) LINE enrichment is a general feature of olfactory receptor gene culsters. Per-class statistics for clusters (ClustCov) are reported in orange and randomizations (MeanRandomCov) are reported in blue.

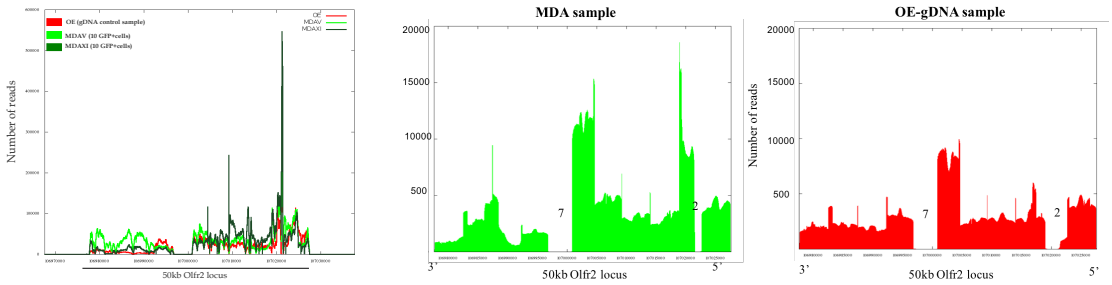
3.2.5 Illumina sequencing

For Illumina sequencing we performed long range PCR of Olfr2 locus on two different MDA biological replicas out of 11 (MDA-V and MDA-XI samples) and on bulk genomic DNA from OE (OE sample). Finally, 13 PCR amplicons for each sample were sequenced for a total of 39 Illumina libraries.

Illumina reads were mapped onto the mouse reference genome MouseGRCm38/mm10 assembly. On average 6.5 million read pairs were mapped onto the genome per individual (see Materials and Methods). We obtained a good coverage of Olfr2 locus, 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.

3.2.6 PacBio sequencing

The best PCR products from each of 11 MDA biological replicas were purified and pooled together for Pac Bio sequencing in parallel with PCR products from bulk OE genomic DNA. 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. Unfortunately, long range PCR amplification failed for amplicon 7, due to the presence of the long IRES-GFP construct (in this study we used B6;129P2Olfr2tm1Mom/MomJ mice that contain a GFP gene inserted at the 3' of the Olfr2 locus in order to identify fluorescent cells that naturally activate the transcription of Olfr2). PCR failure was observed also for amplicon 2 due to high density of repetitive sequences in this amplicon (89% of bp in this amplicon were part of repeated elements) which made very difficult primer designing.



a b

Figure 3.2.5 Sequencing coverage of 50kb *Olfr2* locus. *Olfr2* 50 kb locus Illumina (a) and PacBio (b) coverage in MDA sample and OE samples is represented plotting the number of reads for the 50 kb locus coordinates. OE coverage is represented in red, MDAV coverage is represented in light green and MDA XI coverage is represented in dark green. The numbers "7" and "2" indicate the amplicons which were not correctly amplified.

3.2.7 Identification of non-annotated structural variants from Illumina reads

Due to the high density of TE in this region and the presence of putative endonuclease cutting sites, we expected the presence of non-annotated genomic SVs in the area (Korbel et al., 2007, Huang et al., 2010). Variation discovery on each PCR amplicon was performed with Pindel starting from Illumina, mapped reads. In this way, we certainly missed SV spanning 2 or more amplicons but, giving the fact that no band-size selection was performed for the library preparation, we expected to target the "full spectrum" of SV detectable by Pindel, contained in each amplicon delimited by its specific primers.
It must be stressed that Pindel was chosen as it is known to perform better when the goal is to target medium-sized SVs, especially large deletions. Overall in the Olfr2 locus, we identified almost 2400 deletions, 34 inversions, 407 tandem duplications and 806 insertions covered by at least 5 bp in at least one sample (MDAV, MDAXI, and OE).

	DELETIONS	INVERSIONS	TANDEM DUPLICATIONS	INSERTIONS
MDAV	749	12	116	200
MDA XI	826	15	100	424
MDAV-XI	241	2	76	65
MDAV-OE	62	2	15	10
MDAXI-OE	55	0	6	14
MDAV-MDAXI-OE	238	1	92	58
OE	220	2	2	35
TOTAL	2391	34	407	806

Table 3.2.7 SV detected with Pindel on Illumina reads.

Among all the detected SV, we selected only SVs longer than 50 bp and covered by at least 5 bp for conceptual and technical reasons. Indeed, from an exploratory point of view, we're looking for transposable element mediated SVs, so large rearrangements. Unfortunately, from a technical point of view, we have to deal with the limits of current SV callers: 5 reads appeared to be a good threshold to consider reliable an SV call and 50 bp of minimum SV length was imposed in prevision of the following validation with PacBio reads of SVs detected with Pindel.
Interestingly, in both MDA and OE samples, all the variations supported by Illumina reads were found in a "heterozygous condition" with a small number of reads supporting the non-annotated variant and a high number of reads supporting the annotated reference sequence. The low read coverage level of the non-reference allele, may suggest a low-

level somatic mosaicism for SVs occurring in one or just a few cells, ideally the neurons expressing the receptor.

3.2.8 Genomic deletions in Olfr2 locus

Genomic deletions are the most abundant SVs found in the Olfr2 locus. For this reason, we decided to focus our attention, firstly, on this category.
Deletion length ranged from a maximum of 5141 bp (deletions longer than the amplicon size were discarded, as artifacts), to a minimum of 50 bp (because of the imposed cutoff size).
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), deletion distribution was found to be prominent at the 5' of Olfr2 coding sequence.

	AMPLICONS											
	1	**2.1**	**2.2**	**2.7**	**3**	**4**	**5**	**6**	**8**	**9**	**10**	**11**
MDAV	17	196	9	99	135	94	26	24	9	63	57	20
MDAXI	35	81	6	48	277	285	61	4	15	4	7	3
MDAV+MDAXI	10	75	2	12	68	33	8	5	4	7	15	2
MDAV+OE	21	27	0	4	0	2	3	1	3	1	0	0
MDAXI+OE	17	3	0	6	10	11	5	0	3	0	0	0
MDAV+MDAXI+OE	43	120	0	4	27	13	7	5	11	6	2	0
OE	161	10	0	0	7	1	4	1	35	1	0	0
sum	304	512	17	173	524	439	114	40	80	82	81	25

Table 3.2.8 Deletions detected with Pindel on Illumina reads (length >= 50 bp)

Given that all the amplicons have a similar length ranging from 3500 bp to 5500 bp (only amplicon 2, divided in 3 parts for technical reasons, was much shorter than the others), the fact that the deletions located at the 5' of the coding sequence are not only the most abundant but also, on average, the longest might suggest that SVs upstream the gene may regulate its expression as already suggested by Serizawa and collegues in 2000.
As expected, amplicon 6, which contains the Olfr2 coding sequence, presented a very low number of deleted sequences. Moreover, compared to the other amplicon the deletions of amplicon 6 were also very short (for example, the average deletion length in amplicon 6 is 517 bp, while in amplicon 3 is 2911 bp).

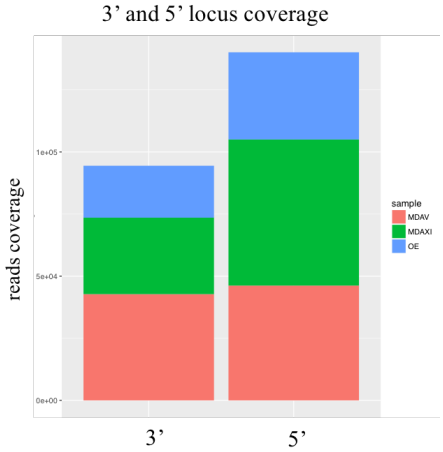


Figure 3.2.8 Illumina coverage of 5' and 3' amplicons with respect to *Olfr2* TSS. 3' amplicons are 6-11 while 5 prime amplicons are 1-5. MDAV (orange); MDAXI (green); OE (blue).

3.2.9 Validation of Illumina supported deletions with Pac Bio long reads

Variant callers are known to be prone to false positive calls due to alignment errors. Such errors may occur when the number of bases in the reads, matching the reference genome, is too few and when the number of reads supporting a SV is small. This problem exacerbates in highly repetitive regions.

For this reason, we decided to use a PacBio read data set, for an in-silico validation of Illumina supported deletions, in order to increase the accuracy of variation prediction in Olfr2 locus, a very repetitive region.

PacBio RS single-molecule technology provides previously unprecedented sequencing read lengths, making it useful for the detection of long SVs even in low complexity regions. Moreover, PacBio long reads by definition are not affected by sequencing amplification bias. Unfortunately, these sequences contain random errors involving 10–15% of nucleotides.

Illumina sequencing, on the other hand, offers stable lengths of short reads with errors most likely to be grouped at the ends of the reads.

Herein, we combined the advantages of the short reads of a second-generation sequencing technology (Illumina reads from this study had an average length of 300 bp) with the long reads of the PacBio platform (PacBio RS reads from this study had a mean length of 2459

bp) in order to increase the accuracy of variation prediction in Olfr2 locus, a very repetitive region.

Each Pindel deletion supported by Illumina reads was checked for the presence of any additional supporting Pac Bio read (see Materials and Methods).

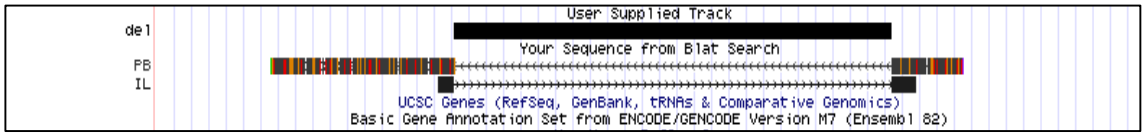


Figure 3.2.9 Pindel deletion supported by Illumina and PacBio reads. In this example, the black bar on the top of the UCSC genome browser screenshot represents the coordinates (bed format) of a deletion detected with Pindel starting from Illumina reads, the lowest bar represents the sequence (fasta format) of an Illumina read supporting the deletion, the middle bar represents the sequence (fasta format) of a PacBio read supporting the same deletion.

After this filtering process, that we consider as a first validation step, the number of deletions dramatically decreased. Only 149 deletions (6%) were supported by at least 1 PacBio read (min PacBio reads supporting a deletion =1, max PacBio reads supporting a deletion = 69).

	AMPLICONS						
	1	**3**	**4**	**5**	**9**	**10**	**11**
MDAV	0	19	13	5	4	2	0
MDA XI	0	31	22	1	0	0	1
MDAV-XI	1	12	6	1	1	1	1
MDAV-OE	0	0	0	1	0	0	0
MDAXI-OE	2	0	0	0	0	0	0
MDAV-MDAXI-OE	9	1	2	0	0	2	0
OE	9	0	0	1	1	0	0
sum	21	63	43	9	6	5	2

Table 3.2.9 Deletions detected with Pindel on Illumina reads supported by PacBio reads.

Looking at the frequency of deletion length distribution, after Pac Bio filtering, we could appreciate that the number of very small and very long deletions was reduced, with the majority of deletions having a length ranging from ca. 2000 to 4000 bp. Deletions in amplicon 2, 6, 8 and 9 did not have any Pac Bio supporting reads.

Interestingly, 19 deletions validated with only MDA PacBio reads were present in both Illumina MDAV, MDAXI replicates but not in OE. These deletions, mostly located upstream the coding sequence, in amplicon 3, may be the putative SV involved in Olfr2 expression.

3.2.10 Repetitive elements in deletions

LINE elements, the most abundant repeats in Olfr2 locus, appeared to be also the repeat class mostly affected by deletions. Looking at the percentage of bases covered by deletions detected with Pindel and supported by Illumina reads we can observe that deletions covered 98% of LINE, 85% of SINE and 79% LTR sequence. This event is even more pronounced looking at the loss of LINE sequences (77%) compared to SINEs and LTRs (46% and 36% of bp covered, respectively) resulting from deletions supported by PacBio reads.

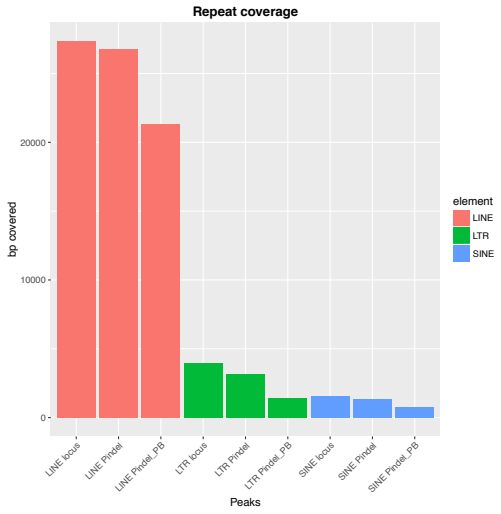


Figure 3.2.10 Repeat coverage for *Olfr2* deletions. Deletions bp coverage for different classes of repetitive elements is shown. LINE coverage is represented in orange, LTR coverage in green and SINE coverage in blue. For each class, the number of bp covered in the locus (first column for each group), the number of bp covered by Pindel deletions (second column), and the number of bp covered by Pindel deletions supported by PB reads (third column).

3.2.11 Deletion clustering

In order to reduce the complexity of the detected deletions we clustered together deletions supported by at least 1 PB and 5 or more Illumina reads in at least one sample (MDAV, MDAXI or OE) if they overlapped the same LINE elements (a minimum of 2 LINE elements should be overlapped by every considered deletion) reported in the reference genome. The clustering resulted in a subset of 125 deletions, intersecting 2 to 5 LINEs, out of 149 original deletions. Only one clustered deletion involved a FL-L1 at one boundary. Although chromosomal breakage could originate anywhere in the genome, it is tempting to speculate that LINE sequences themselves may be involved in break formation (Erwin et al., 2016b). To verify it, a random set of deletions of the same number and length of the 125 clustered deletions was created in the 70kb locus to compare the frequency of observed characteristics in the real deletions versus a randomly selected population.

3.2.12 Inferring mechanisms of deletion formation

The presence and characteristics of particular genomic features associated with the observed deletions suggested us that the same principal mechanism could operate in the formation of all the deletions: microhomology mediated end joining (MMEJ). Homologous sequences ranging from 5 to 25 bp were found in proximity of the breakpoint junctions. Moreover, sequence motifs and transposable elements were found in the breakpoint regions. GC content in the homologous sequences flanking real and random deletions was also evaluated and found to be significantly different.

3.2.12.1 5-25 bp Microhomology between the breakpoints

Microhomology is defined as one or more base pairs (bp) of perfectly matching sequence shared between the proximal and distal reference sequences surrounding the breakpoints. Direct repeated homologous sequences (DRS) ranging from 4 bp up to 17 bp (8 bp on average) were present at the breakpoints. 120 times over 125 the homology was 5 bp or more suggesting MMEJ as the major mechanism of repair to double strand breaks.

92 times one DRS was removed by the deletion process and the other retained (figure 3.2.12.1a, case 1), 13 times both DRS were inside the deletion and therefore removed by the deletion process (case 2), 2 times left and right DRS were deleted but an extra DRS was present in the retained portion (case 3), 14 times both DRS were external to the deletion and so retained (case 4).

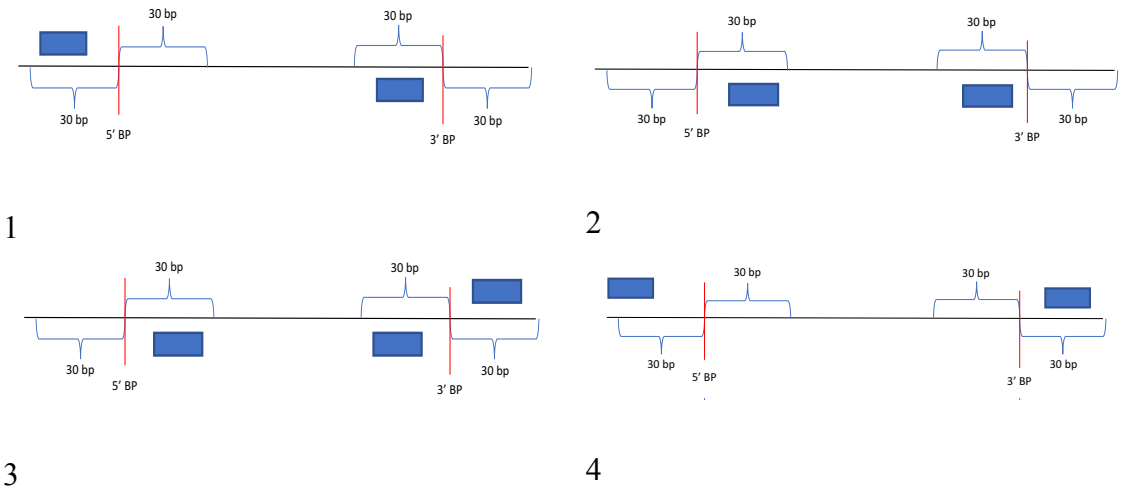
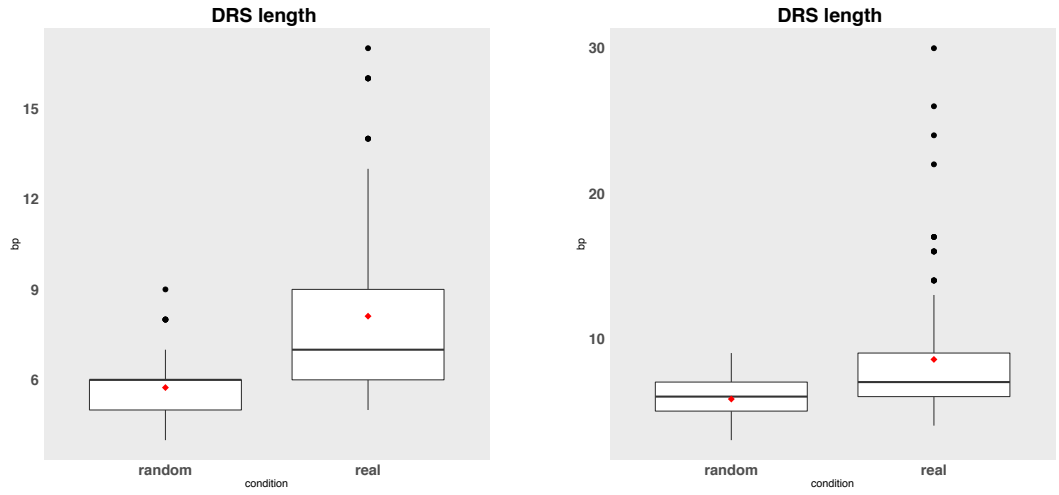


Figure 3.2.12.1a Schematic representation of DRS position at deletion breakpoints.

In order to assess whether these microhomology regions were significantly enriched in our data set we generated a random deletion data set. In the random dataset DRS had an average length of 6 bp. Using a Student's t Test we observed that DRS occurring at the boundaries of real deletions are significantly longer than DRS occurring at the boundaries of random deletions (pvalue < 2 e-16).



a b

Figure 3.2.12.1b DRS length distribution boxplot. a. DRS occurring at the boundaries of real clustered deletions are significantly longer than DRS occurring at the boundaries of random deletions (pvalue < 2 e-16). **b.** This is true also considering the DRS occurring at the boundaries of all the deletions (non clustered, original subset). In the boxplot the average length is represented by the red dot.

3.2.12.2 Presence of known sequence motifs in the DRS

Particular genomic architectural features were found in all breakpoint regions and some of them were known as being significantly associated with structural break resolution. Identified motifs belonged to different categories including meiotic recombination hotspots (genome locations where many recombination events are concentrated), polymerase beta frameshift hotspots (Chuzhanova et al., 2009), hamster and human APRT deletion hotspots (Smith and Adair, 1996), FOXO recognition elements (Alkhatib et al., 2012) , indel super - hotspot motifs (Ball et al., 2005), and immunoglobulin heavy chain class switch repeats (Abeysinghe et al., 2003, Chuzhanova et al., 2009). The presence of these motifs at the breakpoints was checked in the sequenced data set and the randomly generated data sets in order to assess whether these motifs were enriched in our samples.

3.2.12.3 Repetitive elements at the 3' and 5' of the deletions

The well-known capacity of sequence motifs to predispose to DNA breakage led us to analyze the nucleotide context of the breakpoint regions (60 bp around the breakpoints). Most of the deletions (122/125) presented repetitive elements at their 3' and 5' breakpoints harboring the microhomology region (3/125 presented a repetitive element only at one end). These numbers are particularly interesting if compared with the ones of the random deletion dataset. Only 23/125 random deletions harbored repetitive elements at both breakpoints, 51 random deletions presented no repetitive elements at all at the breakpoint region and 51 random deletions involved only one repetitive element.

The Blast2 analysis were performed to determine the percentage of sequence identity between the repetitive elements bridging the deletion boundaries. We found evidence that 27 deletions displayed more than 70% homology between LINE elements present at the left and right breakpoint. This may suggest that LINE elements could be involved in the restoration of DSBs occurring within repeated sequences. When the deletion bridges two LINEs, the transposable elements may help to repair the DNA lesions forming secondary structures which can bring distant DNA segments close to each other and therefore, promoting the recombination between them. On the other hand, LINE sequences themselves may be involved in break formation. Interestingly, even when the homology region between repetitive elements overlapping the deletion breakpoints was very small, it corresponded to the DRS located in the 60 bp surroundings of the left and the right breakpoint.

3.2.12.4 GC-rich microhomology

Reportedly (Verdin et al., 2013), base composition influences MMEJ. In particular, GC rich motifs increase the stability of the annealed pairs during MMEJ (Kent et al., 2015). So, we next asked whether GC content in the real deletions DRS was higher than the

random deletions DRS. The percent of GC of the real DRS (39.5% in all DRS, 44% in DRS longer than 8 bp) resulted to be significantly higher than the ones of the random DRS (34% in all DRS, 36.6% in DRS longer than 8 bp) consistently with previous findings. From this data we could conclude that the observed deletions may be the result of a microhomology-mediated mechanisms of double strand break (DSB) resolution such as MMEJ.

3.2.13 Validation of Pindel deletions with PCR

In parallel to Pac Bio validation, Pindel deletions were validated independently by end-point PCR in order to answer the following questions:

- Is SV calling based only on Illumina reads reliable?
- Is MDA amplification inducing artifacts?
- Is PCR amplification inducing artifacts?
- Are DRS real?

3.2.13.1 Is SV calling based only on Illumina reads reliable?

The in-silico validation with PacBio reads, performed on the deletions detected with Pindel, resulted in a considerable loss of information. Indeed, the 94% of deletions detected with Pindel was discarded because no PacBio read supported them. However, by randomly checking Illumina coverage of deletions with no PacBio coverage we could appreciate a perfect fit between Illumina reads and Pindel calls.

Figure 3.2.13.1a Intersection between deletions detected with Pindel and Illumina supporting reads. UCSC genome browser view of deletions 662 and 314 located on amplicon 6 that lacked PacBio coverage.

So, assuming that Pindel SV calling is reliable, the lack of PacBio coverage of the majority of the detected deletions may have at least two explanations: either most of the deletions detected with Pindel and supported by Illumina reads are artifacts or we are not at saturation with PacBio reads and we require a higher sequencing depth to increase the chances to target rare events.

To investigate this issue, we decided to include in the validation pool also deletions supported by Illumina reads but not by Pacbio reads. Among all the deletions that lacked PacBio support we focused on the ones present in both MDA samples, our samples of interest, and OE sample, the control. Moreover, since most of the deletions were concentrated upstream the Olfr2 TSS we decided to focus on amplicons 1-5. Amplicon 2 was not taken in consideration since it has no PB coverage at all.

Because of the highly repetitive nature of the locus and given the nested deletion pattern, designing the probes appeared to be very challenging. To overcome these difficulties, and maximize the chances of success, we decided to focus on the deletions showing the highest Illumina read coverage and design the primers on the mapped side of a consensus of the reads supporting the deletions at their 5' and 3'.

Overall, we selected for validation 6 deletions: 3 of them without any PacBio coverage. One deletion located in amplicon 1, one located in amplicon 3, two located in amplicon 4 and two located in amplicon 5. PCR validations were performed on the same amplified PCR products (MDA and OE) which were sent for Pac Bio and Illumina sequencing.

	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 3.2.13.1 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 PB reads for sample.

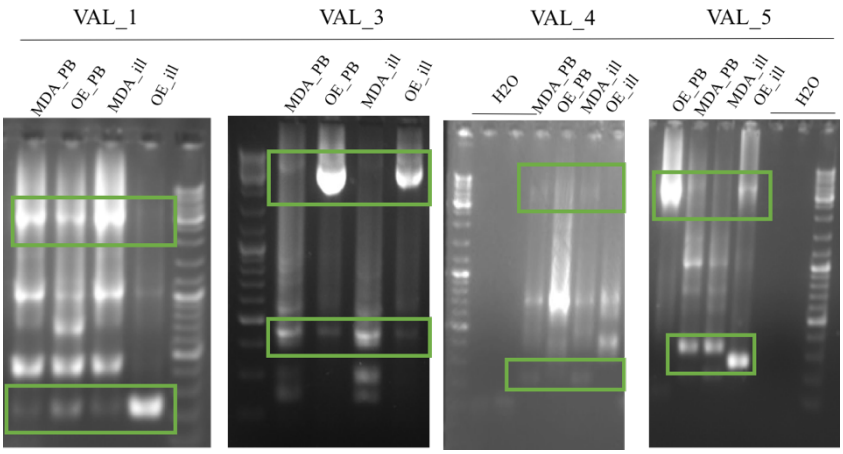


Figure 3.2.13.1b PCR validation results. Green rectangles indicated validated deleted and non-delted band for each validation assay. MDA_PB and OE_PB represent MDA and OE PCR amplicon samples sequenced by Pac Bio; MDA_ill and OE_ill represent MDA and OE PCR amplicon samples sequenced by Illumina.

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. Confirmed

Sanger sequences were intersected with the coordinates of the respective Pindel deletions, presenting perfect match at the boundary.

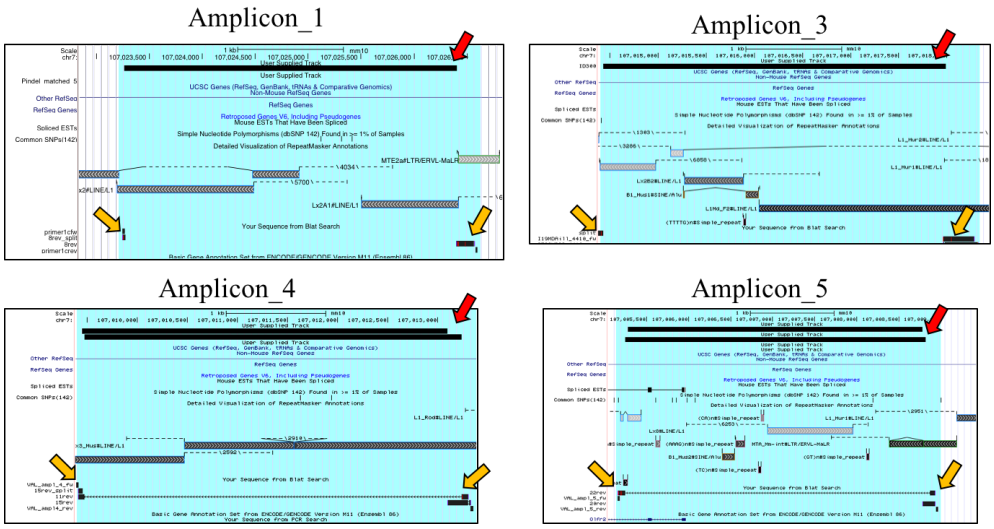


Figure 3.2.13.1c Sanger sequences supporting selected Pindel deletions. UCSC genome browser view of Sanger sequences for amplicons 1, 3, 4 and 5 are indicated by the yellow arrows. Red arrows indicate the Pindel deletions supported by Sanger sequences. Detailed Repeat Masker annotation is activated.

Since we were able to validate also deletions without any PB supporting read, our second hypothesis is confirmed: maybe we need a higher PacBio coverage for an exhaustive description of our locus or we have to relax the parameters orthe alignment PB illumine.

3.2.13.2 Is MDA amplification inducing artifacts?

It is well known that MDA amplification induces artifacts (Lasken and Stockwell, 2007, Treiber and Waddell, 2017) but working with low number of cells requires whole-genome amplification (WGA) techniques to increase DNA starting quantity.

In order to be aware of possible SV false calls, resulting from MDA induced chimeric fragments we adopted a “technical positive control” that did not undergo MDA amplification: OE DNA sample. Deletions present in both MDA samples and OE samples can-not be MDA induced artifacts.

Independent PCR validations confirmed two deletions at the 5’ of Olfr2 TSS, both supported by Pac Bio reads and shared by two MDA replicates. A third deletion, not

supported by PacBio reads nor by OE Illumina reads, unexpectedly, was validated also in the OE sample. On one side this evidence may exclude the possibility of a possible MDA-induced chimeric nature of the deletions but most probably it can be a clue of a lack of sequencing saturation.

3.2.13.3 Is PCR amplification inducing artifacts?

In order to exclude the possibility of considering as real events, deletions that are artifacts resulting from the initial Olfr2 locus PCR amplification, we performed a validation also on total MDA amplified starting material prepared for PCR amplification. Amplicon 3 deletion was successfully validated from both bulk OE DNA and total MDA amplified DNA. This allows us to confidently exclude the presence of a PCR induced artifact.

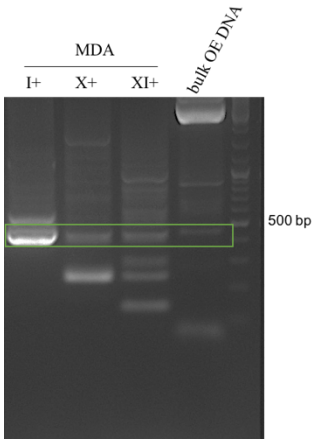


Figure 3.2.13.3 Validation on total MDA amplified starting material for amplicon3-deletion. Validated amplicon 3-deletion bands are shown for three MDA replicates (I, X, XI) and from total gDNA expected from OE (bulk OE DNA) in the green rectangle.

3.2.14 Are DRS real?

All validated deletions were characterized by the presence of DRSs. Interestingly, not only DRSs detected bioinformatically on clustered deletions were recovered, but DRSs were detected also at the borders of validated deletions not covered by PacBio reads and therefore excluded from the clustering.

	Amplicon 1	Amplicon 3	Amplicon 4	Amplicon 4	Amplicon 5	Amplicon 5
ID_Pindel	199	300	159	149	260	256
MDA_PB_reads	1	1	0	0	9	0
OE_PB_reads	0	0	0	0	0	0
DRS motif	TCCCATCCTCCC	ATTTTGAT	GAGAGG	CCTAG	GTACCAATT	GTACCAATT

Table 3.2.14 DRS are present in all the validated deletions. For each deletion validated by PCR all the related information is summarized. ID_Pindel (best match) = ID of Pindel deletion which is supported by Sanger sequence with the highest nucleotide precision; MDA and OE-PB reads= number of supporting PB reads for sample. DRS sequence = microhomology motif flanking the boundaries of the deletions.

Apart from simple self-insertion, investigated with the SPAM technique and described in the previous results section , LINE-1 elements appear to alter the primary structure of the genome also providing material for DNA recombination (Burwinkel and Kilimann, 1998) leading to deletion and duplication of sequence within the repeats.

The abundance of MMEJ mediated deletions, alternative LINE-1 endonuclease target sites and extensive LINE-1 occupancy of the whole locus seem to be predictors of an unstable genomic region, prone to DNA breakage, rearrangements and structural variation where the detected deletions may have arisen from a repair process of existing DNA lesions.

The same dynamics may potentially be an advantageous feature of all Olfr genes: a strategy to create diversity.

The renewal ability of olfactory epithelium (OE), constantly regenerating mature OSNs, makes this tissue an ideal substrate for LINE-1 mobilization. At the same time, the high activity and direct exposure to external environment of OE, result in the accumulation of DNA damage and the loss of genomic integrity consequent to inefficient repair mechanisms.

To obtain further insights into this issue, we profiled DSB distribution in olfactory epithelium.

3.3 Chip Seq analysis of endogenous γ-H2AX in mouse olfactory epithelium and liver

3.3.1 Profiling double strand breaks: cause and effect of structural variants

3.3.1.1 γ-H2AX peak characterization

Phosphorylation of histone H2AX (γ-H2AX), occurs in response to DSB as an early signal to recruit the DNA-damage repair protein machinery . Therefore, γ-H2AX can be a suitable indicator for the presence of DNA-DSBs in different tissues. Most research relying on chromatin immunoprecipitation (ChIP) methods, to understand how γH2AX contributes to double-strand break repair in mammalian cells, starts from artificially induced, often target specific DNA breaks. Little is known about the differential distribution of γ-H2AX marker for DSB throughout the genome at physiological conditions. To address this question, we used the ChIP-seq technique to profile the chromosomal distribution of γ-H2AX in C57BL/6J mice OE (at p6 and 1m) and L (at p6).

In order to obtain reliable ChIP enriched regions for each biological condition (OE and L at p6 and OE at 1 month) we intersected the peak sets of two replicates (A and B). For comparison, a random set of peaks (shuffled peaks) was generated as a control dataset to use in all the analysis.

IP samples	peaks	peak length (average bp)	Replicates IntersectBed
Lp6A	31469	3782	23363 (74.2%)
Lp6B	45791	5875	
OEp6A	28465	5755	16050 (56.4%)
OEp6B	34280	3855	
OE1mA	23110	4360	10298 (44.6%)
OE1mB	23406	4458	

Table 3.3.1.1 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.

Almost 6000 peaks were in common among the three samples, while the number of Lp6 exclusive peaks (almost 1200), doubled the number of OE1m peaks (6000) and was six times higher than the number of OEp6 sample (2000).

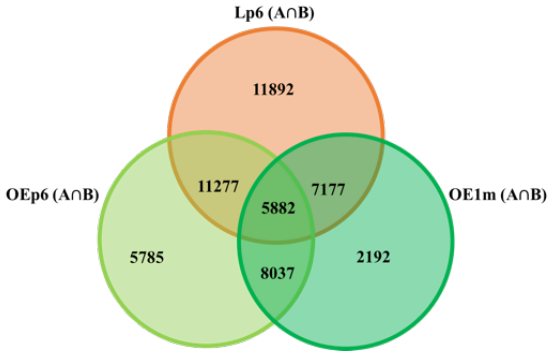


Figure 3.3.1.1 ChIP-seq sample peaks intersection. Venn diagram of peaks intersection among different ChIP-seq sample datasets.

3.3.1.2 γ-H2AX peaks genomic distribution

In order to gain insights about γ-H2AX distribution, we looked at the position in the genome of the detected peaks. From a first glance at the cariotype, the profile looked very similar in the different datasets.

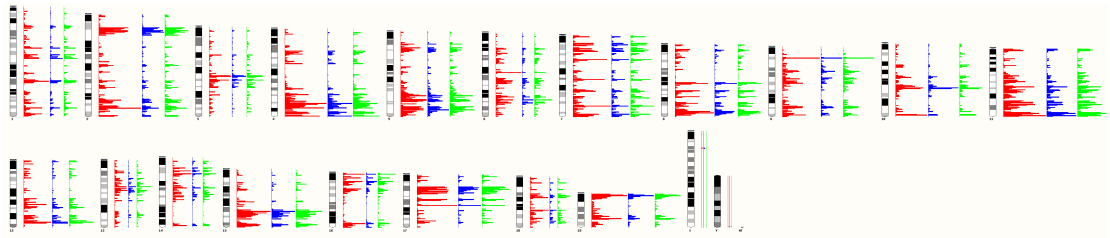


Figure 3.3.1.2a Peaks distribution with respect to the caryotype. Ensembl view of all mouse chromosomes. Chip-seq peaks are indicated with different color per sample: Lp6 (red); OE1m (blue); OEp6 (green)

The samples continued to present similar patterns as we proceeded describing the signal more in detail.

In particular, the enrichment of the peaks in mouse genome was mostly found in correspondence of gene bodies, peaking from 0 to 10kb downstream the transcription start site (**TSS**). This was in sharp contrast with the distribution of shuffled peaks, mainly enriched at the intergenic level.

Moreover, more than 20% of real peaks and 3% of random peaks covered GC-rich regions, namely very sensitive to DNA damage. This result is aligned with the previous ones, since GC-rich regions are considered gene markers in vertebrate genomes (Han and Zhao, 2009).

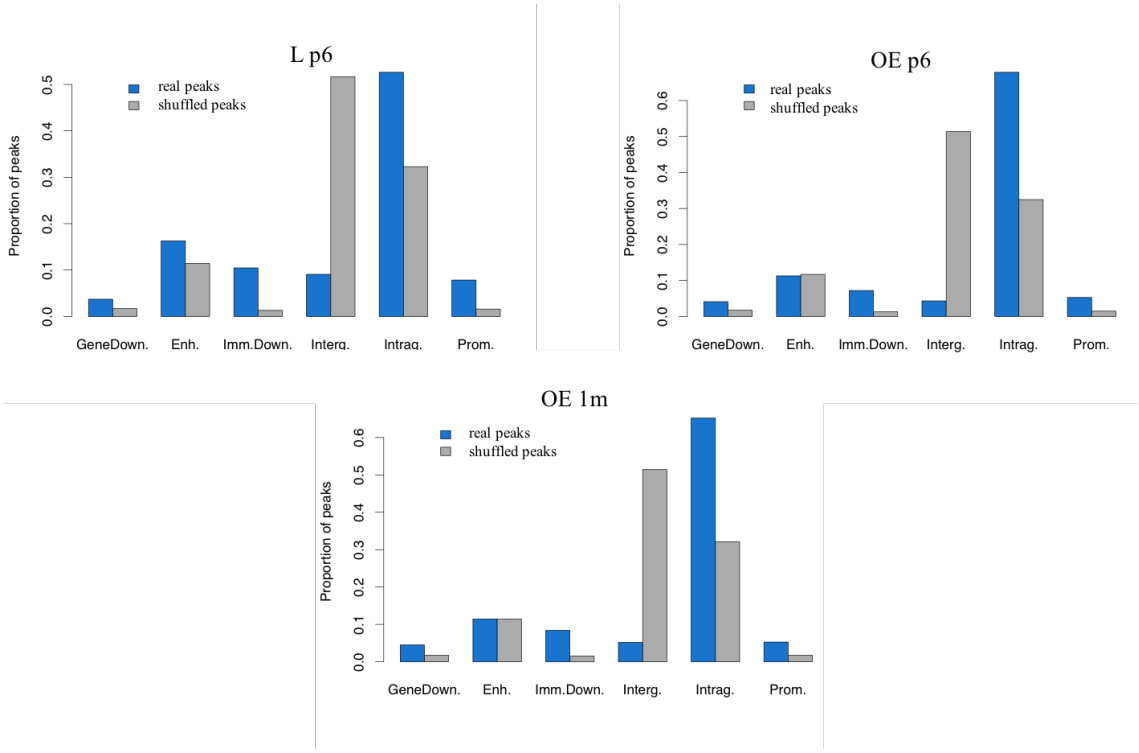


Figure 3.3.1.2b ChIP-seq peaks genome annotation. Peaks genome distribution was performed with the bioinformatics tool NEBULA; the genomic regions were considered with respect to gene start site (TSS). For each sample the proportion of peaks falling in each genomic region is shown. NEBULA legend: Gene 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.

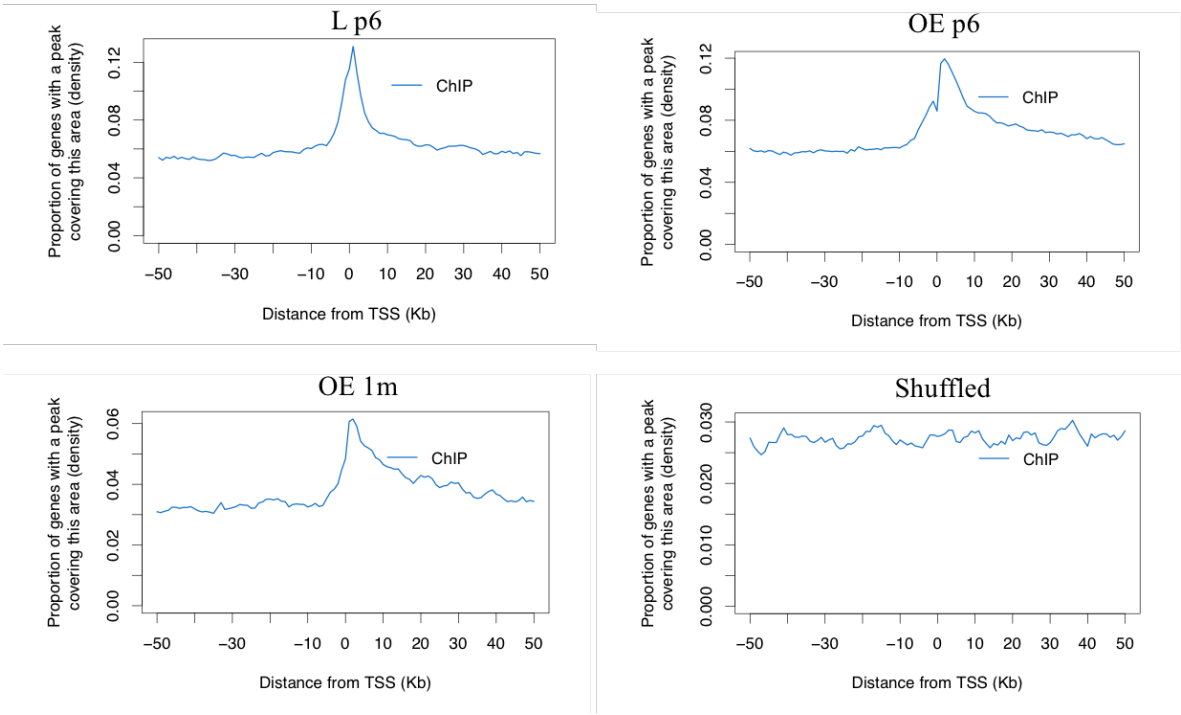


Figure 3.3.1.2c ChIP-Seq peaks distribution around TSS. Peaks genome distribution with respect to annotated TSS was performed with the bioinformatics tool NEBULA. For real and shuffled peak datasets peak density was plotted with the distance from annotated TSSs.

Sample dataset	CpG overlapping peaks (%)
Lp6	21.66
OEp6	24.21
OE1m	20.88
shuffle	3.01

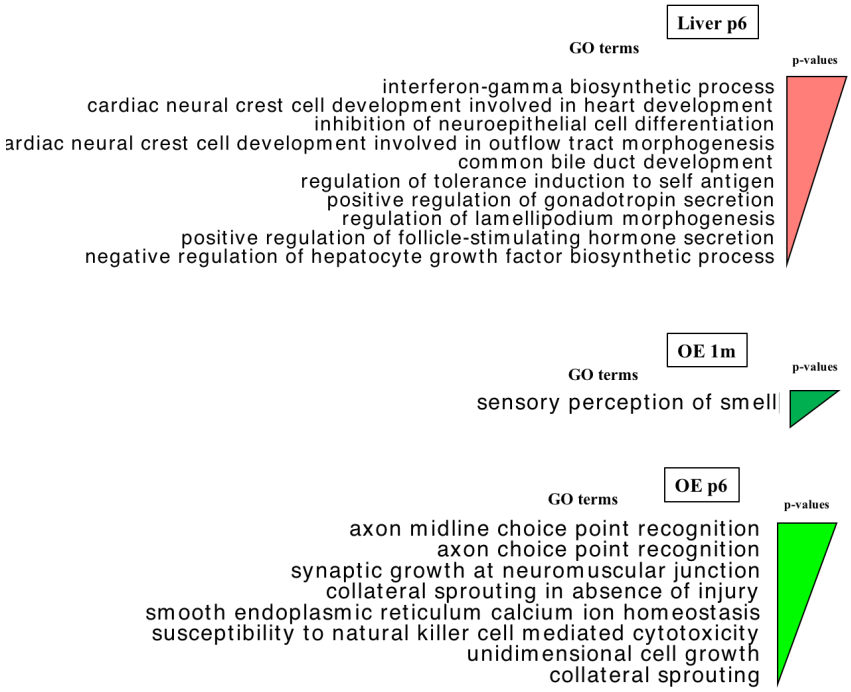
Table 3.3.1.2 Percentage of peaks overlapping CpG islands. For each sample the percentage of peaks overlapping CpG islands is shown.

3.3.1.3 Gene Ontology enrichment analysis

To examine whether our peaks were associated with genes with specific functions, we performed gene ontology analysis using GREAT. We compared the enrichment of each sample dataset (foreground dataset) against total number of sample peaks (background dataset). Interestingly "sensory perception of smell" was the only biological function enriched in olfactory epithelium of older mice (OE1m). While younger mice showed an enrichment of biological functions related processes linked to neuron innervation like "axon choice point recognition" and "collateral sprouting". This is coherent with the fact that at six days after birth (p6) OE is not completely mature and cell proliferation rate is high while at 1 m OSNs innervation pattern is complete and a slower cell proliferation rate parallels a decline in apoptosis and OSNs regeneration.

Among the terms enriched in liver we can list, "common bile tract development" and "negative regulation of hepatocyte growth". These results are consistent with the age of the mice: at birth, biliary epithelium is still immature, and its maturation continues during the first years of life through hepatoblasts differentiation into periportal hepatocytes and adult hepatic progenitor cells (Strazzabosco and Fabris, 2012).

Overall these data may suggest a possible association of γ-H2AX peaks with development.



Figuure 3.3.1.3: 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.

3.3.1.4 Regulatory sites of transcription colocalize with γ-H2AX peaks

Since our results show an enrichment of OE and L peaks around gene TSS and, starting from the hypothesis that γ-H2AX is a marker of DSB, we decided to explore peak distribution with respect to the principal transcription regulatory regions involved in chromatin remodelling: Type II DNA topoisomerases (TOP2), RNA polymerase II (Pol II), CTCF and DNAse I.

Unfortunately, no public dataset was found for TOP2 genomic distribution, so we focused our attention on the other three sites. Moreover, among the available datasets, there was no information for olfactory epithelium. Olfactory bulb, cortex and whole brain tissues, were chosen instead, as they are the closer tissue types we could find data for.

3.3.1.4.1 DNAse I regulatory sites

DNAse I hypersensitive sites are indicators of open chromatin regions where it is possible to find many different regulatory elements including promoters, enhancers, insulators and silencers as well as TSS and regions of early replication (Boyle et al., 2008). According to our results, showing an enrichment of H2AX peaks in genic regions and around the TSS, and in agreement with what obtained in other works (Lensing et al., 2016) we expected to find our peaks distributed in proximity of these regions.
Indeed, about 70-75% of peaks for each sample fell within 1kb of a DNAse I. Looking at shuffled dataset, this was true only for 25% of peaks, confirming our prediction.

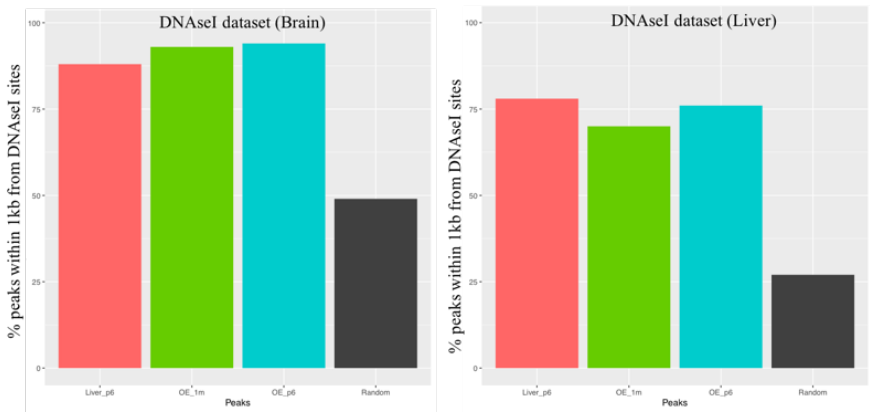


Figure 3.3.4.1a Percentage of peaks overalpping DNAse I sensible sites in different datasets. Lp6 (red); OE1m (green); OEp6 (blue); random peaks (black).

3.3.1.4.2 CTCF regulatory sites

A similar analysis was performed to investigate peak distribution with respect to CTCF binding sites. CTCF proteins participate in transcription mediating the formation of loops aimed to promote interactions between various regulatory regions, such as promoters and enhancers (Ong and Corces, 2014). CTCF peaks are known to be enriched in the surrounding of the TSS of genes that incur DSBs where they co-localize with DNAse I hypersensitivity sites and TOP2B (Uusküla-Reimand et al., 2016).

Similarly to what we observed for DNAse I, the percentage of the peaks falling within 1 kb from a CTCF binding site (almost 30%), was higher than the proportion of random peaks.

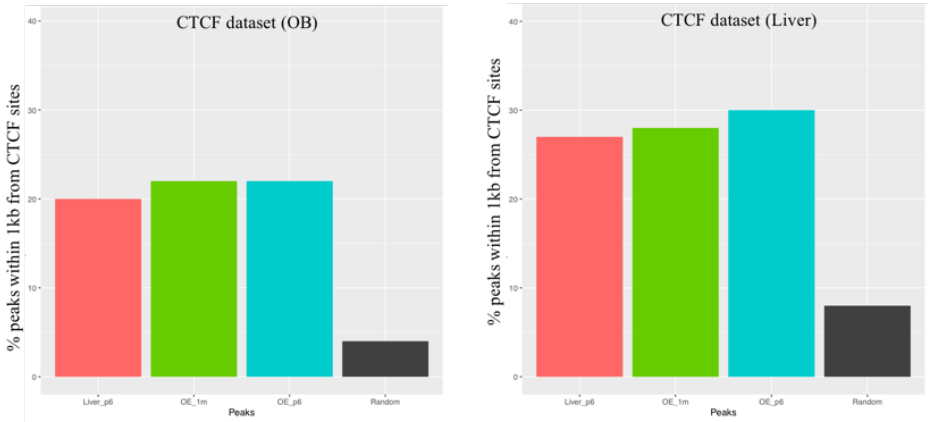


Figure 3.3.1.4.2 Percentage of peaks overalpping CTCF binding sites different datasets. Lp6 (red); OE1m (green); OEp6 (blue); random peaks (black).

3.3.1.4.3 Pol II regulatory sites

In order to test transcription-coupled enrichment of γ-H2AX without external damage we compared the distribution of H2AX peaks with RNA polymerase II binding sites.
The enzyme co-localizes with DNAse I hypersensitivity sites, showing a binding preference for open chromatin regions, and is reported to be in close association with CTCF and topoisomerase proteins at TSS. Endogenous γ-H2AX foci were reported at the TSS of genes undergoing Pol II stalling consequent to hyper-activated transcription (Seo et al., 2012).
About 10-15% of peaks for each dataset fall in a 1 kb interval surrounding Pol II binding sites, suggesting an association between peaks and actively transcribed genes. This motivated us to investigate the correlation between γ-H2AX and expression.

Similar results were obtained comparing the peaks with brain and olfactory bulb datasets.

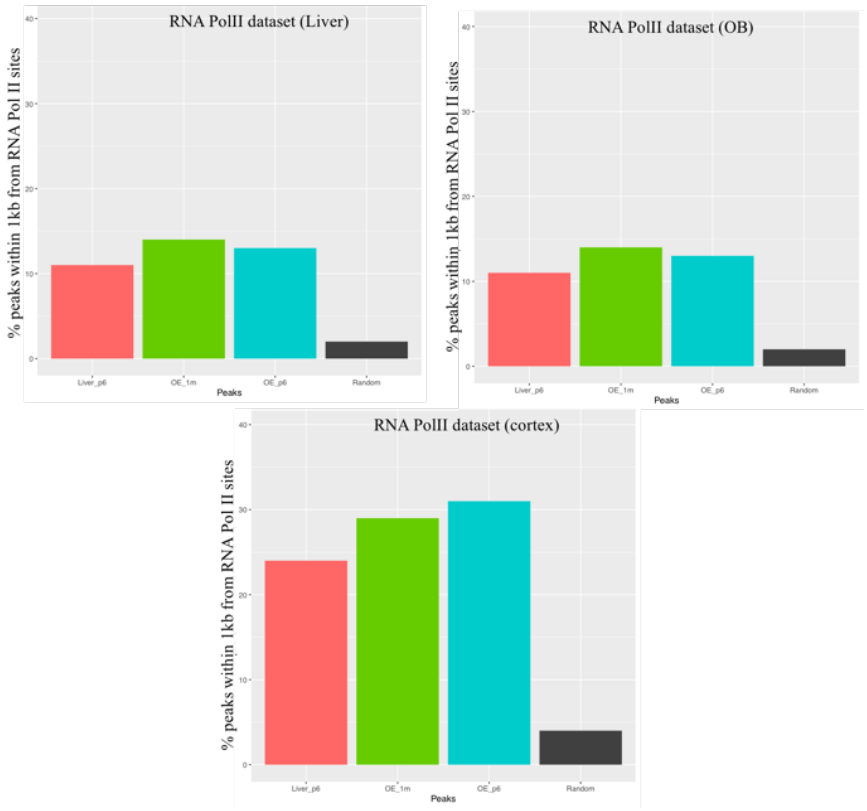


Figure 3.3.1.4.3 Percentage of peaks overalpping Pol II binding sites in different datasets. For each chart, the dataset used is indicated. Lp6 (red); OE1m (green); OEp6 (blue); random peaks (black).

3.3.1.5 γ-H2AX peaks correlation with gene expression

Transcription is, reportedly, one of the most important endogenous agents producing DSB (Kim and Jinks-Robertson, 2012, Schwer et al., 2016). The fact that H2AX signal is enriched in proximity of active chromatin markers supports this hypothesis. In this paragraph, we are going to investigate further the correlation between γ-H2AX peaks and gene expression in OE and L. To accomplish this task, we explored peak distribution within tissue specific, active TSS.

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

Here, we compared the proportion of peaks overlapping both Pol II binding sites and active TSS, for each sample, including the shuffle dataset. As expected most of the peaks which already overlap Pol II binding sites, overlap also an active TSS (about 71 to 75%).

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 (72%)
shuffle	745	447 (60%)	425 (57%)

Table 3.3.1.6 Pol II-overlapping peak correlation with active TSSs. The column "Pol II-overlapping peaks" indicates the number of γ-H2AX peaks overlapping Pol II binidng sites; the column "L_TSSs-overlapping peaks" indicates the Pol II-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 Pol II-associated peaks); similarly the column "OE_TSSs-overlapping peaks" indicates the number of Pol II-associated peaks overlapping one active TSS according to the OE dataset.

3.3.1.7 γ-H2AX peaks correlation with OE 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.

More in general, we noticed that OE-active TSSs in overlap with real peaks show a higher expression level than OE-active TSSs not in overlap with real peaks or in overlap with shuffle peaks. This trend is observable regardless the age or the tissue considered. OE samples presented a slightly higher, but not yet significant, expression level than OE-active TSSs in overlap with liver peaks.

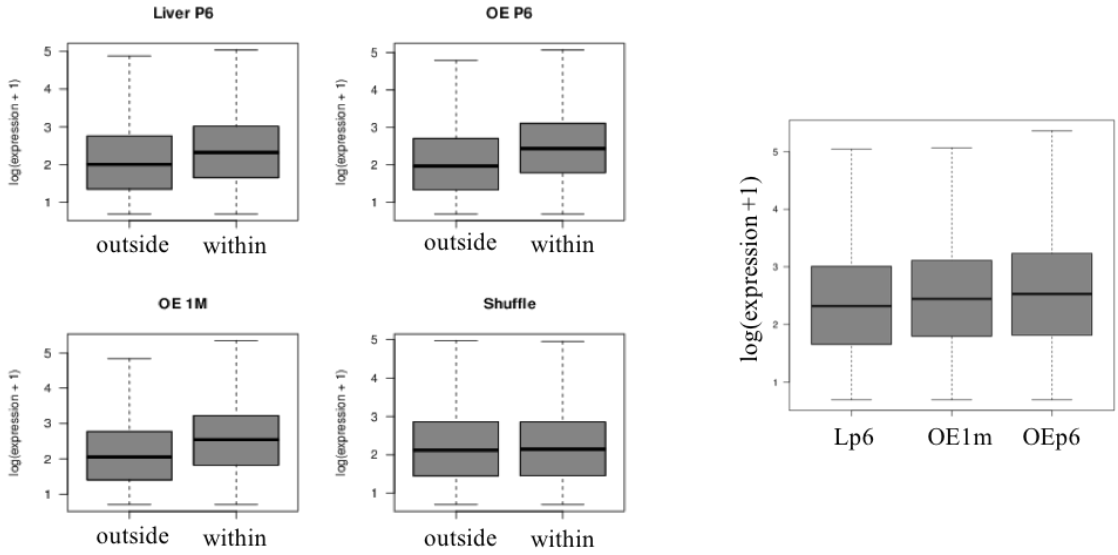


Figure 3.3.1.7a ChIP-seq peaks comparison with active OE-TSS. Left panel: per each sample, the expression values of OE-active TSSs falling within and without the peaks are compared. Right panel: fold induction of OE-active TSSs expression values is shown for each sample.

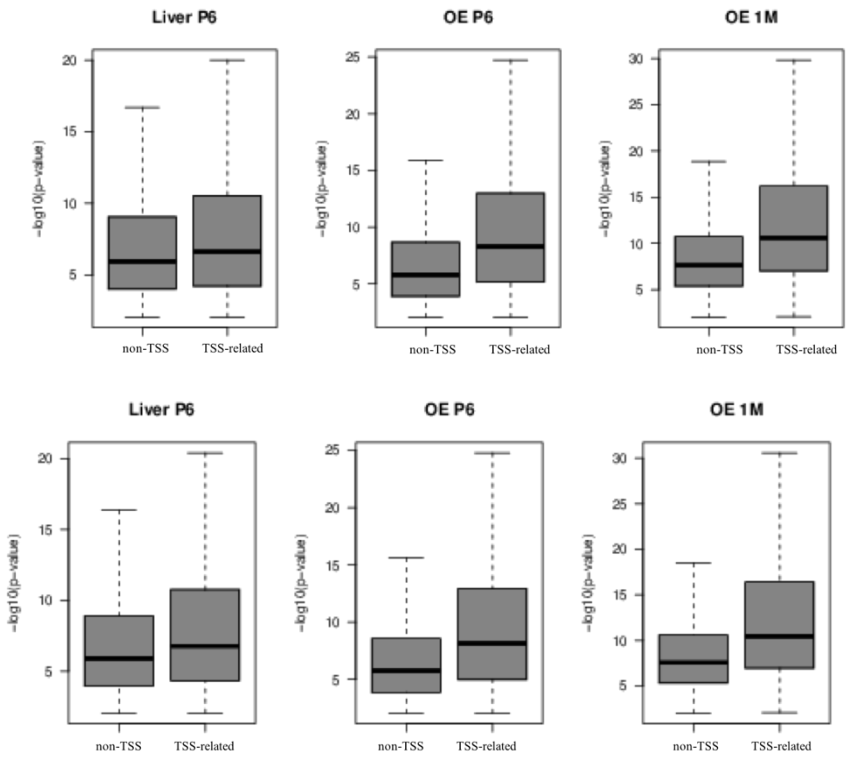


Figure 3.3.1.7b Comparison of peaks p-values with respect to active-TSSs. Top panels are referred to TSSs active in L and bottom panels to TSSs active in OE. For each sample the p-values of peaks associated to a TSS is compared with the p-values of peaks not associated to a TSS.

Given that most of the γ-H2AX peaks that overlap an active TSS in the L dataset also overlap an active TSS in the OE dataset, we hypothesize that often the same expressed gene is overlapped, regardless the tissue.

Sample dataset	L-TSSs overlapping peaks	OE-TSSs overlapping peaks	and OE-TSSs overlapping peaks
Lp6	4040	4490	3541 (87.6%)
OEp6	3119	3630	2802 (89.8%)
OE1m	152	1935	1352 (88.7%)

Table 3.3.1.7 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.

3.3.1.8 Chromatin segmentation

A chromatin segmentation analysis was performed with the aim of explaining the observed data according to different chromatin states of the mouse genome (see details in Methods and Materials).

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

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 3.3.1.8 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.

Chromatin states 2, 4 and 7 presented the strongest enrichment in the three datasets (ChromHMM). 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.

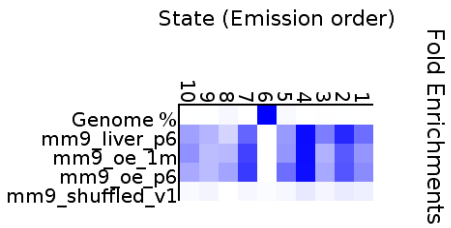


Figure 3.3.1.8 Fold enrichment between samples among different chromatin states. White and blue colored squares represent, respectively, no enrichment and tha maximum enrichment.

3.3.1.9 Peaks overlapping active enhancers

We previously found that γ-H2AX signal is associated with broadly transcribed genes particularly involved in age specific biological processes and it is further enriched near the TSSs of generally expressed genes. Consistent with this notion, we wanted to investigate whether γ-H2AX signal tends to occur within putative enhancers. Since the accepted chromatin signature for active enhancers is the presence of both the H3K4me1 and the H3K27ac marks (Calo and Wysocka, 2013), we extracted the set of peaks overlapping chromatin segments in state 3 and compared their distribution with respect to γ-H2AX real and random peaks.

Indeed, the proportion of real peaks co-localizing with γ-H2AX signal was higher than the proportion of random peaks.

Sample dataset	State 3 overlapping peaks
Lp6	**20.76%**
OEp6	**19.57%**
OE1m	**16.36%**
shuffle	**3.73%**

Table 3.3.1.9 Percentage of peaks overlapping active enhancers. For each sample, the percentage of peaks in overla with state-3 marks (active enhancers) is indicated.

3.3.1.10 γ-H2AX peaks enrichment for different classes of repeats

In order to further describe the genomic context of the detected γ-H2AX signal, we sought to characterize peak distribution with respect to different classes of repeats, including SINEs, LINEs, LTRs, DNA Transposons and satellite repeats. The plots showed the presence of a marked bias towards SINEs, coherent with the SINE gene-centered distribution (Elbarbary et al., 2016). The results are also consistent with the low concentration of LINE-1 elements in highly expressed genes regions (Graham and Boissinot, 2006).

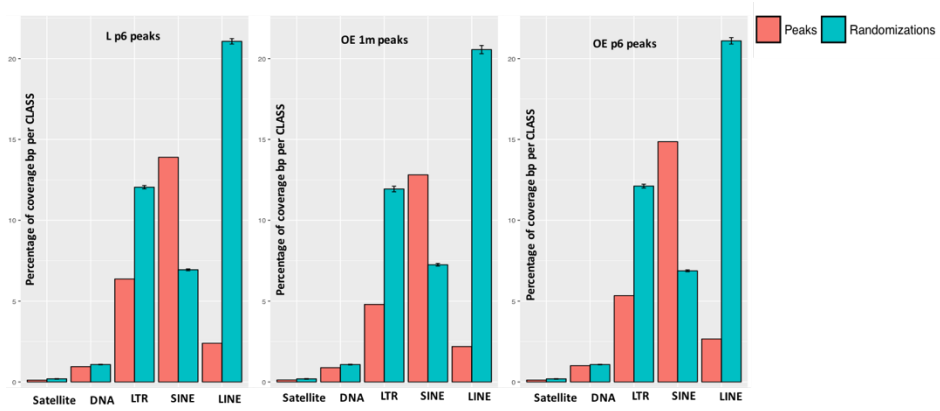


Figure 3.3.1.10 ChIP-seq peaks enrichment for different classes of repetitive elements. For all the samples the percentage of peaks coverage (bp) for each class of repeats is shown. Real peaks are shown in red and random peaks are shown in green. Black lines represent standard deviation.

Given the well-known involvement of the B1 and B2 SINE lineages in segmental duplications within the mouse genome (Jurka et al., 2005) we decided to investigate also γ-H2AX distribution around gene clusters.

3.3.1.11 Do DSBs initiate recombination events?

To summarize, our analysis revealed that γ-H2AX peaks:

- are enriched at the gene body;
- are enriched at the TSS;
- involve broadly expressed genes,
- are linked with the developmental stage of the tissue;
- are associated with DNAse I open chromatin regions;
- are associated with CTCF binding sites;
- are associated with Pol II binding sites;
- overlap active enhancers;
- are enriched for SINE elements.

All together these findings suggest us that γ-H2AX distribution in the genome is far from being random but the peaks are preferentially distributed near specific genomic regions.

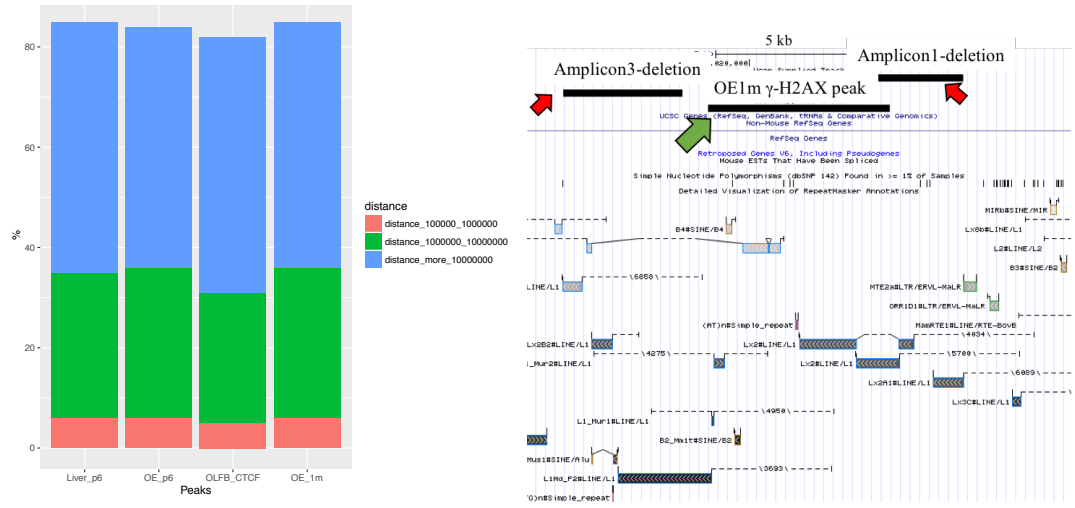
3.3.1.12 γ-H2AX peaks distribution with respect to gene clusters

Clustered genes are thought to be the result of extensive rearrangements on a common precursor such as duplication, unequal crossing-over, transposition, including retrotransposition and mutation.

One characteristic shared by many gene clusters is conserved CTCF sites at 5′ and 3′ of the loci (Kim at al., 2006, Tchurikov et. al 2013).

Comparing the distribution of CTCF peaks (public dataset) and CTCF-overlapping γ-H2AX peaks with respect to gene clusters considering different intervals, within and outside the clusters, we noticed that CTCF peaks were not enriched inside the clusters but distributed outside of them together with CTCF.

As expected, the peaks were depleted also inside *Olfr2* cluster, extensively studied in the previous result section, but preferentially distributed outside it. Interestingly, the only exception was one OE1m γ-H2AX peak falling inside the *Olfr2* locus, in proximity of amplicon 3 and amplicon 1 validated deletions. According to this finding we can speculate that the deletions described in the previous section may have arisen from a repair process of existing DNA lesions.



a b

Figure 3.3.1.11 Peak distribution with respect to gene clusters. (a) CTCF-overlapping γ-H2AX peaks distribution with respect to OR-gene clusters. CTCF-overlapping peaks were intersected with olfactory clusters to see the percentage of 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 plot. (b) 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.

4 Discussion

In this thesis we have presented results that suggest that LINE-1-mediated rearrangements have an important role in the formation of SVs in mammalian genomes. Leveraging next-generation sequencing technologies we investigated SVs resulting from three main mechanisms related to LINE-1 elements: LINE-1 integration in new genomic locations, LINE-1 mediated rearrangements and LINE-1 induced double strand breaks.

4.1 Analysis of FL-L1 elements in the genomes of AD post-mortem brains

In this study we introduced the Splinkerette Analysis of Mobile Elements (SPAM) technique, which was developed in order to unambiguously map only active FL-L1 elements present in the human genome, still able to mobilize and give rise to novel LINE-1 integration sites .

Inspired by the splinkerette PCR (spPCR) protocol, SPAM technique, starts with a PCR enrichment step which allows us to effectively target the boundary of a FL-L1 element and its flanking genomic region. Then it proceeds with an accurate bioinformatics pipeline to unambiguously map amplified LINE-1-containing genomic fragments and estimate their distribution.

In comparison with other similar techniques, SPAM has several advantages. First of all, most studies still focus on the 3' LINE-1 region (Ewing and Kazazian, 2010, Erwin et al., 2016), aiming at the identification of both the integer (1%) and the 5'truncated forms (99%) of LINE-1 elements present in the human genome. Importantly, focusing on the very 5' LINE-1 region, we significantly reduce the complexity of the targetable repeats

and we analyze only the small fraction of LINE-1s that retain their potential to impact genomic structure and gene expression: FL-L1 elements. Second, SPAM is a very precise technique, that allows LINE-1 Integration Sites (L1-IS) discovery with the single base precision. Third, SPAM is also a very efficient technique: we identified nearly 100% of the FL-L1 reference integration sites detectable with our primers with 0 mismatches. The fourth advantage of our technique is its scalability. Here we present SPAM as a valuable method to detect FL-L1 elements in the human genome, but, embedded in the methodology is the flexibility to perform the same technique in different organisms and considering different classes of TE.

After developing and optimizing the SPAM technique, we used it to discover annotated (AIS), polymorphic (PIS), and non-annotated FL-L1 integration sites in the human genome of Alzheimer's disease affected patients (AD) and controls (CTRL).

SPAM analysis was performed on a brain tissue (frontal cortex) and an extra brain tissue (kidney) of 4 AD patients and 4 CTRLs. This choice was motivated by the desire to study the effect of FL-L1 retrotransposition in a tissue severely affected by the disease, where active retrotransposition was already demonstrated to occur, and a control tissue, stable from the retrotransposition point of view and not directly related to the disease.

This study identified many novel LINE-1 insertions, the majority of which were private, single tissue insertions (72%). Anyway, in relation with the disease, the most interesting IS appeared to be the less abundant ones (3%): PIS. Polymorphic Integration Sites are recent insertions events present in a restricted number of individuals and therefore not annotated in the human reference genome, arising from currently active, mobile LINE-1 elements (Burns and Boeke, 2012). Accordingly, PIS, presented many characteristics in common with AIS compared to PIS: from MapFragments coverage, to genomic distribution.

The most intriguing result about PIS emerged from the differential integration analysis between AD and CTRL samples. Statistically significant differences were detected in differential coverage of MapFragments per specific MapClusters and MapFragments per Gene. Considering an FDR-adjusted p-value <0.1, 18 AIS, 42 PIS and 0 NIS presented a significantly different coverage in terms of MapFragments in the comparison between AD and CTRLs (frontal cortex and kidney together). Among the PIS, the IS showing the most significant result (since present in 3 out of 4 AD samples and none of the CTRLs)

was located in the intergenic region between the HLA-DRB1 and the HLA-DQA1 genes, inside the MHC class II locus, the most variable region in the human genome. This PIS (IS-HLA) corresponds to a known FL-L1 polymorphism of the human population and is reported to be present in the MANN and DBB MHC haplotypes (Horton et al., 2008). Unfortunately, the validation, performed on 410 AD samples and 239 CTRLs, actually did not show a different incidence of the IS-HLA, whose allelic frequency in AD samples was 0.138 and in CTRL samples 0.126. However, both the abovementioned PIS and an IS very close to it were found able to negatively influence the expression of its 4 closest genes and of a non-coding expressed anti-sense RNA. A possible explanation resides in the genomic position of the ISs. The FL-L1 insertion in the HLA locus, during its integration, disrupted an LTR sequence which was previously demonstrated to be a transcribed enhancer region (Thurman et al., 2012, Andersson et al., 2014) with a possible role as gene expression modulator. Transcription factor binding sites for TBP, TAF1 and Pol II in correspondence of the IS, further suggest a possible regulatory role of the disrupted transposable element.

In order to assess the implications of FL-L1 IS on gene function we examined the set of NIS and PIS associated genes with respect to GO functional category classifications. No significant enrichment was appreciable for PIS while some interesting hints about FL-L1 preferential IS sites emerged looking at the enrichments of NIS associated genes. In agreement with what observed in other works, our results suggests that retrotransposition effectively tags genes associated to neural differentiation, learning and memory (Temtamy et al., 2008). Interestingly, in the FC of AD patients, the most targeted loci are associated to signal transduction pathways and receptor activity, already known to be compromised in the elderly (Lu et al., 2004, Mobley et al., 2014). This is intriguing and suggestive of a pathogenic role of FL-L1 insertions in the brain. Signaling dysfunctions resulting from insertions in these regions, are reported in relation with Schizophrenia, and might promote the disease (Guffanti et al., 2016).

Looking more in general at FL-L1 IS distribution, regardless from the disease, unexpected outcomes emerge. The first surprising outcome of our analysis is the discovery of 45 FL-L1 IS in the mitochondrial genome: 20 IS (for a total of 61 mf) in AD affected patients and 42 IS (for a total of 144 mf) in CTRLs, most of them located in the gene body. The prevalence of gene associated IS is not surprising giving the high concentration of genes in the small mitochondrial (mt) genome. The unexpected finding was the presence of a

germinal IS (present in both the FC and the K of the same individual) that we managed to validate. We cannot exclude that we are dealing with Nuclear Mitochondrial DNA sequences (NUMTs) integrated in the human genome in proximity of a LINE-1 transposable element (Ju et al., 2015).

The second surprising outcome of this study is the incredibly high number of NIS detected in the kidney of AD and CTRL patients, almost five times higher than the number of NIS detected in the FC. The unexpectedly high level of somatic retrotransposition in this tissue is not correlated with AD. It may be the consequence of aging, of other pathologies affecting the individuals object of our study or it may be linked to the activity of the staminal niches that maintain and preserve the renal tissue throughout life (Rinkevich et al., 2014, Hishikawa et al., 2015).
Several lines of evidence suggest that LINE-1 elements are active in the normal somatic tissues of the brain and in cancers (Shukla et al., 2013, Doucet-O'Hare et al., 2015), but the extent of somatic activity in other normal tissues is still largely unexplored. This is mostly due to the technical complexity of confidently detecting (and validating) in bulk tissue non-reference LINE-1 retrotransposition events unique to a single cell. With our method, we were able to discover thousands of unpredicted FL-L1 non-annotated IS. But, since our technique is not quantitative, this number may be just suggestive of an incredibly high level of unexplored LINE-1 mosaicism in human tissues. A potential perspective to determine the real rate of somatic transposition would be to adapt targeted LINE-1 discovery methods, such as SPAM, to single cell sequencing technologies (Evrony et al., 2012, Upton et al., 2015). Here, to increase the chances to validate very rare events (e.g. somatic NIS covered by 2 or a few MapFragments) in bulk genomic DNA, we took advantage of Digital Droplet PCR (ddPCR). Unfortunately, even employing this remarkably sensitive instrument, we were not able to validate any somatic IS.

Thus, in order to evaluate the content of potentially active LINE-1s in different tissues of AD and CTRL individuals we performed a TaqMan based CNV analysis. The experiment, performed on different brain tissues from different cohorts of individuals (Spanish, and Brazilian) as well as in the kidney, revealed the presence of a lower amount of FL-L1s in the AD frontal cortex, cerebellum and kidney tissues as compared to controls, while, surprisingly, no differences could be observed in the hippocampus, which

is known to be heavily compromised by the disease. Again, unexpectedly, kidney resulted to be unstable from the retrotransposon-mobilization point of view.

To understand if the loss of FL-L1 in the tissues of AD affected patients could be attributable to a loss of larger genomic fragments resulting from genomic rearrangements, we took advantage of high density arrays to compare the occurrence of FL-L1 elements in correspondence of genomic variations detected in AD and CTRL patients. The analysis, performed using the Illumina Infinium high-density chip revealed a considerable amount of deleted genome sequence in the brain. In the FC of the Spanish cohort in particular, we were able to detect statistically significant differences in the coverage of the putative FL-L1 elements targeted by the Taqman assay, which resulted to be significantly enriched in heterozygous deletions in AD patients, as compared to CTRL samples. Unfortunately, we didn't reach the significance level in the Brazilian cohort were the same trend was observed. Since age is a very important variable in a pathology like LOAD, our concern is that it may have influenced the result. Indeed, while Brazilians AD patients and CTRLs are aged matched (~80 years old), Spanish CTRLs are younger (70 ± 8 years old) than Spanish AD patients (~80 years old).

To summarize, our study revealed that FL-L1 polymorphisms can be a relevant source of structural variants associated to AD risk, and suggested that somatic LINE-1 retrotransposition might occur more broadly than previously appreciated. A key goal for the future will be to understand exactly how these polymorphisms affect LINE-1-mediated diseases, including Alzheimer.

4.2 Analysis of LINE-1 mediated SVs in Olfr2 locus

While recent advances in sequencing technology facilitated the process of SV discovery in multiple genomes, the phenotypic impact of most of these SVs remains unclear. For example, one possible function of DNA rearrangements could be controlling OR gene choice in mouse olfactory epithelium. Mouse OSN express, from a family of more than 1000 genes, one single OR gene, mono-allelically (Chess et al., 1994). Various hypotheses have been made about the control mechanism of OR gene expression but so far, little is known about it. In our opinion genomic organization is an important determinant of OR gene choice via DNA rearrangements in the olfactory neurons (Kratz et al., 2002, Clowney et al., 2012, Hoppe et al., 2006). In the present work, therefore, we investigated the involvement of locus-specific mechanisms for the control of OR gene expression.

To perform this study, we characterized the SV profile of a locus distinguished by a very high abundance of repetitive elements, especially LINE-1 retrotransposons: mouse Olfr2 locus, in order to find the putative SV that regulates the expression of the receptor. Olfr2 was chosen among other receptors due to the availability of a transgenic mouse line (B6;129-Olfr2-GFP mice) where a GFP gene was inserted at the 3' of the Olfr2 gene. This made cells that naturally activated the transcription of Olfr2 fluorescent, and easy to detect. A total of 10 cells expressing the receptor (GFP+) were collected and their material was amplified via MDA, to increase the DNA starting quantity. As a control, we used bulk genomic DNA that did not undergo MDA amplification.

Then, we sequenced 50 kb of sequence around Olfr2 transcription start site (TSS) starting from cells expressing the receptor (that we called MDA sample) and from bulk genomic DNA (that we called OE sample) combining PacBio single molecule sequencing for reliable mapping across repeat expansions with a complementary high-fidelity paired-end Illumina sequencing for accurate identification of breakpoints in a locus where a very high repeat concentration made read mapping really challenging. Indeed, inaccurate mapping in repetitive regions and poor coverage, can often induce variant callers to return false positive results. In order to overcome the risk of dealing with false SV calls produced by an Illumina based tool such as Pindel, we retained only SVs longer than 50 bp, covered by at least 5 reads and among them, SVs supported by at least one PacBio read.

Surprisingly, the analysis revealed hundreds of heterozygous structural variants in the vicinity of the locus, among which, deletions were the most abundant. This gives us hint of the incredible complexity of the region. Interestingly, the number of reads supporting the non-annotated variant was always much lower than the number of reads supporting the annotated reference sequence. This may suggest the presence of somatic mosaicism for deletions occurring in one or just a few cells. Potentially, among the somatic genome variations detected in the neurons expressing the receptor (MDA samples), there are also the ones which explain the functional diversity of the cells. Because of the filters applied (see methods), the vast majority of deletions detected with Pindel was discarded. Relaxing the stringency is a way to improve the sensitivity; however, this may come at the cost of specificity. Considering that we were able to validate also deletions without PB coverage and supported by a number of Illumina reads lower than the imposed threshold of 5, we realized that relaxing the stringency of the imposed parameters (e.g. decrease blastn word size and increase the e-value in order to increase the number of matching PacBio and Illumina reads) would perhaps result in a more accurate profile of this region. While, single cell sequencing, probably would help to reduce the complexity of the SV pattern in the locus.

In order to be sure that potentially interesting SVs detected in the samples expressing the receptor are not merely the result of amplification induced artifacts we took some precautions. First, to exclude MDA induced artifacts (Lasken and Stockwell, 2007, Treiber and Waddell, 2017), we focused on deletions present in both MDA samples and OE sample (our technical and biological control); deletions present in OE samples cannot be MDA induced artifacts because no MDA amplification was necessary for bulk gDNA. Second, to exclude artifacts resulting from the initial Olfr2 locus PCR amplification, we performed a validation also on total MDA amplified starting material prepared for PCR amplification. Deletions passing these filters, like the one present in amplicon three, represent good candidate regulatory features.

Gross genomic rearrangements, such as the ones examined in this thesis, are often the result of other mechanisms mediated by genomic structural features. The high number of retrotransposable element in the locus and the presence of matching repeats in the immediate vicinity of the sequenced breakpoints made us think about a possible LINE involvement in deletion formation. Repetitive elements have been implicated in genome

rearrangements (Han et al., 2008), and LINE elements themselves, providing large regions of sequence similarity, serve as recombination templates.

On the other hand, it has previously been suggested that olfactory receptor genes flourish in repeat rich and rearrangement prone regions, like sub-telomeres and peri-centromeres, (Linardopoulou et al., 2005). Interestingly, LINE enrichment isn't just a peculiarity of our locus. We fond this trend persisting in monoallelically expressed genes. In particular OR-clusters, show a peculiar enrichment for this class of retrotransposons compared with other classes of repeats.

Our findings suggest also an important role played by MMEJ for the generation of the deletions. Microhomology-mediated end joining is an alternative non-homologous DSB repair pathway that relies on microhomologies (5–25 bp) on either side of the break to join and stabilize the broken DNA. As a consequence of this repair mechanism, the DNA sequence between the microhomologies is often deleted and typical direct repetitive sites (DRS) can be retrieved at the boundaries of the SV. Although chromosomal breakage could originate anywhere in the genome, it is tempting to speculate that LINE sequences themselves may be involved in break formation (Erwin et al., 2016b). Indeed, MMEJ repair resolves most of DSBs left by the endonuclease cutting activity during LINE-1 mobilization events, and was recently proposed as the mechanism responsible for the formation of kb long deletions in the brain (Erwin et al., 2016a). Due to the high density of LINE-1 elements in Olfr2 region, we started wondering if the rearrangements observed in Olfr2 locus could be the result of this repair mechanism as well.
In order to elucidate it, we further reduced the complexity of the deletion pattern, clustering together SVs overlapping the same LINE-1 elements annotated in the reference genome. Interestingly, the 98% of the clustered deletions overlapped a repetitive element at both their 3' and 5' breakpoints. Among them, one fourth displayed more than 70% homology between LINE elements present at the left and right breakpoint. This is in sharp contrast with what observed at the breakpoints of random deletions distributed in the same chromosome. In fact, only the 18% of random deletions harbored repetitive elements at both breakpoints. This may suggest a scenario where LINE elements restore DSBs occurring within repeated sequences, bringing distant DNA segments close to each other and promoting their recombination. LINE sequences themselves therefore, may be at the same time involved in break formation and break resolution.

It must be highlighted that most LINE-1 elements annotated in Olfr2 locus object of our study are truncated, and among the clustered deletions, only one involves a FL-L1 at one boundary. For this reason, we propose that MMEJ could be the DSB resolving mechanism of chromosomal breaks and rearrangements mediated by recombination events occurring between truncated retrotransposons (Han et al., 2008).

The presence of typical hallmarks associated with the real deletions but not with the random deletions, suggested us that the same principal mechanism is operating in the formation of all the deletions. First of all, direct repeated homologous sequences (DRS) ranging from 4 bp up to 17 bp were present at the breakpoints of the clustered deletions (4 bp up to 17 bp) and were significantly longer than DRS occurring at the boundaries of random deletions (pvalue < 2 e-16). Second, some of the DRS motifs at the breakpoints are already known for being significantly associated with structural break resolution. Examples are: meiotic recombination hotspots, polymerase beta frameshift hotspots (Chuzhanova et al., 2009), hamster and human APRT deletion hotspots (Smith and Adair, 1996), FOXO recognition elements (Alkhatib et al., 2012) , indel super - hotspot motifs (Ball et al., 2005), and immunoglobulin heavy chain class switch repeats (Abeysinghe et al., 2003, Chuzhanova et al., 2009). Third, GC content of the real deletions DRS was significantly higher than GC content of the random deletions DRS. Base composition at the breakpoint regions is an important hallmark of MMEJ since GC rich motifs increase the stability of the pairs during the annealing process, facilitating it (Verdin et al., 2013, Kent et al., 2015). Fourth, DRS were validated after Sanger re-sequencing at the borders of all validated deletions, even the ones lacking PacBio coverage. This result extends the above result to the hundreds of deletions only supported by Illumina reads that we temporarily set aside.

The existence of a control mechanism which regulates Olfr2 expression remains an open question which would need to be followed up in further experiments. Nevertheless, we have discovered a number of strong associations between deletions, repetititive elements and MMEJ DSB resolving mechanism that led to an improved picture of Olfr2 locus.

4.3 Chip Seq analysis of endogenous γ-H2AX in mouse olfactory epithelium and liver

Reading this dissertation, becomes evident how the genome is an unstable place, prone to DNA breakage and rearrangements where SVs continuously arise as a result of recombination events and repair process of existing DNA lesions. Endogenous agents producing DSBs such as active transposable elements (McConnell et al., 2013, Cai et al., 2014), transcription (Kim and Jinks-Robertson, 2012), replication stress (Zeman and Cimprich, 2014), and oxidative stress (Woodbine et al., 2011) continuously attack DNA integrity. This occurs ubiquitously in all the cells of an organism but there moments when the damage becomes more severe (Jung and Pfeifer, 2015), there are tissues which are more exposed, and there are genomic regions which are more fragile than others (Linardopoulou et al., 2005). In particular, the high activity and direct exposure to external environment of OE, result in the progressive accumulation of DNA damage and the loss of genomic integrity consequent to inefficient repair mechanisms. Constantly regenerating mature OSNs, OE is an ideal substrate for LINE-1 mobilization. This, potentially, increases even more the risk of DSBs in the tissue. To further investigate this concept, we performed a genome-wide profile of endogenous mouse DNA DSBs. DNA DSBs represent a major threat to genomic stability leading to dangerous mutagenic rearrangements or apoptosis (Lieber, 2010). In spite of this, our knowledge of endogenous DSBs is still limited. For this reason, understanding the sensitivity of the genome to the various DNA insults is as much as important as understanding how, potentially harmful events such as DSBs become, instead, instrumental to genome regulation.

To this purpose, we performed a chromatin immunoprecipitation and sequencing (ChIP-Seq) analysis of γ-H2AX. The phosphorylation of the histone occurs in response to DSB as an early signal to increase DNA accessibility and recruit the different repair proteins necessary to initiate the repair of DSBs (Turinetto and Giachino, 2015). It should be noted that this marker, is not an exclusive indicator of DSB. DSB-independent background foci may be caused by ATR-mediated H2AX phosphorylation in growing cells with dis-regulated DNA metabolism and in response to heat (Wang et al., 2014). Although, it is still the best marker based on its cell phase-independent formation, tight correlation with repair kinetics and repair pathway independence. Importantly, as γ-H2AX is formed de

novo, it is the a more reliable DSB marker than other DNA repair proteins that are present in the cell even when there is no DNA damage. For this reason, this marker became the choice of most researchers aiming to profile the effects of DSB inducing agents. Little is known about the differential distribution of γ-H2AX marker for DSB throughout the genome at physiological conditions. To address this question, we used the ChIP-seq technique to profile the chromosomal distribution of γ-H2AX in C57BL/6J mice OE (at p6 and 1m) and L (at p6). Liver tissue was chosen as a control since it is characterized by different cell types with different expression pattern. Moreover, a wide range of public datasets is available for this extensively studied tissue.

Regarding the distribution of the peaks, the three samples presented overall a similar behaviour. In agreement with other works, our results show that the peaks in mouse genome are mostly found in correspondence of gene bodies, peaking from 0 to 10kb downstream active transcription start sites (**TSS**) and enriched in GC-rich regions. This distribution pattern, suggests the presence of transcription associated DSBs in proximity of actively transcribed genes. DSBs at gene promoters have been proposed to reduce supercoiling and induce chromatin relaxation to facilitate transcription initiation and full expression of long genes (Madabhushi et al., 2015b). This finding is further bolstered by the fact that most of the peaks co-localize with the principal transcription regulatory regions involved in chromatin remodelling: RNA polymerase II (Pol II), CTCF and DNAse I, and are enriched in proximity of histone modifications found on active promoters (H3K4me3, H3K9ac) and gene bodies (H3K36me3, H3K79me2). Transcription is, reportedly, one of the most important endogenous agents producing DSBs (Kim and Jinks-Robertson, 2012, Schwer et al., 2016). The fact that H2AX signal is enriched in proximity of active chromatin markers supports our hypothesis.
Moreover, characterizing peak distribution with respect to different classes of repeats, we could appreciated a peak enrichment towards SINE elements. This is coherent with the SINE gene-centered distribution (Elbarbary et al., 2016). Therefore, given the well-known involvement of the B1 and B2 SINE lineages in segmental duplications within the mouse genome (Jurka et al., 2005) we decided to investigate also γ-H2AX distribution around gene clusters, especially OR clusters. OR reside in repeat rich regions, resulting from extensive genomic rearrangements on a common precursor, where the patchwork of repeats still retains the potential to induce DSB and recombination. One characteristic shared by many gene clusters is the presence of conserved CTCF sites at 5′ and 3′ of the

loci (Kim at al., 2006, Tchurikov et. al 2013). CTCF peaks are known to be enriched in the surrounding of the TSS of genes that incur DSBs where they co-localize with DNAse I hypersensitivity sites and TOP2B (Uusküla-Reimand et al., 2016, Madabhushi et al., 2015). Comparing the distribution of CTCF peaks (public dataset) and CTCF-overlapping γ-H2AX peaks with respect to gene clusters considering different intervals, within and outside the clusters, we noticed that CTCF peaks were not enriched inside the clusters but distributed outside of them together with CTCF. As expected, the peaks were preferentially distributed also outside *Olfr2* cluster, extensively described in the previous Results section, and depleted on the inside. This is in agreement with previous findings suggesting a possible role of DSB in transcription regulation. In particular, CTCF proteins facilitate transcription mediating the formation of loops aimed to promote interactions between various regulatory regions, such as promoters and enhancers (Ong and Corces, 2014). A link to tissue specific expression emerged from the GO functional annotation. 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 while among the most significantly enriched terms in L were related to biliary epithelium differentiation and development. Overall, the GO results appeared also always consistent with the age of the mice: suggesting a possible association of γ-H2AX peaks with development.

In conclusion, alogether our results provide strong evidence that the observed DSBs are mostly related to gene expression under physiological conditions. On the other hand, we did not observe the expected abundance of DSBs in OE compared to liver. We suppose that LINE-1 mediated DSBs are obscured by other sources of DNA damage, like transcription, that have to be at work as well in the examined tissues.

5 Conclusion

We have presented the effects on genome stability of the most abundant retrotransposon family in mammals: LINE-1 elements. Our goal was to determine how LINE-1 mediated SV may impact health and disease.

Concerning Alzheimer's disease, at the LINE-1 insertion level, we identified a FL-L1 polymorphism which may act as a gene expression modulator with potential implications in AD susceptibility. While, concerning the regulation of OR genes, the high density of LINE-1 elements in OR-loci let us speculate about a possible regulatory function of retrotransposable elements for the expression of OR genes.

Expanding our scope from LINE-1 direct insertion to retrotransposition-independent LINE-1 mediated SV, we observed a significant decrease of FL-L1s in the tissues of AD affected patients, probably consequent to deletions involving sequences containing LINE-1 fragments. LINE-1 related heterozygous deletions appeared to be also the predominant SV observed in Olfr2 locus.

Overall, thousands of somatic LINE-1 insertions have been recovered in human FC and K and hundreds of LINE-1 mediated deletions have been recovered in mouse OE, confirming LINE-1 activity in metabolically active tissues where extensive DNA damage results from active transcription. This may suggest a scenario where LINE-1 elements may be at the same time involved in break formation and break resolution. On one side promoting the formation of DSBs and on the other side restoring DSB through recombination.

Taken together, our results suggest that both active and non-active LINE-1s play an important role in shaping the chromatin landscape and regulating gene expression under physiological and pathological conditions.

Bibliography

Abeysinghe, S.S., Chuzhanova, N., Krawczak, M., Ball, E.V., Cooper, D.N., 2003. Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs. Hum. Mutat. 22, 229–244. doi:10.1002/humu.10254

Al-Gubory, K.H., 2014. Environmental pollutants and lifestyle factors induce oxidative stress and poor prenatal development. Reprod. Biomed. Online 29, 17–31. doi:10.1016/j.rbmo.2014.03.002

Alkan, C., Coe, B.P., Eichler, E.E., 2011. Genome structural variation discovery and genotyping. Nat. Rev. Genet. 12, 363–376. doi:10.1038/nrg2958

Alkhatib, A., Werner, M., Hug, E., Herzog, S., Eschbach, C., Faraidun, H., Köhler, F., Wossning, T., Jumaa, H., 2012. FoxO1 induces Ikaros splicing to promote immunoglobulin gene recombination. J. Exp. Med. 209, 395–406. doi:10.1084/jem.20110216

Allen, E., Horvath, S., Tong, F., Kraft, P., Spiteri, E., Riggs, A.D., Marahrens, Y., 2003. High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes. Proc. Natl. Acad. Sci. U. S. A. 100, 9940–9945. doi:10.1073/pnas.1737401100

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. doi:10.1016/S0022-2836(05)80360-2

Andersen, J.K., 2004. Oxidative stress in neurodegeneration: cause or consequence? Nat. Med. 10 Suppl, S18-25. doi:10.1038/nrn1434

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F.O., Jørgensen, M., Andersen, P.R., Bertin, N., Rackham, O., Burroughs, A.M., Baillie, J.K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C.J., Meehan, T.F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C.O., Heutink, P., Hume, D.A., Jensen, T.H., Suzuki, H., Hayashizaki, Y., Müller, F., Forrest, A.R.R., Carninci, P., Rehli, M., Sandelin, A., 2014. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461. doi:10.1038/nature12787

Antony, J.M., DesLauriers, A.M., Bhat, R.K., Ellestad, K.K., Power, C., 2011. Human endogenous retroviruses and multiple sclerosis: Innocent bystanders or disease determinants? Biochim. Biophys. Acta BBA - Mol. Basis Dis., Molecular Basis of Multiple Sclerosis 1812, 162–176. doi:10.1016/j.bbadis.2010.07.016

Aravin, A.A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Toth, K.F., Bestor, T., Hannon, G.J., 2008. A piRNA Pathway Primed by Individual Transposons Is Linked to De Novo DNA Methylation in Mice. Mol. Cell 31, 785–799. doi:10.1016/j.molcel.2008.09.003

Arner, E., Daub, C.O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drabløs, F., Lennartsson, A., Rönnerblad, M., Hrydziuszko, O., Vitezic, M., Freeman, T.C., Alhendi, A.M.N., Arner, P., Axton, R., Baillie, J.K., Beckhouse, A., Bodega, B., Briggs, J.,

Brombacher, F., Davis, M., Detmar, M., Ehrlund, A., Endoh, M., Eslami, A., Fagiolini, M., Fairbairn, L., Faulkner, G.J., Ferrai, C., Fisher, M.E., Forrester, L., Goldowitz, D., Guler, R., Ha, T., Hara, M., Herlyn, M., Ikawa, T., Kai, C., Kawamoto, H., Khachigian, L.M., Klinken, S.P., Kojima, S., Koseki, H., Klein, S., Mejhert, N., Miyaguchi, K., Mizuno, Y., Morimoto, M., Morris, K.J., Mummery, C., Nakachi, Y., Ogishima, S., Okada-Hatakeyama, M., Okazaki, Y., Orlando, V., Ovchinnikov, D., Passier, R., Patrikakis, M., Pombo, A., Qin, X.-Y., Roy, S., Sato, H., Savvi, S., Saxena, A., Schwegmann, A., Sugiyama, D., Swoboda, R., Tanaka, H., Tomoiu, A., Winteringham, L.N., Wolvetang, E., Yanagi-Mizuochi, C., Yoneda, M., Zabierowski, S., Zhang, P., Abugessaisa, I., Bertin, N., Diehl, A.D., Fukuda, S., Furuno, M., Harshbarger, J., Hasegawa, A., Hori, F., Ishikawa-Kato, S., Ishizu, Y., Itoh, M., Kawashima, T., Kojima, M., Kondo, N., Lizio, M., Meehan, T.F., Mungall, C.J., Murata, M., Nishiyori-Sueki, H., Sahin, S., Nagao-Sato, S., Severin, J., de Hoon, M.J.L., Kawai, J., Kasukawa, T., Lassmann, T., Suzuki, H., Kawaji, H., Summers, K.M., Wells, C., FANTOM Consortium, Hume, D.A., Forrest, A.R.R., Sandelin, A., Carninci, P., Hayashizaki, Y., 2015. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science 347, 1010–1014. doi:10.1126/science.1259418

Arnold, C., Hodgson, I.J., 1991. Vectorette PCR: a novel approach to genomic walking. PCR Methods Appl. 1, 39–42.

Arun, R., Dhivya, S., Abraham, S.K., Premkumar, K., 2016. Low-dose chemotherapeutic drugs induce reactive oxygen species and initiate apoptosis-mediated genomic instability. Toxicol. Res. 5, 547–556. doi:10.1039/C5TX00391A

Ayarpadikannan, S., Kim, H.-S., 2014. The Impact of Transposable Elements in Genome Evolution and Genetic Instability and Their Implications in Various Diseases. Genomics Inform. 12, 98–104. doi:10.5808/GI.2014.12.3.98

Bailey, J.A., Church, D.M., Ventura, M., Rocchi, M., Eichler, E.E., 2004. Analysis of Segmental Duplications and Genome Assembly in the Mouse. Genome Res. 14, 789–801. doi:10.1101/gr.2238404

Bakunina, N., Pariante, C.M., Zunszain, P.A., 2015. Immune mechanisms linked to depression via oxidative stress and neuroprogression. Immunology 144, 365–373. doi:10.1111/imm.12443

Balato, A., Raimondo, A., Balato, N., Ayala, F., Lembo, S., 2016. Interleukin-33: increasing role in dermatological conditions. Arch. Dermatol. Res. 308, 287–296. doi:10.1007/s00403-016-1638-7

Ball, E.V., Stenson, P.D., Abeysinghe, S.S., Krawczak, M., Cooper, D.N., Chuzhanova, N.A., 2005. Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. Hum. Mutat. 26, 205–213. doi:10.1002/humu.20212

Basu, U., Franklin, A., Schwer, B., Cheng, H.-L., Chaudhuri, J., Alt, F.W., 2009. Regulation of Activation-Induced Cytidine Deaminase DNA Deamination Activity in B Cells by Serine-38 Phosphorylation. Biochem. Soc. Trans. 37, 561–568. doi:10.1042/BST0370561

Beck, C.R., Garcia-Perez, J.L., Badge, R.M., Moran, J.V., 2011. LINE-1 Elements in Structural Variation and Disease. Annu. Rev. Genomics Hum. Genet. 12, 187–215. doi:10.1146/annurev-genom-082509-141802

Belancio, V.P., Deininger, P.L., Roy-Engel, A.M., 2009. LINE dancing in the human genome: transposable elements and disease. Genome Med. 1, 97. doi:10.1186/gm97

Bitanihirwe, B.K.Y., Woo, T.-U.W., 2011. Oxidative Stress in Schizophrenia: An Integrated Approach. Neurosci. Biobehav. Rev. 35, 878–893. doi:10.1016/j.neubiorev.2010.10.008

Bochman, M.L., Paeschke, K., Zakian, V.A., 2012. DNA secondary structures: stability and function of G-quadruplex structures. Nat. Rev. Genet. 13, 770–780. doi:10.1038/nrg3296

Bochukova, E.G., Roscioli, T., Hedges, D.J., Taylor, I.B., Johnson, D., David, D.J., Deininger, P.L., Wilkie, A.O.M., 2009. Rare mutations of FGFR2 causing apert syndrome: identification of the first partial gene deletion, and an Alu element insertion from a new subfamily. Hum. Mutat. 30, 204–211. doi:10.1002/humu.20825

Boeva, V., Lermine, A., Barette, C., Guillouf, C., Barillot, E., 2012. Nebula--a web-server for advanced ChIP-seq data analysis. Bioinformatics 28, 2517–2519. doi:10.1093/bioinformatics/bts463

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. doi:10.1093/bioinformatics/btu170

Bollati, V., Galimberti, D., Pergoli, L., Dalla Valle, E., Barretta, F., Cortini, F., Scarpini, E., Bertazzi, P., Baccarelli, A., 2011. DNA methylation in Repetitive Elements and Alzheimer disease. Brain. Behav. Immun. 25, 1078–1083. doi:10.1016/j.bbi.2011.01.017

Bonner, W.M., Redon, C.E., Dickey, J.S., Nakamura, A.J., Sedelnikova, O.A., Solier, S., Pommier, Y., 2008. γH2AX and cancer. Nat. Rev. Cancer 8, 957–967. doi:10.1038/nrc2523

Bourc'his, D., Bestor, T.H., 2004. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature 431, 96–99. doi:10.1038/nature02886

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., Crawford, G.E., 2008. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell 132, 311–322. doi:10.1016/j.cell.2007.12.014

Braak, H., Braak, E., 1995. Staging of Alzheimer's disease-related neurofibrillary changes. Neurobiol. Aging 16, 271-278-284.

Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H., Moran, J.V., Kazazian, H.H., 2003. Hot L1s account for the bulk of retrotransposition in the human population. Proc. Natl. Acad. Sci. 100, 5280–5285. doi:10.1073/pnas.0831042100

Budworth, H., McMurray, C.T., 2013. A Brief History of Triplet Repeat Diseases. Methods Mol. Biol. Clifton NJ 1010, 3–17. doi:10.1007/978-1-62703-411-1_1

Bundo, M., Toyoshima, M., Okada, Y., Akamatsu, W., Ueda, J., Nemoto-Miyauchi, T., Sunaga, F., Toritsuka, M., Ikawa, D., Kakita, A., Kato, M., Kasai, K., Kishimoto, T., Nawa, H., Okano, H., Yoshikawa, T., Kato, T., Iwamoto, K., 2014. Increased L1 Retrotransposition in the Neuronal Genome in Schizophrenia. Neuron 81, 306–313. doi:10.1016/j.neuron.2013.10.053

Burns, K.H., 2017. Transposable elements in cancer. Nat. Rev. Cancer 17, 415–424. doi:10.1038/nrc.2017.35

Burns, K.H., Boeke, J.D., 2012. Human Transposon Tectonics. Cell 149, 740–752. doi:10.1016/j.cell.2012.04.019

Burwinkel, B., Kilimann, M.W., 1998. Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease1. J. Mol. Biol. 277, 513–517. doi:10.1006/jmbi.1998.1641

Cai, X., Evrony, G.D., Lehmann, H.S., Elhosary, P.C., Mehta, B.K., Poduri, A., Walsh, C.A., 2014. Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. Cell Rep. 8, 1280–1289. doi:10.1016/j.celrep.2014.07.043

Calo, E., Wysocka, J., 2013. Modification of enhancer chromatin: what, how and why? Mol. Cell 49. doi:10.1016/j.molcel.2013.01.038

Campos-Sánchez, R., Cremona, M.A., Pini, A., Chiaromonte, F., Makova, K.D., 2016. Integration and Fixation Preferences of Human and Mouse Endogenous Retroviruses Uncovered with Functional Data Analysis. PLoS Comput. Biol. 12. doi:10.1371/journal.pcbi.1004956

Carvalho, C.M.B., Lupski, J.R., 2016. Mechanisms underlying structural variant formation in genomic disorders. Nat. Rev. Genet. 17, 224–238. doi:10.1038/nrg.2015.25

Casacuberta, E., González, J., 2013. The impact of transposable elements in environmental adaptation. Mol. Ecol. 22, 1503–1517. doi:10.1111/mec.12170

Chen, J.-M., Chuzhanova, N., Stenson, P.D., Férec, C., Cooper, D.N., 2005. Complex gene rearrangements caused by serial replication slippage. Hum. Mutat. 26, 125–134. doi:10.1002/humu.20202

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

Chouliaras, L., Mastroeni, D., Delvaux, E., Grover, A., Kenis, G., Hof, P.R., Steinbusch, H.W.M., Coleman, P.D., Rutten, B.P.F., van den Hove, D.L.A., 2013. Consistent decrease in global DNA methylation and hydroxymethylation in the hippocampus of Alzheimer's disease patients. Neurobiol. Aging 34, 2091–2099. doi:10.1016/j.neurobiolaging.2013.02.021

Chuzhanova, N., Chen, J.-M., Bacolla, A., Patrinos, G.P., Férec, C., Wells, R.D., Cooper, D.N., 2009. Gene conversion causing human inherited disease: evidence for involvement of non-B-DNA-forming sequences and recombination-promoting motifs in DNA breakage and repair. Hum. Mutat. 30, 1189–1198. doi:10.1002/humu.21020

Clowney, E.J., LeGros, M.A., Mosley, C.P., Clowney, F.G., Markenskoff-Papadimitriou, E.C., Myllys, M., Barnea, G., Larabell, C.A., Lomvardas, S., 2012. Nuclear Aggregation of Olfactory Receptor Genes Governs Their Monogenic Expression. Cell 151, 724–737. doi:10.1016/j.cell.2012.09.043

Cooper, G.M., Coe, B.P., Girirajan, S., Rosenfeld, J.A., Vu, T.H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., Abdel-Hamid, H., Bader, P., McCracken, E., Niyazov, D., Leppig, K., Thiese, H., Hummel, M., Alexander, N., Gorski, J., Kussmann, J., Shashi, V., Johnson, K., Rehder, C., Ballif, B.C., Shaffer, L.G., Eichler, E.E., 2011. A copy number variation morbidity map of developmental delay. Nat. Genet. 43, 838–846. doi:10.1038/ng.909

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Marchetto, M.C.N., Muotri, A.R., Mu, Y., Carson, C.T., Macia, A., Moran, J.V., Gage, F.H., 2011. Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. Proc. Natl. Acad. Sci. U. S. A. 108, 20382–20387. doi:10.1073/pnas.1100273108

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Yeo, G.W., Mu, Y., Lovci, M.T., Morell, M., O'Shea, K.S., Moran, J.V., Gage, F.H., 2009. L1 retrotransposition in human neural progenitor cells. Nature 460, 1127–1131. doi:10.1038/nature08248

Deininger, P., 2011. Alu elements: know the SINEs. Genome Biol. 12, 236. doi:10.1186/gb-2011-12-12-236

DISTECHE, C.M., BERLETCH, J.B., 2015. X-chromosome inactivation and escape. J. Genet. 94, 591–599.

Donley, N., Stoffregen, E.P., Smith, L., Montagna, C., Thayer, M.J., 2013. Asynchronous Replication, Mono-Allelic Expression, and Long Range Cis-Effects of ASAR6. PLOS Genet. 9, e1003423. doi:10.1371/journal.pgen.1003423

Doucet-O'Hare, T.T., Rodić, N., Sharma, R., Darbari, I., Abril, G., Choi, J.A., Young Ahn, J., Cheng, Y., Anders, R.A., Burns, K.H., Meltzer, S.J., Kazazian, H.H., 2015. LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. Proc. Natl. Acad. Sci. U. S. A. 112, E4894–E4900. doi:10.1073/pnas.1502474112

Downs, J.A., Nussenzweig, M.C., Nussenzweig, A., 2007. Chromatin dynamics and the preservation of genetic information. Nature 447, 951–958. doi:10.1038/nature05980

Eggert, H., Bergemann, K., Saumweber, H., 1998. Molecular screening for P-element insertions in a large genomic region of Drosophila melanogaster using polymerase chain reaction mediated by the vectorette. Genetics 149, 1427–1434.

Elbarbary, R.A., Lucas, B.A., Maquat, L.E., 2016. Retrotransposons as regulators of gene expression. Science 351, aac7247. doi:10.1126/science.aac7247

Ernst, J., Kellis, M., 2012. ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods 9, 215–216. doi:10.1038/nmeth.1906

Erwin, J.A., Marchetto, M.C., Gage, F.H., 2014. Mobile DNA elements in the generation of diversity and complexity in the brain. Nat. Rev. Neurosci. 15, 497–506. doi:10.1038/nrn3730

Erwin, J.A., Paquola, A.C.M., Singer, T., Gallina, I., Novotny, M., Quayle, C., Bedrosian, T.A., Alves, F.I.A., Butcher, C.R., Herdy, J.R., Sarkar, A., Lasken, R.S., Muotri, A.R., Gage, F.H., 2016a. L1-associated genomic regions are deleted in somatic cells of the healthy human brain. Nat. Neurosci. 19, 1583–1591. doi:10.1038/nn.4388

Evrony, G.D., Cai, X., Lee, E., Hills, L.B., Elhosary, P.C., Lehmann, H.S., Parker, J.J., Atabay, K.D., Gilmore, E.C., Poduri, A., Park, P.J., Walsh, C.A., 2012. Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain. Cell 151, 483–496. doi:10.1016/j.cell.2012.09.035

Ewing, A.D., Kazazian, H.H., 2011. Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. Genome Res. 21, 985–990. doi:10.1101/gr.114777.110

Ewing, A.D., Kazazian, H.H., 2010. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res. 20, 1262–1270. doi:10.1101/gr.106419.110

Ewing, G.B., Jensen, J.D., 2014. Distinguishing neutral from deleterious mutations in growing populations. Front. Genet. 5. doi:10.3389/fgene.2014.00007

Feng, S., Lo, C.-C., Li, P.-E., Chain, P.S.G., 2016. ADEPT, a dynamic next generation sequencing data error-detection program with trimming. BMC Bioinformatics 17. doi:10.1186/s12859-016-0967-z

Feschotte, C., Pritham, E.J., 2007. DNA Transposons and the Evolution of Eukaryotic Genomes. Annu. Rev. Genet. 41, 331–368.

doi:10.1146/annurev.genet.40.110405.090448

Forman, H.J., Torres, M., 2001. Redox signaling in macrophages. Mol. Aspects Med. 22, 189–216.

Franchitto, A., 2013. Genome Instability at Common Fragile Sites: Searching for the Cause of Their Instability. BioMed Res. Int. 2013. doi:10.1155/2013/730714

Gasior, S.L., Wakeman, T.P., Xu, B., Deininger, P.L., 2006. The Human LINE-1 Retrotransposon Creates DNA Double-strand Breaks. J. Mol. Biol. 357, 1383–1393. doi:10.1016/j.jmb.2006.01.089

Gerdes, P., Richardson, S.R., Mager, D.L., Faulkner, G.J., 2016. Transposable elements in the mammalian embryo: pioneers surviving through stealth and service. Genome Biol. 17. doi:10.1186/s13059-016-0965-5

Gogvadze, E., Buzdin, A., 2009. Retroelements and their impact on genome evolution and functioning. Cell. Mol. Life Sci. CMLS 66, 3727–3742. doi:10.1007/s00018-009-0107-2

Goodier, J.L., Kazazian, H.H., 2008. Retrotransposons Revisited: The Restraint and Rehabilitation of Parasites. Cell 135, 23–35. doi:10.1016/j.cell.2008.09.022

Goodier, J.L., Zhang, L., Vetter, M.R., Kazazian, H.H., 2007. LINE-1 ORF1 Protein Localizes in Stress Granules with Other RNA-Binding Proteins, Including Components of RNA Interference RNA-Induced Silencing Complex. Mol. Cell. Biol. 27, 6469–6483. doi:10.1128/MCB.00332-07

Graham, T., Boissinot, S., 2006. The Genomic Distribution of L1 Elements: The Role of Insertion Bias and Natural Selection. J. Biomed. Biotechnol. 2006. doi:10.1155/JBB/2006/75327

Gu, W., Zhang, F., Lupski, J.R., 2008. Mechanisms for human genomic rearrangements. PathoGenetics 1, 4. doi:10.1186/1755-8417-1-4

Guffanti, G., Gaudi, S., Klengel, T., Fallon, J.H., Mangalam, H., Madduri, R., Rodriguez, A., DeCrescenzo, P., Glovienka, E., Sobell, J., Klengel, C., Pato, M., Ressler, K.J., Pato, C., Macciardi, F., 2016. LINE1 insertions as a genomic risk factor for schizophrenia: Preliminary evidence from an affected family. Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet. 171, 534–545. doi:10.1002/ajmg.b.32437

Ha, H., Hwang, I.-A., Park, J.H., Lee, H.B., 2008. Role of reactive oxygen species in the pathogenesis of diabetic nephropathy. Diabetes Res. Clin. Pract. 82 Suppl 1, S42-45. doi:10.1016/j.diabres.2008.09.017

Haigis, M.C., Yankner, B.A., 2010. The Aging Stress Response. Mol. Cell 40, 333–344. doi:10.1016/j.molcel.2010.10.002

Han, K., Lee, J., Meyer, T.J., Remedios, P., Goodwin, L., Batzer, M.A., 2008. L1 recombination-associated deletions generate human genomic variation. Proc. Natl. Acad. Sci. 105, 19366–19371. doi:10.1073/pnas.0807866105

Han, L., Zhao, Z., 2009. CpG islands or CpG clusters: how to identify functional GC-rich regions in a genome? BMC Bioinformatics 10, 65. doi:10.1186/1471-2105-10-65

Hancks, D.C., Kazazian, H.H., 2016. Roles for retrotransposon insertions in human disease. Mob. DNA 7, 9. doi:10.1186/s13100-016-0065-9

Hasin, Y., Olender, T., Khen, M., Gonzaga-Jauregui, C., Kim, P.M., Urban, A.E., Snyder, M., Gerstein, M.B., Lancet, D., Korbel, J.O., 2008. High-resolution copy-number

variation map reflects human olfactory receptor diversity and evolution. PLoS Genet. 4, e1000249. doi:10.1371/journal.pgen.1000249

Hassan, M.I., Waheed, A., Yadav, S., Singh, T.P., Ahmad, F., 2009. Prolactin inducible protein in cancer, fertility and immunoregulation: structure, function and its clinical implications. Cell. Mol. Life Sci. CMLS 66, 447–459. doi:10.1007/s00018-008-8463-x

Hastings, P.J., Ira, G., Lupski, J.R., 2009a. A Microhomology-Mediated Break-Induced Replication Model for the Origin of Human Copy Number Variation. PLoS Genet. 5. doi:10.1371/journal.pgen.1000327

Hastings, P.J., Lupski, J.R., Rosenberg, S.M., Ira, G., 2009b. Mechanisms of change in gene copy number. Nat. Rev. Genet. 10, 551–564. doi:10.1038/nrg2593

Heck, D.E., Vetrano, A.M., Mariano, T.M., Laskin, J.D., 2003. UVB light stimulates production of reactive oxygen species: unexpected role for catalase. J. Biol. Chem. 278, 22432–22436. doi:10.1074/jbc.C300048200

Hedges, D.J., Deininger, P.L., 2007. Inviting Instability: Transposable elements, Double-strand breaks, and the Maintenance of Genome Integrity. Mutat. Res. 616, 46–59. doi:10.1016/j.mrfmmm.2006.11.021

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., Glass, C.K., 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589. doi:10.1016/j.molcel.2010.05.004

Helmrich, A., Ballarino, M., Tora, L., 2011. Collisions between Replication and Transcription Complexes Cause Common Fragile Site Instability at the Longest Human Genes. Mol. Cell 44, 966–977. doi:10.1016/j.molcel.2011.10.013

Hishikawa, K., Takase, O., Yoshikawa, M., Tsujimura, T., Nangaku, M., Takato, T., 2015. Adult stem-like cells in kidney. World J. Stem Cells 7, 490–494. doi:10.4252/wjsc.v7.i2.490

Hoppe, R., Breer, H., Strotmann, J., 2006. Promoter motifs of olfactory receptor genes expressed in distinct topographic patterns. Genomics 87, 711–723. doi:10.1016/j.ygeno.2006.02.005

Horst, J., Griese, E.-U., Kleihauer, E., Kohne, E., 1984. α-Globin gene deletion causes α-thalassemia syndromes in two German families. Hum. Genet. 68, 260–263. doi:10.1007/BF00418398

Horton, R., Gibson, R., Coggill, P., Miretti, M., Allcock, R.J., Almeida, J., Forbes, S., Gilbert, J.G.R., Halls, K., Harrow, J.L., Hart, E., Howe, K., Jackson, D.K., Palmer, S., Roberts, A.N., Sims, S., Stewart, C.A., Traherne, J.A., Trevanion, S., Wilming, L., Rogers, J., de Jong, P.J., Elliott, J.F., Sawcer, S., Todd, J.A., Trowsdale, J., Beck, S., 2008. Variation analysis and gene annotation of eight MHC haplotypes: The MHC Haplotype Project. Immunogenetics 60, 1–18. doi:10.1007/s00251-007-0262-2

Hu, M., Deng, K., Qin, Z., Dixon, J., Selvaraj, S., Fang, J., Ren, B., Liu, J.S., 2013. Bayesian Inference of Spatial Organizations of Chromosomes. PLoS Comput. Biol. 9. doi:10.1371/journal.pcbi.1002893

Huang, C.R.L., Schneider, A.M., Lu, Y., Niranjan, T., Shen, P., Robinson, M.A., Steranka, J.P., Valle, D., Civin, C.I., Wang, T., Wheelan, S.J., Ji, H., Boeke, J.D., Burns, K.H., 2010. Mobile Interspersed Repeats Are Major Structural Variants in the Human Genome. Cell 141, 1171–1182. doi:10.1016/j.cell.2010.05.026

Huang, X., Madan, A., 1999. CAP3: A DNA Sequence Assembly Program. Genome Res. 9, 868–877.

Ibarra-Soria, X., Levitin, M.O., Saraiva, L.R., Logan, D.W., 2014. The olfactory transcriptomes of mice. PLoS Genet 10, e1004593.

Isik, A.T., 2010. Late onset Alzheimer's disease in older people. Clin. Interv. Aging 5, 307–311. doi:10.2147/CIA.S11718

Jern, P., Coffin, J.M., 2008. Effects of Retroviruses on Host Genome Function. Annu. Rev. Genet. 42, 709–732. doi:10.1146/annurev.genet.42.110807.091501

Ju, Y.S., Tubio, J.M.C., Mifsud, W., Fu, B., Davies, H.R., Ramakrishna, M., Li, Y., Yates, L., Gundem, G., Tarpey, P.S., Behjati, S., Papaemmanuil, E., Martin, S., Fullam, A., Gerstung, M., Nangalia, J., Green, A.R., Caldas, C., Borg, Å., Tutt, A., Lee, M.T.M., van't Veer, L.J., Tan, B.K.T., Aparicio, S., Span, P.N., Martens, J.W.M., Knappskog, S., Vincent-Salomon, A., Børresen-Dale, A.-L., Eyfjörd, J.E., Flanagan, A.M., Foster, C., Neal, D.E., Cooper, C., Eeles, R., Lakhani, S.R., Desmedt, C., Thomas, G., Richardson, A.L., Purdie, C.A., Thompson, A.M., McDermott, U., Yang, F., Nik-Zainal, S., Campbell, P.J., Stratton, M.R., 2015. Frequent somatic transfer of mitochondrial DNA into the nuclear genome of human cancer cells. Genome Res. 25, 814–824. doi:10.1101/gr.190470.115

Jung, M., Pfeifer, G.P., 2015. Aging and DNA methylation. BMC Biol. 13, 7. doi:10.1186/s12915-015-0118-4

Jurka, J., 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc. Natl. Acad. Sci. U. S. A. 94, 1872–1877.

Jurka, J., Kohany, O., Pavlicek, A., Kapitonov, V.V., Jurka, M.V., 2005. Clustering, duplication and chromosomal distribution of mouse SINE retrotransposons. Cytogenet. Genome Res. 110, 117–123. doi:10.1159/000084943

Kassiotis, G., 2014. Endogenous retroviruses and the development of cancer. J. Immunol. Baltim. Md 1950 192, 1343–1349. doi:10.4049/jimmunol.1302972

Katsube, T., Mori, M., Tsuji, H., Shiomi, T., Wang, B., Liu, Q., Nenoi, M., Onoda, M., 2014. Most hydrogen peroxide-induced histone H2AX phosphorylation is mediated by ATR and is not dependent on DNA double-strand breaks. J. Biochem. (Tokyo) 156, 85–95. doi:10.1093/jb/mvu021

Kazazian, H.H., Wong, C., Youssoufian, H., Scott, A.F., Phillips, D.G., Antonarakis, S.E., 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332, 164–166. doi:10.1038/332164a0

Kent, T., Chandramouly, G., McDevitt, S.M., Ozdemir, A.Y., Pomerantz, R.T., 2015. Mechanism of Microhomology-Mediated End-Joining Promoted by Human DNA Polymerase Theta. Nat. Struct. Mol. Biol. 22, 230–237. doi:10.1038/nsmb.2961

Kent, W.J., 2002. BLAT—The BLAST-Like Alignment Tool. Genome Res. 12, 656–664. doi:10.1101/gr.229202

Khanna, K.K., Jackson, S.P., 2001. DNA double-strand breaks: signaling, repair and the cancer connection. Nat. Genet. 27, 247–254. doi:10.1038/85798

Khodosevich, K., Lebedev, Y., Sverdlov, E., 2002. Endogenous Retroviruses and Human Evolution. Comp. Funct. Genomics 3, 494–498. doi:10.1002/cfg.216

Kim, N., Jinks-Robertson, S., 2012. Transcription as a source of genome instability. Nat. Rev. Genet. 13, 204–214. doi:10.1038/nrg3152

Kines, K.J., Sokolowski, M., deHaro, D.L., Christian, C.M., Belancio, V.P., 2014. Potential for genomic instability associated with retrotranspositionally-incompetent L1 loci. Nucleic Acids Res. 42, 10488–10502. doi:10.1093/nar/gku687

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., Taillon, B.E., Chen, Z., Tanzer, A., Saunders, A.C.E., Chi, J., Yang, F., Carter, N.P., Hurles, M.E., Weissman, S.M., Harkins, T.T., Gerstein, M.B., Egholm, M., Snyder, M., 2007. Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. Science 318, 420–426. doi:10.1126/science.1149504

Kratz, E., Dugas, J.C., Ngai, J., 2002. Odorant receptor gene regulation: implications from genomic organization. Trends Genet. TIG 18, 29–34.

Lamprecht, B., Walter, K., Kreher, S., Kumar, R., Hummel, M., Lenze, D., Köchert, K., Bouhlel, M.A., Richter, J., Soler, E., Stadhouders, R., Jöhrens, K., Wurster, K.D., Callen, D.F., Harte, M.F., Giefing, M., Barlow, R., Stein, H., Anagnostopoulos, I., Janz, M., Cockerill, P.N., Siebert, R., Dörken, B., Bonifer, C., Mathas, S., 2010. Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. Nat. Med. 16, 571–579, 1p following 579. doi:10.1038/nm.2129

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe,

T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J., Szustakowki, J., International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921. doi:10.1038/35057062

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. Bioinforma. Oxf. Engl. 23, 2947–2948. doi:10.1093/bioinformatics/btm404

Lavie, L., Maldener, E., Brouha, B., Meese, E.U., Mayer, J., 2004. The human L1 promoter: Variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. Genome Res. 14, 2253–2260. doi:10.1101/gr.2745804

Leboyer, M., Tamouza, R., Charron, D., Faucard, R., Perron, H., 2013. Human endogenous retrovirus type W (HERV-W) in schizophrenia: a new avenue of research at the gene-environment interface. World J. Biol. Psychiatry Off. J. World Fed. Soc. Biol. Psychiatry 14, 80–90. doi:10.3109/15622975.2010.601760

Ledbetter, D.H., Riccardi, V.M., Airhart, S.D., Strobel, R.J., Keenan, B.S., Crawford, J.D., 1981. Deletions of Chromosome 15 as a Cause of the Prader–Willi Syndrome. N. Engl. J. Med. 304, 325–329. doi:10.1056/NEJM198102053040604

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L.J., Lohr, J.G., Harris, C.C., Ding, L., Wilson, R.K., Wheeler, D.A., Gibbs, R.A., Kucherlapati, R., Lee, C., Kharchenko, P.V., Park, P.J., 2012. Landscape of Somatic Retrotransposition in Human Cancers. Science 337, 967–971. doi:10.1126/science.1222077

Lee, J.A., Carvalho, C.M.B., Lupski, J.R., 2007. A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. Cell 131, 1235–1247. doi:10.1016/j.cell.2007.11.037

Lee, Y., Lee, J.S., Lee, K.J., Turner, R.S., Hoe, H.-S., Pak, D.T.S., 2017. Polo-like kinase 2 phosphorylation of amyloid precursor protein regulates activity-dependent amyloidogenic processing. Neuropharmacology 117, 387–400. doi:10.1016/j.neuropharm.2017.02.027

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Li, W., Jin, Y., Prazak, L., Hammell, M., Dubnau, J., 2012. Transposable Elements in TDP-43-Mediated Neurodegenerative Disorders. PLoS ONE 7. doi:10.1371/journal.pone.0044099

Lieber, M.R., 2010. The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End Joining Pathway. Annu. Rev. Biochem. 79, 181–211. doi:10.1146/annurev.biochem.052308.093131

Linardopoulou, E.V., Williams, E.M., Fan, Y., Friedman, C., Young, J.M., Trask, B.J.,

2005. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. Nature 437, 94–100. doi:10.1038/nature04029

Liou, G.-Y., Storz, P., 2010. Reactive oxygen species in cancer. Free Radic. Res. 44. doi:10.3109/10715761003667554

Liu, P., Carvalho, C.M.B., Hastings, P.J., Lupski, J.R., 2012. Mechanisms for recurrent and complex human genomic rearrangements. Curr. Opin. Genet. Dev. 22, 211–220. doi:10.1016/j.gde.2012.02.012

Lu, T., Pan, Y., Kao, S.-Y., Li, C., Kohane, I., Chan, J., Yankner, B.A., 2004. Gene regulation and DNA damage in the ageing human brain. Nature 429, 883–891. doi:10.1038/nature02661

Madabhushi, R., Gao, F., Pfenning, A.R., Pan, L., Yamakawa, S., Seo, J., Rueda, R., Phan, T., Yamakawa, H., Pao, P.-C., Stott, R.T., Gjoneska, E., Nott, A., Cho, S., Kellis, M., Tsai, L.-H., 2015a. Activity-Induced DNA Breaks Govern the Expression of Neuronal Early-Response Genes. Cell 161, 1592–1605. doi:10.1016/j.cell.2015.05.032

Madabhushi, R., Gao, F., Pfenning, A.R., Pan, L., Yamakawa, S., Seo, J., Rueda, R., Phan, T.X., Yamakawa, H., Pao, P.-C., Stott, R.T., Gjoneska, E., Nott, A., Cho, S., Kellis, M., Tsai, L.-H., 2015b. Activity-Induced DNA Breaks Govern the Expression of Neuronal Early-Response Genes. Cell 161, 1592–1605. doi:10.1016/j.cell.2015.05.032

Mah, L.-J., El-Osta, A., Karagiannis, T.C., 2010. γH2AX: a sensitive molecular marker of DNA damage and repair. Leukemia 24, 679–686. doi:10.1038/leu.2010.6

Manoharan, S., Guillemin, G.J., Abiramasundari, R.S., Essa, M.M., Akbar, M., Akbar, M.D., 2016. The Role of Reactive Oxygen Species in the Pathogenesis of Alzheimer's Disease, Parkinson's Disease, and Huntington's Disease: A Mini Review. Oxid. Med. Cell. Longev. 2016. doi:10.1155/2016/8590578

Massaad, C.A., Klann, E., 2011. Reactive Oxygen Species in the Regulation of Synaptic Plasticity and Memory. Antioxid. Redox Signal. 14, 2013–2054. doi:10.1089/ars.2010.3208

Matthews, A.J., Zheng, S., DiMenna, L.J., Chaudhuri, J., 2014. Regulation of Immunoglobulin Class-Switch Recombination: Choreography of Noncoding Transcription, Targeted DNA Deamination, and Long-Range DNA Repair. Adv. Immunol. 122, 1–57. doi:10.1016/B978-0-12-800267-4.00001-8

Mattson, M.P., Magnus, T., 2006. Aging and Neuronal Vulnerability. Nat. Rev. Neurosci. 7, 278–294. doi:10.1038/nrn1886

Maynard, S., Schurman, S.H., Harboe, C., de Souza-Pinto, N.C., Bohr, V.A., 2009. Base excision repair of oxidative DNA damage and association with cancer and aging. Carcinogenesis 30, 2–10. doi:10.1093/carcin/bgn250

McClintock, B., 1951. Chromosome Organization and Genic Expression [WWW Document]. McClintock Barbara Chromosome Organ. Genic Expr. Cold Spring Harb. Symp. Quant. Biol. 16 1951 13-47 Artic. 35 Images. URL https://profiles.nlm.nih.gov/ps/retrieve/ResourceMetadata/LLBBBJ (accessed 8.22.17).

McConnell, M.J., Lindberg, M.R., Brennand, K.J., Piper, J.C., Voet, T., Cowing-Zitron, C., Shumilina, S., Lasken, R.S., Vermeesch, J., Hall, I.M., Gage, F.H., 2013. Mosaic Copy Number Variation in Human Neurons. Science 342, 632–637. doi:10.1126/science.1243472

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger,

A.M., Bejerano, G., 2010. GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. 28, 495–501. doi:10.1038/nbt.1630

McNeely, K.C., Cupp, T.D., Little, J.N., Janisch, K.M., Shrestha, A., Dwyer, N.D., 2017. Mutation of Kinesin-6 Kif20b causes defects in cortical neuron polarization and morphogenesis. Neural Develop. 12. doi:10.1186/s13064-017-0082-5

McVey, M., Lee, S.E., 2008. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet. TIG 24, 529–538. doi:10.1016/j.tig.2008.08.007

Mehta, A., Haber, J.E., 2014. Sources of DNA Double-Strand Breaks and Models of Recombinational DNA Repair. Cold Spring Harb. Perspect. Biol. 6. doi:10.1101/cshperspect.a016428

Miné, M., Chen, J.-M., Brivet, M., Desguerre, I., Marchant, D., de Lonlay, P., Bernard, A., Férec, C., Abitbol, M., Ricquier, D., Marsac, C., 2007. A large genomic deletion in the PDHX gene caused by the retrotranspositional insertion of a full-length LINE-1 element. Hum. Mutat. 28, 137–142. doi:10.1002/humu.20449

Mir, A.A., Philippe, C., Cristofari, G., 2015. euL1db: the European database of L1HS retrotransposon insertions in humans. Nucleic Acids Res. 43, D43-47. doi:10.1093/nar/gku1043

Mirkin, E.V., Mirkin, S.M., 2007. Replication Fork Stalling at Natural Impediments. Microbiol. Mol. Biol. Rev. MMBR 71, 13–35. doi:10.1128/MMBR.00030-06

Mobley, A.S., Rodriguez-Gil, D.J., Imamura, F., Greer, C.A., 2014. Aging in the olfactory system. Trends Neurosci. 37, 77–84. doi:10.1016/j.tins.2013.11.004

Moran, J.V., Holmes, S.E., Naas, T.P., DeBerardinis, R.J., Boeke, J.D., Kazazian, H.H., 1996. High frequency retrotransposition in cultured mammalian cells. Cell 87, 917–927.

Morrish, T.A., Gilbert, N., Myers, J.S., Vincent, B.J., Stamato, T.D., Taccioli, G.E., Batzer, M.A., Moran, J.V., 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat. Genet. 31, 159–165. doi:10.1038/ng898

Mortusewicz, O., Herr, P., Helleday, T., 2013. Early replication fragile sites: where replication–transcription collisions cause genetic instability. EMBO J. 32, 493–495. doi:10.1038/emboj.2013.20

Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M.R., Brown, D.G., Brown, S.D., Bult, C., Burton, J., Butler, J., Campbell, R.D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A.T., Church, D.M., Clamp, M., Clee, C., Collins, F.S., Cook, L.L., Copley, R.R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K.D., Deri, J., Dermitzakis, E.T., Dewey, C., Dickens, N.J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D.M., Eddy, S.R., Elnitski, L., Emes, R.D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G.A., Flicek, P., Foley, K., Frankel, W.N., Fulton, L.A., Fulton, R.S., Furey, T.S., Gage, D., Gibbs, R.A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T.A., Green, E.D., Gregory, S., Guigó, R., Guyer, M., Hardison, R.C., Haussler, D., Hayashizaki, Y., Hillier, L.W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D.B., Johnson, L.S., Jones, M., Jones, T.A., Joy, A., Kamal, M., Karlsson, E.K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W.J.,

Kirby, A., Kolbe, D.L., Korf, I., Kucherlapati, R.S., Kulbokas, E.J., Kulp, D., Landers, T., Leger, J.P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D.R., Mardis, E.R., Matthews, L., Mauceli, E., Mayer, J.H., McCarthy, M., McCombie, W.R., McLaren, S., McLay, K., McPherson, J.D., Meldrim, J., Meredith, B., Mesirov, J.P., Miller, W., Miner, T.L., Mongin, E., Montgomery, K.T., Morgan, M., Mott, R., Mullikin, J.C., Muzny, D.M., Nash, W.E., Nelson, J.O., Nhan, M.N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M.J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K.H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C.S., Poliakov, A., Ponce, T.C., Ponting, C.P., Potter, S., Quail, M., Reymond, A., Roe, B.A., Roskin, K.M., Rubin, E.M., Rust, A.G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M.S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J.B., Slater, G., Smit, A., Smith, D.R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J.P., Von Niederhausern, A.C., Wade, C.M., Wall, M., Weber, R.J., Weiss, R.B., Wendl, M.C., West, A.P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R.K., Winter, E., Worley, K.C., Wyman, D., Yang, S., Yang, S.-P., Zdobnov, E.M., Zody, M.C., Lander, E.S., 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562. doi:10.1038/nature01262

Mügge, A., 1998. The role of reactive oxygen species in atherosclerosis. Z. Kardiol. 87, 851–864.

Muñoz-López, M., García-Pérez, J.L., 2010. DNA Transposons: Nature and Applications in Genomics. Curr. Genomics 11, 115–128. doi:10.2174/138920210790886871

Muotri, A.R., Chu, V.T., Marchetto, M.C.N., Deng, W., Moran, J.V., Gage, F.H., 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435, 903–910. doi:10.1038/nature03663

Muotri, A.R., Zhao, C., Marchetto, M.C.N., Gage, F.H., 2009. Environmental influence on L1 retrotransposons in the adult hippocampus. Hippocampus 19, 1002–1007. doi:10.1002/hipo.20564

Niimura, Y., Nei, M., 2005. Comparative evolutionary analysis of olfactory receptor gene clusters between humans and mice. Gene 346, 13–21. doi:10.1016/j.gene.2004.09.025

Novikova, O., 2009. Chromodomains and LTR retrotransposons in plants. Commun. Integr. Biol. 2, 158–162.

O'Donnell, K.A., Burns, K.H., 2010. Mobilizing diversity: transposable element insertions in genetic variation and disease. Mob. DNA 1, 21. doi:10.1186/1759-8753-1-21

Olgiati, S., Thomas, A., Quadri, M., Breedveld, G.J., Graafland, J., Eussen, H., Douben, H., Klein, A. de, Onofrj, M., Bonifati, V., 2015. Early-onset parkinsonism caused by alpha-synuclein gene triplication: Clinical and genetic findings in a novel family. Parkinsonism Relat. Disord. 21, 981–986. doi:10.1016/j.parkreldis.2015.06.005

Ong, C.-T., Corces, V.G., 2014. CTCF: An Architectural Protein Bridging Genome Topology and Function. Nat. Rev. Genet. 15, 234–246. doi:10.1038/nrg3663

Ostertag, E.M., DeBerardinis, R.J., Goodier, J.L., Zhang, Y., Yang, N., Gerton, G.L., Kazazian, H.H., 2002. A mouse model of human L1 retrotransposition. Nat. Genet. 32, 655–660. doi:10.1038/ng1022

Payer, L.M., Steranka, J.P., Yang, W.R., Kryatova, M., Medabalimi, S., Ardeljan, D.,

Liu, C., Boeke, J.D., Avramopoulos, D., Burns, K.H., 2017. Structural variants caused by Alu insertions are associated with risks for many human diseases. Proc. Natl. Acad. Sci. 114, E3984–E3992. doi:10.1073/pnas.1704117114

Penzkofer, T., Jäger, M., Figlerowicz, M., Badge, R., Mundlos, S., Robinson, P.N., Zemojtel, T., 2017. L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. Nucleic Acids Res. 45, D68–D73. doi:10.1093/nar/gkw925

Perrat, P.N., DasGupta, S., Wang, J., Theurkauf, W., Weng, Z., Rosbash, M., Waddell, S., 2013. Transposition-driven genomic heterogeneity in the Drosophila brain. Science 340, 91–95. doi:10.1126/science.1231965

Piccolo, S.R., Frey, L.J., 2008. Somatic Mutation Signatures of Cancer. AMIA. Annu. Symp. Proc. 2008, 202–206.

Pickeral, O.K., Makałowski, W., Boguski, M.S., Boeke, J.D., 2000. Frequent Human Genomic DNA Transduction Driven by LINE-1 Retrotransposition. Genome Res. 10, 411–415.

Popa-Wagner, A., Mitran, S., Sivanesan, S., Chang, E., Buga, A.-M., 2013. ROS and Brain Diseases: The Good, the Bad, and the Ugly. Oxid. Med. Cell. Longev. 2013. doi:10.1155/2013/963520

Pugh, T.J., Delaney, A.D., Farnoud, N., Flibotte, S., Griffith, M., Li, H.I., Qian, H., Farinha, P., Gascoyne, R.D., Marra, M.A., 2008. Impact of whole genome amplification on analysis of copy number variants. Nucleic Acids Res. 36, e80. doi:10.1093/nar/gkn378

Quinlan, A.R., Hall, I.M., 2012. Characterizing complex structural variation in germline and somatic genomes. Trends Genet. TIG 28, 43–53. doi:10.1016/j.tig.2011.10.002

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

Ramel, C., 1997. Mini- and microsatellites. Environ. Health Perspect. 105, 781–789.

Ray, P.D., Huang, B.-W., Tsuji, Y., 2012. Reactive oxygen species (ROS) homeostasis and redox regulation in cellular signaling. Cell. Signal. 24, 981–990. doi:10.1016/j.cellsig.2012.01.008

Redon, C.E., Dickey, J.S., Bonner, W.M., Sedelnikova, O.A., 2009. γ-H2AX as a biomarker of DNA damage induced by ionizing radiation in human peripheral blood lymphocytes and artificial skin. Adv. Space Res. Off. J. Comm. Space Res. COSPAR 43, 1171–1178. doi:10.1016/j.asr.2008.10.011

Refsland, E.W., Harris, R.S., 2013. The APOBEC3 Family of Retroelement Restriction Factors. Curr. Top. Microbiol. Immunol. 371, 1–27. doi:10.1007/978-3-642-37765-5_1

Reilly, M.T., Faulkner, G.J., Dubnau, J., Ponomarev, I., Gage, F.H., 2013. The role of transposable elements in health and diseases of the central nervous system. J. Neurosci. Off. J. Soc. Neurosci. 33, 17577–17586. doi:10.1523/JNEUROSCI.3369-13.2013

Rhoads, A., Au, K.F., 2015. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics 13, 278–289. doi:10.1016/j.gpb.2015.08.002

Richards, R.I., Holman, K., Kozman, H., Kremer, E., Lynch, M., Pritchard, M., Yu, S., Mulley, J., Sutherland, G.R., 1991. Fragile X syndrome: genetic localisation by linkage mapping of two microsatellite repeats FRAXAC1 and FRAXAC2 which immediately flank the fragile site. J. Med. Genet. 28, 818–823.

Rinkevich, Y., Montoro, D.T., Contreras-Trujillo, H., Harari-Steinberg, O., Newman,

A.M., Tsai, J.M., Lim, X., Van-Amerongen, R., Bowman, A., Januszyk, M., Pleniceanu, O., Nusse, R., Longaker, M.T., Weissman, I.L., Dekel, B., 2014. In vivo Clonal Analysis Reveals Lineage-Restricted Progenitor Characteristics in Mammalian Kidney Development, Maintenance and Regeneration. Cell Rep. 7, 1270–1283. doi:10.1016/j.celrep.2014.04.018

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., Bennetzen, J.L., 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science 274, 765–768.

Sattler, M., Winkler, T., Verma, S., Byrne, C.H., Shrikhande, G., Salgia, R., Griffin, J.D., 1999. Hematopoietic growth factors signal through the formation of reactive oxygen species. Blood 93, 2928–2935.

Schumacher, B., Garinis, G.A., Hoeijmakers, J.H.J., 2008. Age to survive: DNA damage and aging. Trends Genet. TIG 24, 77–85. doi:10.1016/j.tig.2007.11.004

Schwer, B., Wei, P.-C., Chang, A.N., Kao, J., Du, Z., Meyers, R.M., Alt, F.W., 2016. Transcription-associated processes cause DNA double-strand breaks and translocations in neural stem/progenitor cells. Proc. Natl. Acad. Sci. U. S. A. 113, 2258–2263. doi:10.1073/pnas.1525564113

Scott, E.C., Devine, S.E., 2017. The Role of Somatic L1 Retrotransposition in Human Cancers. Viruses 9. doi:10.3390/v9060131

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., Pai, D., Zhang, R., Lee, Y.-H., Hicks, J., Spence, S.J., Lee, A.T., Puura, K., Lehtimäki, T., Ledbetter, D., Gregersen, P.K., Bregman, J., Sutcliffe, J.S., Jobanputra, V., Chung, W., Warburton, D., King, M.-C., Skuse, D., Geschwind, D.H., Gilliam, T.C., Ye, K., Wigler, M., 2007. Strong Association of De Novo Copy Number Mutations with Autism. Science 316, 445–449. doi:10.1126/science.1138659

Segal, Y., Peissel, B., Renieri, A., de Marchi, M., Ballabio, A., Pei, Y., Zhou, J., 1999. LINE-1 elements at the sites of molecular rearrangements in Alport syndrome-diffuse leiomyomatosis. Am. J. Hum. Genet. 64, 62–69. doi:10.1086/302213

Sen, S.K., Han, K., Wang, J., Lee, J., Wang, H., Callinan, P.A., Dyer, M., Cordaux, R., Liang, P., Batzer, M.A., 2006. Human Genomic Deletions Mediated by Recombination between Alu Elements. Am. J. Hum. Genet. 79, 41–53.

Sen, S.K., Huang, C.T., Han, K., Batzer, M.A., 2007. Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. Nucleic Acids Res. 35, 3741–3751. doi:10.1093/nar/gkm317

Seo, J., Kim, S.C., Lee, H.-S., Kim, J.K., Shon, H.J., Salleh, N.L.M., Desai, K.V., Lee, J.H., Kang, E.-S., Kim, J.S., Choi, J.K., 2012. Genome-wide profiles of H2AX and γ-H2AX differentiate endogenous and exogenous DNA damage hotspots in human cells. Nucleic Acids Res. 40, 5965–5974. doi:10.1093/nar/gks287

Serizawa, S., Ishii, T., Nakatani, H., Tsuboi, A., Nagawa, F., Asano, M., Sudo, K., Sakagami, J., Sakano, H., Ijiri, T., Matsuda, Y., Suzuki, M., Yamamori, T., Iwakura, Y., Sakano, H., 2000. Mutually exclusive expression of odorant receptor transgenes. Nat. Neurosci. 3, 687–693. doi:10.1038/76641

Shapiro, J.A., von Sternberg, R., 2005. Why repetitive DNA is essential to genome function. Biol. Rev. 80, 227–250. doi:10.1017/S1464793104006657

Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U., Pertz, L.M., Clark, R.A., Schwartz, S., Segraves, R., Oseroff, V.V., Albertson, D.G., Pinkel, D., Eichler, E.E., 2005. Segmental Duplications and Copy-Number Variation in the Human Genome. Am. J. Hum. Genet. 77, 78–88.

Sheen, F.M., Sherry, S.T., Risch, G.M., Robichaux, M., Nasidze, I., Stoneking, M., Batzer, M.A., Swergold, G.D., 2000. Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. Genome Res. 10, 1496–1508.

Shi, Y., Ding, Y., Lei, Y.-P., Yang, X.-Y., Xie, G.-M., Wen, J., Cai, C.-Q., Li, H., Chen, Y., Zhang, T., Wu, B.-L., Jin, L., Chen, Y.-G., Wang, H.-Y., 2012. Identification of novel rare mutations of DACT1 in human neural tube defects. Hum. Mutat. 33, 1450–1455. doi:10.1002/humu.22121

Shpyleva, S., Melnyk, S., Pavliv, O., Pogribny, I., Jill James, S., 2017. Overexpression of LINE-1 Retrotransposons in Autism Brain. Mol. Neurobiol. doi:10.1007/s12035-017-0421-x

Shukla, R., Upton, K.R., Muñoz-Lopez, M., Gerhardt, D.J., Fisher, M.E., Nguyen, T., Brennan, P.M., Baillie, J.K., Collino, A., Ghisletti, S., Sinha, S., Iannelli, F., Radaelli, E., Dos Santos, A., Rapoud, D., Guettier, C., Samuel, D., Natoli, G., Carninci, P., Ciccarelli, F.D., Garcia-Perez, J.L., Faivre, J., Faulkner, G.J., 2013. Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. Cell 153, 101–111. doi:10.1016/j.cell.2013.02.032

Singer, T., McConnell, M.J., Marchetto, M.C.N., Coufal, N.G., Gage, F.H., 2010. LINE-1 Retrotransposons: Mediators of Somatic Variation in Neuronal Genomes? Trends Neurosci. 33, 345–354. doi:10.1016/j.tins.2010.04.001

Skourti-Stathaki, K., Proudfoot, N.J., 2014. A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. Genes Dev. 28, 1384–1396. doi:10.1101/gad.242990.114

Slotkin, R.K., Martienssen, R., 2007. Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. 8, 272–285. doi:10.1038/nrg2072

Smith, B.F., Yue, Y., Woods, P.R., Kornegay, J.N., Shin, J.-H., Williams, R.R., Duan, D., 2011. An intronic LINE-1 element insertion in the dystrophin gene aborts dystrophin expression and results in Duchenne-like muscular dystrophy in the corgi breed. Lab. Investig. J. Tech. Methods Pathol. 91, 216–231. doi:10.1038/labinvest.2010.146

Smith, D.G., Adair, G.M., 1996. Characterization of an apparent hotspot for spontaneous mutation in exon 5 of the Chinese hamster APRT gene. Mutat. Res. 352, 87–96. doi:10.1016/0027-5107(96)00007-3

Strazzabosco, M., Fabris, L., 2012. Development of the Bile Ducts: Essentials for the Clinical Hepatologist. J. Hepatol. 56, 1159–1170. doi:10.1016/j.jhep.2011.09.022

Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E., Inouye, M., 1966. Frameshift Mutations and the Genetic Code. Cold Spring Harb. Symp. Quant. Biol. 31, 77–84. doi:10.1101/SQB.1966.031.01.014

Streva, V.A., Jordan, V.E., Linker, S., Hedges, D.J., Batzer, M.A., Deininger, P.L., 2015. Sequencing, identification and mapping of primed L1 elements (SIMPLE) reveals significant variation in full length L1 elements between individuals. BMC Genomics 16. doi:10.1186/s12864-015-1374-y

Strietholt, S., Maurer, B., Peters, M.A., Pap, T., Gay, S., 2008. Epigenetic modifications in rheumatoid arthritis. Arthritis Res. Ther. 10, 219. doi:10.1186/ar2500

Szyf, M., Pakneshan, P., Rabbani, S.A., 2004. DNA demethylation and cancer: therapeutic implications. Cancer Lett. 211, 133–143. doi:10.1016/j.canlet.2004.04.009

Tanzi, R.E., 2012. The Genetics of Alzheimer Disease. Cold Spring Harb. Perspect. Med. 2. doi:10.1101/cshperspect.a006296

Temtamy, S.A., Aglan, M.S., Valencia, M., Cocchi, G., Pacheco, M., Ashour, A.M., Amr, K.S., Helmy, S.M.H., El-Gammal, M.A., Wright, M., Lapunzina, P., Goodship, J.A., Ruiz-Perez, V.L., 2008. Long interspersed nuclear element-1 (LINE1)-mediated deletion of EVC, EVC2, C4orf6, and STK32B in Ellis–van Creveld syndrome with borderline intelligence. Hum. Mutat. 29, 931–938. doi:10.1002/humu.20778

Teytelman, L., Özaydın, B., Zill, O., Lefrançois, P., Snyder, M., Rine, J., Eisen, M.B., 2009. Impact of Chromatin Structures on DNA Processing for Genomic Analyses. PLoS ONE 4. doi:10.1371/journal.pone.0006700

The 1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. Nature 526, 68–74. doi:10.1038/nature15393

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T.K., Diegel, M., Dunn, D., Ebersol, A.K., Frum, T., Giste, E., Johnson, A.K., Johnson, E.M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Safi, A., Sanchez, M.E., Sanyal, A., Shafer, A., Simon, J.M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M.O., Hansen, R.S., Navas, P.A., Stamatoyannopoulos, G., Iyer, V.R., Lieb, J.D., Sunyaev, S.R., Akey, J.M., Sabo, P.J., Kaul, R., Furey, T.S., Dekker, J., Crawford, G.E., Stamatoyannopoulos, J.A., 2012. The accessible chromatin landscape of the human genome. Nature 489, 75–82. doi:10.1038/nature11232

Trask, B.J., Friedman, C., Martin-Gallardo, A., Rowen, L., Akinbami, C., Blankenship, J., Collins, C., Giorgi, D., Iadonato, S., Johnson, F., Kuo, W.L., Massa, H., Morrish, T., Naylor, S., Nguyen, O.T., Rouquier, S., Smith, T., Wong, D.J., Youngblom, J., van den Engh, G., 1998. Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. Hum. Mol. Genet. 7, 13–26.

Treiber, C.D., Waddell, S., n.d. Resolving the prevalence of somatic transposition in Drosophila. eLife 6. doi:10.7554/eLife.28297

Tremblay, A., Jasin, M., Chartrand, P., 2000. A double-strand break in a chromosomal LINE element can be repaired by gene conversion with various endogenous LINE elements in mouse cells. Mol. Cell. Biol. 20, 54–60.

Turinetto, V., Giachino, C., 2015. Multiple facets of histone variant H2AX: a DNA double-strand-break marker with several biological functions. Nucleic Acids Res. 43, 2489–2498. doi:10.1093/nar/gkv061

Tutar, Y., 2012. Pseudogenes. Int. J. Genomics 2012, e424526. doi:10.1155/2012/424526

Upton, K.R., Gerhardt, D.J., Jesuadian, J.S., Richardson, S.R., Sánchez-Luque, F.J., Bodea, G.O., Ewing, A.D., Salvador-Palomeque, C., van der Knaap, M.S., Brennan, P.M., Vanderver, A., Faulkner, G.J., 2015. Ubiquitous L1 Mosaicism in Hippocampal Neurons. Cell 161, 228–239. doi:10.1016/j.cell.2015.03.026

Uren, A.G., Mikkers, H., Kool, J., van der Weyden, L., Lund, A.H., Wilson, C.H., Rance, R., Jonkers, J., van Lohuizen, M., Berns, A., Adams, D.J., 2009. A high-throughput

splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. Nat. Protoc. 4, 789–798. doi:10.1038/nprot.2009.64

Uusküla-Reimand, L., Hou, H., Samavarchi-Tehrani, P., Rudan, M.V., Liang, M., Medina-Rivera, A., Mohammed, H., Schmidt, D., Schwalie, P., Young, E.J., Reimand, J., Hadjur, S., Gingras, A.-C., Wilson, M.D., 2016. Topoisomerase II beta interacts with cohesin and CTCF at topological domain borders. Genome Biol. 17, 182. doi:10.1186/s13059-016-1043-8

Verdin, H., D'haene, B., Beysen, D., Novikova, Y., Menten, B., Sante, T., Lapunzina, P., Nevado, J., Carvalho, C.M.B., Lupski, J.R., De Baere, E., 2013. Microhomology-mediated mechanisms underlie non-recurrent disease-causing microdeletions of the FOXL2 gene or its regulatory domain. PLoS Genet. 9, e1003358. doi:10.1371/journal.pgen.1003358

Walsh, T., McClellan, J.M., McCarthy, S.E., Addington, A.M., Pierce, S.B., Cooper, G.M., Nord, A.S., Kusenda, M., Malhotra, D., Bhandari, A., Stray, S.M., Rippey, C.F., Roccanova, P., Makarov, V., Lakshmi, B., Findling, R.L., Sikich, L., Stromberg, T., Merriman, B., Gogtay, N., Butler, P., Eckstrand, K., Noory, L., Gochman, P., Long, R., Chen, Z., Davis, S., Baker, C., Eichler, E.E., Meltzer, P.S., Nelson, S.F., Singleton, A.B., Lee, M.K., Rapoport, J.L., King, M.-C., Sebat, J., 2008. Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia. Science 320, 539–543. doi:10.1126/science.1155174

Walters, R.D., Kugel, J.F., Goodrich, J.A., 2009. InvAluable Junk: The Cellular Impact and Function of Alu and B2 RNAs. IUBMB Life 61, 831–837. doi:10.1002/iub.227

Wang, H., Adhikari, S., Butler, B.E., Pandita, T.K., Mitra, S., Hegde, M.L., 2014. A Perspective on Chromosomal Double Strand Break Markers in Mammalian Cells. Jacobs J. Radiat. Oncol. 1.

Wang, J., Song, L., Grover, D., Azrak, S., Batzer, M.A., Liang, P., 2006. dbRIP: A Highly Integrated Database of Retrotransposon Insertion Polymorphisms in Humans. Hum. Mutat. 27, 323–329. doi:10.1002/humu.20307

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F.A., Hakonarson, H., Bucan, M., 2007. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 17, 1665–1674. doi:10.1101/gr.6861907

White, T.B., McCoy, A.M., Streva, V.A., Fenrich, J., Deininger, P.L., 2014. A droplet digital PCR detection method for rare L1 insertions in tumors. Mob. DNA 5, 30. doi:10.1186/s13100-014-0030-4

Wilhelm, T., Ragu, S., Magdalou, I., Machon, C., Dardillac, E., Técher, H., Guitton, J., Debatisse, M., Lopez, B.S., 2016. Slow Replication Fork Velocity of Homologous Recombination-Defective Cells Results from Endogenous Oxidative Stress. PLoS Genet. 12. doi:10.1371/journal.pgen.1006007

Wood, J.G., Helfand, S.L., 2013. Chromatin structure and transposable elements in organismal aging. Front. Genet. 4. doi:10.3389/fgene.2013.00274

Woodbine, L., Brunton, H., Goodarzi, A.A., Shibata, A., Jeggo, P.A., 2011. Endogenously induced DNA double strand breaks arise in heterochromatic DNA regions and require ataxia telangiectasia mutated and Artemis for their repair. Nucleic Acids Res. 39, 6986–6997. doi:10.1093/nar/gkr331

Xing, J., Zhang, Y., Han, K., Salem, A.H., Sen, S.K., Huff, C.D., Zhou, Q., Kirkness,

E.F., Levy, S., Batzer, M.A., Jorde, L.B., 2009. Mobile elements create structural variation: analysis of a complete human genome. Genome Res. 19, 1516–1526. doi:10.1101/gr.091827.109

Xu, H., Luo, X., Qian, J., Pang, X., Song, J., Qian, G., Chen, J., Chen, S., 2012. FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. PLOS ONE 7, e52249. doi:10.1371/journal.pone.0052249

Xu, S., Grullon, S., Ge, K., Peng, W., 2014. Spatial Clustering for Identification of ChIP-Enriched Regions (SICER) to Map Regions of Histone Methylation Patterns in Embryonic Stem Cells. Methods Mol. Biol. Clifton NJ 1150, 97–111. doi:10.1007/978-1-4939-0512-6_5

Yalcin, B., Wong, K., Agam, A., Goodson, M., Keane, T.M., Gan, X., Nellåker, C., Goodstadt, L., Nicod, J., Bhomra, A., Hernandez-Pliego, P., Whitley, H., Cleak, J., Dutton, R., Janowitz, D., Mott, R., Adams, D.J., Flint, J., 2011. Sequence-based characterization of structural variation in the mouse genome. Nature 477, 326–329. doi:10.1038/nature10432

Yamamori, T., Yasui, H., Yamazumi, M., Wada, Y., Nakamura, Y., Nakamura, H., Inanami, O., 2012. Ionizing radiation induces mitochondrial reactive oxygen species production accompanied by upregulation of mitochondrial electron transport chain function and mitochondrial content under control of the cell cycle checkpoint. Free Radic. Biol. Med. 53, 260–270. doi:10.1016/j.freeradbiomed.2012.04.033

Yang, D., Elner, S.G., Bian, Z.-M., Till, G.O., Petty, H.R., Elner, V.M., 2007. Pro-inflammatory Cytokines Increase Reactive Oxygen Species through Mitochondria and NADPH Oxidase in Cultured RPE Cells. Exp. Eye Res. 85, 462–472. doi:10.1016/j.exer.2007.06.013

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., Ning, Z., 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25, 2865–2871. doi:10.1093/bioinformatics/btp394

Young, M.D., Wakefield, M.J., Smyth, G.K., Oshlack, A., 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 11, R14. doi:10.1186/gb-2010-11-2-r14

Zamudio, N., Bourc'his, D., 2010. Transposable elements in the mammalian germline: a comfortable niche or a deadly trap? Heredity 105, 92–104. doi:10.1038/hdy.2010.53

Zeman, M.K., Cimprich, K.A., 2014. Causes and consequences of replication stress. Nat. Cell Biol. 16, 2–9. doi:10.1038/ncb2897

Zhao, J., Bacolla, A., Wang, G., Vasquez, K.M., 2010. Non-B DNA structure-induced genetic instability and evolution. Cell. Mol. Life Sci. CMLS 67, 43–62. doi:10.1007/s00018-009-0131-2

Zorov, D.B., Juhaszova, M., Sollott, S.J., 2014. Mitochondrial Reactive Oxygen Species (ROS) and ROS-Induced ROS Release. Physiol. Rev. 94, 909–950. doi:10.1152/physrev.00026.2013