SISSA
Scuola
Internazionale
Superiore di
Studi Avanzati

Neuroscience Area – PhD course in

Functional and Structural Genomics

Investigation of mosaicism sources in Alzheimer's disease. A focus on nucleotide variants and LINE-1 copy number variants

Candidate:

Gabriele Leoni

Advisors:

Prof. Remo Sanges

Prof. Stefano Gustincich

Academic Year 2020-2021

Table of contents

Abstract

Technological advancements, in the form of SNPs arrays and whole genome sequencing provided high-throughput capability in investigating both the quantity and the sequence of cellular DNA. Throughout their extensive applications, several types of variants have been identified to generate mosaicism, the phenomenon characterized by the presence of cells with genetic differences within an organism. In humans, mosaicism has been found to be highly pervasive in both healthy and impaired brains, although its roles and its potential pathological effects are not yet fully understood. Among all the types of somatic variants found to concur in mosaicism, single nucleotide variants and insertions of retrotransposons are of particular interest. Single nucleotide variants (SNVs) are mutations that affect single positions of the genome, and although are predominantly physiological, due to the intrinsic error rate of the DNA replication process [McCulloch and Kunkel, 2008], have been found to cause several brain-related diseases, such as malformations of the brain [Gleeson et al., 2000; Rivière et al., 2012] and severe epileptic brain malformation [Lee et al., 2012; Poduri et al., 2012]. Retrotransposons instead, are a class of repetitive elements which can mobilize within the genome and increase, as a consequence of the process, their copy number. Through this, retrotransposons can shape the human genome by generating structural variants and possibly lead to gene function alterations. Among all the retrotransposons, the LINE-1 family (L1) is the only thought to be still active in humans, and therefore able to concur to mosaicism.

In Alzheimer's disease (AD), which is the most common neurodegenerative disorder characterized by the accumulation of plaques composed of amyloid β, neurofibrillary tangles containing Tau, synaptic loss and neuronal death, mosaicism has been detected. A recent publication demonstrated that a pathogenic SNVs in PIN1 gene, that result in the loss-of-function mutation of the protein, can lead to tau phosphorylation and aggregation, suggesting therefore a possible link between SNVs and the appearance of tau pathology in AD brains [Park et al., 2019]. Moreover, multiple observation linked AD-key proteins (such as Tau and TDP-43) with the reactivation of retrotransposons [Krug et al., 2017; Saleh et al., 2019]. However, the real impact of SNVs and retrotransposons in AD still remain largely unknown.

Motivated by this lack of knowledge, I decided to investigate SNVs abundance and retrotransposon copy number (CNV), mainly L1's, using AD post-mortem tissue samples. Giving the current technological limitations that affect mosaicism detection, the dataset was studied by coupling two different strategies and by developing a new targeted sequencing approach. In order to call the highest number of SNVs with the highest quality possible, the most dense SNP array available to date (*i.e.* with the highest number of different SNP probes) was applied upon cerebellum, frontal cortex and kidney samples that belonged to the same individuals. Despite arrays were also used to assess retrotransposon CNV content, they are ineffective in the detection of new retrotransposition events. For this reason, and to further expand SNVs detection to the whole genome, it was applied, as a second strategy, short-reads high coverage (~100x) whole genome sequencing additionally extending the analyses to temporal cortex and hippocampus tissues. Thanks to this approach, it was also unveiled, for the first time to my knowledge, the presence of multi-nucleotide somatic variants in brain (MNVs), a class of variants characterized by two nearby SNVs within the same haplotype.

Finally, although whole genome sequencing strategies were proven to be successfully in retrotransposition genotyping, being able to uniquely map short reads originated from repetitive elements to specific genomic regions is currently problematic. Therefore, CNVs, polymorphism and structural variants in overlap with repetitive elements may remain undetected, an aspect that would be even more exacerbated for somatic variants. To improve mapping specificity and resolution, we developed a targeted sequencing approach designed to specifically amplify and sequence both the genomic upstream and the 5' region of a subset of ~3000 full-length L1. We named this technology LIFE-seq from LINE-1 Five prime End sequencing. I contributed by developed an analysis pipeline able to genotype sequenced loci, testing it upon a subset of the AD dataset.

List of figures

List of tables

Chapter I: General introduction

Genomic mosaicism and its implications in diseases

It has been estimated that the human body contains 37.2 trillion of cells [Bianconi et al., 2013]. From the earliest studies on *the* genetic material in 1952, it was erroneously assumed that all the somatic cells of an individual shared the very same DNA sequence. Some dermatological disorders, however, historically provided the very first evidences that, within an individual, genomes can be different from cell to cell. Color variegation that follows the *lines of Blaschko* (lines of normal cell development in the skin) or *heterochromia irides* (a condition in which an individual has two iris with different colors), for example, were indeed readily recognized as representing genetic differences [Biesecker and Spinner, 2013].

The condition of having two or more population of cells with distinct genotypes is called *Mosaicism* and takes its names from the intricate images created by craftsmen from small pieces of colored tiles, or glass [Strachan and Read, 2018]. Depending on which parts of the body harbor the variant cells, and the potential for transmission to offspring, mosaicism can be classified as *germinal*, *somatic* or *gonosomal* (a combination of germinal and somatic) [Biesecker and Spinner, 2013].

With the advent of cytogenetics in the 70', researchers started to have the tools to study mosaicism in more details, by directly assessing the DNA of cells both quantitatively and qualitatively. At first, mosaicism was strictly observed in the context of pathologies, which were mainly tumors. It was found that mosaicism was primarily imputed to chromosomal aberration, as aneuploidy and segmental aneuploidy. However, subsequent technological improvements such as molecular cytogenetics, microarray-based and sequencing-based strategies, deeply increased the resolution limits and permitted the discovery of other types of variants implied in mosaicism. These were copy number variants (CNVs) and single nucleotide variants (SNVs). On the other hand, the massive application of these new technologies clarified that mosaicism is not only present in defective cells, but that it is also the natural condition of all somatic tissues [Behjati et al., 2014; Frumkin et al., 2005; Lynch, 2010].

1.1 Somatic single nucleotide variants are a physiological phenomenon that may lead to pathologies

Mosaicism can be generated by two main types of events: constitutive limitations to the fidelity of the genetic material inheritance, such as the intrinsic error rate of DNA replication [McCulloch and Kunkel, 2008], and contingent alterations of the genetic material [Lynch, 2010] promoted by environmental factors. Single nucleotides variants, which are mutation that affect one single nucleotide, have been discovered to be a consistent source of mosaicism mainly thanks to technological advancement that allow researchers to reach single nucleotide resolutions. SNVs generation is highly physiological, and mainly occurs during replication, with an *in vivo* estimated frequency of 10^{-9} per replicative cycle [Lynch, 2010; McCulloch and Kunkel, 2008]. Therefore, three new mutations are generated in the daughters of any dividing human cell. Considering only base substitutions in coding regions, it has been estimated that each of the ~40 trillion cells of the human body would accumulate 100-1,000 *de novo* mutations during the first 15 years of life [Lynch, 2010], and this rate can be higher taking into account other sources of variants generation, such as the effect of mutagens.

This high level of point mutations per single cell led to the hypothesis that SNVs may be associated to diseases while next generation sequencing (NGS) technologies allowed their identifications on a large scale of samples. Among them, non-overgrowth mosaic disorders include benign keratinocytic epidermal nevi, which have been demonstrated to be caused by mutations in fibroblast growth factor receptor 3 (FGFR3), phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha (PIK3CA) and different RAS family members [Hafner et al., 2006, 2007, 2012]. A series of mosaic overgrowth disorders was molecularly delineated, beginning with Proteus syndrome, which was shown to be due to AKT1 mutations [Lindhurst et al., 2011], and followed by several other disorders. These include asymmetrical neuronal migration abnormalities and hemimegencephaly caused by mutations in PIK3CA, AKT3, mammalian target of rapamycin (MTOR) and phosphoinositide-3-kinase, regulatory subunit 2 (PIK3R2) [Lee et al., 2012; Poduri et al., 2012; Rivière et al., 2012], and non-CNS fibroadipose overgrowth and CLOVES syndrome, which are both caused by PIK3CA mutations [Kurek et al., 2012; Lindhurst et al., 2012].

Different types of nevus sebaceous syndromes are caused by mosaic HRAS and KRAS mutations [Groesser et al., 2012], and Waldenström macroglobulinaemia are caused by a mutation in myeloid differentiation primary response gene 88 (MYD88) [Treon et al., 2012].

Most of these overgrowth disorders are caused by mutations that are lethal in a constitutional state and, in addition, are found in tumors. These phenotypes have in common that they are hyperplastic or hypertrophic abnormalities that are related to the growth-promoting effects of these mutations, which can thus manifest macroscopically even when occurring late in development. It is hypothesized that these mutations are lethal when constitutional, because the aberration of growth regulation should severely disrupt early embryonic development.

However, some growth-promoting mutations, such as activating mutations in FGFR3 associated with achondroplasia, are compatible with viability and can be inherited in an autosomal dominant pattern [Shiang et al., 1994; Rousseau et al., 1994]. Furthermore, prezygotic *de novo* FGFR3 mutations confer a substantial survival advantage to the male germ line, thus increasing the degree of germline mosaicism and also the frequency of transmission to affected offspring [Goriely and Wilkie, 2012].

While being extensively studied in tumors, SNVs were also found in neuronal tissues. Human brain is composed by about 86 billion neurons in postmitotic states [Azevedo et al., 2009] and it has been estimated that each of these continues to accumulate about 23 SNVs per year, approximately linearly with age [Lodato et al., 2018]. Although roles and effects of SNVs in neurons are currently under investigation, some findings, better elucidated in chapter 1.3, suggest mosaicism as a potential source of brain-related disease.

1.1.1 Multi-nucleotides variants are an underestimated source of variations

Multi-nucleotides variants (MNVs), defined as two or more nearby variants existing on the same haplotype in an individual, are a clinically and biologically important class of genetic variation (figure 1-A). The analyses of genome data from the 1000 Genomes Project [Auton et al., 2015] (2,504 individuals) and exome data from the Exome Aggregation Consortium [Lek et al., 2016] (60,706 individuals) led to the identification of over 10,000 germinal MNVs altering protein sequences, demonstrating the pervasive nature of MNV [Wang et al., 2020]. Studies of newly occurring (de novo) MNVs have also been performed using trio data sets. As part of the Deciphering Developmental Disorders (DDD) study [The Deciphering Developmental Disorders Study et al., 2015], Kaplanis and colleagues analyzed exome-sequencing data from over 6000 trios to quantify the pathogenic impact of MNVs in developmental disorders, showing that such variants are substantially more likely to be deleterious than SNVs. Moreover, they further clarified the mutational mechanisms at the bases of their generation, indicating that MNVs signatures can be ascribed to polymerase zeta, an error prone translesion polymerase, and APOBEC DNA deaminases proteins [Kaplanis et al., 2019]. Overall, these analyses have also provided estimates of the germline MNV rate per generation, falling into a consistent range of 1–3% of the SNV rate. Although these studies have provided valuable information about the mutational origins and functional impact of MNVs, to date there has been few analyses that investigated MNVs across the entire genome (including noncoding regions) in many thousands of deeply sequenced individuals, limiting our understanding of the genome-wide profile and complete frequency distribution of this class of variation.

Additionally, despite their potentially high impact on protein functions, it must be noted that most existing variant callers softwares still fail to correctly annotate MNVs. MNVs are usually reported as multiple adjacent SNVs which often results in incorrect amino acid predictions [Wei et al., 2015]. This software limitation is particularly affecting our understanding of somatic MNVs, which started to be described only recently and only from tumor samples [Srinivasan et al., 2020]. Notably, MNVs mis-annotations has been observed to lead to incorrect conclusions on variants effects [Srinivasan et al., 2020]. This is easily understandable when a MNVs hit a single codon, where the overall impact may differ from the functional consequences of the individual variants (figure 1-B) [Lek et al., 2016]. Taken together, all these aspects make MNVs a clinically underestimated source of variations.

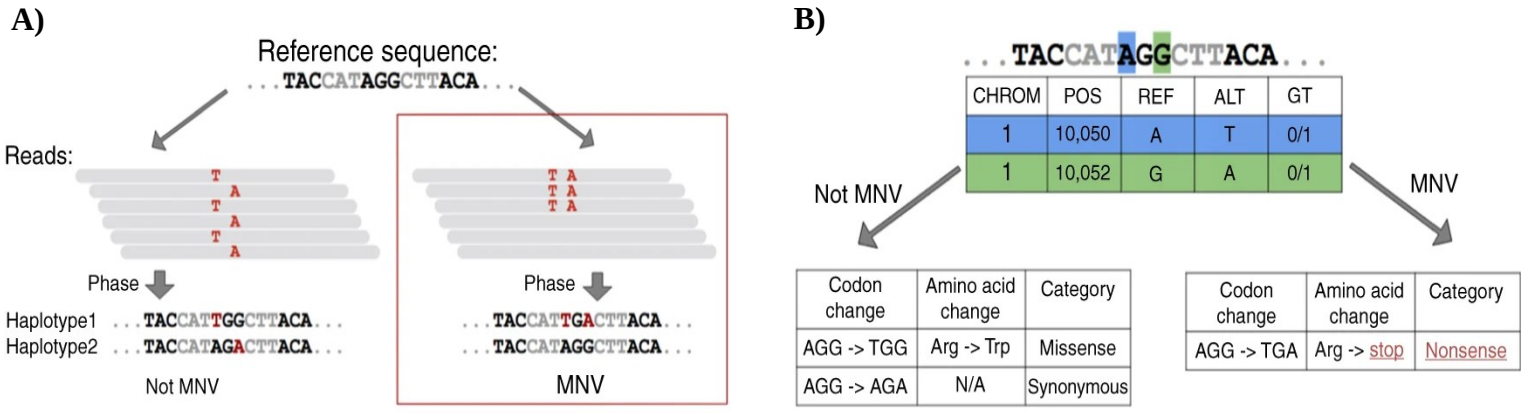


Figure 1. A) Example of MNVs. MNVs are defined as multiple and nearby variants that exist on the same haplotype and are thus present within the same read (right panel vs left panel); **B)** Example of MNVs impact in coding regions with respect to MNVs mis-annotations. Mis-annotated SNVs (left table) may result in missense and/or synonymous variants that hide the real functional impact of the correct MNV annotation (right table). Taken from [Wang et al., 2020].

1.1.2 Mutational signature analyses

Nucleotide variants are a consequence of multiple mutational processes that can have both physiological and pathogenic origin. Each single mutagen process creates a single and specific pattern of nucleotide substitutions, termed as “mutational signature”, that historically was classified using a six-class spectrum. The six classes are generally represented as couples of complementary nucleotides (*i.e.* C·G, one nucleotide for each DNA strand), followed by their mutated form (*i.e.* > A·T), and comprise: C·G > A·T, C·G > G·C, C·G > T·A, T·A > A·T, T·A > C·G, and T·A > G·C. However, pioneering works of Alexandrov L.B. from Stratton’s group [Alexandrov et al., 2013], changed this classification by combining the six possible SNV classes together with their trinucleotide contexts (*i.e.* the triplet of nucleotides that is generated when considering also the upstream and downstream nucleotides with respect to the SNVs), thus increasing the number of possible substitutions to 96. This classification resulted in precise nucleotides substitution matrices that, through a non-negative matrix factorization (NNMF) approach, can be converted into specific mutational signatures. Ultimately, signatures can be also associated with known mutagen processes. In this regard, the identification of

tobacco smoking signature from cancer samples can be a valid explanatory example. It was known for almost 60 years that smoking tobacco was one of the most avoidable risk factors for cancer. However, the detailed mechanisms by which tobacco smoke damages the genome and creates the mutations that ultimately cause cancer were still not fully understood. Alexandrov and colleagues in 2016 examined mutational signatures in over 5000 genome sequences from 17 different cancer types linked to smoking [Alexandrov et al., 2016]. By applying a NNMF approach to the trinucleotide substitution matrices, they found a complex pattern of mutational signatures. One of these signatures, called signature 4 (SBS4) and characterized by C·G > A·T mutations, was mainly found in cancers derived from tissues directly exposed to tobacco smoke. Moreover, SBS4 was found to be very similar to the mutational signature induced in vitro by exposing cells to benzo[a]pyrene (cosine similarity = 0.94), a tobacco smoke carcinogen. With this, authors were able to assign SBS4 to the direct mutational consequence of misreplication of DNA damage induced by tobacco carcinogens.

Besides tobacco's signature, the NNMF approach has been widely used, and resulted in the identification of more than 30 different driver mutational signatures (http://cancer.sanger.ac.uk/cosmic/signatures) [Nik-Zainal et al., 2016] at the origin of SNVs generation. Some of theme, were associated to exogenous (*e.g.*, tobacco smoking and UV-exposure) or endogenous processes (*e.g.*, APOBEC over-activity, deficiency in double strand break repair, and polymerase slippage), while others still remain with unknown origin.

In the last 15 years, the identification of the known mutational signatures, became important in both research and clinical approaches to cancer. However, this method is poorly exploited in other contexts. The reasons behind this must be searched in the requirements and in the limitations of this technique. The firsts *de-novo* mutational signature analyses would not be possible without a large series of whole genome sequencing and whole exome sequencing (WES) studies that comprised prevalently cancer studies, such as The Cancer Genome Atlas (TCGA) [Cancer Genome Atlas Research Network, 2008], Welcome Trust Sanger Institutes's Cancer Genome Project [Pleasance et al., 2010] and the International Cancer Genome Consortium (ICGC) [International Cancer Genome Consortium et al., 2010]. Therefore, current deposited signatures came mostly from tumor samples. Moreover, even when focusing on cancer, obtaining evidence that the proposed etiology of a signatures is a specific mutational process, is not straightforward. This is complicated by the lack of complete catalogs of true pathogenic driver variants and missing information on the environmental exposure history of the patient cohort. Additional complexities can be found then in the heterogeneous landscape of mutational

processes that is typically identified in individual cancers. Furthermore, the detected somatic mutations are the result of a balance between mutation-inducing and DNA repair processes, which are not fully independent, and mechanisms may vary between tissues. Last but not least, in tumors the detection of somatic SNVs is facilitated by clonal expansion of cells, a feature that is lacking in *i.e.* neurodegenerative diseases or in physiological conditions.

Despite the above-mentioned limitations, mutational signature analyses have the potential to provide information regarding the ethology of various diseases. Fortunately, whole genome sequencing (WGS) and whole exome sequencing (WES) are currently starting to be largely applied to growing cohorts of different diseases, facilitating future applications of mutational signature analyses and thus the *de-novo* discovery of signatures related to the most diverse disease not limited to cancer.

1.2 Copy Number Variants concur in the generation of mosaicism and diseases

Another important source of mosaicism consists in copy number variants (CNVs). They are defined as a difference in the dosage of genomic segments, ranging in size from one kilobase to several megabases, when compared to a reference human genome [Shaikh, 2017]. CNVs can result from structural variations within the genome including deletions, duplications, insertions, unbalanced translocations and inversions, which can lead to either a loss or gain of genomic segments. Overall, they account for more of the inter-individual variability between genomes in terms of total number of bases involved that all the single nucleotide variations and small insertion-deletions combined [Sudmant et al., 2015; Lupski, 2015]. The ability to detect CNVs strictly depends on the dimension of the variant and on the resolution of the adopted technology. These aspects are exacerbated in mosaicism, since not all cells manifest the same variant. It is therefore not surprising that solid pieces of evidences started to accumulate only with the advent of high throughput approaches (more details on mosaicism detection can be found in chapter 1.4).

Firstly, evidences of somatic CNVs were found from disease studies and in particular were made with earlier chromosomal microarray experiments that relied on aCGH technologies. Briefly, these consist in arrays coated in nucleotide probes that span the entire human genome and that can bind to normal DNA sequences by complementarity. Differentially labeled test DNA and normal reference DNA are hybridized simultaneously on the chip and the hybridization is detected with two different fluorochromes. Regions of gain or loss of DNA sequences, such as deletions, duplications or amplifications, are therefore seen as changes in the ratio of the intensities of the two fluorochromes along the target chromosomes [Pinkel et al., 1998]. Ballif and colleagues, for example, in 2006 found that mosaic CNVs represented 8% of the results from a diagnostic laboratory focused on developmental abnormalities [Ballif et al., 2006]. Interestingly, this estimation has increased with technological advancements. The advent of SNP arrays, which represented an improvement respect aCGH technologies, that consist in arrays able to genotype millions of SNPs and CNVs across the genome (further details in paragraph 1.4.2), led to the discovery of several types of mosaicism. These included: 1. monosomies and trisomies, as shown by the characteristic changes in probe intensity in combination with the altered genotype frequencies across a whole chromosome; 2. mosaicism for some types of

biparental and uniparental chromosomal regions, for the characteristic changes in genotype frequencies without an accompanying change in intensity, indicating a copy-number neutral change, such as loss of heterozygosity (LOH) [Conlin et al., 2010; Rodríguez-Santiago et al., 2010].

In the cytogenomics laboratory at the Children's Hospital of Philadelphia, analysis of pediatric individuals referred for genomic copy number analysis for various congenital and developmental anomalies has revealed a potentially pathogenic genomic variant in 22% of patients; 17% of which were mosaic [Conlin et al., 2010]. Other studies have reported that in individuals with pediatric disorders that warrant cytogenomic testing, mosaic abnormalities occured in ~0.5–2.0% of the cases [Conlin et al., 2010; Ballif et al., 2006; Cheung et al., 2007]. In addition to individuals with pediatric presentation of clinical abnormalities, mosaicism originated from CNVs has also been observed in adults who were under diagnosis of various diseases, most commonly cancer. The frequency of mosaic abnormalities was found to increase with age in a study of >50,000 individuals enrolled in the Gene–Environment Association Studies (GENEVA) consortium. The diagnoses under study included several types of cancer (including melanoma, lung and prostate) as well as other lung disease, cleft lip and palate, addiction, blood disorders, dental caries and glaucoma. The frequency of individuals with detectable clonal mosaicism for genomic anomalies larger than 50 kb was less than 0.5% from birth to 50 years of age but rose quickly after age 50 to 2–3% [Laurie et al., 2012; Jacobs et al., 2012]. However, an independent study of 1,991 individuals with bladder cancer, found mosaic genomic abnormalities in 1.7% of samples, which were present in both the blood and bladder tissue, suggesting that somatic CNVs can also have early origin [Rodríguez-Santiago et al., 2010]. Evidences of somatic CNVs are not restricted to disease cohorts. As an example, induced pluripotent stem cells derived from skin fibroblasts were found to contain an average of two CNVs [Abyzov et al., 2012]. Although these CNVs were not initially detected in the parental fibroblasts, a subsequent more sensitive genomic analyses confirmed that at least half of these CNVs preexisted at a low frequency in those cells [Abyzov et al., 2012]. These data suggest that the extreme heterogeneity of the tissue and technical difficulties in assessing single-cell genomes severely limit the detection of somatic CNVs thus suggesting their frequency is much higher than current estimates. Therefore, the mosaicism that is routinely detected is probably the tip of the iceberg and our understanding of its true extension is still biased by current technical limitation. Copy number variation has also been analyzed across tissues from the same individual, confirming substantial variation across tissues [Piotrowski et al., 2008]. These findings have clear relevance for disorders that might be caused by tissue-specific alterations.

Mutations limited to the affected tissue pose technical challenges for diagnosis since they will not be identified from tests on other, more accessible tissues. Finally, variation has also been identified between identical twins and may explain discordant phenotypes in monozygotic twins [Bruder et al., 2008; Breckpot et al., 2012].

1.2.1 Retrotransposons activity results in mosaicism

The human genome is abundantly composed by repetitive elements (REs), DNA sequences that exist in multiple copies. Reports estimate that REs represent about 45% of the human genome [Lander et al., 2001]. These sequences are highly heterogeneous and can be classified in two main categories: tandem repeats and interspersed repeats. In humans, interspersed repeats, also known as transposable elements or mobile elements, account for about 40% of the genome. Their peculiarity relies on the ability to amplify the number of their copies and/or change their position within the genome. Depending on the nature of the intermediate used in their mobilization, they can be classified in class I (as for RNA) and class II (as for DNA) transposons. However, the majority of mobile elements are believed to be inactive in humans, whit the exception of some class I transposons *[Ostertag and Kazazian Jr, 2001]*. Class I transposons, also called retrotransposons, mobilize throughout a strategy called retrotransposition, or copy-and-paste mechanism. The final effect of this process, as the name may suggest, is the generation, and insertion, of a newly synthesized copy at a different genomic location. Given their ability to mobilize, they increase their copy number through insertions and therefore they can be a source of mosaicism within the human genome. Moreover, it has been recently found that some of these elements can undergo somatic deletions, expanding therefore the known mechanisms through which they can account for somatic variants generation [Erwin et al., 2016].

Depending on the presence or absence of long terminal repeats *(LTRs)* at the edges of the element, class I transposons can be further divided in LTR and non-LTR.

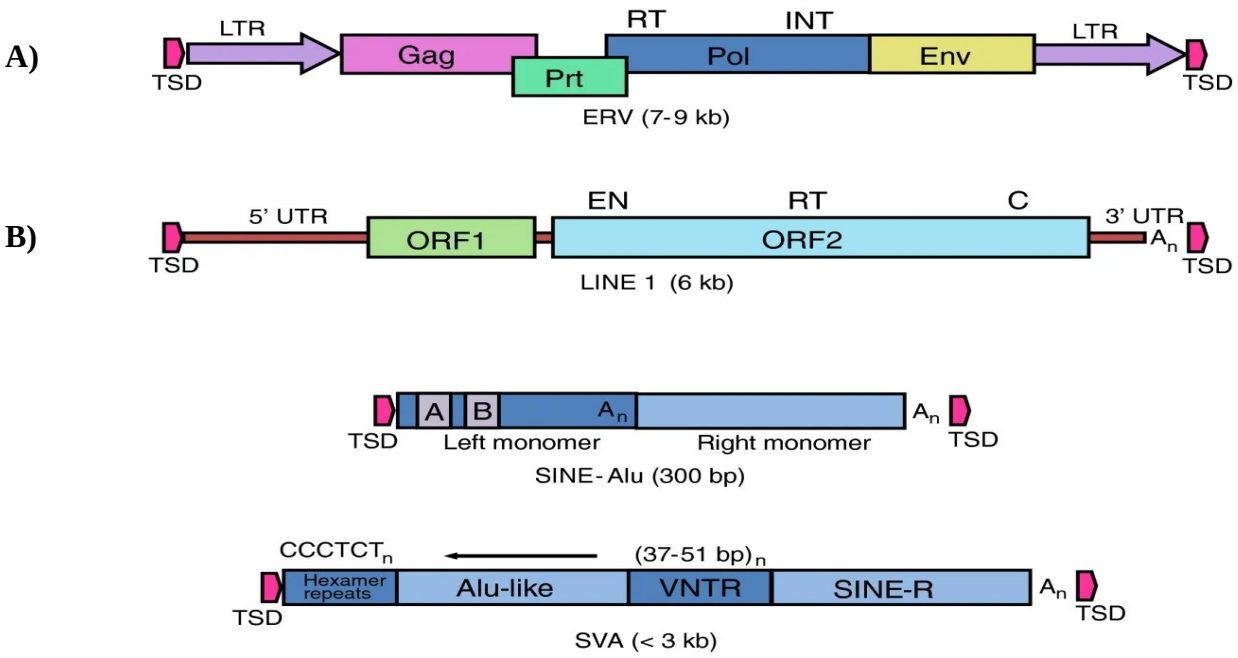


Figure 2: Class I transposons. **A)** LTR retrotransposons structure. Gag, specific group gene; Pol, RNA-dependent DNA polymerase with a reverse transcriptase (RT) domain and an integrase (INT) domain; Env envelope; LTR, long terminal repeats; **B)** Non-LTR retrotransposons structure. L1, Alu and SVA elements are represented. TSD, target-site duplication; LINE: EN, endonuclease domain; RT, reverse transcriptase domain; C, zinc knuckle domain; An, poly(A). SINE: A/B, Box A and Box B Pol III promoters. Taken from [Kazazian, 2011].

1.2.1.1 LTR retrotransposons

LTR elements are composed by two LTR regions, which are identical repeat sequences containing promoters and termination signals that flank the coding region (figure 2-A). Between these, LTR retrotransposons contain different enzymatic and structural genes that are required for their mobilization: the specific group antigen (GAG) that encodes for the monomers required to the generation of the virus-like particle (VLP); a *polymerase* (POL) required for the replication; a reverse transcriptase (RT) used to generate a complementary DNA (cDNA) from a RNA template; a Rnase H that cleaves the RNA in RNA/DNA hybrids and an integrase (INT) that serves for the target DNA cleavage [Burke et al., 2002; Prak and Kazazian, 2000]. Much of what is known about the mechanism of LTR retrotransposition (figure 3-A) comes from works on yeast retrotransposons [Voytas and Boeke, 2002; Sandmeyer et al., 2002]. The RNA transcript of LTR retrotransposons contains a region repeated at either end (R), a 5′ unique segment (U5) and a segment only included at the 3′ end (U3). The 3′ end of a cellular tRNAs serves as primer for the reverse transcription by hybridizing to the primer binding site (PBS) few nucleotides after the end of the 5′ LTR. After the LTR at 5′ has been copied into first-strand cDNA, the Rnase H activity of reverse transcriptase (RT) degrades the complementary RNA, and the elongating cDNA is transferred to the 3′ end of the retrotransposon transcript hybridizing to the R region. The remaining RNA is partially degraded by Rnase H, leaving behind primers for the second-strand cDNA synthesis. After a second transfer event, first- and second-strand synthesis can be completed to result in a full-length, double-stranded retroviral DNA that can be integrated into a new genomic position by the integrase (figure 3-B) [Schorn et al., 2017].

A) **B)**

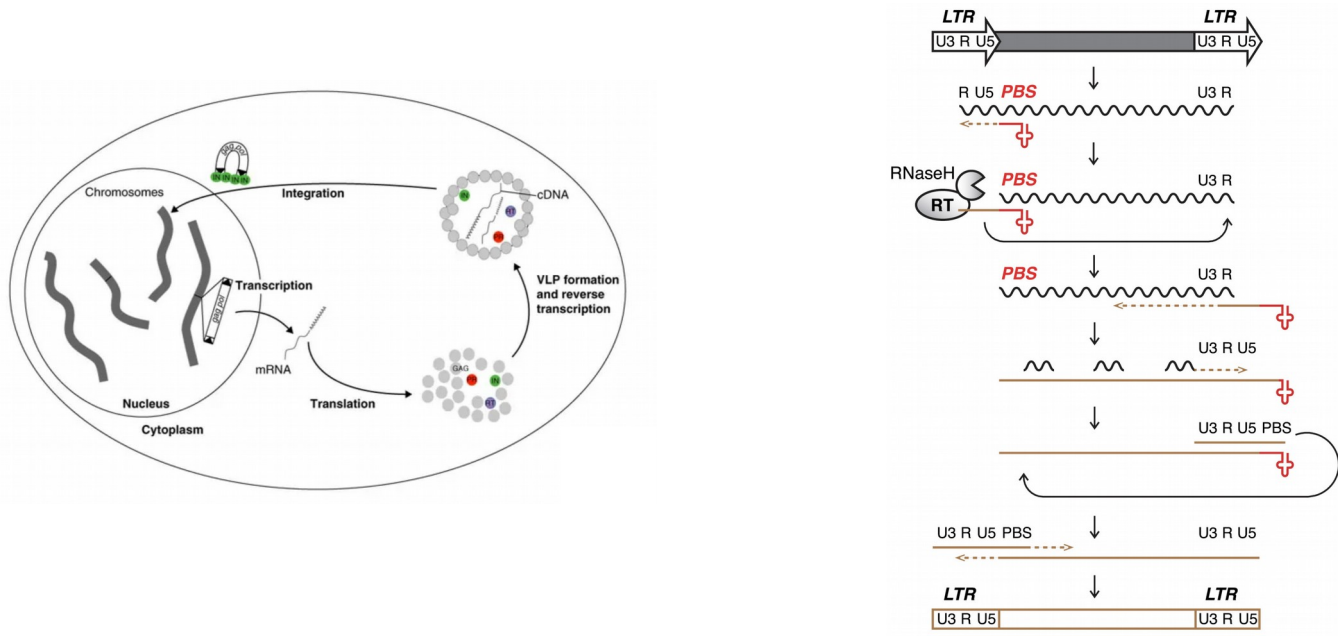


Figure 3: A) LTR-retrotransposons life cycle. IN, integrase; PR, protease; RT, reverse transcriptase; VLP, virus like particle, Black triangles represent the LTRs. Adapted from [Havecker et al., 2004]; **B)** Synthesis of LTR-retrotransposons. Adapted from [Schorn et al., 2017]

Endogenous retroviruses (ERVs) are believed to derive from LTR retrotransposons. Although a clear boundary between ERVs and LTRs has not been yet drawn, the main characterizing feature of LTR retrotransposon, in contrast to ERVs, is the lack in the *envelope* (ENV) gene or the presence of a non-functional copy, which do not allow the generation of infectious particles therefore limiting their life cycle.

LTR retrotransposons have been proposed as fundamental for the evolution of mammals and for the acquisition of the viviparity traits, by providing a plethora of LTR retrotransposons-derived genes. In humans it has been found that at least 30 genes derive from LTR retrotransposons, like the sushi-ichi retrotransposon homologue (SIRH) genes, which play essential roles in placenta formation and maturation [Kaneko-Ishino et al., 2017]. Despite being extremely important for human evolution, it is not yet fully understood whereas there are still active copies in the genome and thus whether they may account for mosaicism. Massive sequencing programs are relatively recent, therefore covering a very little percentage of the human population, and studies that focus on several tissues of the same

individuals are rare. What is known mainly derives from mouse, in which it has been observed that an Etn/MusD family of LTR elements remains insertionally active [Zhang et al., 2008]. Among all the thousands of human ERVs (HERVs) sequences that populates the human genome, the HERV-K family is the only one that has been active since the divergence of humans and chimpanzees. Furthermore, the number of polymorphic elements of this family is not significantly different from that predicted by a standard population genetic model that assumes constant activity until present. Active elements are likely to have inserted in the genome very recently, thus implying a very low allele frequency [Belshaw et al., 2005]. Very recently, several unfixed HERV-K elements were identified across human genomes including an intact insertion that maintains its infectivity and it is prone to generate mosaicism [Wildschutte et al., 2016]. Several hypothesis are involving HERVs activity in diseases, particularly of the nervous system, but their causative effects are still missing [Christensen, 2016].

1.2.1.2 Non-LTR retrotransposons

Contrary to LTR retrotransposons, Non-LTR retrotransposons do not present flanking LTR sequences. They contain the only classes of retrotransposons proved to be currently active in the human genome and represent an important source of mosaicism both through insertions and deletions. Depending on their length, Non-LTR retrotransposons are divided in: Long interspersed nuclear elements (LINEs) and Short interspersed nuclear elements (SINEs). LINEs are considered autonomous elements because they encode for the retrotransposition machinery required for their own mobilization. SINEs instead, are non-autonomous elements that thus depend on other retrotransposons machinery for their own spreading.

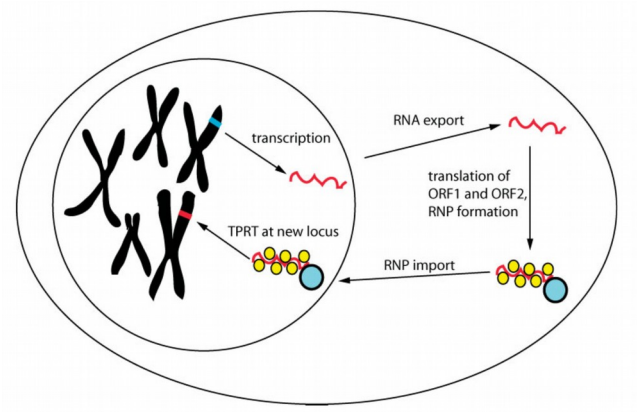
1.2.1.2.1 Autonomous retrotransposons: LINEs

LINE elements are a family of autonomous non-LTR retrotransposons. Among these, the LINE-1 (L1) subclass contains the only autonomous elements currently active in the human genome. At present, there are ~ 500.000 L1 copies, that represent 17% of the human genome. Despite their high numbers, L1 sequences are mostly inactive due to point mutations and/or truncations. However, there is still a set of about 80-100 full-length L1 copies that, to date, are potentially active in the human genome [Ostertag and Kazazian Jr, 2001; Brouha et al., 2003; Sassaman et al., 1997].

Intact, active L1 elements are 6 Kb in length and their transcripts are composed by both 5′ and 3′ untranslated regions (UTRs) and three open reading frames (ORFs) (figure 2-B). The 5′ UTRs of functional LINEs contains a primate-specific antisense promoter which gives birth to a transcript with a small ORF named ORF0 transcribed in antisense to the sense of canonical L1 retrotransposons and do not codify for any known protein. ORF0 as two splice donor sites, which permit the formation of fusion proteins with downstream exons [Nigumann et al., 2002; Speek, 2001] and that has been proposed to have been at the basis of the generation of some novel human specific non-coding genes [Uesaka et al., 2014]. The canonical L1 is transcribed by a sense promoter always within the 5′ UTR of the element [Becker et al., 1993; Wong and Choo, 2004]. From this promoter a polycistronic transcript is transcribed and formed by two ORFs named ORF1 and ORF2. ORF1 protein (ORF1p) is thought to have RNA binding activity [Dawson et al., 1997; Hohjoh and Singer, 1997; Kolosha and Martin, 2003] and nucleic acid chaperone activity [Martin and Bushman, 2001] while ORF2 encodes for a 150 kDa protein (ORF2p) presenting endonuclease (EN) [Feng et al., 1996] and reverse transcriptase (RT) activity [Mathias et al., 1991] indispensable for the retrotransposition [Feng et al., 1996; Moran et al., 1996]. The N-terminal may also contain a cysteine-rich domain that can function as a zinc-knuckle domain [Fanning and Singer, 1987], however their exact function needs to be elucidated. In addition, the 3′ UTRs of L1 elements contains a weak polyadenylation signal that can be bypassed by the RNA polymerase II. This sometimes can cause the continuing of the transcription outside the L1 element, resulting in the inclusion, within the polycistronic transcript, of a part of the adjacent 3′ genomic region, which will therefore be copied and inserted in a new genomic location. This phenomenon is named transduction [Moran et al., 1999].

As previously mentioned, retrotransposons rely on the copy and paste mechanism, also known as target-site primed reverse transcription (TPRT) [Luan et al., 1993], for their replication cycle. Retrotransposition starts with the transcription of an active full-length copy of the element whose mRNA is exported from the nucleus to the cytoplasm. In the cytosol ORF1 and ORF2 are translated and the produced proteins bound to their mRNA and other ribonucleoprotein (RNPs) [Hohjoh and Singer, 1996; Martin, 1991; del Carmen Seleme et al., 2005] (figure 4-A). The poly(A) stretch located at the 3′ end of L1 RNA acts as a substrate for poly(A) binding proteins (PABPs) containing RNA recognition motifs (RRMs). These proteins, in particular the poly(A) binding protein C1 (PABPC1), mediate the interaction with the proteins produced by the translation of the LINE mRNA and are required for an efficient retrotransposition [Dai et al., 2012]. This step is crucial since it can be predated from non-autonomous retrotransposons transcripts to transpose themselves [Kramerov and Vassetzky, 2011]. The RNP protein-mRNA complex is transported into the nucleus, where the ORF2p EN domain generates a nick into the insertion target site [Kinsey, 1990; Kubo et al., 2006]. The exposed 3′ hydroxyl DNA, which is protruding from the nick, binds to the mRNA polyA tract, acting then as primer for the ORF2p RT domain [Cost et al., 2002; Feng et al., 1996]. The next steps are currently not completely known but it is thought that a second strand cleavage creates a primer for the second strand DNA synthesis. This process results in the integration of a new L1 copy at a new genomic location site. Moreover, because the action of the endonuclease can lead to staggered DNA break, the integrated element is flanked by target site duplications (TSDs) that measure, on average, 7-20 bp in length [Han, 2010] (figure 4-B). Since RT lacks proofreading (3′ to 5′ endonuclease) activity, it can introduce mutation into the new copy with a rate of ~1 each 6,500 bases [Gilbert et al., 2005]. RT is often incapable to complete first strand synthesis, resulting in 5′ truncation of the newly formed copy [Szak et al., 2002] and it has been estimated that only ~30% of the transpositions events result in an inserted full-length element [Myers et al., 2002; Richardson et al., 2015]

A)



B)

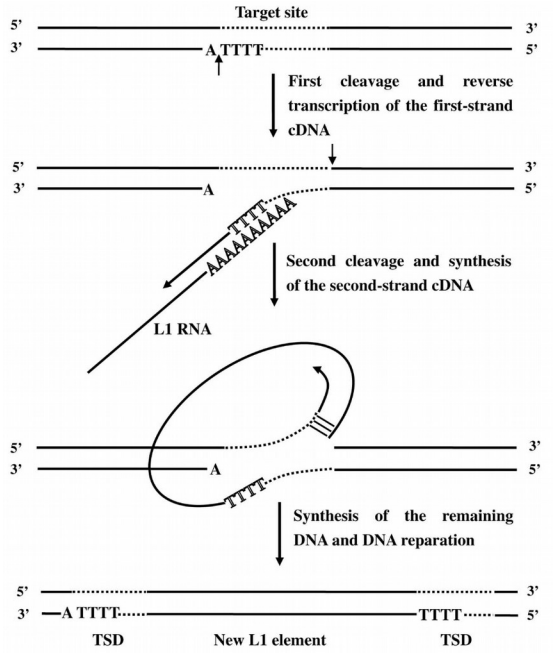


Figure 4: A) L1 life cycle. Taken from [Han et al., 2005]. **B)** L1 target primed reverse transcription (TPRT) mechanism. Taken from [Ding et al., 2006].

NGS technologies have fostered the accumulation of experimental evidences that somaticism due to L1 insertions may be more common than previously appreciated [Kano et al., 2009; Kazazian, 2011; Muotri et al., 2005]. Of particular interest, L1 mosaicism has been proposed to be pervasive in brain including the hippocampus and caudate nucleus [Baillie et al., 2011; Upton et al., 2015; Evrony et al., 2012]. However, the rate of new insertions per single neuron is currently subject of strong debate. Estimates range from less than 0.04 to 13.7 L1 insertions per cell [Erwin et al., 2016; Evrony et al., 2016]. Furthermore, the effects and functions (if any) of somatic retrotransposition in the human neurons are largely unknown. It has been suggested that retrotransposon reactivation during neurogenesis might have been positively selected by contributing to functional neuronal diversity [Kuwabara et al., 2009; Kurnosov et al., 2015] and being related to cognitive capabilities [Baillie et al., 2011]. However, it was also shown that depending on onset time, L1 mosaicism may lead to different neurodevelopmental and/or neurodegenerative disorders, such as epileptic brain malformation [Poduri et al., 2012] and schizophrenia [Bundo et al., 2014], better elucidated in chapter 1.2.1.4. L1 mosaicism is not restricted to only retrotransposition. Somatic deletions caused by L1 endonuclease cutting

activity were also recently observed [Erwin et al., 2016] further expanding the way in which L1s may shape neuronal genomes. These discoveries have contributed to position genomic mosaicism among the most interesting biological question in current research.

1.2.1.2.2 Non-autonomous retrotransposons: SINEs and SVA

SINEs are a large family of fragments long from 75 to 700 bp that are generally composed by a 5′ head, a middle body specific of each SINE family and a 3′ terminal tail [Kramerov and Vassetzky, 2011]. In the human genome they rely on the L1 retrotransposition protein machinery for their own spreading through the genome [Deininger et al., 2003]. In more details, SINEs retrotransposition starts with the displacement of L1 mRNA from the L1 ORF2 protein just translated in the cytoplasm. Indeed, the poly(A) tail of SINEs can compete with the poly(A) tail of LINEs for the binding to L1 proteins. When the SINEs mRNA binds to the retrotransposition protein machinery, the SINE element is reverse transcribed and transposed instead of the L1s [Kramerov and Vassetzky, 2011].

Alus, the most abundant family of SINEs in human, account for about 11% of the human genome with more than 1 million copies and represent the most abundant non-autonomous element in humans. Alus are derived from the 7SL RNA component of the signal recognition particle [Ullu and Tschudi, 1984] and are constituted of two similar monomers: left and right (figure 2-B). The left monomer contains the sequences required for the transcription [Chu et al., 1995], as the Box A and Box B Pol III promoters, whereas the right contains the sequence deputed for the binding with the retrotransposition machinery and the poly(A) tail, which in some elements can act as the recognized sequence [Dewannieux and Heidmann, 2005; Roy-Engel et al., 2002]. In contrast to L1s, most Alus are full-length. However, 5′-truncated Alu elements have been identified in human genomes [Wildschutte et al., 2015] as *de novo* insertions resulting in disease [Hancks and Kazazian, 2016]. In this regard, it has been proposed that Alu retrotransposition within mitochondrial genes may be also implicated in neurodegenerative diseases wherein mitochondrial dysfunction has been identified, including Alzheimer’s, Parkinson’s, Huntington’s diseases and myotrophic lateral sclerosis [Larsen et al., 2017]. Despite their potential implications in diseases, Alu elements have directly influenced human evolution by facilitating genome innovation through novel gene formation, elevated transcriptional diversity, long non-coding RNA and microRNA evolution (including circular RNAs), transcriptional regulation, and creation of novel

response elements [Chen and Yang, 2017; Jeck et al., 2013; Lehnert et al., 2009]. Moreover, Alus are known to alter the three-dimensional architecture and spatial organization of genomes by defining the boundaries of chromatin interaction domains (*i.e.*, topologically associating domains (TADs)) [Dixon et al., 2012]. Genome architecture has a direct influence on biological function, and the observation that Alus are enriched within both TADs and super-enhancer domains (SEDs) supports the hypothesis that Alus directly influence a wide range of critically important processes across multiple levels, from overall genome stability to tissue-specific gene regulation [Huda et al., 2009; Dixon et al., 2012; Glinsky, 2018]. Interestingly, Alu elements are also involved in neurogenesis and in the proper formation and function of the brain connectome [Oliver and Greene, 2011; Li and Church, 2013; Bitar and Barry, 2018]. In brain, it was also discovered that Alus serve as primary target for adenosine-to-inosine (A-to-I) RNA editing, which plays a significant role in mediating neuronal gene expression pathways [Tariq and Jantsch, 2012; Behm and Öhman, 2016]. Beyond RNA editing mechanisms, human neuronal gene pathways are also regulated by non-coding RNAs originating from Alu elements (e.g., BC200 and NDM29) and specific Alu subfamilies contain retinoic acid response elements which help to regulate neural patterning, differentiation, and axon outgrowth [Vansant and Reynolds, 1995; Maden, 2007; Castelnuovo et al., 2010; Smalheiser, 2014]. There is a deep connection between Alus and the formation and function of neurological networks, and this has led to the hypothesis that Alu elements were essential for development of the transcriptional diversity and regulation required for the genesis of human cognitive functions [Oliver and Greene, 2011; Li and Church, 2013].

SINE-VNTR-Alu (SVA) elements represent the youngest active human retrotransposon. They are 2 Kb in length, hominid-specific, non-coding composite sequences [Ostertag et al., 2003; Han et al., 2007; Wang et al., 2005]. Generally, SVA structure is composed, from its 5′ to its 3′ end, by a CCCTCT repeat that ranges from a few copies up to a hundred, an Alu-like domain derived from two Alu antisense fragments, a variable number of very GC-rich tandem repeats (VNTR), a SINE-R domain that shares sequence homology to the ENV gene of a HERV-K and a polyA tail similar to L1s [Shen et al., 1994; Damert, 2015; Hancks and Kazazian, 2010] (figure 2-B). The VNTR region accounts for most of the element-to-element sequence variation, which is higher than L1s and Alu [Damert, 2015; Hancks and Kazazian, 2010; Damert et al., 2009]. There are approximately 2,700 SVA elements in the human genome reference sequence [Wang et al., 2005]. However, due to its more recent discovery relative to L1 and Alu elements, less is known about their biology.

SVAs requires L1 ORF2p for retrotranposition [Ostertag et al., 2003; Hancks et al., 2011; Raiz et al., 2012] whereas it is currently unclear the requirement of L1 ORF1p. Although being recently discovered, SVA elements were also associated to diseases. For instance, it is known that a SVA element insertion in the fukutin (FKTN) gene cause Fukuyama muscular dystrophy [Kobayashi et al., 1998; Taniguchi-Ikeda et al., 2011].

Finally, through improvement in sequencing technology it was possible to observe that, in addition to L1, also Alu and SVA generates genomic mosaicism in hippocampus and caudate nucleus [Baillie et al., 2011].

1.2.1.3 Retrotransposons effects on the host genome

Retrotransposon can shape the host genome structure and modify gene expression in various ways. The vast majority of evidences were collected from L1, Alu and SVA elements, which alone contributed with ~750 million bases (Mb) to the human genome sequence [Lander et al., 2001]. Structural alterations of the genome can originate by:

1- Insertional mutagenesis.

The most straightforward way in which a retrotransposon can impact genome function is by inserting into protein-coding gene or regulatory regions, resulting in direct phenotypic consequences (figure 5-A). Due to this immediate effect in many human genetic disorders, insertional mutagenesis was the first retrotransposon structural effect to be detected [Kazazian et al., 1988]. In addition to L1 elements themselves, other protein coding mRNA can also be integrated into the genome through the L1-mediated retrotransposition, leading to the formation of processed pseudogenes (PPs). PPs are characterized by the lack of introns and the presence of a 3′ polyA tract flanking direct repeats. Moreover, PPs are unable to encode a functional protein and have accumulated frameshift mutations and premature stop codons during evolution, but few of them are transcriptionally active [Ding et al., 2006].

2- Insertion-mediated deletions.

Insertional events can result in the concomitant deletion of the adjacent genomic sequences, ranging in size from 1 bp to possibly > 130Kb [Gilbert et al., 2002]. Their mechanism, apparently relies on both endonuclease-dependent and endonuclease-independent processes, which are involved in repairing dsDNA nicks generated by L1 during its integration (figure 5-B) [Gilbert et al., 2005].

3- Non-allelic homologous recombinations.

Due to their extremely high copy numbers, L1 and Alu elements can create structural genomic variation at the post-insertion stage, through recombination between non-allelic homologous (NAHR) elements. This process can result in various types of genomic rearrangements such as deletion, duplication and invertions (figure 5-C). The amount of structural variations caused by NAHR is significant and accounts for more than 0.3% of human genetic diseases [Belancio et al., 2008]

4- 3′ and 5′ transduction.

During the retrotransposition process, L1s and SVA elements can carry with them upstream or downstream flanking genomic sequences (termed as 5′ and 3′ transduction, respectively) (figure 5-D). In 3′ transduction, the RNA transcription machinery skips the weak retrotransposon polyadenylation signal and terminates transcription by using an alternative signal located in the 3′ downstream flanking sequence. Similarly, 5′ transduction occurs when the promoter of a transcript upstream to a retrotransposon is used to transcribe the mobile element sequence. 3′ transduction has been shown to occur frequently in the human genome, in about 10% of L1s and SVA insertions [Cordaux and Batzer, 2009], while 5′ transduction appears to be much less common. However, 5′ transduction could be underestimated giving that it can be observed by only examining the full-length retrotransposon sequence [Beck et al., 2011].

5- Heterochromization.

Retrotransposition of L1 elements was observed to alter chromatin states. First speculations were made by Lyon in 1998, which suggested how L1 could act as a booster element to promote the spreading of heterochromatin formation during chromosome X inactivation [Lyon, 1998]. The mechanism, which was validated with experiments performed on embryonic stem cells, first implies that silent L1s tightly packaged in heterochromatin, facilitate nucleation of a silent heterochromatic compartment into which genes are recruited. Then, a subset of active L1s, expressed during X-chromosome inactivation (XCI), participate in local propagation of XCI to genes that otherwise would be prone to escape [Chow et al., 2010].

6- Transposition-mediated toxicity.

Retrotransposon expression, L1s in particular, can have direct and fatal consequences on cells. For instance, they can induce apoptosis and promote cell cycle arrest [Gasior et al., 2006; Belgnaoui et al., 2006]. The endonuclease activity of L1s ORF2p was speculated to be related to this deleterious effect by creating large excess of DNA double strand brakes, which are known to lead to apoptosis and senescence [Gire et al., 2004; Wallace et al., 2008]. On support to this hypothesis, there is evidence that lack of DNA double strand break repairing enzymes causes defective neurogenesis manifested by extensive apoptotic death of newly generated postmitotic neuronal cells in mice [Gao et al., 1998].

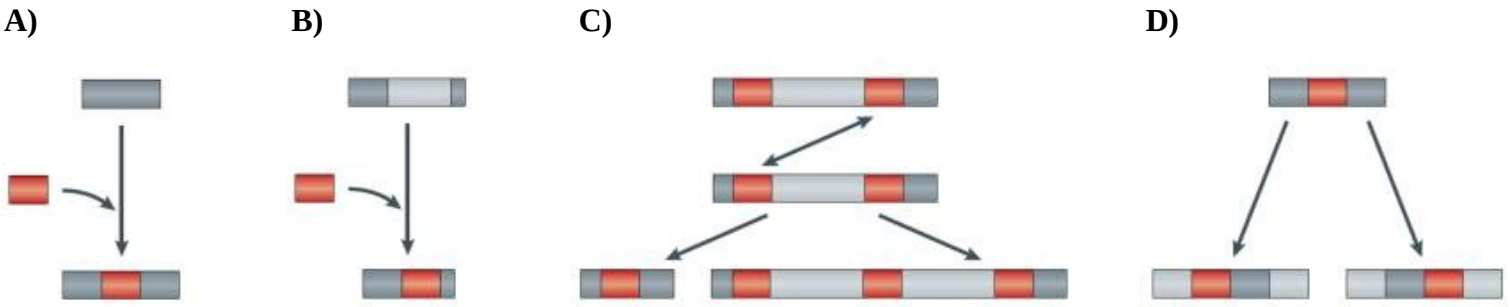


Figure 5: retrotransposons effect on the host genome. Retrotransposons are represented with red boxes. **A)** Insertional mutagenesis; **B)** Insertion-mediated deletions. The insertion of the mobile element result in the deletion of a genomic flanking region (light grey box); **C)** Non-allelic homologous recombinations. NAHR may result in the generation of structural variants (deletions or duplications) at the post-insertion site; **D)** 3′ and 5′ transduction. Genomic regions flanking the 3′ or the 5′ extremity of the mobile element can be carried during the integration process. Adapted from [Cordaux and Batzer, 2009].

The direct consequence of the aforementioned genome alterations is the modulation of cellular gene expression and transcription. L1 element, for instance, can be a source of promoters, alternative splice sites and polyadenilation signals, that depending on the integration site can generate new reorganized transcription units [Faulkner et al., 2009]. Similarly, the impact of retrotransposons on gene expression is dictated by their insertion site. Intergenic or intronic insertions can often have no detectable effects on genes. However, it is known that intronic insertions of L1s can induce exon skipping or exonization, both able to provoke alternative splicing of protein coding genes. Exon skipping occurs when a functional acceptor splice site is disrupted by an L1 insertion. The acceptor site situated on the following intron is then used for correct splicing, leading to the skipping of the exon included between the two spliced introns. Exonization instead, is determined whenever the retrotransposon element contains both functional donor and acceptor splice sites, resulting in the recruitment of the element as an exon, with its integration in the gene. The presence of multiple donor and acceptor splice sites within L1s sequences, combined with the potential generation of new ones through their activity, enhance the complexity of the combinatorial usage of alternative splicing sites within cells [Zemojtel et al., 2007]. Finally, insertions in exons or regulatory sequences have the potential to profoundly alter gene expression and function, by disrupting coding- or cis-regulating sequences [Viollet et al., 2014].

The potential treat represented by uncontrolled retrotransposition is balanced by the presence of several cellular mechanisms deputed to its restriction. Many of these mechanisms act in the cytoplasm to degrade retroelement RNA or inhibit its translation while other factors act in the nucleus and involve DNA repair enzymes or epigenetic processes of DNA methylation and histone modification [Goodier, 2016]. For example, one of the firsts anti-retrotransposon restriction factors identified were the Apolipoprotein B editing complex enzymes (APOBEC). These are an evolutionarily conserved, vertebrate-specific family of cytidine deaminases, originally identified as enzymes that edits mRNA species by deaminating cytosine to uracil. From later studies, it has been shown that the APOBEC family has members able to act also on cytosines in DNA [Koito and Ikeda, 2013]. Some members are represented by the APOBEC3 proteins, which have been shown to “lethally edit” the reverse transcripts of retroviral DNA [Mangeat et al., 2003] and human papillomavirus DNA [Vartanian et al., 2008] suggesting important roles in intrinsic response to viral infections. Most important, all APOBEC3 proteins has been observed to inhibit LINE-1 retrotransposition to varying degrees, with APOBEC3A and APOBEC3B being most effective [Lovšin and Peterlin, 2009]. Unexpectedly, catalytically inactive APOBEC3s still inhibit non-LTR retrotransposons, and several investigations found scant genomic evidence for L1 editing by cytidine deamination [Stenglein and Harris, 2006]. Deamination-independent mechanisms of APOBEC action were therefore proposed, including sequestration of retrotransposon RNPs in high molecular weight cytoplasmic complexes and their targeting to SGs and PBs for possible degradation by RNAi silencing [Bogerd et al., 2006; Chiu et al., 2006].

1.2.1.4 The role of L1 in neurological diseases

A plethora of human diseases was found to be caused by the direct effect of L1 retrotransposition [Hancks and Kazazian, 2016]. Neurological diseases, such as Rett syndrome, Ataxia telangiectasia, Schizophrenia and Huntington's disease among others, have been recently observed to have mis-regulation of L1 retrotransposition, which could contribute to some pathological aspects.

Rett syndrome (RTT) is a neurodevelopmental disorder caused by mutations in the MeCP2 gene [Amir et al., 1999] which is characterized by autism, loss of speech, hand-wringing, anxiety, and eventual motor deterioration. Although MeCP2 role in neurons and how its mutations contribute to RTT pathology are still under investigation, its protein product was shown to negatively regulate L1 elements [Muotri et al., 2010]. MeCP2 is known to be involved in global DNA methylation [Skene et al., 2010] and particularly, in neural stem cells, it has been found in association with the CpG-methylated promoter of L1, where it forms a repressive complex with HDAC1 that inhibit L1 expression [Muotri et al., 2010; Coufal et al., 2009]. RTT patients, carrying MeCP2 mutations, show hypomethylation of L1 promoter, support 2.5-fold higher retrotransposition and display greater L1 genomic DNA copies compared with age-matched controls [Muotri et al., 2010]. However, the functional impact of L1 overexpression and its increased retrotransposition on Rett syndrome is unknown and may be a consequence, rather than a cause of the disease.

Ataxia telangiectasia (AT) is a rare, hereditary neurodegenerative disorder caused by mutations in the Ataxia Telangiectasia Mutated (ATM) gene. Individuals affected by AT are characterized by loss of motor function, dilation of their capillaries and severe complications that result in their premature death. ATM is a serine/threonine protein kinase able to senses and responds to DNA damage. It detects DNA double-strand breaks initiating a signaling cascade by phosphorylating its multiple substrates that result in activating a DNA-damage checkpoint, leading to the arrests of the cell cycle, until the damage is repaired [Bar-Shira et al., 2002]. In cells deficient in ATM function, double-strand breaks in DNA can go unnoticed, leading to an increase in DNA mutagenesis at each cell cycle. Coufal and colleagues, in 2011 demonstrated that the brain of both ATM knock-out mice and patients with AT display significantly greater levels of L1 retrotransposition and longer L1 insertions compared to the controls. These results led to the hypothesis that ATM normally participates in the detection of the reverse transcription activity of L1 ORF2 protein and signal to inhibit the insertion step, while the lack of

functional ATM could delay the DNA damage response during L1 integration, finally leading to neurodegeneration [Coufal et al., 2011].

L1 retrotransposition was found also in schizophrenia. In 2014, Bundo and colleague demonstrated the presence of an increased L1 copy number in neurons from the prefrontal cortex of affected patients and in induced pluripotent stem cells-derived neurons containing 22q11 deletion (one of the highest risk factors for schizophrenia). Whole genome sequencing revealed brain-specific L1 insertion, localized preferentially in synapse- and schizophrenia-related genes. Further experiments on animal models aimed at identifying the causes of this L1 copy number alteration, suggested that hyperactive retrotransposition of L1 in neurons triggered by environmental and/or genetic risk factors may contribute to the susceptibility and pathophysiology of schizophrenia [Bundo et al., 2014].

Finally, mobilization of L1 elements was observed to promote neuron apoptosis in Huntington's disease (HD) mouse models but not in wild-type mice, suggesting a possible link between retrotransposon and HD progression. Defective autophagy pathways have been shown to facilitate HD progression, while activation of AMPK alpha, a critical regulator of autophagy, can ameliorate HD disease conditions in cells [Walter et al., 2016; Vázquez-Manrique et al., 2016]. In 2018, Tan and colleagues, elucidated a possible link between L1s and AMPK alpha. First, they found increased L1 copy numbers and L1 transcripts in HD mouse brain tissue as opposed to the control mouse brain tissue. Importantly, similar genomic alterations were not found in matched liver tissues, indicating that changes were limited to brain tissues. Second, they showed that increased expression of L1 ORF transcripts decreased AMPK alpha expression. Finally, through in vivo experiments they further demonstrated that overexpressed L1 ORF2 transcripts can decrease phosphorylation of at least 13 different AKT target proteins, widely indicated as pro-survival proteins, overall suggesting that L1 transposition, AMPK alpha and autophagy all function in a common pathway [Tan et al., 2018].

1.3 Mosaicism in neurodegenerative diseases, a focus on Alzheimer's disease

The best-known clinical implication of mosaicism is cancer. The accumulation of hundreds to thousands of somatic alterations combined with the developmental timing and cell lineage can potentially convert a cell from normal into malignant. However, it has been pointed out by several studies that, given the long lifespan of neurons and their central role in neural circuits and behavior, somatic mosaicism could represents a potential mechanism that may contribute to neuronal diversity and the etiology of numerous brain-related diseases [Bushman and Chun, 2013]. Lots of efforts has been made to unveil these hypothetical links and, as a result, somatic mutations and retrotranspositions have been found both in healthy and dysfunctional brains [Gleeson et al., 2000; Rivière et al., 2012; Lee et al., 2012; Poduri et al., 2012]. These findings led the way to more studies in the field, most of which started recently and/or are still ongoing. The Brain Somatic Mosaicism Network, for instance, was established in 2017 to investigate the effects of mosaicism in Neuropsychiatric disorders, such as schizophrenia, autism spectrum disorders, bipolar disorder, Tourette syndrome and epilepsy [McConnell et al., 2017]. Neurodegenerative disorders are also under investigation giving that their molecular etiologies are largely unknown.

Within the plethora of neurodegenerative diseases, Alzheimer's disease (AD) represents the most common dementia, impacting an estimated 5.4 million people in the United States alone (1 in 10 people over the age of 65) and 50 million people worldwide [Bright Focus Foundation, 2019]. AD neuropathology includes the accumulation of plaques composed of amyloid β (Aβ), neurofibrillary tangles containing Tau, synaptic loss and neuronal death in several brain regions, including the hippocampus, and frontal and entorhinal cortices, leading to progressive cognitive decline. Familial AD makes up ~5% of all cases, with early onset (<60 years) caused by inherited autosomal dominant mutations or CNVs that affect predominantly 3 genes: Amyloid precursor protein (APP) on chromosome 21, which is cleaved by ɣ-secretase to form the Aβ peptide found in amyloid plaques; presenilin-1 (PSEN1) on chromosome 14, and presenilin-2 (PSEN2) on chromosome 1, which contribute to the catalytic activity of the Aβ-cleaving enzyme. γ-Secretase can produce different lengths of Aβ. Most of the Aβ produced is the variant of 40 residues (Aβ40), even though a longer form of 42

residues (Aβ42) can also be produced. This last variant is more hydrophobic than the shorter one and form aggregates easier than the Aβ40 form [Cavallucci et al., 2012].

Late-onset (>60 years) sporadic AD arises from a less understood set of genetic, epigenetic, and environmental risk factors [Kingsbury et al., 2006; Gatz et al., 2006; Bertram et al., 2010], but shares the same neuropathology with familial cases.

1.3.1 Genetics of early-onset Alzheimer's disease

As mentioned above, mutations in three different genes are known to cause early-onset Alzheimer's disease (EO-AD): amyloid beta precursor protein, presenilin 1, and presenilin 2. Clinical features and pathology vary depending on the mutation's locus and position within each gene [Ridge et al., 2013].

In the case of APP gene, duplications (as in Down's syndrome patients) are sufficient in many cases to cause early-onset familiar Alzheimer's disease (EO-FAD), due to increased Aβ42 production and deposition. Nonetheless, mutations in this gene were found to account for 13–16% of all EO-FAD cases [Raux et al., 2005; Sleegers et al., 2006].

More than 180 AD causing mutations were found in PSEN1 gene [Cruts and Broeckhoven, 1998]. Different PSEN1 mutations lead to EO-FAD forms with substantial variation in age at onset (mean 45.5 years old), rate of progression, and severity of disease (average survival after diagnosis 8.4 years) [Heckmann et al., 2004]. Since PSEN1 is a component of γ-secretase, its mutations can change the secretase activity and increase the ratio of Aβ42 to Aβ40 leading to more aggregates. In general, mutations in this gene can be divided into two groups: before protein position 200 and after. Pathology resulting from mutations before position 200 resembles the pathology found in sporadic AD cases, whereas mutations at subsequent positions in the protein result in more severe amyloid deposits in the arteries of the brain that can lead to strokes [Ryan and Rossor, 2010].

EO-FAD causing mutations in PSEN2 are relatively rare compared to PSEN1. They appear to have a more variable penetrance, higher age of onset (53.7 years old) and patients live longer after diagnosis. To date, 38 PSEN2 mutations are known and only 17 are predicted to be disease-causing mutations [Cai et al., 2015]. While the exact function of PSEN2 is unknown, it is believed to have a similar

function to PSEN1, and to cause AD pathology by increasing Aβ42 levels [Ridge et al., 2013]. Besides APP, PSEN1 and PSEN2, mutations in other three genes have been identified as possible causes of EO-FAD. A missense mutation in the tau gene was reported to be tightly linked to AD in a Belgian family [Rademakers et al., 3]. EO-FAD in a Dutch family was also linked to polymorphisms present in the chromosome 7 [Rademakers et al., 2005]. Finally, the gene for PEN2, encoding the γ-secretase component, pen-2, was reported to harbor a missense mutation in an AD family [Sala Frigerio* et al., 2005]. Clearly, all three of these additional EO-FAD candidate genes will require further investigations [Tanzi, 2012].

1.3.2 Genetics of late-onset Alzheimer's disease

Family history is the second strongest risk factor for AD, following advanced age. Twin and family studies indicate that genetic factors are estimated to play a role in at least 80% of AD cases [Tanzi, 2012]. For many years, only one genetic risk factor, the APOE ε4 allele, was firmly implicated in late-onset and early-onset AD, but technological advances, such as large-scale genome-wide association studies (GWAS), that allow to analyze millions of polymorphisms in thousands of subjects, revealed new genes associated to late-onset AD (LOAD) risk [Bettens et al., 2013]. Apolipoprotein E (APOE) is the strongest risk factor for LOAD. APOE gene is located on chromosome 19 and encodes three alleles (ε2, ε3, ε4). APOEε4 is associated to an increased AD risk, in particular: one APOEε4 allele increases AD risk 3-fold, and two APOEε4 alleles increase AD risk by 12-fold, with a decrease in age at onset. Conversely, APOEε2 is associated with decreased risk for AD and later age at onset [Karch et al., 2014]. APOE is a regulator of lipoprotein metabolism and plays several important roles in the central nervous system, such as cholesterol transport, neuroplasticity, and inflammation [Kim et al., 2009a]. In particular, APOE is able to bind Aβ, influencing the clearance of soluble Aβ and Aβ aggregation [Verghese et al., 2013]. Neuropathologic and neuroimaging studies demonstrated that APOEε4 carriers exhibit accelerated and more abundant Aβ deposition than APOEε4-negative individuals [Morris et al., 2010]. Since 2009, European and international genome-wide association collaborations allowed the discovery of at least nine new risk loci for AD, involved in lipid metabolism, inflammatory response and endocytosis [Karch and Goate, 2015].

Besides APOE, variants of other two genes involved in cholesterol metabolism were found to be associated to AD: clusterin (CLU) and the ATP binding cassette subfamily A member 7 (ABCA7).

Clusterin (CLU), an apolipoprotein, is located on chromosome 8, and encodes three alternative transcripts [Rizzi et al., 2009]. Several SNPs have been identified in CLU that confer protection against LOAD [Schrijvers et al., 2011]. Clusterin likely influences Aβ clearance, amyloid deposition, and neuritic toxicity and it seems to modulate the membrane attack complex, where it inhibits the inflammatory response associated with complement activation [DeMattos et al., 2004].

ABCA7 is located on chromosome 19 and alternative splicing can generate two transcripts, both expressed in the brain [Kim et al., 2008]. Several SNPs near and inside the ABCA7 gene were identified by GWAS in LOAD as risk alleles [Vasquez et al., 2013]. ABCA7 functions in the efflux of lipids from cells into lipoprotein particles. In vitro, ABCA7 stimulates cholesterol efflux and inhibits Aβ secretion, and its increased expression has been demonstrated to also increase microglial phagocytosis of apoptotic cells, synthetic substrates, and Aβ [Kim et al., 2006]. APP transgenic mice that are ABCA7-deficient have increased Aβ deposition compared with singly transgenic animals [Kim et al., 2013].

GWAS studies allowed the identification of gene variants associated to LOAD, involved also in neuroinflammation and dysregulation of the immune response, which are central features of AD [Holtzman et al., 2011]. Complement receptor 1 (CR1) encodes for a protein that is involved in the complement response. Several SNPs in this gene are strongly associated with AD risk [Liu and Niu, 2009]. Moreover, CR1 encodes high-expression and low-expression alleles: individuals who are homozygous for the low-expression CR1 allele have <200 copies of CR1 per cell, whereas individuals who are homozygous for the high-expression allele express nearly 1400 copies per cell [Krych-Goldberg et al., 2002]. Higher CR1 protein expression is associated with a higher clearance rate of immune complexes and this dampening of the complement response is associated to a lower risk of developing AD pathology [Rogers et al., 2006]. SNPs identified near to CD33, encoding for a member of the sialic acid-binding Ig-like lectin family of receptors, seem to reduce LOAD risk [Malik et al., 2013]. Since CD33 expression is specifically increased in microglia, and Aβ fagocytosis has been demonstrated to be inhibited in immortalized microglial cells expressing CD33, genetic variations of this gene may play an important role in Aβ clearance and other neuroinflammatory pathways mediated by microglia in the brain [Griciuc et al., 2013]. MS4A is a locus that contains several genes associated

with the inflammatory response: MS4A4A, MS4A4E, and MS4A6E. SNPs localized near to these genes have been associated to both an increase and a decrease in LOAD risk [Karch et al., 2012]. Finally, TREM2 is a receptor expressed in microglia that stimulates phagocytosis and suppresses inflammation. Rare, missense mutations in TREM2 have been reported to increase LOAD risk [Reitz and Mayeux, 2013].

Genes associated with endocytosis and synaptic function were identified in several GWAS of LOAD risk. Bridging integrator 1 (BIN1) for example, is involved in regulating endocytosis and trafficking, immune response, calcium homeostasis, and apoptosis [Ren et al., 2006]. Specific SNPs in BIN1 gene have been linked to increased risk for LOAD [Chapuis et al., 2013]. SNPs 5′ to another gene, the phosphatidylinositol binding clathrin assembly protein (PICALM) predominantly expressed in neurons and involved in APP trafficking in vitro and in vivo, have been associate to a reduced LOAD risk, while few SNPs found in the gene coding for the CD2-associated protein (CD2AP), that is a scaffolding protein involved in cytoskeletal reorganization and intracellular trafficking, are associated to an increased LOAD risk [Karch et al., 2014]. Also SNPs located near to the gene EPH Receptor A1 (EPHA1) and sortilin-related receptor L1 (SORL1), involved in intercellular signaling and vesicle trafficking respectively, have been associated to reduced LOAD risk [Martínez et al., 2005; Rogaeva et al., 2007]. Even if through an unknown mechanism, rare coding variants in the phospholipase D3 gene (PLD3) seem to confer risk for LOAD [Cruchaga et al., 2014]. The GWAS approach involves genotyping common ancestral polymorphisms that usually occur in >5% of the general population.

The risk effects exerted by the GWAS-derived genes discussed above are tiny, that is they confer only a ∼0.10- to 0.15-fold increase or decrease in AD risk in carriers versus non-carriers of the associated alleles, as compared with a four- to 15-fold increase in AD risk owing to the inheritance of APOEε4. This means that probably a substantial proportion of the genetic variance of LOAD remains unexplained by the currently known susceptibility genes [Tanzi, 2012]. With regard to rare variants conferring large effects on risk for LOAD, two rare mutations in the ADAM10 gene were recently reported, that caused AD at an average age of 70 years in seven of 1000 LOAD families tested [Kim et al., 2009b]. ADAM10 encodes for the major α-secretase in the brain, that cleaves within the Aβ domain of APP to preclude the formation of β-amyloid. The two novel ADAM10 LOAD mutations, are located in the prodomain region, and dramatically impair the ability of ADAM10 to carry out α-secretase cleavage of APP. Thus, the two mutations in ADAM10 would appear to be strong candidates for the first rare, highly penetrant pathogenic mutations genetically associated with LOAD [Kim et al., 2009b].

1.3.3 Mosaicism in Alzheimer’s disease

Even if the very first observation that linked mosaicism to sporadic early-onset AD dates to 2004 [Beck et al., 2004], evidences that suggest mosaicism as a potential AD trigger started to accumulate only recently, when technological improvement allowed the study of single neuron genomes and the identification of somatic CNVs and SNVs. Bushman and colleagues, in 2015, found that neurons from Alzheimer's disease post-mortem brains contained more DNA (on average, hundreds of millions of DNA base pairs more) and more copies of the APP gene, with some neurons containing up to 12 copies [Bushman et al., 2015]. Furthermore, Parcerisas and colleagues in 2014 showed that 38 genes in overlap with hippocampus-specific SNVs (hs-SNVs) were present in more than 6 AD individuals out of 17. Interestingly, some of these hs-SNVs were in genes previously related to AD (*e.g.*, CSMD1, LRP2). The most frequent genes with hs-SNVs were associated with neurotransmission, DNA metabolism, neuronal transport, and muscular function. Interestingly, 19 recurrent hs-SNVs were common to 3 AD patients [Parcerisas et al., 2014]. More significantly, in 2019, applying whole exome-sequencing (at ~ 584x coverage) to hippocampal formations, Jun and colleagues showed that SNVs accumulates with increasing age in AD [Park et al., 2019]. SNVs were found in PI3K-AKT, MAPK, and AMPK pathway genes, known to contribute to hyperphosphorylation of the tau protein. Furthermore, a pathogenic brain somatic mutation in PIN1 gene was observed to lead to a loss-of-function (LOF) mutation. It is known through in vitro mimicking that haploinsufficiency of PIN1 aberrantly increases tau phosphorylation and aggregation. Therefore, authors demonstrated for the first time that SNVs can be implicated in the appearance of tau pathology in AD brains [Park et al., 2019]. In addition to SNVs and CNVs, the role of mobile elements somaticism in AD is under intense investigation. As discussed in the previous paragraph, somaticism generated by mobile elements activity is pervasive in human brain, even if its role is not yet fully understood. Recently, multiple observation linked AD-key proteins (Tau, TDP-43) with the reactivation of retrotransposons and brain inflammation [Guo et al., 2018; Krug et al., 2017; Saleh et al., 2019]. For instance, by analyzing human brain transcriptomes, Guo and colleagues identified differential retrotransposon expression signatures in association with neurofibrillary tangle burden along with evidence for widespread activation of selected TE clades, including L1s. To establish specificity, they turned to Drosophila transgenic models, revealing that Tau is sufficient to activate

numerous TEs. Moreover, this activation was found to be further enhanced by aging and with a mutant form of Tau associated with increased neurotoxicity. Finally, they proposed a model in which Tau modulates transcriptional activity at TE loci, possibly via chromatin remodeling, leading to neuronal dysfunction and/or loss [Guo et al., 2018].

New insertions could also impact the normal translocation of proteins between the outer and inner mitochondrial membrane, by affecting the structure of TOMM40 β-barrels [Larsen et al., 2017]. The disruption of mitochondrial protein trafficking across TOMM40 is known to be implicated in several neurodegenerative diseases, such as Alzheimer's disease, Parkinson's disease and Huntington's [Richards et al., 2016]. Interestingly, it has been reported that at least one primate-specific Alu insertion in antisense orientation within TOMM40 intron 6 is associated with late-onset Alzheimer's disease, further suggesting a role for retrotransposons elements that requires further studies [Larsen et al., 2017].

Overall, from the initial associations with tumors, mosaicism has been found to be associated to several pathologies, most importantly brain-related diseases, which affect a consistent fraction of the population. These findings are relatively recent and still have many open questions that need to be addressed. In particular, in Alzheimer's disease there are currently no sufficient evidences to state that mosaicism is one of the major phenomenon that acts as a trigger for the pathology and therefore more studies are required to support these observations.

1.4 Technological advancements and current limitations in mosaicism detection

The detection of mosaicism in human disease has been historically challenging both because it requires analysis of single cells within a given tissue and because mosaicism may be tissue-specific or tissue-limited. Moreover, the detection of mosaicism requires analysis of multiple tissues within an individual. In some cases, the choice of tissues is suggested by the diagnosis of a specific disease or phenotype, such as in the case of the detection of patchy pigmentation. In the absence of phenotypic clues to trigger the search for mosaicism, its detection relies on using sensitive genotyping techniques, such as single-nucleotide polymorphism (SNP) microarrays or next-generation sequencing (NGS) [Conlin et al., 2010; Gottlieb et al., 2010]. In some cases, a vigilant clinician will request analysis of multiple tissues to rule out low-level mosaicism: blood, skin, saliva and any particular affected tissues being the most common. We can list three major class of technologies that permit mosaicism investigation at different level of resolution. These are cytogenetic analysis, microarrays and next generation sequencing.

1.4.1 Cytogenetic techniques

Cytogenetic techniques were the first methods developed to study structural and numerical abnormalities of chromosomes. These mainly include banding and molecular cytogenetics techniques. Banding techniques, such as the analysis of G-banded chromosomes, were developed starting from the 70' [Caspersson et al., 1970], and consist in staining chromosomes with a fluorochrome and examining them with fluorescence microscopy. These staining have a very limited resolution that also depends on the type of microscope. In optimal conditions, banding can detect only large chromosomal aberrations in the range of 5-10 Megabases [Cui et al., 2016]. Molecular cytogenetics, such as fluorescent *in situ* hybridization (FISH) [Pinkel et al., 1986] and comparative genomic hybridization (CGH) [Kallioniemi et al., 1992], started to be available only at the end of the 80'. They rely on fluorescent-labelled probes that can pair with specific nucleic acid sequences in morphologically normal chromosomes. They brought an improvement in the resolution limits which is now in the range of 100-200 Kilobases [Cui

et al., 2016]. Mosaicism has been detected since the earliest usage of cytogenetic techniques in the sixties, when cells within an individual were found to have divergent chromosome contents, such as monosomic or trisomic chromosomes in the karyotype of one cell and a normal karyotype in another [Hirschhorn et al., 1960]. Although mosaic aneuploidy was traditionally the most detected form of mosaicism due to the low resolution of the banding techniques, the development of molecular cytogenetics and with an increase in resolution, led to the discovery of a wide variety of chromosome abnormalities implied in mosaicism, such as insertions, deletions and duplications [Erickson, 2010]. However, cytogenetic detection of low-level mosaicism is challenging, as a sufficient number of cells must be analyzed [Hook, 1977]. Moreover, current resolution is limited to Kilobase scale variants that therefore do not permit SNVs detection.

1.4.2 Microarrays

Beginning in 2005, microarray-based techniques began to replace cytogenetic testing, with the introduction first of array-based comparative genomic hybridization (aCGH), which can analyze genomic copy number variants, followed by genome-wide single nucleotide polymorphism (SNP) arrays. SNPs arrays are a particular type of chip platform originally developed to simultaneously genotype human DNA at thousands of SNPs across the genome. A SNP is defined as a single DNA nucleotide variation with respect to the reference genome within individuals of a population, wherein it has an abundance of 1% or higher. SNPs arrays rely on the biochemical principle that nucleotide bases bind to their complementary partners in Watson–Crick base pairs. Current arrays contain hundreds to millions of unique nucleotide probe sequences, also known as markers, designed to bind to a target DNA sequence (figure 6-A and 6-B). Hybridization is detected through specialized equipment that can measure the intensity signal associated with each probe and its target after pairing. Extensive processing and analysis of these raw intensity measures yield SNP genotype inferences with an accuracy that has been estimated to be over 99% [Affymetrix Genome-Wide Human SNP Array 6.0 data sheet, 2007] (figure 6-C). Since intensity signals depends upon target DNA quantity, as well as the affinity between target and probe, they can then be used to infer CNVs, such as deletions and duplications, and loss of heterozygosity (LOH) [Zhao et al., 2004]. As mentioned, recent platforms can provide even millions of different markers, thus allowing an extensive sample genotyping at SNPs loci.

However, probe number alone is not sufficient to yield good CNV calls. Increasing probe densities in known CNV regions of the genome, in combination with a sufficient genome-wide backbone of probes, generally leads to more detection power. However, if the backbone coverage is not sufficient or regions such as gene deserts are devoid of probes, the design may not detect even some relatively large CNVs. Large CNVs in gene deserts may still be biologically relevant, for example they may be potentially associated with molecular and phenotypic effects that could be transmitted by changes in chromatin conformation and/or regulatory regions. Haraksingh and colleagues demonstrated this concept in 2017 by comparing high-density chips enriched for additional exome content with the same base array types. Moreover, they showed that the additional exome content present in ultra high-density chips (for instance on Illumina Omni arrays) had no clear benefit for CNV discovery. In fact, when comparing the exome-enriched arrays with their relative standard types, they observed a dramatic increase in non-validated CNV calls without a corresponding gain in validated ones [Haraksingh et al., 2017]. Overall, SNPs arrays provided an unprecedented genotyping power, which is only currently overpowered by next generation sequencing technologies. However, CNV detection, especially with ultra high-density arrays, has still limitations. Nevertheless, by using more than one CNV calling algorithm during data analysis, coupled with appropriate extensive experimental validation with orthogonal techniques, the validity of CNV detection could be maximized.

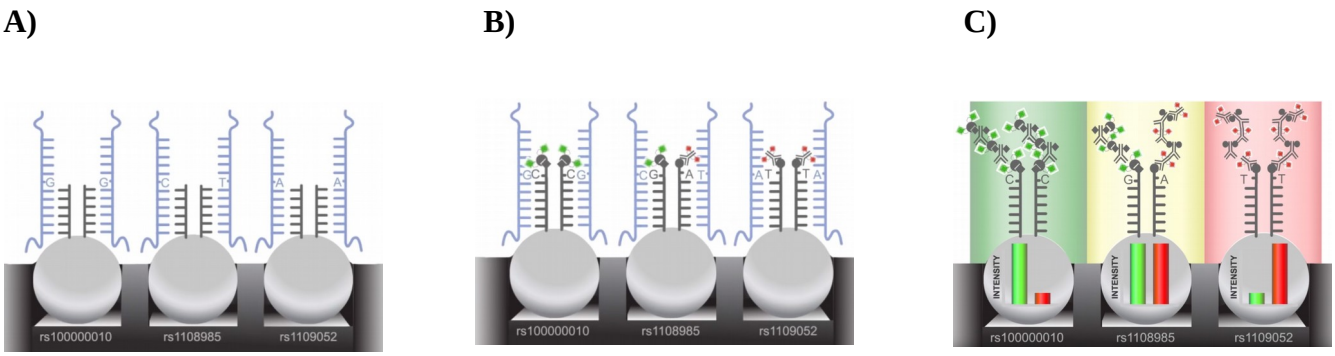


Figure 6: SNPs array technology. **A)** SNPs probes (identified with the "rs" ids) are physicaly attached to the chip. Test DNA sequences (in blue) pairs to the SNPs probes by Watson–Crick base complementary; **B)** Depending on the nucleotide complementarity, one of four labeled bases is inserted; **C)** Labeled nucleotides excited by laser emits signals that are detected through scanners and converted into precise intensity signals. The interpretation of intensity signals yield genotyping calls. When two signals are similarly detected from a single locus, heterozygosity can be called (yellow background). On contrary, when a single signal is detected homozygosity can be called instead (green and red backgrounds).

The advantages of array-based testing (which typically analyses DNA extracted from whole blood) for mosaicism detection mainly consist in its high throughput conferred by the thousands to millions of genomic sites that can be tested at the same time. Furthermore, many cells and different cell types are analyzed simultaneously while samples do not require culturing, which might itself cause mutations. Several studies using aCGH tested the ability of this technique to identify mosaicism, and it was demonstrated that mosaicism could be identified when variant cells constituted >10% of the total cell population. SNP arrays are much more sensitive than aCGH for mosaicism detection, and mosaicism involving <5% of cells has been detected using these arrays [Conlin et al., 2010; Rodríguez-Santiago et al., 2010]. In addition to detecting mosaicism at lower levels, SNP arrays are also able to genotype the SNPs, thus aiding the analysis of the genetic mechanism by which the mosaicism has occurred, (see below) [Conlin et al., 2010]. However, microarray-based techniques have intrinsic limitations mainly due to probes design. Resolution strictly depends on the number and on the mapping position of the different probes available on the chip, which directly affect the size of the potential CNVs to be detected. Moreover, SNVs can be identified only if there is a specific probe designed for. Unique probes design is challenging as a consistent fraction of the genome is composed by repetitive elements, therefore this restricts CNVs calling for transposable elements to only large variants or unique regions in the genome. Other limitations then consist in the type of CNVs that may be detected. Insertions can remain undetected or be confused with duplications, depending on the absence or the presence of probes within the inserted sequence. Moreover, not all individuals with uniparental disomy (UPD) can be identified by the SNP arrays. In patients who have constitutional heterodisomy, a normal genotype pattern is indeed detected, and the fact that the two chromosomes are inherited from one parent can be detected only if parental samples are analyzed in conjunction with their child (trios).

1.4.3 Next Generation Sequencing

The massive parallel sequencing technology known as next generation sequencing (NGS) has revolutionized the biological sciences, outperforming previous Sanger-based sequencing strategies. The term “massive parallel” mostly derives from the NGS instruments capability to perform both the enzymology and data acquisition in an orchestrated and stepwise fashion, enabling sequence data to be generated from tens to thousands to billions of DNA template molecules simultaneously. To date, NGS has the highest level of resolution in both SNVs and CNVs detection potentially being able to detect also *de-novo* sequence variants [Bras et al., 2012; Mardis, 2008]. Within the NGS technologies there are two major paradigms: short-read sequencing and long-read sequencing. Short-read sequencing approaches provide lower-cost, higher-accuracy data that are useful for population-level research and clinical variant discovery. By contrast, long-read approaches provide read lengths that are well suited for *de-novo* genome assembly applications and full-length isoform sequencing [Goodwin et al., 2016]. The most applied NGS technology is currently represented by the short-reads Illumina sequencing, which is based on the concept of sequencing by synthesis (SBS). This method is characterized by 1) chemical or physical fragmentation of the genome or other DNA/RNA sample to be sequenced, 2) the generation of a sequencing “library” that results from the attachment of universal adaptors (synthesized oligonucleotides of known sequence) at each end of the template fragments to be sequenced, and 3) on-surface template amplification. This is made possible by virtue of library fragment hybridization to covalently attached oligonucleotides that have sequence complementarity to the synthetic adaptors. Every oligonucleotide provides a free 3′ OH for enzymatic extension on the templates which is coupled with on-instrument data detection (figure 7). Amplification is made on-surface establishing a fixed X–Y coordinate for each template that, in turn, permits data from nucleotide incorporation steps to be assigned to a specific template. Because SBS method is detecting nucleotide incorporation from a population of amplified template molecules at each X–Y coordinate, and because of the increasing noise over sequential incorporation and imaging cycles, SBS is ultimately limited in its length of sequence read (“read length”) that currently can reach 300 nucleotides in length. Although improvements on the enzymology and nucleotide chemistry or synthesis, as well as more sensitive detectors, have yielded improved signal-to-noise over time permitting increased read lengths, SBS read lengths remain shorter than Sanger and alternative “long-reads” strategies should be put in place to obtain reads longer than 800 nucleotides [McCombie et al., 2019].

Finally, SBS methods can provide both single-end or paired-end reading (figure 8). In single-end reading, the sequencer reads a fragment from only one end to the other, generating the sequence of base pairs. In paired-end reading instead, it starts at one read, finishes this direction at the specified read length, and then starts another round of reading from the opposite end of the fragment. Paired-end reading improves the ability to identify the relative positions of various reads in the genome during the downstream bioinformatic analyses, making it much more effective than single-end reading in resolving structural rearrangements such as gene insertions, deletions, or inversions.

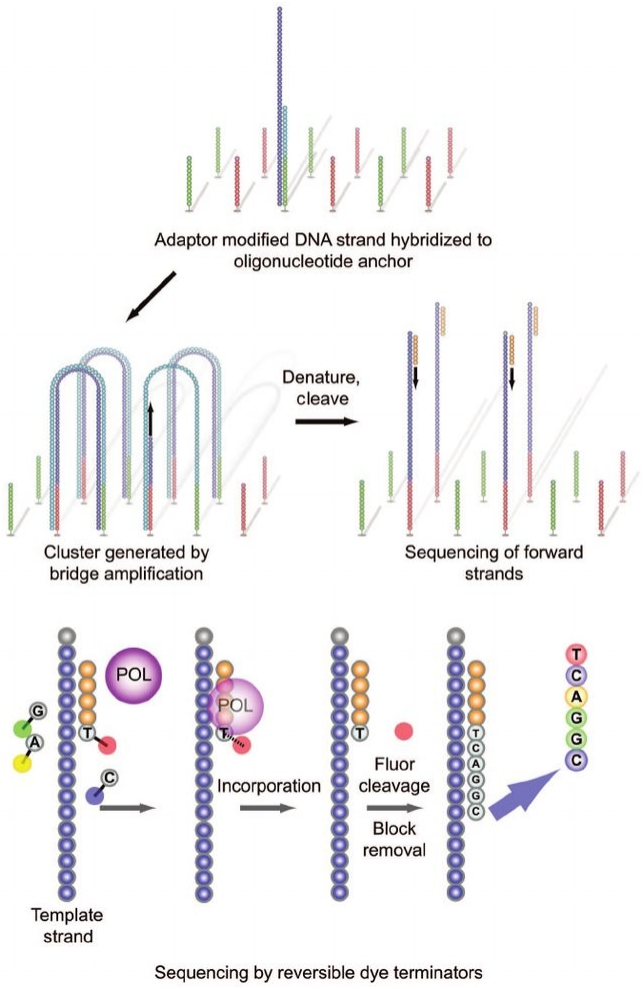


Figure 7: Illumina single-reads sequencing schema. After template fragmentation, adapter-modified, single-stranded DNA is added and immobilized by hybridization on physical supports. Bridge amplification generates clonally amplified clusters. Clusters are denatured and cleaved; sequencing is initiated with addition of primer, polymerase (POL) and 4 reversible dye terminators. Postincorporation fluorescence is recorded. The fluor and block are removed before the next synthesis cycle. Taken from [Voelkerding et al., 2009].

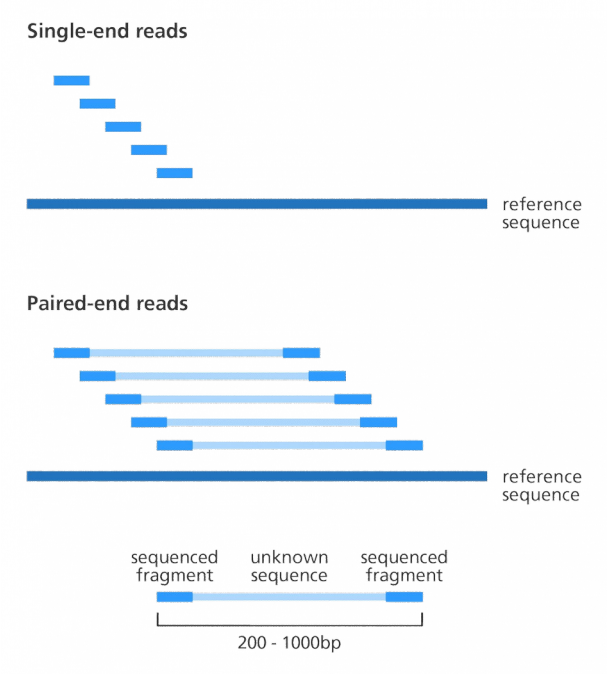


Figure 8: Single-end vs Paired-end reading. Paired-end reads can provide more precise alignments when reads are mapped to the reference genome, in particular within repetitive elements.

When applied to mosaicism detection, NGS approaches show considerable strength with respect to Sanger-based and SNPs array technologies. This consists in their high processivity and quantitative nature, based on reporting reads counts (a quantitative measure) for each allele as integer counts on the sequences of multiple different molecules for every nucleotide in the genome. These data are amenable to statistical analysis, which can distinguish mosaicism from sequencing errors, and that are continuously improved through software and protocol developments.

The massive throughput of NGS has another benefit, since it does not require genetic mapping data to identify causative genetic variants. For this reason, subtractive informatic approaches (in which two samples are informatically processed to identify only those positions in the genome that differ) can be coupled to NGS to identify mosaic alterations. Trio sequencing (see above) has been used to identify *de-novo* alterations as a cause of non-mosaic heritable disorders [Veltman and Brunner, 2012]. Furthermore, intra-patient subtraction can be used to identify mosaic disorders, as it has been done in cancer genomics, by comparing DNA samples from affected and unaffected tissues from the same

patient. This approach has a measurable false-positive rate, which is due to a combination of molecular and informatics errors, yet it has been successfully used to identify numerous mosaic disorders as reported by Conlin and colleagues in 2010, and cited in chapter 1.2. NGS recently started to be applied also at single cells, pushing the limits of mosaicism detection to intra-tissue levels [Cai et al., 2014; Lodato et al., 2018].

Next generation sequencing technologies are not exempt from limitations. Current sequencing costs are forcing the majority of scientists to adopt short reads approaches on bulk tissues and to use low/medium depth of coverages (at about 30x) when sequencing whole genomes. The depth of coverage is defined as the number of times a nucleotide is read during sequencing. The ability to detect somaticism depends on the frequency of cells affected within a sample and on the depth of coverage. For instance, low frequency mosaicism would be difficult to be observed with low/medium coverages from bulk tissues and this is explained by the combination of the small number of sequenced cells (low coverage) with their low probability of including the somatic variant (low frequency). Although decreasing in sequencing costs may overcome this issue making high coverage analyses a common practice, further limitations would still affect mosaicism detection. Short-reads sequencing approaches have intrinsic mapping limitations, that particularly affect repetitive regions (more details in the following chapter). An alternative strategy can rely on long reads sequencing approaches. These technologies are more expensive, but they allow the mapping of longer fragments to the reference genome, improving therefore CNVs discovery. However, they have higher error rate with respect to short-reads sequencing approaches (estimated to be 15% and 0.1% respectively) which make them less suitable for reliable SNVs calling. To overcome costs, exome sequencing and targeted sequencing approaches can be a valid alternative to whole genome sequencing. The latter consists in sequencing only portions of the genome through capture by customizable probes [Pagnamenta et al., 2012; Jamuar et al., 2014]. However, although very high coverages can be achieved with relatively little efforts, mapping issues may still arise depending on the targeted regions posing the limiting factor to the probe design.

These aspects clearly shed light on the fact that a single technology, hardly would result in detecting all sources of mosaicism at once, thus leaving the decision to which scan first to researchers considering the tissue and disease of interest.

1.4.3.1 Retrotransposons detection through sequencing approaches

It has been almost 20 years since the publication of the first human genome reference gave researchers a genome-wide view of human transposable element content [Lander et al., 2001]. From that point, the identification of transposable element, and in particular retrotransposons, mainly resulted from the available high-throughput methods coupled with bioinformatics strategies.

The majority of the WGS data nowadays comes from Illumina platforms and generally consists of millions to billions of 100-150 bp reads in pairs, where each read in a pair represents the end of a longer fragment generally of the length of about 500 bp. Small variants are detected through accurate alignment to the reference genome followed by examination of the deviations from the reference sequence. However, structural variants are much more complex to call, principally because the presence of rearrangements must be inferred from short reads that generally do not span the entire interval affected by the variants. Retrotransposition events are even more difficult to detect, since the repetitive nature of the sequences can result in reads that not map in unique regions within the reference human genome. To detect structural variants and retrotransposon insertions by short-reads sequencing techniques, two approaches result useful: inference from discordant read-pair mappings and clustering of ‘split’ reads that shares common alignment junctions.

Discordant reads-pair mapping approaches relies on the information yielded by discordant read pairs. A discordant read-pair is one pair of reads whose mapping is inconsistent with the sequencing library preparation parameters. During library preparation, genomic DNA is sheared physically or chemically, and fragments of a specific size are selected. Given an expected fragment size distribution, anything significantly outside of that range may be considered discordant. What is significantly outside of the expected range of fragment sizes can be determined only after sequencing and mapping on the reference genome, based on the distribution of distances between the mapping of the two reads in a pair. Additionally, given the library preparation method and/or sequencing platform used, the expected orientation of the ends of the read-pair may be known. For instance, Illumina read pairs are ‘forward-reverse’ meaning that two reads in a pair must have opposite orientation with respect to the reference genome. Reads inconsistent with this pattern may be considered discordant. Finally, reads pairs mapping in which the two reads do not map in the same chromosome are clearly also considered discordant. When using discordant read pairs to inform structural retrotransposon discovery, typically

one end will be ‘anchored’ to an unique sequence of the genome while the other may map to multiple distal locations located within various copies of a repeat element throughout the genome. Once ‘one-end-repeat’ reads have been identified, the non-repeat ends of the read pairs are clustered by genomic coordinates, and possibly filtered by various criteria concerning mapping quality, consistency in read orientations, underlying genomic features, and others (figure 9). Discordant reads mapping alone does not yield exact junctions between the insertion and the reference sequence, therefore sites localized by discordant read mapping are typically refined through local sequence assembly or most commonly through split-read mapping, which, as the name may suggest, relies on split reads. Split reads are defined as reads that contains one segment mapping to a location of the reference genome, while the remaining segment to one or more locations distal from the first. In analogy with discordant reads-pairs, retrotransposons discovery relies on reads segments ‘anchored’ in unique sequences while the other segment may, on may not, map in the repetitive regions of the genome (figure 9). The ability to map the segment to a repetitive region, in fact, mainly depends on the length of the fragment, which confers mapping specificity. Split reads approaches are powerful. They can identify the exact insertion location at base-pair resolution. However, the application of such approaches required specific aligner that can provide alternative mapping locations. Moreover, current reads length may limit mapping quality and provide lower sensitivity and specificity with respect to discordant read-pair mapping [Ewing, 2015]. Several softwares have been designed to detect retrotransposition events. They usually rely on one of the two aforementioned strategies or on combinations of the two. However, it has been demonstrated that most of the available methods still do not produce highly concordant results [Ewing, 2015] requiring further developments in both mobile elements detection methods and sequencing technologies. Finally, in 2017, the Mobile Element Locator Tool (MELT) was released, a software specifically developed as part of the 1000 Genomes Project, able to perform mobile element insertions (MEIs) discovery on a population scale [Gardner et al., 2017]. It was designed to rely on both discordant reads pairs as well as split reads, from Illumina WGS data, to identify precise breakpoints and target site duplication at candidate MEI sites. It also performs genotyping across samples for both novel and reference mobile element copies to provide a comprehensive map of polymorphic MEIs in a given genome. MELT was proven to outperform existing MEI discovery tools in terms of speed, scalability, specificity, and sensitivity, while also detecting a broader spectrum of MEI-associated features such as 5′ inversions and 3′ transductions. Thus, it is becoming common practice to implement

its application in several pipelines and studies aimed to investigate mobile elements from whole genome sequencing data.

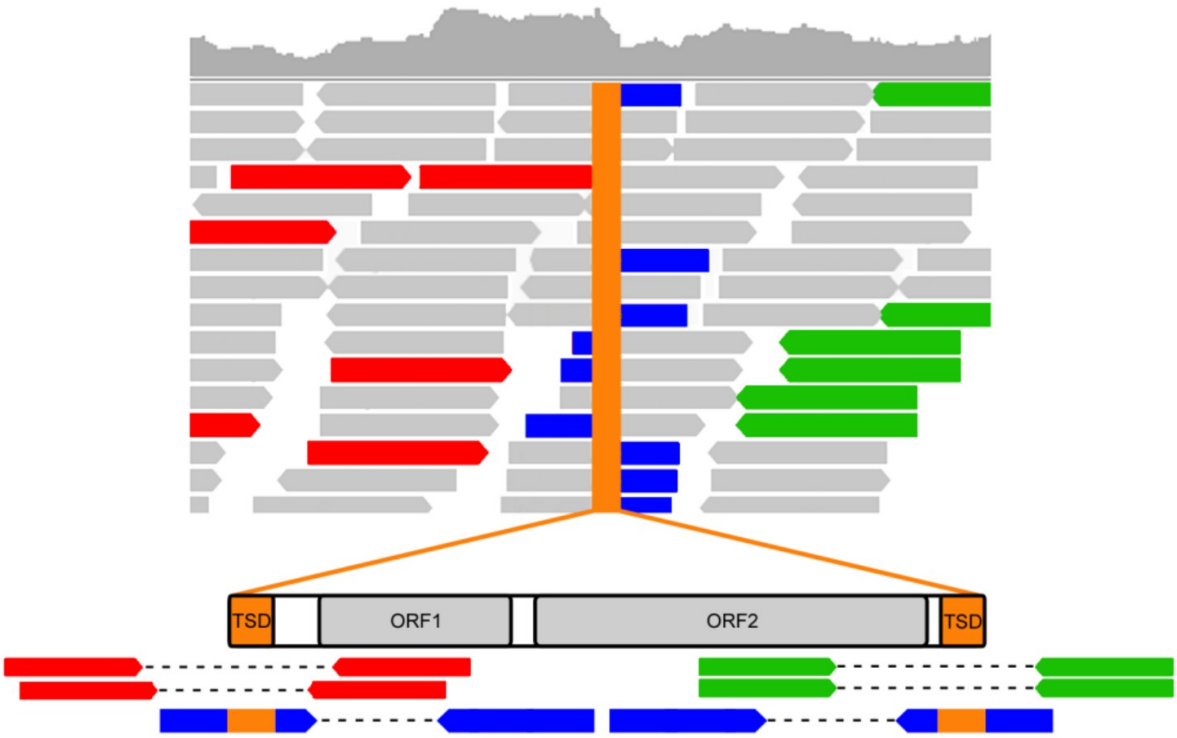


Figure 9. Discordant read pairs and split reads strategies in retrotransposon detection. The picture represent a L1 integration within the genome. Mapped reads are shown as colored arrows under the coverage curve (grey top curve). Discordant read pairs (in red and green), unlike split reads (in blue), cannot provide nucleotide breaking point resolution. Adapted from [Gardner et al., 2017].

1.5 Research aims and objectives

Mosaicism has been proven to be a crucial phenomenon in cancer while its role in other diseases, especially those of the nervous system, is under evaluation. So far, findings have been strictly associated to technological advancements, due to the improvements in the capability of investigating both the quantity and the sequence of the cellular DNA. High throughput approaches, such as SNPs arrays and sequencing, have been elected as one of the most powerful technologies in mosaicism detection, primarily thanks to the high level of achievable resolution. Interestingly, in both healthy and impaired brains, it was recently demonstrated that the pervasive activity of retrotransposons act as a source of mosaicism, while their pathogenic role remain unclear.

Inspired by this lack of knowledge, I investigated SNVs abundance and retrotransposon copy number, mainly L1's, in an AD post-mortem tissues cohort. This is composed by five different tissues belonging to the same donors, from both post-mortem control individuals and AD patients. Tissues consisted in cerebellum, frontal cortex, temporal cortex, hippocampus and kidney, which allowed us to investigate mosaicism in different human brain regions. Giving the current technological limitations that affect mosaicism detection, the dataset was studied by coupling two different strategies and by developing a new targeted sequencing approach.

In order to call the highest number of SNVs with the highest quality possible, the most dense SNP array available to date (*i.e.* with the highest number of different SNP probes) was applied upon cerebellum, frontal cortex and kidney tissues. Despite arrays were also used to assess CNV content, they are ineffective in the detection of new retrotransposition events. Therefore, a second strategy was applied taking advantage of whole genome sequencing with short-reads high and high coverage (~100x), further extending the analyses to temporal cortex and hippocampus tissues. This approach was also aimed to further expand SNVs detection to the whole genome.

The recent discovery of somatic deletions mediated by L1 elements expanded the way in which such class of retrotransposons may generate mosaicism. Whole genome sequencing strategies were proven to be successfully in retrotransposition detection. However, uniquely mapped short reads that belongs to repetitive elements into specific genomic regions remains problematic. Therefore, for instance, low levels of somatic deletions that only span repetitive sequences may remain undetected. To improve mapping specificity and resolution, we developed a targeted sequencing approach designed to

specifically amplify and sequence a genomic region upstream and spanning the 5’ end of a subset of ~ 3000 full-length L1. We named this technology LIFE-seq from LINE-1 Five prime End sequencing. As part of my contribution, I’ve developed and applied an analysis pipeline to genotype sequenced loci in a subset of the AD dataset.

Finally, by taking advantage of both SNPs array and WGS data, I’ve also started to investigate the potential presence of variants within a subset of genes known to be associated with AD.

1.5.1 The AD cohorts composition

The human genomic DNA samples used in both SNPs arrays and sequencing experiments were extracted from tissues of two different cohorts of patients.

The Spanish cohort, received from Prof. Isidro Ferrer (Bellvitge Neuropathology Institute, Barcelona), comprised samples of frontal cortex (FC) from 15 AD patients at the final Braak stages V-VI (severe AD), 9 patients at Braak stages I-II (mild AD), and 9 healthy controls.

The Brazilian cohort, provided by Prof. Lea Grinberg (Brain Bank of Sao Paulo), comprised samples of frontal cortex, temporal cortex (TC), hippocampus (HIP), cerebellum (CER) and an extra-nervous tissue: the kidney (KID), taken from 8 AD patients at Braak stages IV-VI and 9 healthy patients.

1.5.2 Reference retrotransposons databases

Depending on the experiment, different sources of retrotransposon annotations were used in this study. In order to test genomic CNVs for L1 content, L1 sequences were at first retrieved from three L1base databases in genome build hg38 (Full-length L1: FLI-L1, L1 with intact ORF2: ORF2-L1 and non-intact L1 longer than 4,500 bp: FlnI-L1). I’ve converted their coordinates by aligning the sequences to genome build hg19 and then, from the results we kept only coordinates that presented full coverage and 100% of sequence identity with the reference genome. Additionally, from UCSC table, RepeatMasker

annotation were retrieved in genome build hg19. Finally, I've generated a subset composed by only L1 sequences longer than 4,500 bps from the RepeatMasker annotations.

During sequencing experiments instead, I've relied on Alu, L1 and SVA consensus sequences obtained from the 1000 Genomes project [Sudmant et al., 2015]. Finally, retrotransposon annotations were collected from the RepeatMasker tracks provided by the *MELT* software [Gardner et al., 2017].

Chapter II

SNP array suggest increased mosaicism due to SNVs in AD brain tissues

2.1 Introduction

Recent evidences suggested a potential involvement of somatic SNVs and retrotransposon mobilization in Alzheimer's disease [Park et al., 2019; Guo et al., 2018]. Here, I aim to further investigate this hypothesis by taking advantage of post-mortem tissues of two AD cohorts. The Brazilian cohort comprises cerebellum, frontal cortex, and kidney tissues of the same individual for 8 AD and 9 CTRs. The Spanish cohort is composed by frontal cortex tissues obtained from 24 AD individuals and 9 CTRs. SNPs array experiment were performed on these cohorts, taking advantage of the most dense platform available on the market. This resulted in the highest possible resolution in both SNPs and genomic CNVs calling. I've first investigated the genomic CNVs properties, such as the total number and the total lengths of CNVs per sample, to test for the presence of differences between AD and CTRs samples. I then evaluated the effects of genomic CNVs in impacting L1 copy number by relying on several repositories of L1 annotations. Next, I focused my attention on somatic SNVs. In particular, studying the Brazilian cohort, I investigated the presence of intra-individual genotype differences among tissues that can be ascribed to early onset somatic variants. Somatic SNVs were specifically defined as early onset (in opposition to late onset) when they could be defined as germline variants when investigated in a single tissue, meaning that they were present in the majority of the cells of the given tissue. Subsequently, I aimed to decipher the mutagen processes behind somatic SNVs formation by applying mutational signature analyses.

2.2 Materials and Methods

2.2.1 Sample collection

Samples used in the SNPs array experiments were selected from the whole Spanish cohort composed by 9 CTRs and 24 AD and from a set of the Brazilian cohort. This comprised only frontal cortex, cerebellum and kidney tissues of 9 CTRs and 8 AD samples (Table 1).

2.2.2 The Illumina Infinium ultra high-density chip assay

Illumina® Infinium OMNI 5 arrays, composed by more than $4.3x10^6$ SNPs probes, were chosen to carry out an SNPs array experiment upon genomic DNA extracted from our collection of samples. Assays were conducted according to manufacturer's protocols (Illumina Infinium LCG Quad Assay). Intensity signals were converted into CNVs calls using the *PennCNV* tool (version 2014 May 07, parameters: '-confidence') [Wang et al., 2007]. The PFB file (Population Frequency of B allele), which contains the population frequencies, the genomic coordinates of all the markers of the array and is required to reconstruct SNPs genotypes, was downloaded from the *PennCNV* site (http://penncnv.openbioinformatics.org/en/latest/user-guide/download/YALE_Merged_PFB_hg19.pfb on April 2017, v. 2014 Aug 18 from Szatkiewicz *et al.*, 2015). By strictly following the *PennCNV* workflow, 91,193 markers that lacked in PFB informations and two low quality samples (C01_S005_A_FC and C09_S018_A) were discarded. *Plink1.9* (version 1.90b3.31, subcommands *--neighbor , --check-sex*) [Purcell et al., 2007] and *SNPhylo* (version 20180901, default parameters) [Lee et al., 2014] were used to reconstruct sample's relationship and to check for the correctness of the gender metadata.

CNVs with confidence scores below 30 where then discarded. Finally, CNVs that were probably split during the calling process were joined using the PennCNV's *clean.pl* script (default parameters) resulting in the identification of a total of 5,170 CNVs.

SAMPLE ID	QUALITY	GENDER	AGE	COHORT	SAMPLE TYPE	TISSUE	BRAAK NTF
C01_S001_A_CER	PASS	M	72	Brazilian	AD	CER	6
C01_S001_A_FC	PASS	M	72	Brazilian	AD	FC	6
C01_S001_A_K	PASS	M	72	Brazilian	AD	KID	6
C01_S002_A_CER	PASS	M	80	Brazilian	AD	CER	3
C01_S002_A_FC	PASS	M	80	Brazilian	AD	FC	3
C01_S002_A_K	PASS	M	80	Brazilian	AD	KID	3
C01_S003_A_CER	PASS	M	82	Brazilian	AD	CER	6
C01_S003_A_FC	PASS	M	82	Brazilian	AD	FC	6
C01_S003_A_K	PASS	M	82	Brazilian	AD	KID	6
C01_S004_A_CER	PASS	M	87	Brazilian	AD	CER	4
C01_S004_A_FC	PASS	M	87	Brazilian	AD	FC	4
C01_S004_A_K	PASS	M	87	Brazilian	AD	KID	4
C01_S005_A_CER	PASS	F	80	Brazilian	AD	CER	4
C01_S005_A_FC	LOW_QUALITY	F	80	Brazilian	AD	FC	4
C01_S005_A_K	PASS	F	80	Brazilian	AD	KID	4
C01_S006_A_CER	PASS	F	83	Brazilian	AD	CER	6
C01_S006_A_FC	PASS	F	83	Brazilian	AD	FC	6
C01_S006_A_K	GENDER_ERROR	F	83	Brazilian	AD	KID	6
C01_S007_A_CER	PASS	F	90	Brazilian	AD	CER	4
C01_S007_A_FC	PASS	F	90	Brazilian	AD	FC	4
C01_S007_A_K	PASS	F	90	Brazilian	AD	KID	4
C01_S008_A_CER	PASS	F	92	Brazilian	AD	CER	4
C01_S008_A_FC	PASS	F	92	Brazilian	AD	FC	4
C01_S008_A_K	PASS	F	92	Brazilian	AD	KID	4
C01_S001_C_CER	PASS	M	75	Brazilian	CTR	CER	2
C01_S001_C_FC	PASS	M	75	Brazilian	CTR	FC	2
C01_S001_C_K	PASS	M	75	Brazilian	CTR	KID	2
C01_S002_C_CER	PASS	M	79	Brazilian	CTR	CER	0
C01_S002_C_FC	PASS	M	79	Brazilian	CTR	FC	0
C01_S002_C_K	PASS	M	79	Brazilian	CTR	KID	0
C01_S004_C_CER	PASS	M	85	Brazilian	CTR	CER	2
C01_S004_C_FC	PASS	M	85	Brazilian	CTR	FC	2
C01_S004_C_K	PASS	M	85	Brazilian	CTR	KID	2
C01_S005_C_CER	PASS	F	61	Brazilian	CTR	CER	0
C01_S005_C_FC	PASS	F	61	Brazilian	CTR	FC	0
C01_S005_C_K	PASS	F	61	Brazilian	CTR	KID	0
C01_S006_C_CER	PASS	F	75	Brazilian	CTR	CER	2
C01_S006_C_FC	PASS	F	75	Brazilian	CTR	FC	2
C01_S006_C_K	PASS	F	75	Brazilian	CTR	KID	2
C01_S007_C_CER	PASS	F	76	Brazilian	CTR	CER	1
C01_S007_C_FC	PASS	F	76	Brazilian	CTR	FC	1
C01_S007_C_K	PASS	F	76	Brazilian	CTR	KID	1

SAMPLE ID	QUALITY	GENDER	AGE	COHORT	SAMPLE TYPE	TISSUE	BRAAK NTF
A08_00153	PASS	F	74	Spanish	AD	FC	5
C04_S004_A	PASS	M	79	Spanish	AD	FC	5
C04_S006_A	PASS	M	78	Spanish	AD	FC	5
C04_S009_A	PASS	F	85	Spanish	AD	FC	5
C04_S012_A	PASS	F	81	Spanish	AD	FC	5
C04_S013_A	PASS	M	87	Spanish	AD	FC	5
C04_S014_A	PASS	M	84	Spanish	AD	FC	5
C04_S018_A	PASS	F	77	Spanish	AD	FC	6
C09_S001_A	PASS	M	84	Spanish	AD	FC	4
C09_S002_A	PASS	M	75	Spanish	AD	FC	6
C09_S003_A	PASS	F	75	Spanish	AD	FC	4
C09_S004_A	PASS	F	79	Spanish	AD	FC	4
C09_S005_A	PASS	F	96	Spanish	AD	FC	5
C09_S006_A	PASS	M	84	Spanish	AD	FC	4
C09_S007_A	PASS	M	75	Spanish	AD	FC	5
C09_S008_A	PASS	F	83	Spanish	AD	FC	4
C09_S009_A	PASS	F	81	Spanish	AD	FC	4
C09_S011_A	PASS	M	77	Spanish	AD	FC	5
C09_S013_A	PASS	F	86	Spanish	AD	FC	6
C09_S015_A	PASS	F	81	Spanish	AD	FC	5
C09_S017_A	PASS	F	81	Spanish	AD	FC	4
C09_S018_A	LOW_QUALITY	M	79	Spanish	AD	FC	4
C09_S019_A	PASS	M	89	Spanish	AD	FC	4
C09_S020_A	PASS	M	86	Spanish	AD	FC	5
C02_S001_C	PASS	F	65	Spanish	CTR	FC	0
C02_S002_C	PASS	M	67	Spanish	CTR	FC	0
C02_S003_C	PASS	F	69	Spanish	CTR	FC	0
C02_S004_C	PASS	M	85	Spanish	CTR	FC	0
C02_S006_C	PASS	M	66	Spanish	CTR	FC	0
C02_S005_C	PASS	M	78	Spanish	CTR	FC	0
C02_S007_C	PASS	M	61	Spanish	CTR	FC	0
C09_S020_C	PASS	F	66	Spanish	CTR	FC	0
C09_S036_C	PASS	M	75	Spanish	CTR	FC	0

Table 1: Cohorts composition with metadata information. Samples discarded are highlighted in red.

2.2.3 CNVs annotation and bioinformatics analyses

Genomic CNVs were grouped by sample, tissue and type of CNV. Dissimilarities in the distributions of CNVs lengths and CNVs counts were evaluated for each of the CNV group using the R statistical software (version 3.3.2; 2016-Oct-31) [R Development Core Team, 2008]. L1 collections used in assessing L1 content in genomic CNVs were generated from different sources. L1 annotations were retrieved from the three L1base databases in genome build hg38 (L1 Full-length: FLI-L1 , L1 with intact ORF2: ORF2-L1 and non-intact L1 longer than 4500 bp: FLnI-L1; Downloaded on April 2017, latest update: 2016-07-09) [Penzkofer et al., 2017] and then converted to genome build hg19 by aligning them using a local *megablast* (version 2.4.0, 2016-Aug-5, parameters: '-perc_identity 100'). From the results only coordinates that presented full coverage and 100% of sequence identity with the reference genome were kept. From UCSC table [Karolchik et al., 2004], RepeatMasker annotation were retrieved in genome build hg19 (downloaded on April 2017). Finally, a subset composed by only L1 sequences longer than 4,500 bps was generated from the RepeatMasker annotations.

BEDtools suit (version 2.25) [Quinlan and Hall, 2010] was used to calculate the coverage and the number of overlaps between each CNV and L1 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 L1 element over its total length. For each sample, we also calculated a total coverage by using the sum of the CNV lengths and the sum of the L1 fragments in overlap with them, and we used these values to draw the boxplots and perform the statistical analyses. All statistical differences were evaluated by performing T-tests with FDR correction using the R statistical software.

2.2.4 SNVs calling and signature analyses

SNPs from the Brazilian samples were converted to the forward strand of the reference genome build GRCh37 (version GCA_000001405.1) with Plink1.9 (option --keep-allele-order --a2-allele nucleodite annotation file). Nucleotide annotation file was obtained from the Illumina Infinium Omni5 array (version 1.3) from https://www.well.ox.ac.uk/~wrayner/strand/RefAlt.html. Converted SNPs were then filtered with *Plink1.9* (options --geno 0.1, --mind 0.01, --maf 0.05) to select only high-quality SNPs calls. Genotypes of tissues belonging to the same individuals were next compared in pairs (CER vs FC, FC vs KID, CER vs KID). Loci for which pairs of tissues shown different genotypes (SNVs) were selected and counted. Subsequent mutational signature analysis was performed using SNVs called after having set KID as reference tissue (after Plink1.9 filtering). *Helmsmann* software (version 1.4.2) [Carlson et al., 2018] was applied for the matrix generation (parameters --length 3, –decomp nfm), while *DesonstructSign* R package (version 1.8.0) [Rosenthal et al., 2016] was used for the signature assignation. COSMIC signatures database (version 3) was selected as reference repository of signatures (http://cancer.sanger.ac.uk/cosmic/signatures). Prior to assignation, matrix normalization was performed with the number of times that each trinucleotide context was observed in the genome (option tri.counts.method = 'genome').

All statistical differences were evaluated by performing Student T-tests with FDR correction using the R statistical software.

2.3 Results

2.3.1 SNP array experiments

Illumina® Infinium OMNI 5 arrays were applied on the Brazilian and Spanish cohorts (Table 1) due to their exceptional SNPs probes density and ultra-high coverage of the whole Human genome (> $4.3x10^6$ probes). A total of 24 samples from 8 individuals with AD and 27 samples from 9 individuals from controls were analyzed from the Brazilian cohort while 24 AD samples and 9 control samples were studied from the Spanish cohort. 2 samples showed bad quality during intensity signals detection (C01_S005_A_FC and C09_S018_A) and were thus discarded. The remaining sample's intensity signals were converted into SNPs calls, which were further used to check sample's relationship and genders with *Plink1.9* analyses. My investigations indicated that the Brazilian sample C01_S006_K was not related to the other two tissues collected from the same individual, that means its gender was not coherent with its metadata, and was therefore discarded. SNPs data were additionally used to test clustering of samples through a *SNPhylo* analysis. A good clustering of samples that belonged to the same individuals was observed, confirming previous quality checks. Moreover, it was observed that Brazilian and Spanish samples did not generate two separated groups (figure 10).

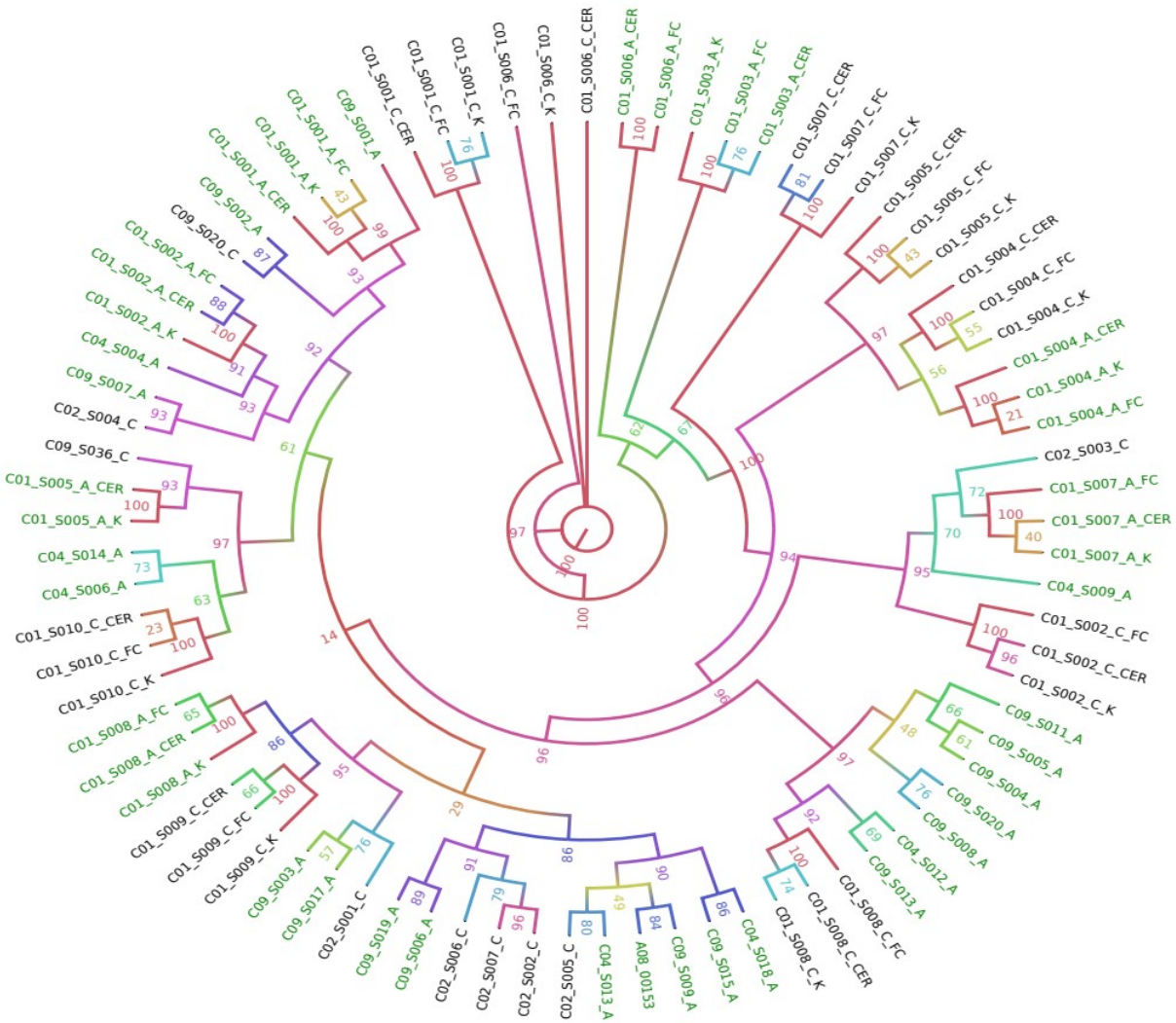


Figure 10: SNPhylo tree representation of SNPs array data relationships. AD samples are highlighted in green while CTRs in black. Brazilian samples are characterized by the presence of a suffix that refers to the tissue, see Table 1 for further ID info. Bootstraps values are represented with different color scales.

2.3.2 CNV analyses on the whole AD cohort

Intensity signals of the SNPs array experiment performed on both Brazilian and Spanish cohorts were used to call a total of 5,170 genomic CNVs with respect to the reference genome build hg19. These consisted in: 138 homozygous deletions (homo-del), 4,290 heterozygous deletions (hetero-del) and 742 duplications (dupl). CNVs Intrinsic properties, as total CNVs counts and total CNVs lengths per sample were then investigated. No significant differences between AD and CTRs were found (figure 11).

Genomic CNVs were next investigated for their content in L1 elements. I could not find any overlap between the identified CNVs and the 146 full-length intact forms of L1s elements (FLI-L1) nor the Full-length L1s with non-intact ORF2 (ORF2-L1) annotated in L1base database. Analysis was further extended to the entire non-intact full-length L1s group (FLnI-L1) consisting of ~11,000 elements mapped on hg19 from the 13,418 L1s longer than 4,500 bp from L1base annotation of the hg38. Taking into account heterozygous deletions, results indicated that, on average, there is a higher content of FLnI-L1 in deletions of AD patients with respect to CTRs. This is just a tendency for the 3 tissues of the small Brazilian cohort (CER, FC and KID), while the difference reaches a significant value for the larger Spanish cohort ($p = 8.5x10^{-3}$) (figure 12). The same analysis performed using all L1 fragments of any length annotated in the human genome did not show any significant difference.

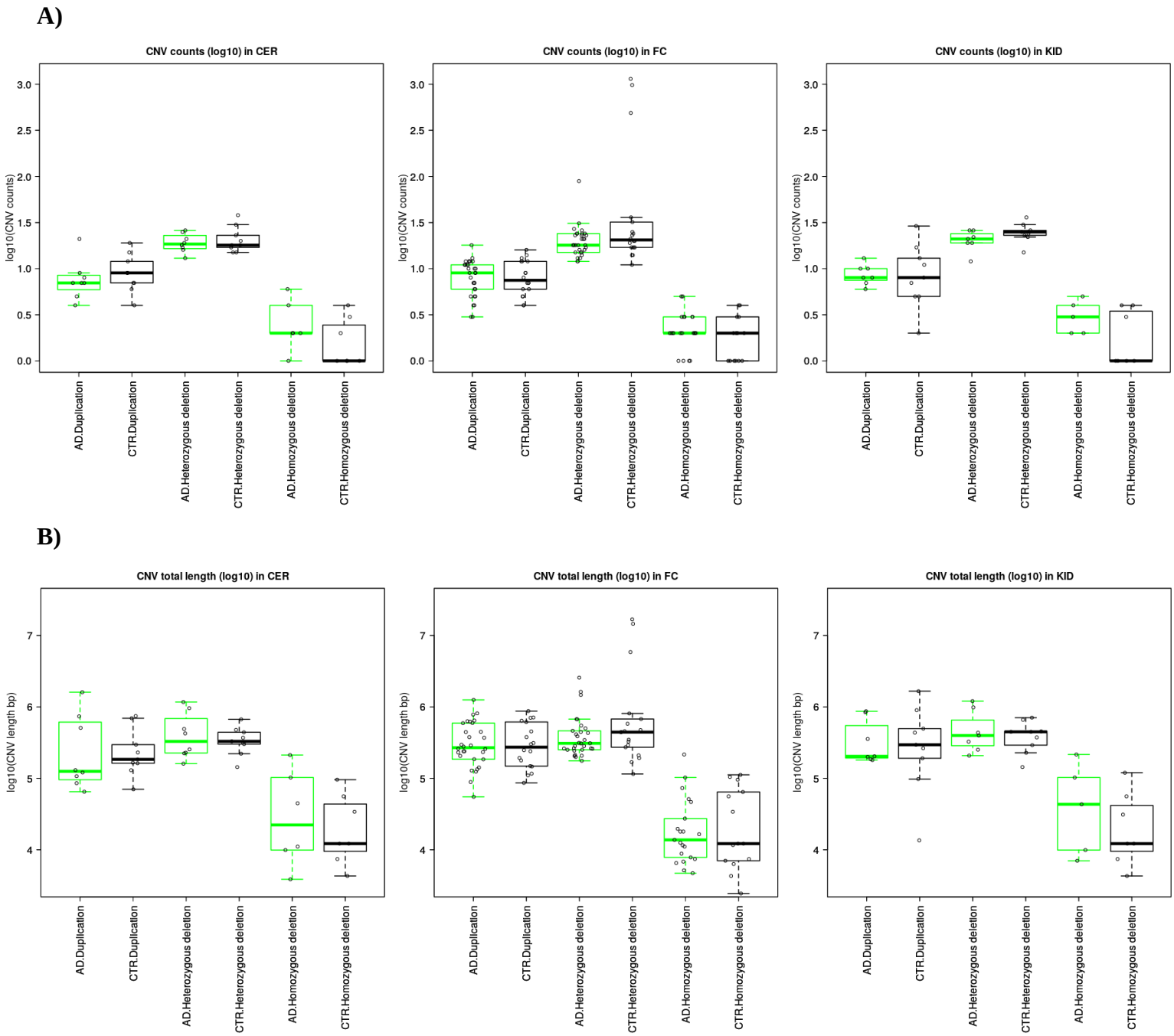
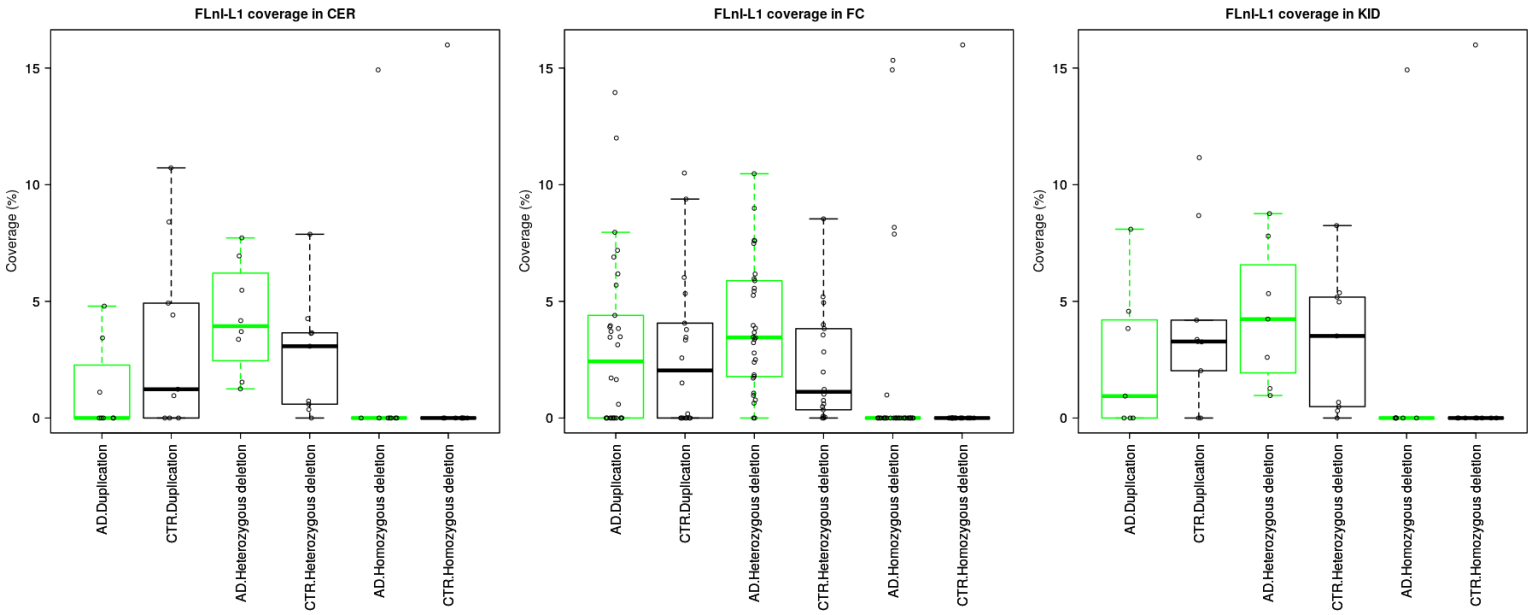


Figure 11: CNVs intrinsic properties. **A)** Distribution of CNVs counts per sample grouped by tissue; **B)** Distribution of CNVs total length per sample grouped by tissue. AD samples in green and CTR in black.

A)



B)

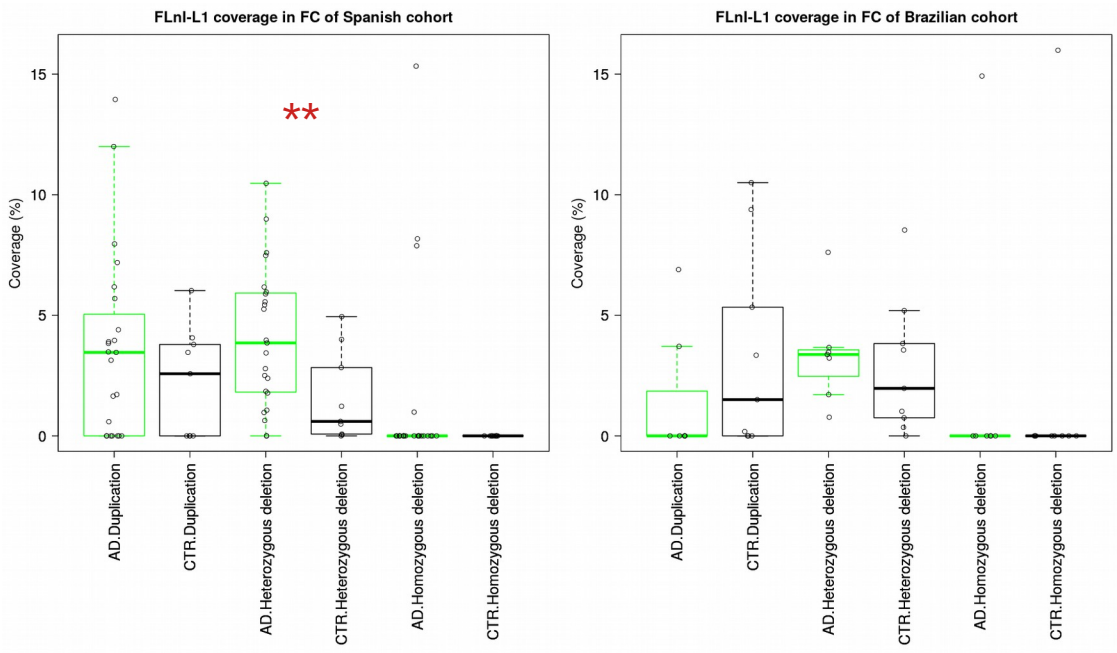


Figure 12: FlnI-L1 coverage distributions. **A)** FlnI-L1 coverage distributions grouped by tissue; **B)** FlnI-L1 coverage distributions in FC of Spanish and Brazilian samples. A significant difference in FlnI-L1 coverage within heterozygous deletions was identified between the FC of Spanish AD and CTRs. AD samples in green and CTR in black.

2.3.3 SNVs exploration on the Brazilian cohort

I then focused my attention on the Brazilian cohort of samples to investigate the possible presence of variants among tissues belonging to the same individuals. I started from the genotyping calls made with SNPs array experiments. Several quality filters (see methods) were applied to focus on a high quality subset of variants, which comprised 1,804,067 different loci per sample. I identified more than 900,000 loci in homozygosity for the reference allele, ~500,000 loci in heterozygosity and more than 300,000 loci in homozygosity for the alternative allele in each sample (table 2).

Genotyping data of tissues belonging to the same individuals were next compared in pairs (CER vs KID, CER vs FC and FC vs KID), and differences that appeared at the same locus, defined as early onset somatic SNVs, were counted (table 3). I found a total of 17,014 redundant SNVs loci: 12,707 from CER vs KID, 10,194 from FC vs CER and 14,802 from FC vs KID. Between FC and CER of single individuals, I observed higher counts of SNVs in AD samples, that reach a statistical significance of 0.022, while no dissimilarities were shown from the comparison of other tissues (figure 13). I noticed that both FC vs CER and CER vs KID AD distributions were bimodal, indicating the potential presence of two subgroups of patients. However, this observation could not be associated to differences in genders, ages or to the pathology grade and needs the analysis of additional clinical data.

SAMPLE	TISSUE	COHORT	TYPE	0/0	0/1	'1/1'	./.
C01_S001_A_CER	CER	Brazilian	AD	924554	559582	319359	572
C01_S001_A_FC	FC	Brazilian	AD	924107	559305	319405	1250
C01_S001_A_KID	KID	Brazilian	AD	924651	559664	319353	399
C01_S002_A_CER	CER	Brazilian	AD	922377	559318	321914	458
C01_S002_A_FC	FC	Brazilian	AD	922443	559077	321778	769
C01_S002_A_KID	KID	Brazilian	AD	922478	559324	321831	434
C01_S003_A_CER	CER	Brazilian	AD	922785	559439	321053	790
C01_S003_A_FC	FC	Brazilian	AD	923072	559044	321094	857
C01_S003_A_KID	KID	Brazilian	AD	922763	559387	320910	1007
C01_S004_A_CER	CER	Brazilian	AD	940992	520176	341915	984
C01_S004_A_FC	FC	Brazilian	AD	941315	520336	341820	596
C01_S004_A_KID	KID	Brazilian	AD	941334	520285	341752	696
C01_S005_A_CER	CER	Brazilian	AD	928334	554943	319801	989
C01_S005_A_KID	KID	Brazilian	AD	928365	555142	319742	818
C01_S006_A_CER	CER	Brazilian	AD	905555	593112	304447	953
C01_S006_A_FC	FC	Brazilian	AD	905558	593131	304433	945
C01_S007_A_CER	CER	Brazilian	AD	910505	579084	313539	939
C01_S007_A_FC	FC	Brazilian	AD	910560	579105	313495	907
C01_S007_A_KID	KID	Brazilian	AD	910489	578933	313572	1073
C01_S008_A_CER	CER	Brazilian	AD	908827	577772	316551	917
C01_S008_A_FC	FC	Brazilian	AD	908890	577828	316567	782
C01_S008_A_KID	KID	Brazilian	AD	908946	577540	316669	912
C01_S001_C_CER	CER	Brazilian	CTR	914674	568385	320264	744
C01_S001_C_FC	FC	Brazilian	CTR	914408	568378	320320	961
C01_S001_C_KID	KID	Brazilian	CTR	914746	568020	320524	777
C01_S002_C_CER	CER	Brazilian	CTR	912976	575965	314733	393
C01_S002_C_FC	FC	Brazilian	CTR	912968	575901	314644	554
C01_S002_C_KID	KID	Brazilian	CTR	912967	575834	314781	485
C01_S004_C_CER	CER	Brazilian	CTR	960223	478218	364979	647
C01_S004_C_FC	FC	Brazilian	CTR	960367	478223	364958	519
C01_S004_C_KID	KID	Brazilian	CTR	960287	478118	365027	635
C01_S005_C_CER	CER	Brazilian	CTR	904666	591504	306923	974
C01_S005_C_FC	FC	Brazilian	CTR	904760	591541	306918	848
C01_S005_C_KID	KID	Brazilian	CTR	904523	591518	306957	1069
C01_S006_C_CER	CER	Brazilian	CTR	901056	600033	301994	984
C01_S006_C_FC	FC	Brazilian	CTR	901124	600034	301977	932
C01_S006_C_KID	KID	Brazilian	CTR	900876	599918	302078	1195
C01_S007_C_CER	CER	Brazilian	CTR	909143	582242	311863	819
C01_S007_C_FC	FC	Brazilian	CTR	909214	582222	311801	830
C01_S007_C_KID	KID	Brazilian	CTR	909254	582217	311768	828
C01_S008_C_CER	CER	Brazilian	CTR	916267	571326	315582	892
C01_S008_C_FC	FC	Brazilian	CTR	916355	571366	315598	748
C01_S008_C_KID	KID	Brazilian	CTR	916432	571288	315558	789
C01_S009_C_CER	CER	Brazilian	CTR	911747	575712	315637	971
C01_S009_C_FC	FC	Brazilian	CTR	911901	575744	315660	762
C01_S009_C_KID	KID	Brazilian	CTR	911895	575556	315567	1049
C01_S010_C_CER	CER	Brazilian	CTR	912395	577097	313816	759
C01_S010_C_FC	FC	Brazilian	CTR	912359	577125	313809	774
C01_S010_C_KID	KID	Brazilian	CTR	912264	576771	313825	1207

Table 2: SNP array genotyping data. 0/0 : Homozygous calls for reference alleles; 0/1 : heterozygous calls; 1/1 : homozygous calls for alternative alleles; ./. : ungenotyped loci per sample.

INDIVIDUAL	COHORT	TYPE	CER vs FC	CER vs KID	FC vs KID
C01_S001_A	Brazilian	AD	2466	645	2419
C01_S002_A	Brazilian	AD	2097	702	2006
C01_S003_A	Brazilian	AD	2359	824	2242
C01_S004_A	Brazilian	AD	2181	1964	1723
C01_S005_A	Brazilian	AD	NA	1769	NA
C01_S006_A	Brazilian	AD	589	NA	NA
C01_S007_A	Brazilian	AD	668	1987	1896
C01_S008_A	Brazilian	AD	737	1942	1852
C01_S001_C	Brazilian	CTR	669	1975	2173
C01_S002_C	Brazilian	CTR	489	1786	1690
C01_S004_C	Brazilian	CTR	629	1826	1712
C01_S005_C	Brazilian	CTR	636	1907	1846
C01_S006_C	Brazilian	CTR	626	2092	2055
C01_S007_C	Brazilian	CTR	608	1802	1774
C01_S008_C	Brazilian	CTR	688	1810	1771
C01_S009_C	Brazilian	CTR	737	1837	1800
C01_S010_C	Brazilian	CTR	629	1939	1938

Table 3: Counts of somatic SNVs per individual and tissue comparison.

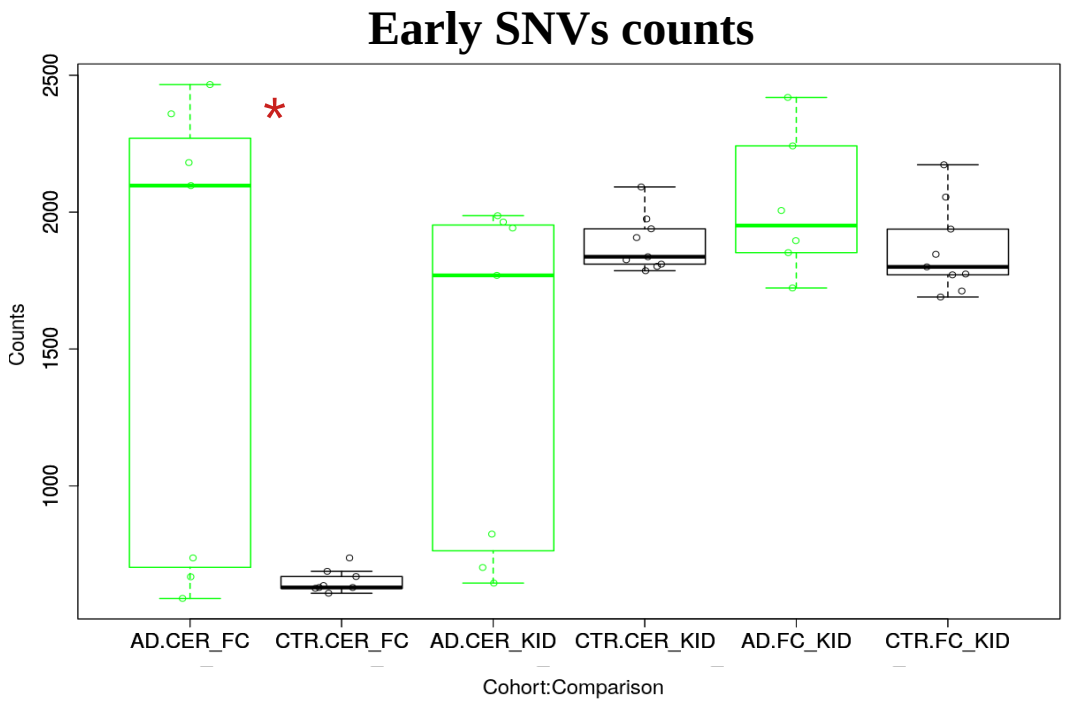


Figure 13: Early onset somatic SNVs distributions grouped by tissue comparisons. AD in green and CTRs in black.

In order to decipher the molecular process behind the origin of the identified SNVs, I carried out a mutational signature analysis following the pioneering work of Alexandrov and colleagues [Alexandrov et al., 2013]. To gain signatures of both brain tissues, I used the KID tissue as reference. Next, I selected only early onset SNVs with respect to the KID tissues, finding that about 50% of SNVs were substitution of the Cytosine residues, in favor to Thymines (C>T), and vice versa (T>C) (figure 14). Next, I searched for already deposited signatures on the COSMIC database. Signature analyses showed the presence of multiple single base signatures (SBS) (figure 15). SBS6, SBS11, SBS15, SBS16, SBS23, SBS29, SBS32 and SBS 46 showed to be present in a limited set of samples and to present extremely low frequencies (< 0.05), suggesting to be false positive calls due to the limited set of variants used as input. From the results, SBS1, SBS24, SBS39 and an unknown signature were also observed. SBS1 (figure 16) was found to describe about the 40% of the pattern of mutations being equally present in AD and CTRs. SBS24 instead (figure 16), showed a higher trend in AD FC with respect to the CTRs and was not observed from the CER. Then, SBS39 (figure 16) was found to present higher trends in CTR, in both CER and FC. Finally, signature analysis was not capable to classify about 30% of variants per sample, that were called as SBS unknown. This was reconnected to our approach, that relying on already deposited SBSs within the COSMIC database, impeded us to gain the mutational pattern relative to unknown signatures.

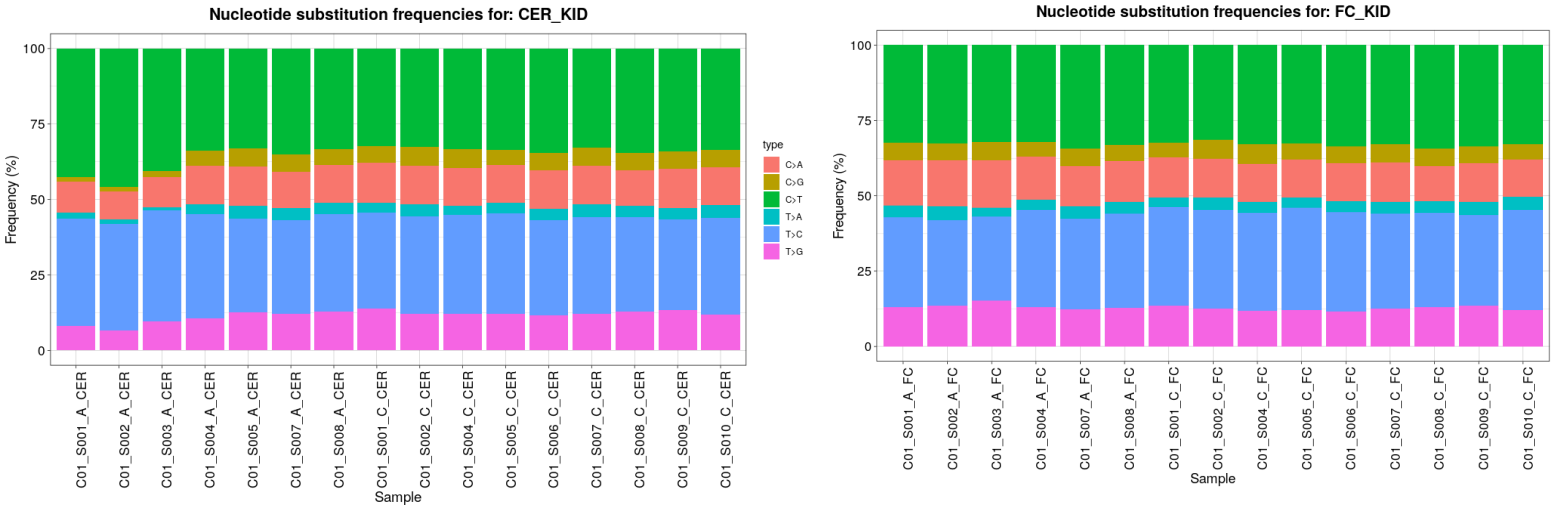


Figure 14: Nucleotides substitutions frequencies for CER vs KID early onset SNVs (left plot) and FC vs KID early onset SNVs (right plot). Frequencies were obtained by normalizing the counts of each type of nucleotide substitutions (C>A,C>G,C>T,T>A,T>C,T>G) with the total counts of early onset SNVs per sample and then displayed as cumulative percentages.

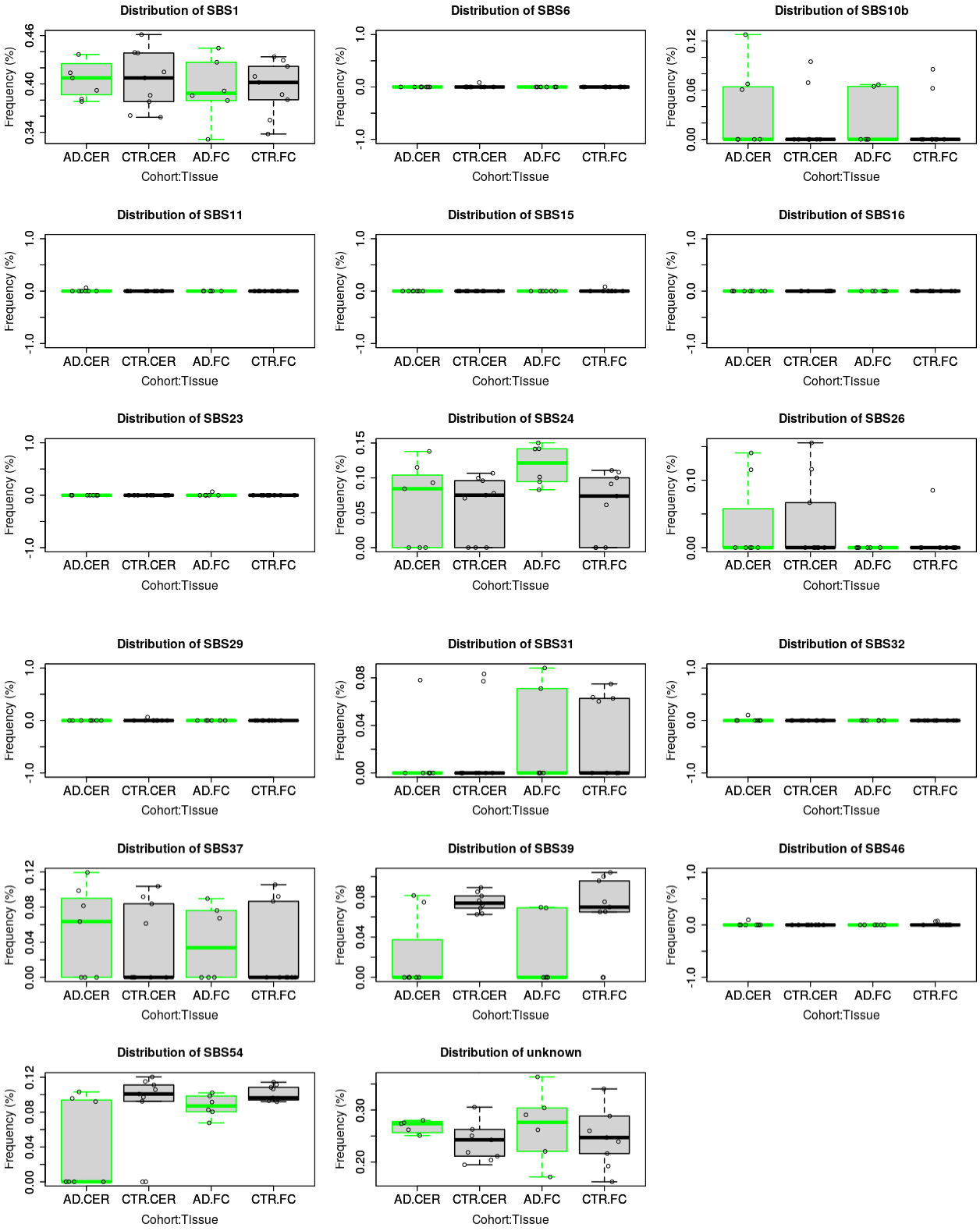


Figure 15: Signature analyses results. Frequencies distributions are reported for each signature identified across tissues with respect to KID.

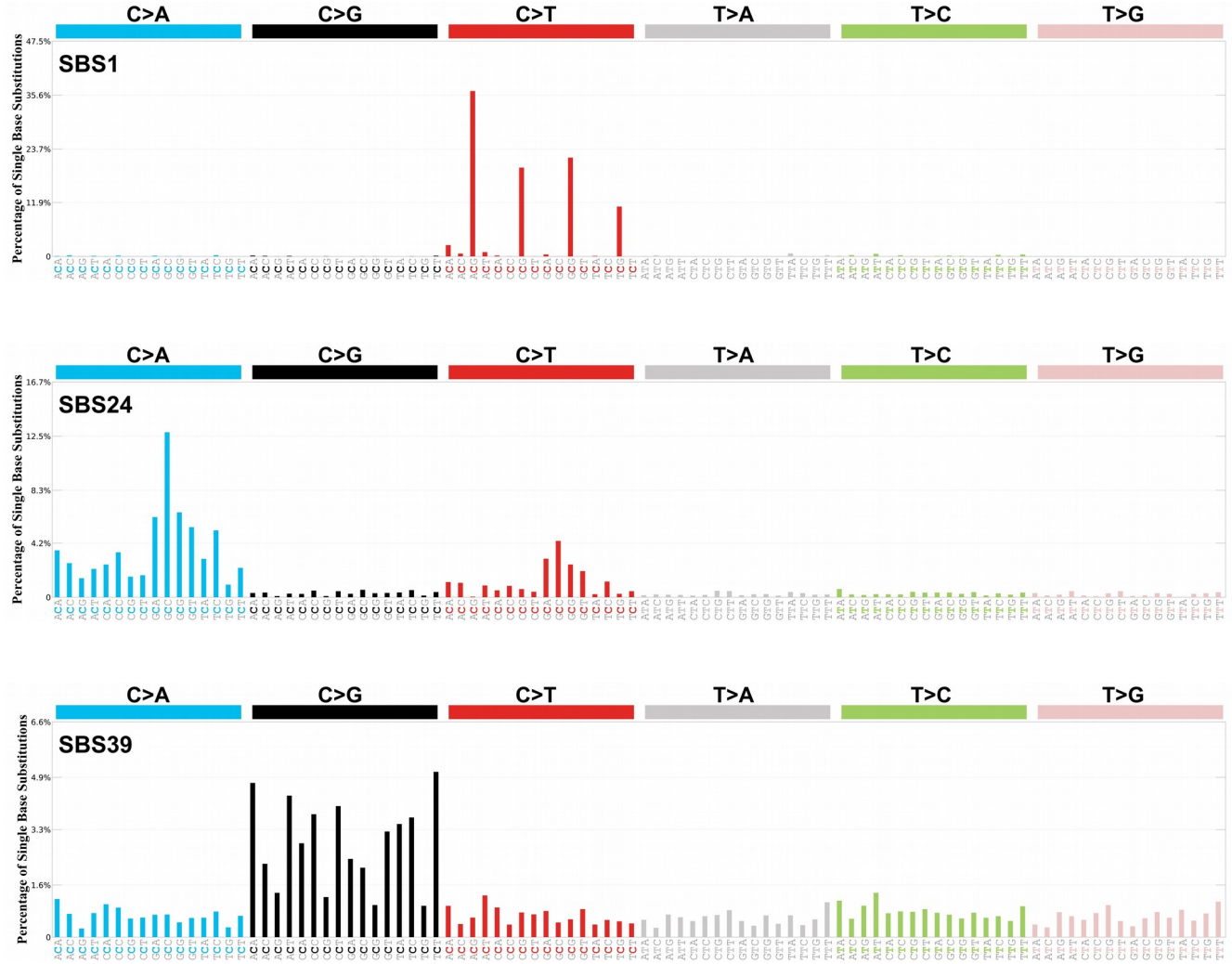


Figure 16: Signature profiles for SBS1, SBS24 and SBS39. Obtained from the COSMIC v.3 database.

2.4 Discussion and Conclusion

Here I presented the first study that, to my knowledge, investigated early onset somatic SNVs and L1 content in genomic CNVs from an AD cohort using ultra high-density Illumina SNP array.

Results indicated that quantitatively, genomic CNVs of ADs were not dissimilar from CTRs ones. Nevertheless, I found that the full-length form of L1 elements were enriched in the heterozygous deletions of FC of AD samples. This data may indicate alterations in L1 copy number between AD and CTRs. SNPs arrays are limited in the size of the CNVs to be called, especially in repetitive regions, for which specific markers are difficult to generate. Moreover, since repetitive elements are depleted in SNPs probes (more details in chapter VI), the presence of additional CNVs that are not yet detected may represent a current limitation of the reported study. Finally, retrotransposon CNVs is highly dictated by active retrotransposition, which was demonstrated to be potentially pervasive in human brain tissues and to concur to the generation of mosaicism. SNPs arrays lack in the ability to detect retrotransposon insertions, thus providing only a part of the retrotransposons CNV informations.

Furthermore, I investigated potential early onset somatic SNVs in different tissues belonging to the same individuals. In this regard I found that CER and FC of AD patients harbored higher SNVs with respect to CTRs suggesting possible alterations in biological processes. To unveil the mutagen actors behind SNVs, I set KID genotypes as reference, identified SNVs with respect to the other tissues and performed molecular signature analyses relying on already known signatures, despite being aware that the low amounts of variants and the small sample size would possibly represent a limitation. From my analyses, SBS1 was found to compose about 40% of the mutational spectra without being differentially present between AD and CTRs. This signature is proposed to be caused by the endogenous mutational process initiated by spontaneous DNA deamination of 5-methylcytosine, a process that hit the genome at constant rate, and that results in accumulation of mutations that are proportional to the chronological age of the sample (also known as *clock-like* signature). Furthermore, also evidences of SBS39 were found. These were noted prevalently in FC and CER of CTRs and observed at lower rates in the same tissues of AD. However, SBS39 aetiology is currently unknown. Interestingly, although not statistically significant, higher frequencies of SBS24 were observed in FC of AD samples, but not in the remaining samples. SBS24 is associated with aflatoxins, a family of fungal toxins mostly produced by *Aspergillus flavus* and *Aspergillus parasiticus.* The same species are found on agricultural crops, such as maize and

peanuts, which can lead people to be exposed to aflatoxins by eating contaminated plant products or by consuming meat or dairy products from animals that ate contaminated feed. Interestingly, they are capable of trespassing the brain blood barrier, possibly damaging its endothelial cells in a time dependent manner [Qureshi et al., 2015]. Additionally, chronic exposure to aflatoxins results in neuroblastoma and potentially to encephalopathy causing cerebral edema with neuronal degeneration (Reye's syndrome) [Ryan et al., 1979]. Most notably, in rats it was recently demonstrated that chronic exposure to aflatoxins led to neurodegeneration. The proposed mechanism relies on ROS release from macrophages and astrocytes in response to the toxin, which then result in cell's apoptosis [Alsayyah et al., 2019], which resembled what observed in neurodegeneration. Taken together, these observations may suggest the presence of a potential link between aflatoxins and Alzheimer's disease, which requires further investigations. However, it is known that such toxins can also damage the kidney tissue through oxidative stress, leading cells to apoptosis [Li et al., 2018]. Therefore, one could assume that aflatoxins are present in both kidney and brain tissues of the same individual at the same time. However, since signature analyses were performed using kidney genotypes as reference, it is not clear why and how aflatoxins signatures may emerge from these tissues comparisons. A simplistic explanation could be that, in our case, all tissues belonging to AD may present mutations due to aflatoxins. However, FC (in this particular case) may be more prone to accumulate variants with respect to kidney and cerebellum, leading to the identification of the signature. Nonetheless, ew should not discard the possibility that due to the low amount of variants used, signature analyses may be leading to false signature assignations.

Finally, we noted that about 30% of variants were ascribed to an undetermined signature. The approach was not a *de-novo* signature discovery, therefore did not provided additional information that would allow me to better characterize the undetermined signature observed. Nonetheless, undetermined signatures may arise in two conditions: when there is a limited mutational signature analysis power (*i.e.* small variants dataset) that results in a not correct reconstruction of the mutational signature patterns by the softwares, or when there is the identification of a newly reported signature. Despite the limited variants dataset, it must be noted that signature analyses has not yet been extensively applied in non-cancer samples, thus current signature databases are limited and biased towards mutagens that represent driver processes in cancer. With current information, it is therefore not clear whether these results indicate softwares limitations or the presence of an unidentified AD-related signature.

Additional *de-novo* signature analyses and a general increase in signature analyses power are thus required.

For the aforementioned observations, it was decided to extend our investigations on both SNVs and retrotransposon CNV by performing high coverage whole genome sequencing of the Brazilian samples set. This would result in whole genome SNVs investigation and therefore to a substantial increase in signature analyses power. Moreover, whole genome data will give me the opportunity to better evaluate retrotransposons CNVs by assessing the impact of insertions being able to specifically identify insertions and integration sites.

Chapter III

Genome-wide analyses of SNVs and retrotransposon CNVs with WGS

3.1 Introduction

Evidences from SNPs array experiments have opened to the possibility that early onset SNVs and full-length L1 content may be linked with AD. However, as discussed, there were several limitations that affected my approach keeping many questions open. For instance, SNPs arrays, lacking the ability to detect L1 insertions, cannot be used to provide a full overview on the CNVs due to retrotransposition processes, reducing the understanding of the impact of L1 copy number in AD. Moreover, although millions of markers were used, the dataset of variants was potentially biased towards already annotated and well-known polymorphic loci (*i.e.* SNPs), which could have affected the signature analyses. A high-coverage whole genome sequencing was therefore performed on the Brazilian cohort, further expanding it with two additional tissues (hippocampus – HIP and temporal cortex – TC) and samples. By using a parallel high-throughput approach, we aimed at validating previously identified early onset SNVs. Furthermore, sequencing at high-depth of coverage would ultimately expand early onset SNVs detection to the whole genome, thus including SNVs not associated to known polymorphisms, as well as saturate the detection of high-frequency variants (*i.e.* present in the majority of the cells of the given tissue). Finally, WGS data were used to assess the contribution of mobile element integrations to retrotransposons CNVs, and to precisely genotype both reference and non-reference retrotransposons annotations.

3.2 Materials and Methods

3.2.1 Whole genome sequencing of the Brazilian cohort

The Brazilian cohort of samples, that consisted in 5 different tissues (cerebellum, frontal cortex, hippocampus, temporal cortex and kidney) from 8 AD patients and 10 CTRs samples (Table 4), was sequenced at high coverage (theoretical 100x depth of coverage) following the Nextera® Flex protocol (https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nextera_dna_flex/nextera-dna-flex-library-prep-reference-guide-1000000025416-07.pdf) using the Illumina® NovaSeq 6000 System PE150X2 and S4 flow-cells with pair-end reads strategy. Reads quality was evaluated with a sequential combination of the f*astQC* (version 0.11.8) [Babraham Bioinformatics] and *MultiQC* (version 1.7) [Ewels et al., 2016] softwares (both with default parameters). Coverage metrics were obtained by applying the *BBtools pileup.sh* script (version 38.22) [sourceforge.net/projects/bbmap/]. Raw fastq.gz files were aligned versus the reference genome build GRCh37 (version GCA_000001405.1) with the *BWA* software (version *0.7.17-r1194-dirty*; subcommand *mem;* default options) [Li and Durbin, 2010]. SAM files were sorted, compressed into BAM files and then indexed with the SAM*tools* suit of tools (version: 1.9; commands: *samtools sort*; *samtools view -hb*; *samtools index* respectively) [Li et al., 2009].

GATK 4 best practices (https://gatk.broadinstitute.org/hc/en-us/articles/360035535912-Data-pre-processing-for-variant-discovery) [Van der Auwera et al., 2013] were strictly followed in order to 1) mark PCR duplicates using *Picard* (version: 2.18.20) [http://broadinstitute.github.io/picard]; 2) recalibrate reads base quality with *GATK 4* (version 4.0.11.0; subcommands *BaseRacall* and *ApplyBQSR*). Metrics for BAM files (as total reads and total mapped reads counts) were obtained using the *samtools stats* command.

ID1	ID2	G.	AGE	COHORT	TYPE	TISSUE	B. NTF	EXPERIMENT	WGS QUAL.	TOT. R. COUNTS	TOT. M.R. COUNTS	M.R. (%)	COV.
929	C01_S005_C_CER	F	61	Brazilian	CTR	CER	0	SNP array; WGS	PASS	2,502,163,048	2,496,981,549	99.79	118.01
929	C01_S005_C_FC	F	61	Brazilian	CTR	FC	0	SNP array; WGS	PASS	2,362,527,316	2,357,820,268	99.8	111.1
929	C01_S005_C_KID	F	61	Brazilian	CTR	KID	0	SNP array; WGS	PASS	2,379,115,104	2,373,937,044	99.78	111.66
929	NA	F	61	Brazilian	CTR	HIP	0	WGS	PASS	2,315,411,092	2,310,206,988	99.78	109.04
929	NA	F	61	Brazilian	CTR	TC	0	WGS	PASS	2,193,835,371	2,190,243,543	99.84	103.59
1282	C01_S002_C_CER	M	79	Brazilian	CTR	CER	0	SNP array; WGS	PASS	2,566,157,901	2,560,266,320	99.77	120.88
1282	C01_S002_C_FC	M	79	Brazilian	CTR	FC	0	SNP array; WGS	PASS	2,237,880,004	2,233,633,336	99.81	105.41
1282	C01_S002_C_KID	M	79	Brazilian	CTR	KID	0	SNP array; WGS	PASS	2,093,128,674	2,088,852,301	99.8	98.53
1282	NA	M	79	Brazilian	CTR	HIP	0	WGS	PASS	1,989,666,769	1,985,463,883	99.79	93.8
1282	NA	M	79	Brazilian	CTR	TC	0	WGS	PASS	2,258,214,798	2,253,579,303	99.79	106.48
2149	C01_S007_C_CER	F	76	Brazilian	CTR	CER	1	SNP array; WGS	PASS	2,373,905,335	2,370,009,011	99.84	111.91
2149	C01_S007_C_FC	F	76	Brazilian	CTR	FC	1	SNP array; WGS	PASS	2,487,959,302	2,483,983,185	99.84	117.04
2149	C01_S007_C_KID	F	76	Brazilian	CTR	KID	1	SNP array; WGS	PASS	2,171,763,201	2,168,338,636	99.84	102.53
2149	NA	F	76	Brazilian	CTR	HIP	1	WGS	PASS	2,280,258,764	2,276,576,942	99.84	106.71
2149	NA	F	76	Brazilian	CTR	TC	1	WGS	PASS	2,296,706,506	2,292,850,254	99.83	108.25
2434	C01_S001_C_CER	M	75	Brazilian	CTR	CER	2	SNP array; WGS	PASS	2,287,374,606	2,280,935,090	99.72	107.72
2434	C01_S001_C_FC	M	75	Brazilian	CTR	FC	2	SNP array; WGS	PASS	2,361,583,907	2,356,835,301	99.8	110.5
2434	C01_S001_C_KID	M	75	Brazilian	CTR	KID	2	SNP array; WGS	PASS	2,148,216,112	2,142,265,381	99.72	101.23
2434	NA	M	75	Brazilian	CTR	HIP	2	WGS	PASS	2,390,650,836	2,383,940,260	99.72	112.77
2434	NA	M	75	Brazilian	CTR	TC	2	WGS	PASS	2,445,845,538	2,440,283,031	99.77	115.35
6868	C01_S009_C_CER	F	89	Brazilian	CTR	CER	2	SNP array; WGS	PASS	3,461,645,703	3,443,205,205	99.47	162.58
6868	C01_S009_C_FC	F	89	Brazilian	CTR	FC	2	SNP array; WGS	PASS	3,149,325,421	3,138,245,183	99.65	147.89
6868	C01_S009_C_KID	F	89	Brazilian	CTR	KID	2	SNP array; WGS	PASS	2,916,388,207	2,911,696,202	99.84	137.53
6868	NA	F	89	Brazilian	CTR	HIP	2	WGS	PASS	2,801,009,322	2,794,862,766	99.78	132
6868	NA	F	89	Brazilian	CTR	TC	2	WGS	PASS	2,329,493,453	2,315,888,840	99.42	109.41
7106	C01_S006_C_CER	F	75	Brazilian	CTR	CER	2	SNP array; WGS	PASS	3,327,929,766	3,322,640,576	99.84	156.93
7106	C01_S006_C_FC	F	75	Brazilian	CTR	FC	2	SNP array; WGS	PASS	2,413,699,851	2,409,259,076	99.82	113.77
7106	C01_S006_C_KID	F	75	Brazilian	CTR	KID	2	SNP array; WGS	PASS	2,718,472,229	2,713,148,597	99.8	128.04
7106	NA	F	75	Brazilian	CTR	HIP	2	WGS	PASS	2,368,491,175	2,363,277,053	99.78	111.88
7106	NA	F	75	Brazilian	CTR	TC	2	WGS	PASS	2,352,129,530	2,348,714,693	99.85	110.94
7711	C01_S004_C_CER	M	85	Brazilian	CTR	CER	2	SNP array; WGS	PASS	2,424,763,648	2,330,631,014	96.12	109.65
7711	C01_S004_C_FC	M	85	Brazilian	CTR	FC	2	SNP array; WGS	PASS	2,414,427,691	2,409,139,925	99.78	114.04
7711	C01_S004_C_KID	M	85	Brazilian	CTR	KID	2	SNP array; WGS	PASS	3,203,923,834	3,197,168,030	99.79	150.95
7711	NA	M	85	Brazilian	CTR	HIP	2	WGS	PASS	2,495,602,444	2,490,003,168	99.78	117.84
7711	NA	M	85	Brazilian	CTR	TC	2	WGS	PASS	2,589,859,229	2,582,532,493	99.72	121.73
9173	NA	M	83	Brazilian	CTR	FC	2	WGS	PASS	2,253,804,144	2,249,308,653	99.8	106.41
9173	NA	M	83	Brazilian	CTR	CER	2	WGS	PASS	2,272,293,709	2,265,306,916	99.69	107.24
9173	NA	M	83	Brazilian	CTR	KID	2	WGS	PASS	2,299,323,155	2,294,478,717	99.79	108.54
9173	NA	M	83	Brazilian	CTR	HIP	2	WGS	PASS	2,292,988,418	2,288,707,374	99.81	108.1
9173	NA	M	83	Brazilian	CTR	TC	2	WGS	PASS	2,062,332,146	2,056,898,365	99.74	97
9269	C01_S008_C_CER	F	88	Brazilian	CTR	CER	3	SNP array; WGS	PASS	2,343,574,912	2,339,703,462	99.83	110.32
9269	C01_S008_C_FC	F	88	Brazilian	CTR	FC	3	SNP array; WGS	PASS	2,160,023,666	2,155,572,713	99.79	101.9
9269	C01_S008_C_KID	F	88	Brazilian	CTR	KID	3	SNP array; WGS	PASS	2,439,377,634	2,433,143,035	99.74	114.84
9269	NA	F	88	Brazilian	CTR	HIP	3	WGS	PASS	3,320,536,891	3,311,635,948	99.73	156.86
9269	NA	F	88	Brazilian	CTR	TC	3	WGS	PASS	3,724,442,885	3,713,889,010	99.72	175.47
9810	C01_S010_C_CER	F	92	Brazilian	CTR	CER	3	SNP array; WGS	PASS	2,462,958,128	2,456,528,333	99.74	116.07
9810	C01_S010_C_FC	F	92	Brazilian	CTR	FC	3	SNP array; WGS	PASS	2,422,689,150	2,416,857,635	99.76	114.04
9810	C01_S010_C_KID	F	92	Brazilian	CTR	KID	3	SNP array; WGS	PASS	2,484,739,860	2,479,833,595	99.8	116.98
9810	NA	F	92	Brazilian	CTR	HIP	3	WGS	PASS	2,596,160,333	2,589,794,127	99.75	122.24
9810	NA	F	92	Brazilian	CTR	TC	3	WGS	PASS	2,175,914,768	2,171,696,534	99.81	102.6

ID1	ID2	G.	AGE	COHORT	TYPE	TISSUE	B. NTF	EXPERIMENT	WGS QUAL.	TOT. R. COUNTS	TOT. M.R. COUNTS	M.R. (%)	COV.
1345	C01_S006_A_CER	F	83	Brazilian	AD	CER	6	SNP array; WGS	PASS	2,281,470,966	2,266,855,416	99.36	107.14
1345	C01_S006_A_FC	F	83	Brazilian	AD	FC	6	SNP array; WGS	PASS	2,446,860,466	2,442,778,181	99.83	115.04
1345	C01_S006_A_KID	F	83	Brazilian	AD	KID	6	SNP array; WGS	GENDER ERROR	2,191,926,351	2,188,958,511	99.86	102.94
1345	NA	F	83	Brazilian	AD	HIP	6	WGS	PASS	2,496,531,237	2,491,393,953	99.79	116.08
1345	NA	F	83	Brazilian	AD	TC	6	WGS	PASS	2,294,422,380	2,273,888,547	99.11	106.98
2682	C01_S003_A_CER	M	82	Brazilian	AD	CER	6	SNP array; WGS	PASS	2,461,721,460	2,456,403,618	99.78	115.97
2682	C01_S003_A_FC	M	82	Brazilian	AD	FC	6	SNP array; WGS	PASS	2,292,449,524	2,287,741,120	99.79	108.04
2682	C01_S003_A_KID	M	82	Brazilian	AD	KID	6	SNP array; WGS	PASS	2,594,833,788	2,589,384,759	99.79	121.54
2682	NA	M	82	Brazilian	AD	HIP	6	WGS	PASS	2,845,295,093	2,838,841,584	99.77	133.9
2682	NA	M	82	Brazilian	AD	TC	6	WGS	PASS	2,439,642,569	2,434,597,251	99.79	115.02
6275	C01_S008_A_CER	F	92	Brazilian	AD	CER	4	SNP array; WGS	PASS	2,287,673,868	2,283,976,349	99.84	106.98
6275	C01_S008_A_FC	F	92	Brazilian	AD	FC	4	SNP array; WGS	PASS	2,140,478,293	2,136,893,554	99.83	100.79
6275	C01_S008_A_KID	F	92	Brazilian	AD	KID	4	SNP array; WGS	PASS	2,157,912,616	2,153,456,156	99.79	101.56
6275	NA	F	92	Brazilian	AD	HIP	4	WGS	PASS	2,849,195,741	2,841,883,943	99.74	134.58
6275	NA	F	92	Brazilian	AD	TC	4	WGS	PASS	2,900,766,113	2,882,020,591	99.35	136.67
7466	C01_S004_A_CER	M	87	Brazilian	AD	CER	4	SNP array; WGS	PASS	2,487,493,151	2,474,948,619	99.5	117.26
7466	C01_S004_A_FC	M	87	Brazilian	AD	FC	4	SNP array; WGS	PASS	2,525,518,691	2,519,467,919	99.76	119.24
7466	C01_S004_A_KID	M	87	Brazilian	AD	KID	4	SNP array; WGS	PASS	2,290,823,883	2,275,154,159	99.32	107.07
7466	NA	M	87	Brazilian	AD	HIP	4	WGS	PASS	2,279,232,853	2,271,856,148	99.68	107.72
7466	NA	M	87	Brazilian	AD	TC	4	WGS	PASS	2,541,809,663	2,536,541,122	99.79	119.84
7660	C01_S001_A_CER	M	72	Brazilian	AD	CER	6	SNP array; WGS	PASS	2,580,285,129	2,574,276,263	99.77	121.52
7660	C01_S001_A_FC	M	72	Brazilian	AD	FC	6	SNP array; WGS	PASS	2,873,261,120	2,867,462,162	99.8	135.26
7660	C01_S001_A_KID	M	72	Brazilian	AD	KID	6	SNP array; WGS	PASS	2,352,110,775	2,347,207,519	99.79	110.2
7660	NA	M	72	Brazilian	AD	HIP	6	WGS	PASS	2,622,573,896	2,614,953,716	99.71	123.19
7660	NA	M	72	Brazilian	AD	TC	6	WGS	PASS	2,173,374,613	2,167,332,582	99.72	102.13
8805	C01_S005_A_CER	F	80	Brazilian	AD	CER	4	SNP array; WGS	PASS	2,223,299,096	2,211,873,361	99.49	104.7
8805	C01_S005_A_FC	F	80	Brazilian	AD	FC	4	SNP array; WGS	PASS	2,449,017,175	2,414,134,223	98.58	114.1
8805	C01_S005_A_KID	F	80	Brazilian	AD	KID	4	SNP array; WGS	PASS	2,318,678,185	2,264,833,751	97.68	107.04
8805	NA	F	80	Brazilian	AD	HIP	4	WGS	PASS	2,258,322,724	2,234,354,778	98.94	105.67
8805	NA	F	80	Brazilian	AD	TC	4	WGS	PASS	2,295,045,867	2,277,285,378	99.23	107.84
9345	C01_S007_A_CER	F	90	Brazilian	AD	CER	4	SNP array; WGS	PASS	2,351,736,180	2,347,624,281	99.83	110.49
9345	C01_S007_A_FC	F	90	Brazilian	AD	FC	4	SNP array; WGS	PASS	2,606,614,945	2,601,184,827	99.79	122.7
9345	C01_S007_A_KID	F	90	Brazilian	AD	KID	4	SNP array; WGS	PASS	3,255,253,363	3,248,797,090	99.8	153.33
9345	NA	F	90	Brazilian	AD	HIP	4	WGS	PASS	2,195,026,294	2,189,635,294	99.75	103.49
9345	NA	F	90	Brazilian	AD	TC	4	WGS	PASS	2,225,113,835	2,220,269,944	99.78	105
10643	C01_S002_A_CER	M	80	Brazilian	AD	CER	3	SNP array; WGS	PASS	2,248,748,210	2,244,479,588	99.81	106.14
10643	C01_S002_A_FC	M	80	Brazilian	AD	FC	3	SNP array; WGS	PASS	2,243,022,083	2,238,549,898	99.8	105.59
10643	C01_S002_A_KID	M	80	Brazilian	AD	KID	3	SNP array; WGS	PASS	2,552,083,339	2,546,947,799	99.8	120.08
10643	NA	M	80	Brazilian	AD	HIP	3	WGS	PASS	2,473,799,144	2,467,430,385	99.74	116.36
10643	NA	M	80	Brazilian	AD	TC	3	WGS	PASS	2,401,172,768	2,397,175,854	99.83	113.33

Table 4: WGS sample metadata. ID1: WGS sample ID; ID2: SNPs array sample ID; G.: gender; B.NTF: braak ntf, WGS QUAL: WGS quality; TOT. T. COUNTs: total reads counts; TOT. M.R. COUNTS: total mapped reads counts; M.R.: mapped reads percentage; COV: coverage. “Experiment” column reports the high-throughput technology applied. The discarded sample is highlighted in red.

3.2.2 Single Nucleotide Variants analyses

Variants at 16,437,462 single nucleotide loci with respect to the reference genome were called using the *GATK 4* germline short variant discovery (SNPs + Indels) pipeline (https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-). Final calls were made by maintaining only SNVs with respect to the reference genome using *BCFtools* (version 1.9 ; subcommand *view*; parameters -v snps) [Li, 2011] and resulted in the identification of 10,862,657 SNVs per sample. *Plink1.9* (version 1.90b3.31, parameters --neighbor and –check-sex) [Purcell et al., 2007] and *SNPhylo* (version 20180901, default parameters) [Lee et al., 2014] were used to reconstruct sample's relationship and to check for the correctness of the gender metadata.

3.2.2.1 Validation of SNPs array observations

Counts of positional overlaps between array markers and WGS calls were obtained by intersecting the markers coordinates and the WGS variants set with *BEDtools* (version 2.27.1, subcommand *intersect*, parameter -wa) [Quinlan and Hall, 2010]. Genotypes obtained through SNPs arrays and WGS approaches were compared with a custom Perl script after having tested reference alleles for their concordance with the reference genome with *BCFtools* (version 1.9; subcommand fixref). Further controls were made by hand checking intensity signals and VCF files of both datasets on a subset of randomly selected SNVs. Integrative genomics viewer (*IGV*, version 2.8.0) [Robinson et al., 2011] was used to manually check random WGS SNVs calls that were in disagreement with SNPs array calls.

3.2.2.2 Discordant genotypes calling and signature analyses

A high quality subset of 7,882,025 variants, was generated by filtering SNVs with *Plink1.9* (parameters --geno 0.1 --maf 0.05). Genotypes data were compared in pairs of tissues, exploiting all the possible combinations (CER vs FC, CER vs HIP, CER vs TC, CER vs KID, FC vs HIP, FC vs TC, FC vs KID, HIP vs TC, HIP vs KID, TC vs KID) for each individual. Two signature analyses were performed. The

first one was performed upon the whole set of calls resulted from the application of the *GATK 4* SNP germline short variant discovery pipeline. Second signature analyses instead, relied on the SNVs resulted from the comparison of pairs of tissues (see above). We used the *Helmsmann* software (version 1.4.2) [Carlson et al., 2018] to generate the nucleotide substitution matrix, followed by the application of the *DeconstructSign* R package (version 1.8.0) [Rosenthal et al., 2016] for the signature assignation, imposing the COSMIC version 3 signatures as database (http://cancer.sanger.ac.uk/cosmic/signatures). Substitution matrix was normalized using the *DeconstructSign* option tri.counts.method = 'genome', in order to take into account the number of times that each trinucleotide context was observed in the genome.

Plots and statistical analyses (Student's t-tests with FDR corrections) were made with the R statistical analysis software (version 4).

3.2.3 Retrotransposons copy number variants analyses

3.2.3.1 Supporting reads analyses

Raw fastq files obtained from sequencing were mapped versus three retrotransposon's consensus sequences (elements: Alu, L1 and SVA) [Sudmant et al., 2015] using the BWA software (version *0.7.17-r1194-dirty;* subcommand *mem*, default parameters). SAM files were sorted, mapped reads were filtered by mapping quality (greater or equal to 60) and finally compressed into BAM files using the *SAMtools* toolkit (version 1.9; subcommands and parameters: *samtools sort*; *samtools view -F 4 -Q 60 -hb*; *samtools index*). Counts of mapped reads were extracted with the *samtools view* command (parameter -c). Raw counts normalization was performed using the total mapped reads counts obtained from the whole genome alignments.

3.2.3.2 Mobile element locator tool (MELT) analyses

MELT software (version 2.2.0) [Gardner et al., 2017] was applied to WGS alignment data. BAM files were preprocessed following the *MELT* user guide (https://melt.igs.umaryland.edu/manual.php#_Preprocessing_.bam_Files_for_MELT). Sub-sequentially

analyses consisted in the usage of *MELT Deletion*, to genotype reference genome retrotransposon annotations (https://melt.igs.umaryland.edu/manual.php#_MELT-Deletion) and *MELT Single*, to detect and genotype newly retrotransposition events (https://melt.igs.umaryland.edu/manual.php#_Running_MELT_Using_MELT-SINGLE). We tested the same three retrotransposons consensus sequences used during the supporting reads analyses (Alu, L1 and SVA). Outputs of both analyses were parsed to obtain reference and alternative allele counts for each sample. While no filters were used for the *MELT Deletion* results, quality "PASS" filter and supporting reads thresholds "LP >=1 & RP >= 1" (Discordant reads pairs that support the Left or the Right breaking Point) were imposed for the *MELT Single* results. Allele counts were next normalized using the total mapped reads counts from the whole genome alignments. Within single individuals, coordinates of alternative loci obtained from different tissues were intersected with the *intervene* software (version 0.6.4; subcommand *venn*) [Khan and Mathelier, 2017]. L1 alternative loci from *MELT Deletion* analyses were considered to be somatic or germinal depending on the number of tissues that displayed the variant call. Alternative loci were called as germinal when all 5 tissues of a single individual reported the same variant call. Somatic loci, instead, where called when variants were found in only one tissue. L1 *MELT Single* alternative loci were considered to be germinal or somatic by combining two approaches. First, *MELT Deletion* intersection strategy was applied to *MELT Single* results, thus integrations were divided in variants private of single tissues or shared among them. Next, distributions of LP, RP and SR (Split Reads that supports both breaking points) supporting reads counts for private and shared variants were compared. A discriminant threshold centered to 15 reads was identified (figure 31) and thus applied to discriminate L1 integrations in germinal (variants with more than 15 LR, 15 RP and 15 SR supporting reads) and somatic sets (variants with less than 15 LR, 15 RP and 15 SR supporting reads). Estimations on net allele counts were obtained for each of the three retrotransposons sequences by summing the reference allele counts from *MELT Deletion* to the alternative allele counts resulted from *MELT Single* analyses. Raw net allele counts were normalized with total mapped reads from whole genome sequencing alignment data.

Plots and statistical analyses (Student's t-tests with FDR corrections) were made with the R statistical analysis software (version 3.6).

3.2.4 Genomic CNVs analyses

Lumpyexpress command from *Lumpy* (version 0.2.13) [Layer et al., 2014] was used to call breaking points from WGS data. Option *-P* was selected to add variants probabilities. Output VCFs files were next submitted to the *SVTyper* tool (version 0.7.1) [Chiang et al., 2015] to both infer CNVs type and genotype. Overlaps between SNPs array CNVs and WGS data were performed with the *intervene* software (version 0.6.4; subcommand *venn*) testing different combination of parameters (option *-f* from 0.1 to 0.9 in combination with option *-r*). Final overlaps were made requiring a 70% of reciprocal overlap (parameters -f 0.7, -r).

Overlaps between WGS CNVs and Illumina Infinium Omni 5 (version 1.3) probe panel were tested with the *BEDtools* suit of tools (version 2.27.1, subcommand *intersect*).

Plots and statistical analyses (Student's t-tests with FDR corrections) were made with the R statistical analysis software (version 4).

3.3 Results

To extend the observations made with SNPs arrays analyses, high-coverage (>100x) WGS upon the same Brazilian samples was performed. A total of 40 samples from 8 individual with AD and 50 samples from 10 individuals from controls were analyzed from the Brasilian cohort. The cohort was further extended with additional tissues (hippocampus – HIP and temporal cortex - TC) and individuals (table 4). On one hand, this allowed to extend SNVs identification to the whole genome, increasing the set of variants and thus signature analyses power. On the other hand, this led me to perform deep retrotransposon's CNVs analyses by assessing both reference alleles and newly integrated ones.

3.3.1 SNVs analyses

By applying a state-of-art SNP calling pipeline (see methods), I was able to genotype more than 16 millions of single nucleotide loci per sample with respect to the reference genome. A set of 7,882,025 loci passed stringent *Plink1.9* quality controls (see methods) and was used for all the following analyses. In each sample, we identified ~4,000,000 loci homozygous for the reference allele, ~ 2,500,000 loci in heterozygosity and ~ 1,100,000 loci homozygous for the alternative allele (figure 17). WGS data were next used to reconstruct sample's relationships, allowing me to validate all but one gender metadata, in agreement with the observations made with the previously described array technology (figure 18).

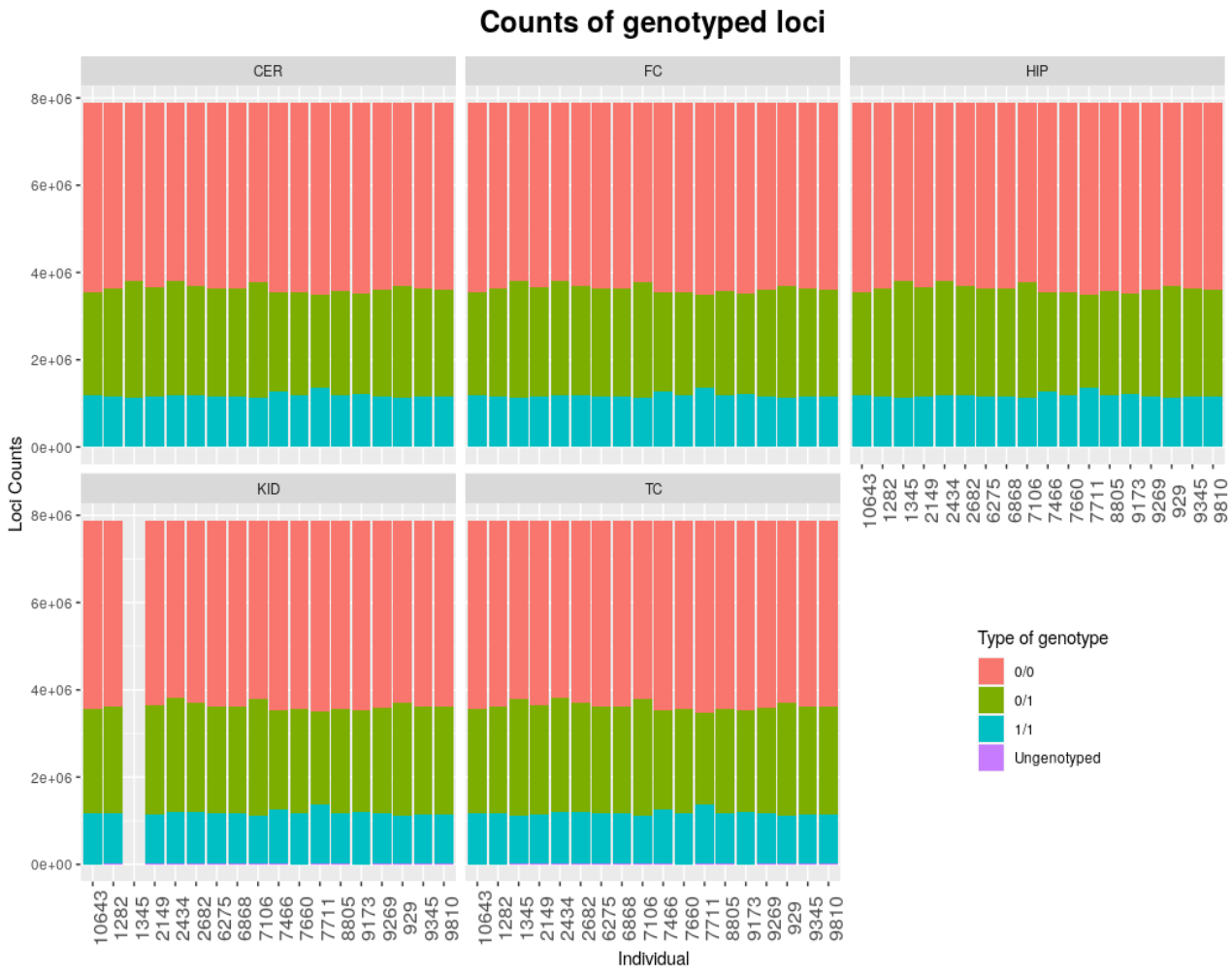


Figure 17: Representation of genotypes counts per sample. Genotype 0/0: homozygous for the reference allele; 1/1 homozygous for the alternative allele; 0/1 heterozygous.

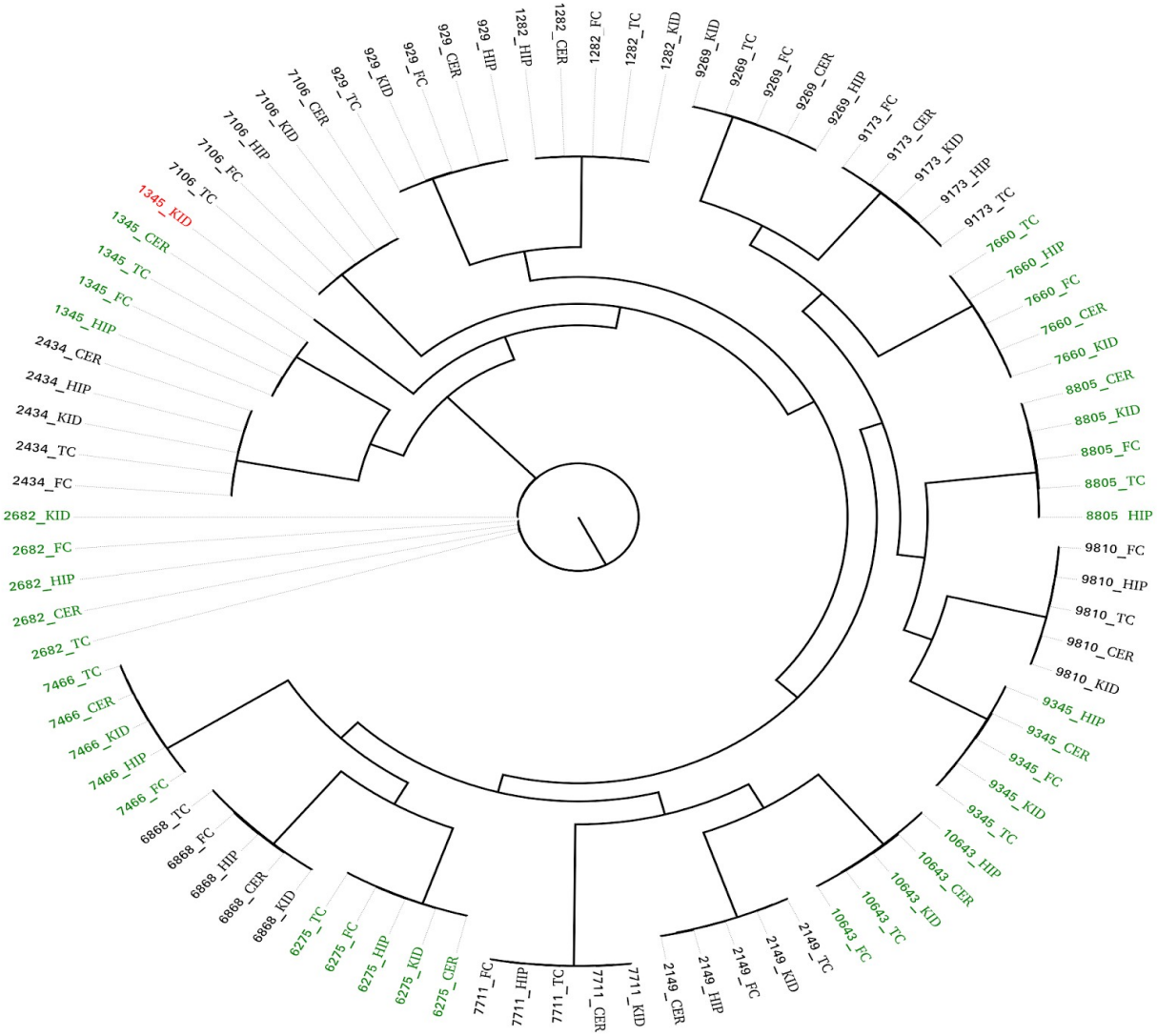


Figure 18: *SNPhylo* tree representation of samples relationships. AD in green, CTRs in black. Sample 1345_KID (in red) showed to not be related with the other 4 tissues of the same individual.

As already mentioned, WGS allowed me to extremely extend the number of genotyped loci with respect of SNPs array. Nonetheless, by merging both datasets together, I reached a total of about 8,000,000 loci per sample. I then observed that about 6,150,000 (77%) loci were genotyped by WGS alone, 1,700,000 (~22%) loci were observed with both technologies and only 88,000 (~1%) loci were found only by SNPs arrays methods (figure 19).

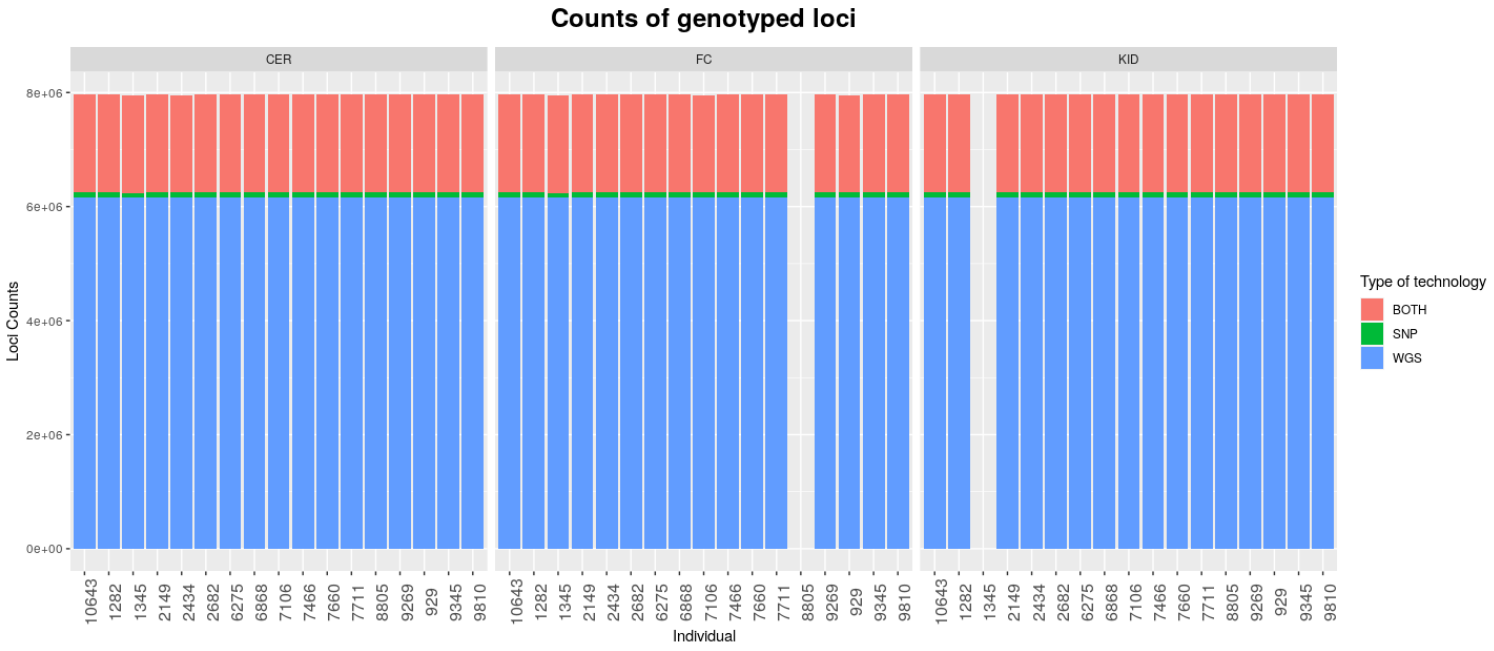


Figure 19: Counts of genotyped loci according to the technology applied. SNP: loci genotyped with SNP array only; WGS: loci genotyped with whole genome sequencing only, BOTH: loci genotyped with both technologies.

I then proceeded to determine the level of concordance between SNPs arrays and WGS variants calls by considering the only set of variants callable by both technologies. From this, I observed that a mean of 1,711,598 SNVs per sample (99.85%) was in agreement for the genotyping data while a mean of 2,670 SNVs per sample was not.

Next, I tested the ability of WGS technology in genotyping the location of the subset of somatic early onset SNVs, the ones that showed different genotypes between neural tissues and kidney from the same individuals, with the Illumina chip assays. From 50 to 80% of early onset SNVs loci (depending on the sample) were genotyped with WGS (figure 20). However, when checking the concordance for these genotypes calls between the SNPs array and the WGS, none of them were in agreement (figure 21).

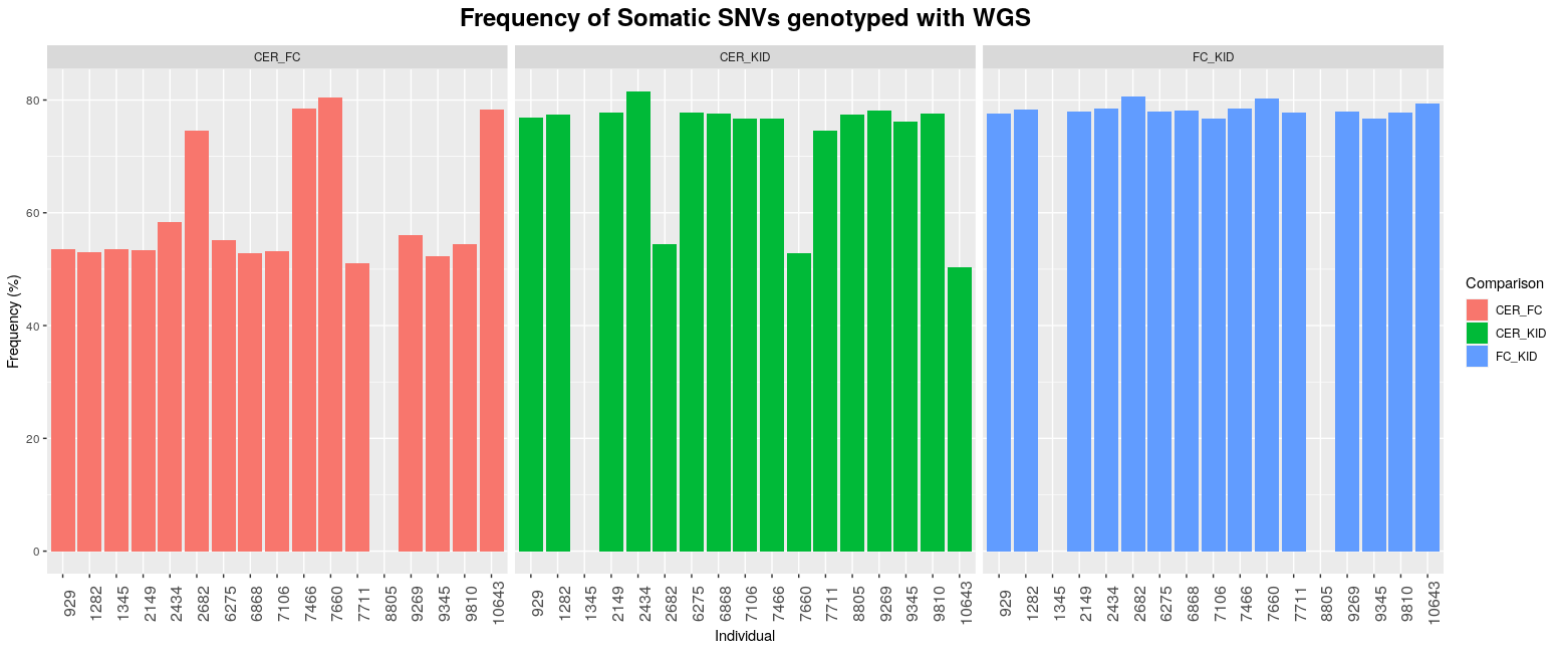


Figure 20: Frequency of early onset somatic SNVs identified with SNPs arrays and genotyped also with WGS.

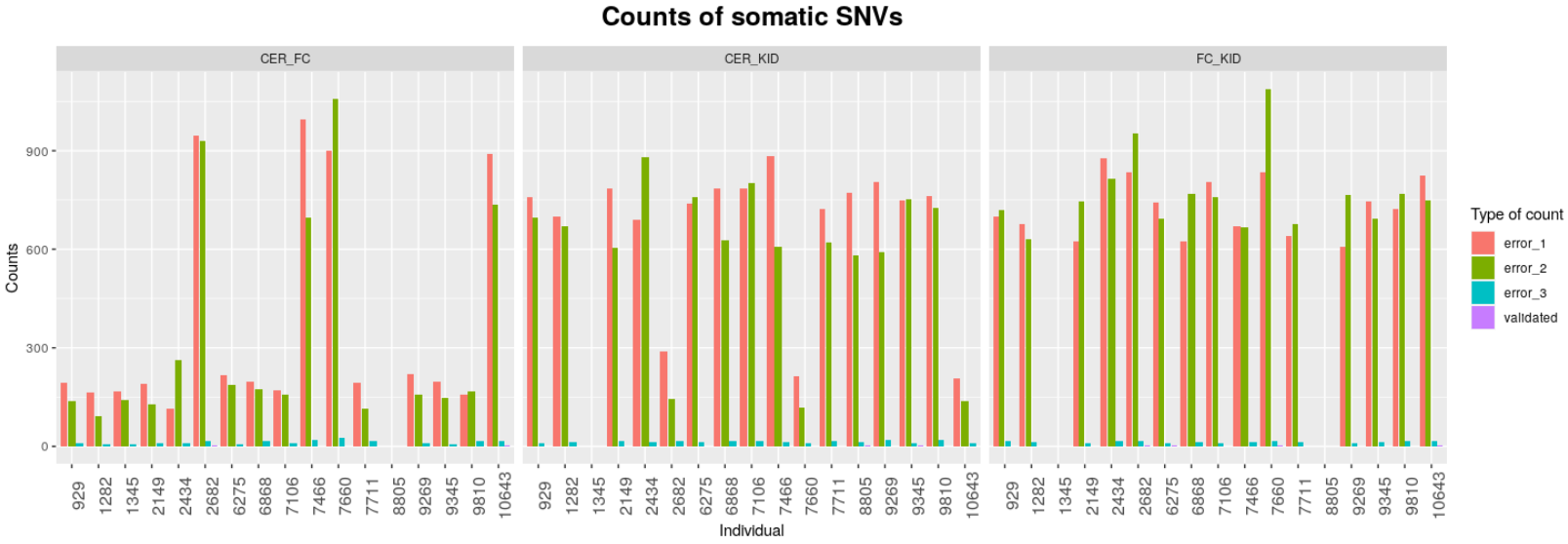


Figure 21: Counts of somatic SNVs genotyped with WGS. SNVs were grouped in 4 classes: Error_1: counts of SNVs for which the genotypes of the first tissue were validated with WGS that however presented different genotypes within the second tissue between SNP and WGS. Error_2: the opposite of error_1, first tissue genotypes that were not validated while second tissue genotyped yes; Error_3: both tissues genotypes were different between SNP and WGS datasets; Validated: validated genotypes.

Further manual checks on some random selected loci were not conclusive since both whole genome calls and SNP array calls were highly supported, thus requiring additional experimental validations (Table 5; figure 22).

SAMPLE_ID	TISSUE	PLATFORM	COORDINATE	SNP_ID	REF	ALT	GENO	B_ALLELE_FREQ	SUPPORTING_READS
6868	CER	SNP_array	chr1:159527480	rs2275674	T	C	0/0	0	NA
6868	FC	SNP_array	chr1:159527480	rs2275674	T	C	0/1	0.5751	NA
6868	CER	WGS	chr1:159527480	rs2275674	T	C	0/0	0*	161 (100%)
6868	FC	WGS	chr1:159527480	rs2275674	T	C	0/0	0*	133 (100%)

Table 5: Example of information used in manual checks on SNPs array and WGS SNV overlapping call. REF: reference nucleotide; ALT: alternative nucleotide; GENO: Genotype call; B_ALLELE_FREQ: Alternative allele frequency, observed from SNPs arrays and evaluated for WGS from supporting reads counts (*); Supporting reads: Counts of reads that supported the reference nucleotide, with the relative frequency with respect to the total reads mapped on the locus. Supporting reads were available for only WGS data. WGS data are represented with *IGV* screen shots in figure 22.

Figure 22: IGV screenshot representing table 5 variant with WGS data (highlighted by a red box). In both 6868_CER and 6868_FC samples (top and bottom track, respectively) no reads were found to support the alternative allele (nucleotide C) as indicated by the absence of colored tiles.

Since I was not able to validate SNPs array evidences, I decided to focus my attention on the whole WGS variants set, which was composed by more that 3 millions of single nucleotide variants per sample with respect to the reference genome (figure 17). Therefore, I first investigated the nucleotide substitutions, identifying, as for SNPs array, that C>T and T>C were the major types of events (figure 23). Next, I performed signature analyses in order to both gain information about the frequencies of nucleotide substitutions normalized by the trinucleotide genome context and to collect data regarding the origin of such variants. Results indicated that the nucleotides differences with respect to the reference genome are predominantly composed by C>T nucleotide substitutions, and assigned to SBS1, the clock-like signature associated with spontaneous deammination of 5-methylcytosine, and hallmark of aging (figure 24). Next, I seek to identify early onset somatic variants using WGS data. Similarly to SNPs array analyses, I compared the genotyping calls of pairs of tissues belonging to the same individual and I defined loci in which genotypes were not concordant as early onset SNVs. From my WGS analyses, I identified a mean of ~95,000 early onset SNVs per sample comparison,which consisted in a 40 times increase with respect to SNPs arrays data. Nonetheless, no differences between AD and CTRs cohorts were observed (figure 25).

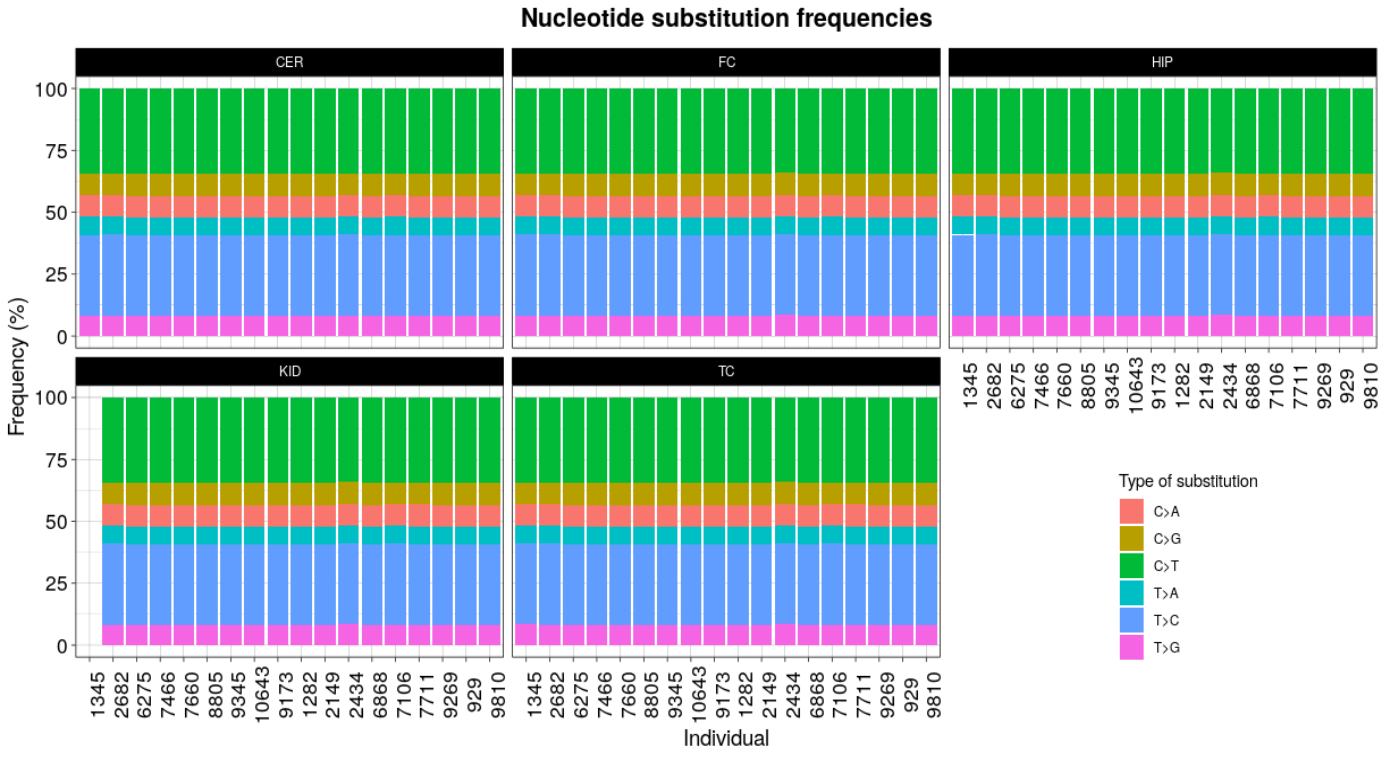


Figure 23: Nucleotide substitutions frequencies for early onset SNVs identified with NGS approach.

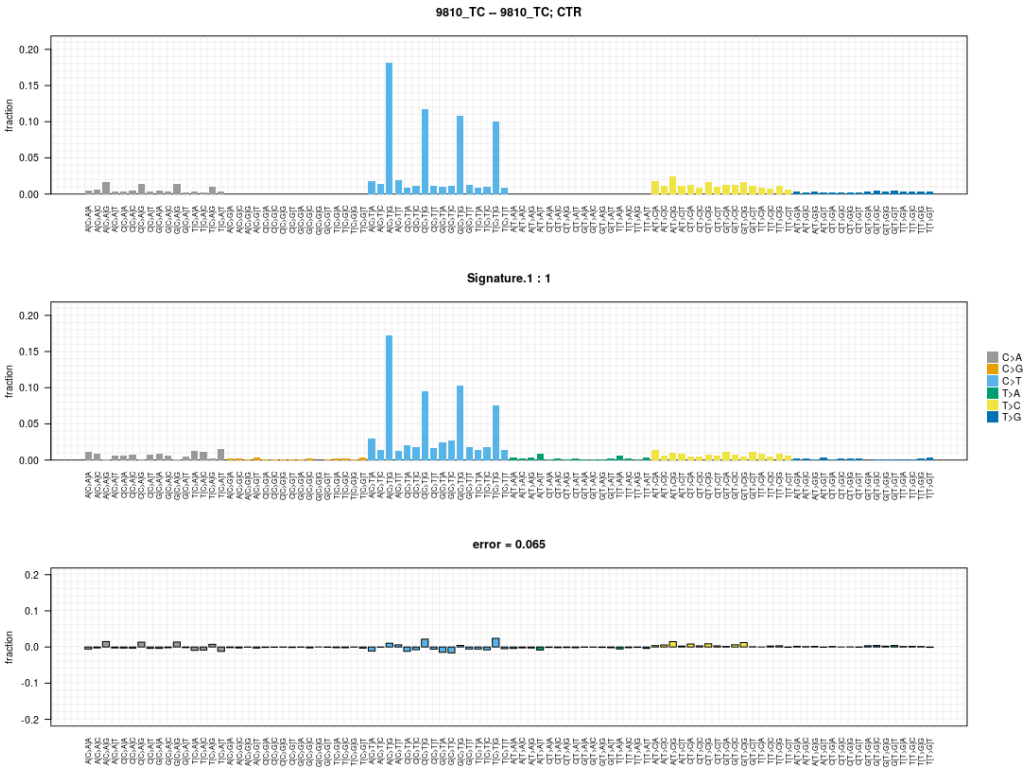


Figure 24: Signature analyses for sample 9810_TC with whole WGS variants set. From top to bottom: Mutational pattern observed with frequencies of nucleotides substitutions; Mutational pattern reconstructed with known signatures; Error rates in signatures reconstructions. Signature.1 was found to reconstruct the whole pattern of nucleotide substitutions with relative small error rates.

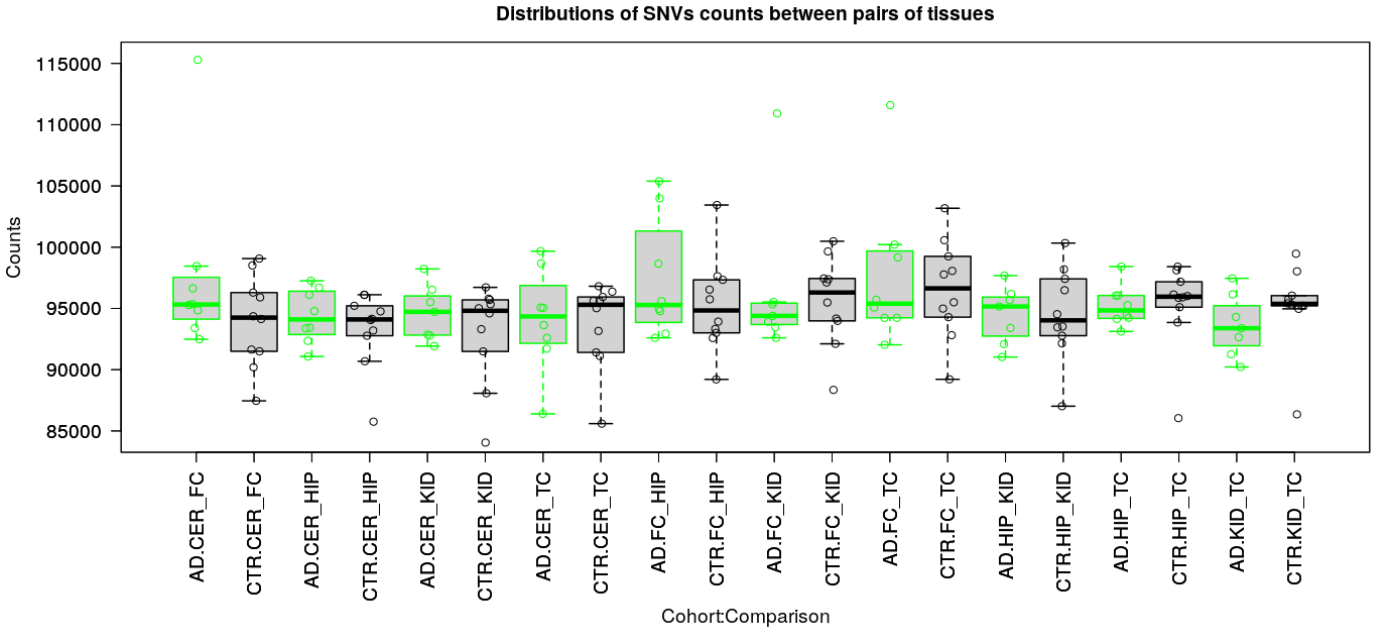


Figure 25: Early onset SNVs counts distributions. AD are displayed in green while CTRs in black.

Next, to identify mutagens processes acting in brain, I focused on only early onset SNVs called from the comparison between brains and KID tissues, collected from the same individuals, and performed additional signature analyses upon this subset. The results demonstrated that, within early onset SNVs with respect to KID, the major source of brain variants were imputable to SBS1, responsible of about 27% of the nucleotide substitution pattern observed (figure 26). Interestingly, the analyses did not confirmed previous SNPs array observations. In particular, SBS24 was not identified, neither at low frequencies, while SBS39 trends were not found (figure 27-A). On the contrary, I was able to unveil the presence of additional signatures: SBS5, SBS6 and SBS10b (figure 27-B), that, however, were not differentially represented in frequencies between AD and CTRs, nor according to tissue classifications. Nonetheless, I found that more than 20% of variants cannot be ascribed to any known signatures (Unknown signature) in agreement with SNPs array results.

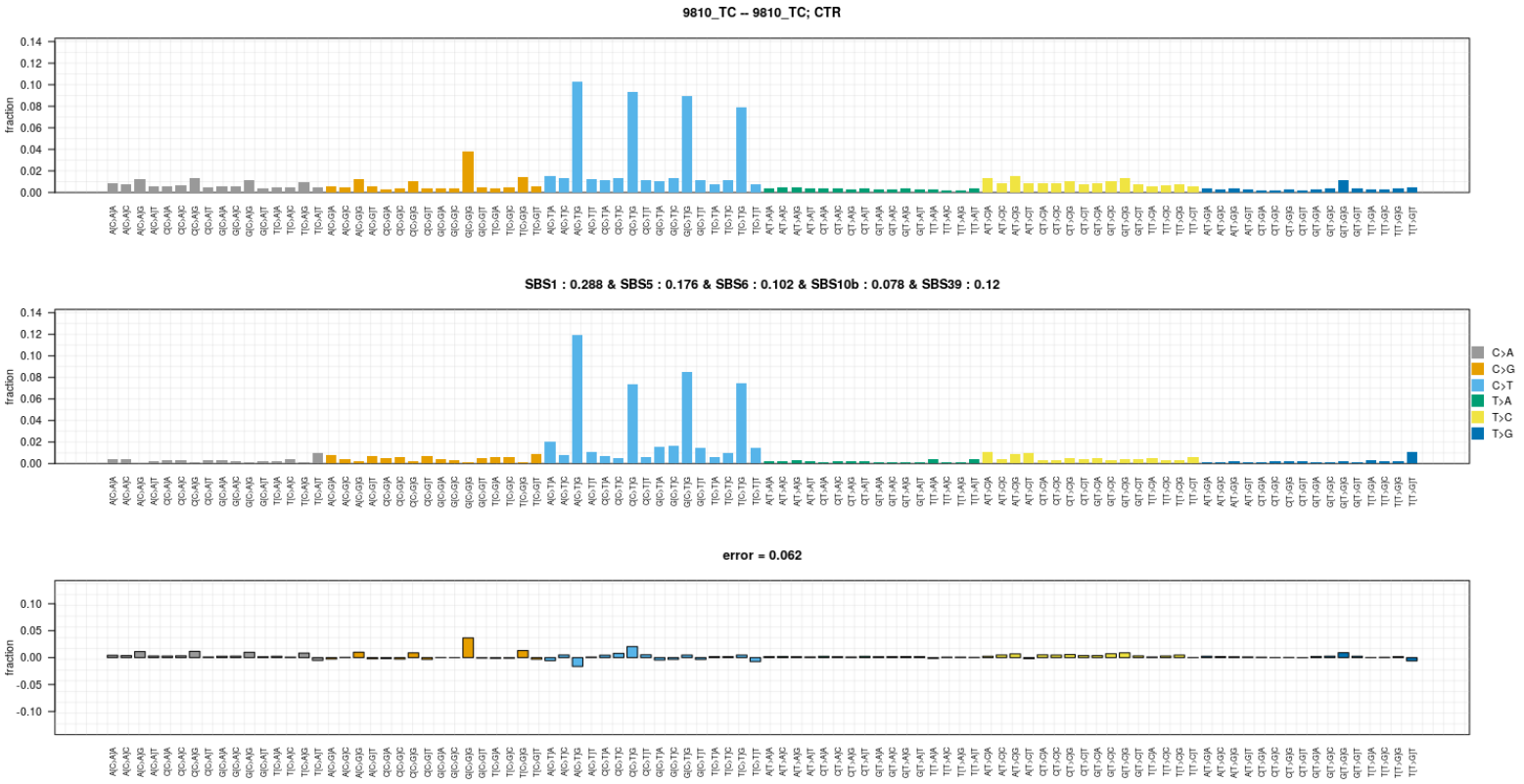


Figure 26: Signature analyses for sample 9810_TC with early onset SNVs called with respect to the KID. From top to bottom: Mutational pattern observed with frequencies of nucleotides substitutions; Mutational pattern reconstructed with known signatures. Frequencies for SBSs are reported within the title; Error rates in signatures reconstructions. Signature.1 was found to reconstruct about 0.29% of the whole pattern of nucleotide substitutions.

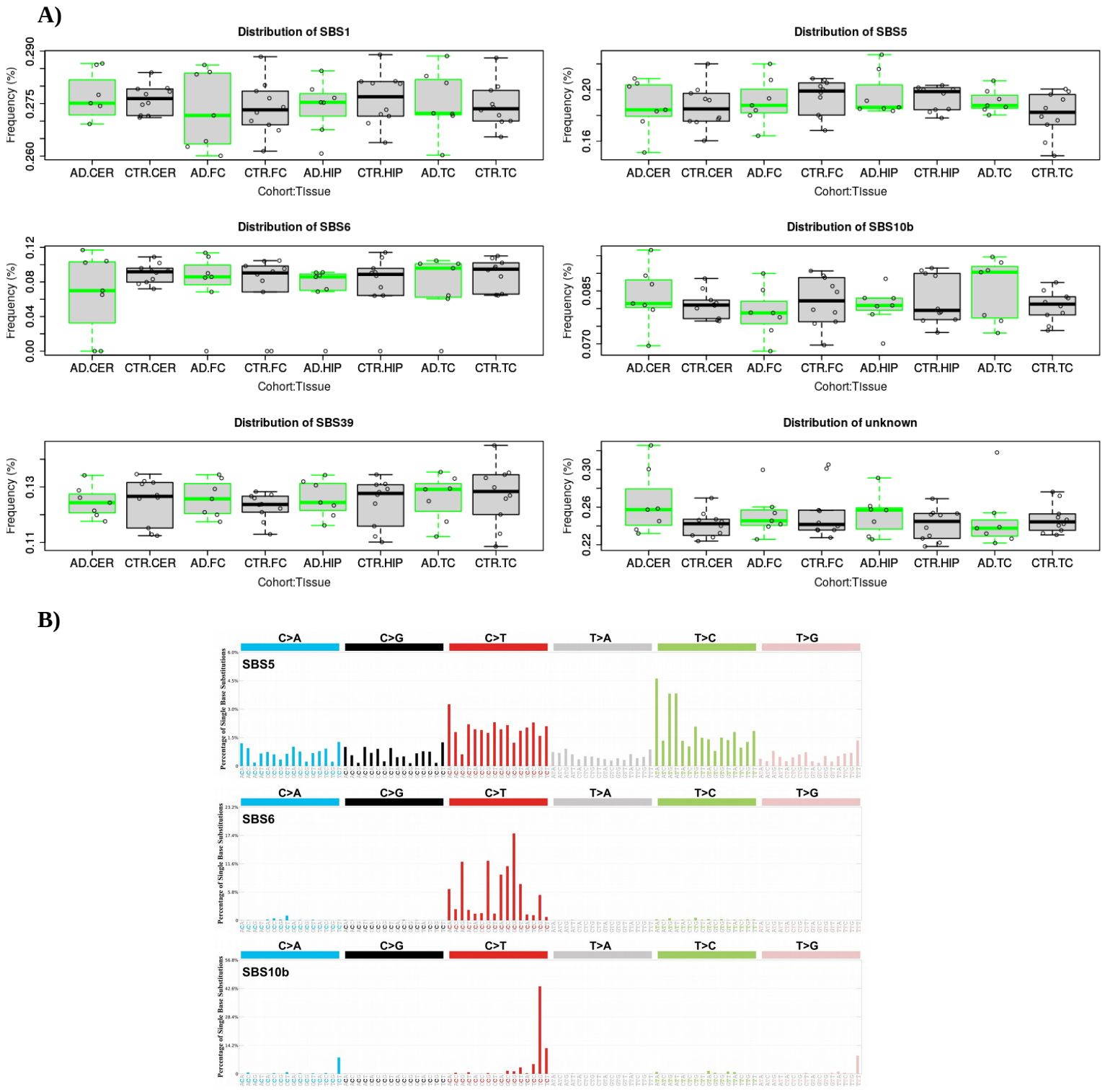


Figure 27: Mutational Signature analyses on WGS early onset SNVs data. **A)** Distributions of identified mutational signatures. AD samples in green; CTRs in black. **B)** Signature spectra from COSMIC v.3 database for SBS5, SBS6 and SBS10b.

3.3.2 Retrotransposons CNV evaluation

3.3.2.1 Supporting reads analyses

To obtain a fast comparison of CNVs associated to retrotransposons, supporting reads analysis was performed. Briefly, this consisted in mapping the raw reads data versus the consensus sequence of a mobile element, followed by a measurement of the number of mapped reads (see methods for further details). This approach relies on the fact that retrotransposon CNVs could impact the number of reads that originate from the selected retrotransposon sequences, modifying its coverage. When comparing two samples with different CNVs, the one with higher CNVs content (*i.e.* more retrotransposon sequences) will present more reads generated from the relative element (higher coverage). *Vice-versa*, the sample with lower CNVs content will present less reads ascribable to that mobile element (lower coverage). We applied this rationale to the three human non-LTR different retrotransposons elements, Alu, L1 and SVA, and imposed several filters before measuring the number of mapped reads. We discarded all duplicated reads that could be generated by PCR amplification, to avoid biases in the calculation of coverages, and reads with mapping quality lower than 60 (the maximum mapping quality for Illumina technology). Although fast and relatively simple, this approach has some limitations. For instance, due their repetitive nature, repetitive elements suffer major mapping problems that can affect the mapped reads counts and that can be only partially resolved by applying mapping quality thresholds. Furthermore, it represents a summary of all the possible CNVs effects, not discriminating between the extent of the *de-novo* integrations and deletions. Although no differences can be extrapolated from the cohorts comparison (figure 28), it cannot be excluded the presence of alterations in the rate of integrations nor in the levels of mobile elements interested by genomic deletions, as observed during the SNPs array experiments (figure 12, chapter 2), or other type of genomic CNVs.

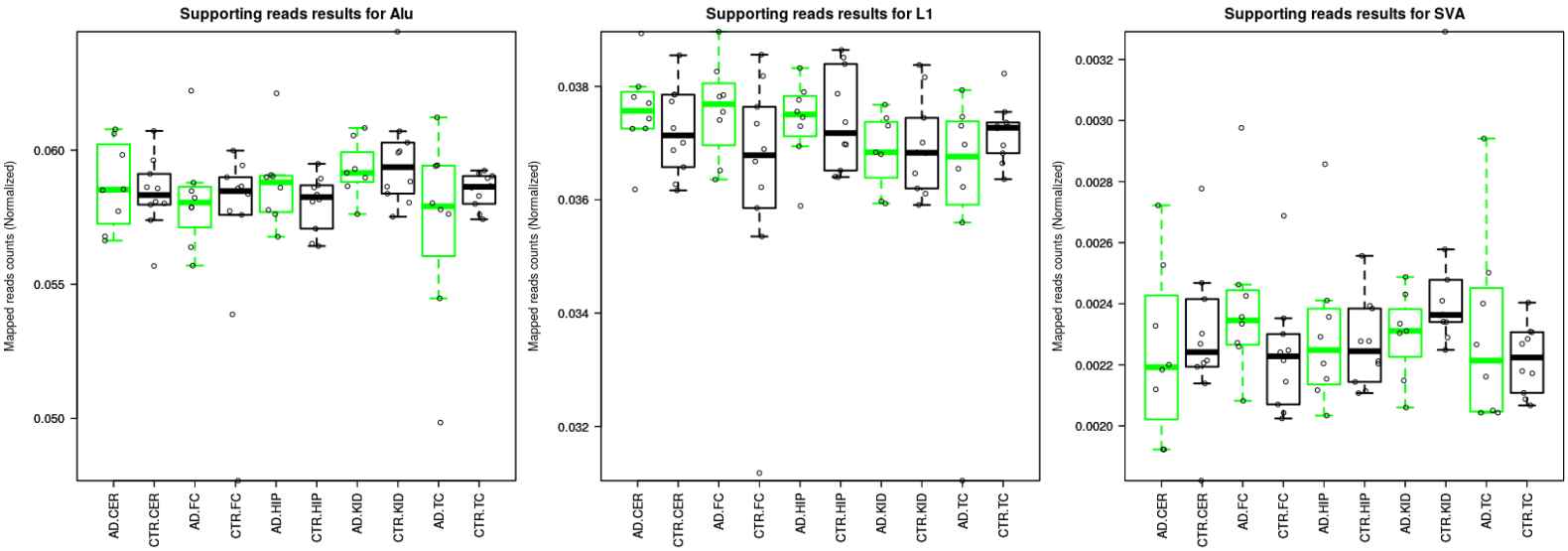


Figure 28: Supporting reads analyses results. From left to right, data reports counts distributions of Alu, L1 and SVA elements, respectively, normalized using the number of mapped reads.

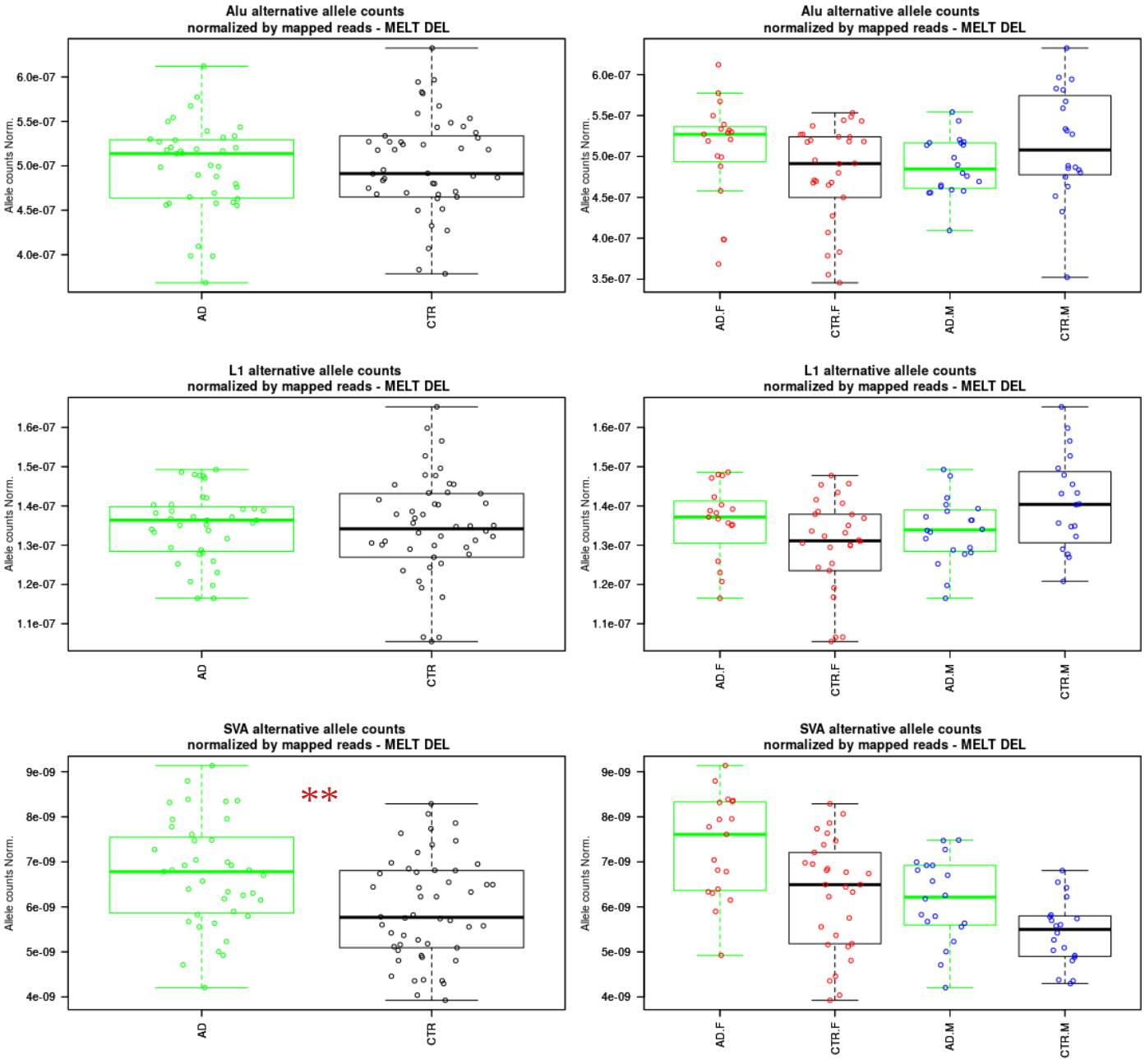
3.3.2.2 MELT analyses

To improve retrotransposons CNVs estimation I then took advantage of *MELT analyses* to genotype both reference retrotransposons annotations as newly integration events with respect to the reference genome. These were composed by *MELT Deletion* and by *MELT Single* analyses. *MELT Deletion* is able to genotype reference annotations of transposable elements, while *MELT Single* is capable to detect and genotype newly integration events with respect to the reference genome. Both analyses resulted in the identification of alternative loci, which represent genomic regions that presented differences with respect to the reference genome. ***These loci that I will call "alternative loci" indicate the absence of the transposable element reference sequences (in MELT Deletion results or novel insertion in MELT Single results.*** From the alternative loci, alternative allele counts (normalized by total mapped reads) per sample were obtained and used as unity of measurement of retrotransposon CNVs.

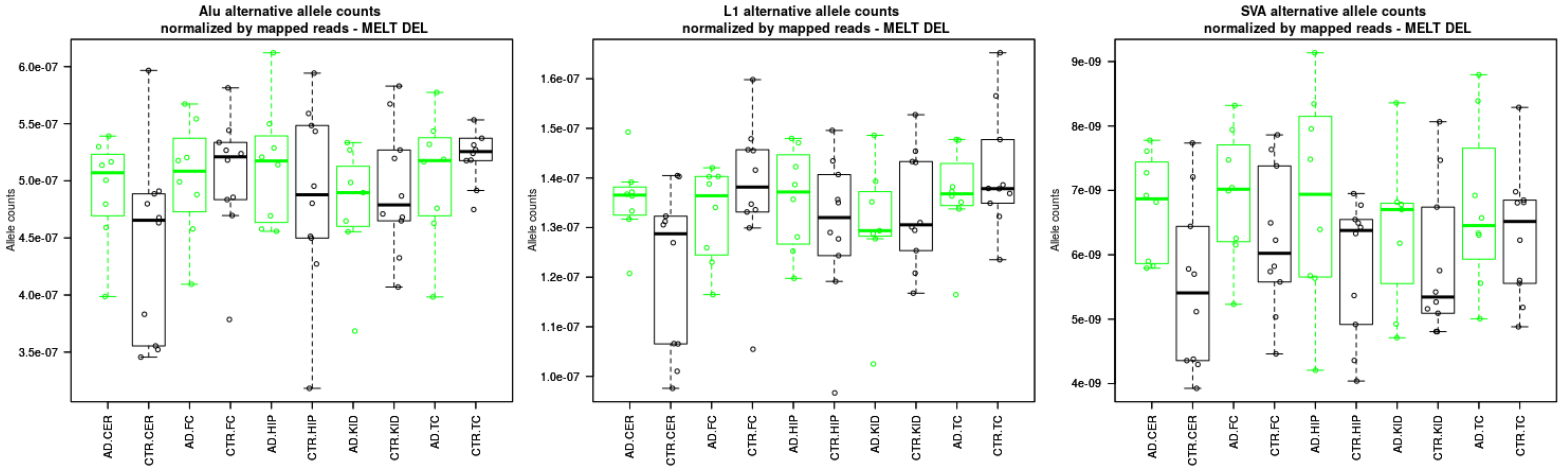
3.3.2.2.1 *MELT Deletion* analyses

From the analyses of the reference annotations (performed with *MELT Deletion*), no statistical significant differences in the levels of alternative alleles for L1 and Alu elements were found (figure 29-A). On the contrary, an higher alternative allele counts for SVA sequences was evident in AD, that reached statistical significance (AD vs CTR p value: 0.002) when not stratifying according to the tissue (figure 29-B). Alternative L1-containing loci were next classified in *germinal* and *somatic* depending on the number of tissues that harbored the variant calls. Loci were called as *germinal* when all 5 tissues supported an alternative locus while loci were defined as *somatic* when variants were found exclusively in one tissue. Although alternative alleles counts from *germinal* alternative loci were not dissimilar between the two cohorts, we found that CER of AD samples presented higher *somatic* alternative allele counts with respect to the CTRs (figure 29-C).

A)


Alu alternative allele counts
normalized by mapped reads - MELT DEL
Allele counts Norm.
6.0e-07
5.5e-07
5.0e-07
4.5e-07
4.0e-07
AD
CTR
Alu alternative allele counts
normalized by mapped reads - MELT DEL
Allele counts Norm.
6.0e-07
5.5e-07
5.0e-07
4.5e-07
4.0e-07
3.5e-07
AD.F
CTR.F
AD.M
CTR.M
L1 alternative allele counts
normalized by mapped reads - MELT DEL
Allele counts Norm.
1.6e-07
1.5e-07
1.4e-07
1.3e-07
1.2e-07
1.1e-07
AD
CTR
L1 alternative allele counts
normalized by mapped reads - MELT DEL
Allele counts Norm.
1.6e-07
1.5e-07
1.4e-07
1.3e-07
1.2e-07
1.1e-07
AD.F
CTR.F
AD.M
CTR.M
SVA alternative allele counts
normalized by mapped reads - MELT DEL
Allele counts Norm.
9e-09
8e-09
7e-09
6e-09
5e-09
4e-09
**
AD
CTR
SVA alternative allele counts
normalized by mapped reads - MELT DEL
Allele counts Norm.
9e-09
8e-09
7e-09
6e-09
5e-09
4e-09
AD.F
CTR.F
AD.M
CTR.M

B)



C)

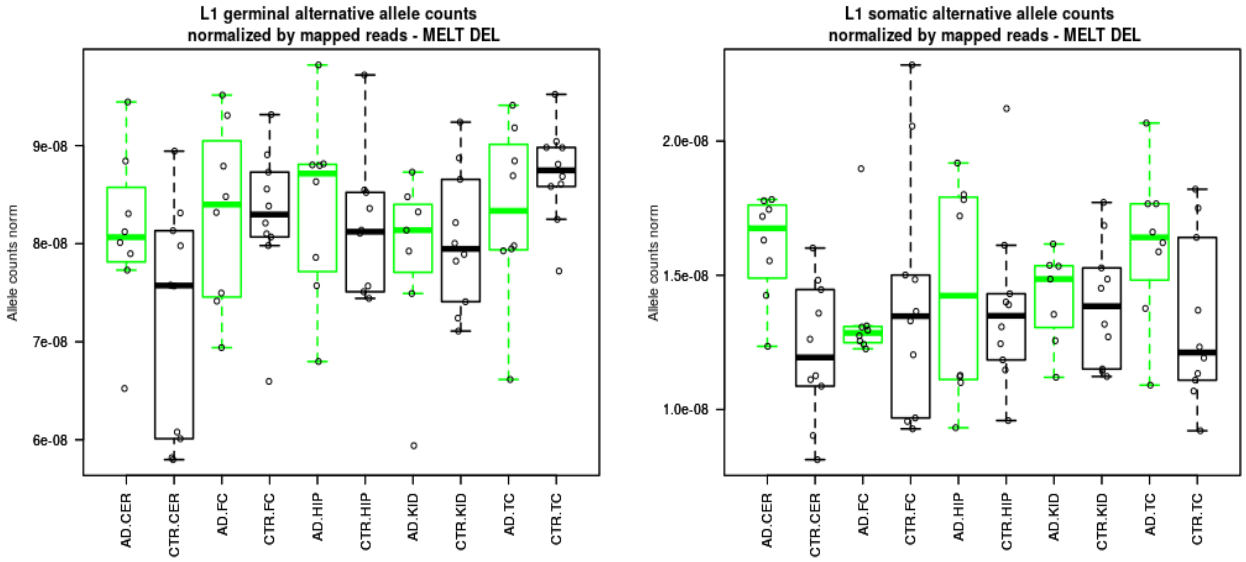
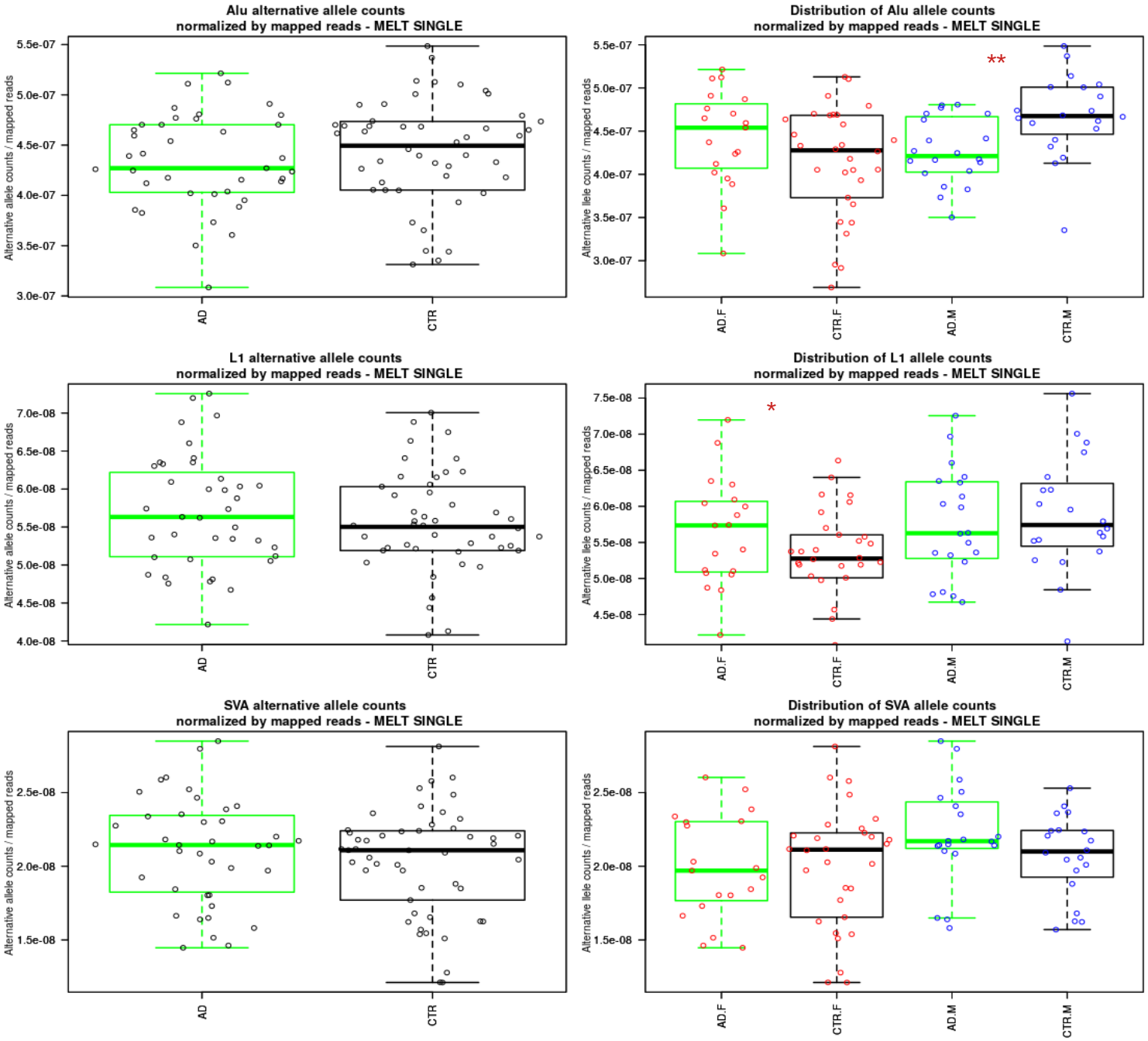


Figure 29: *MELT Deletion* results. **A)** Representation of Alu, L1 and SVA alternative allele counts normalized with total mapped reads. Results are pooled by cohort (left) and then stratified by cohort and gender (right). **B)** Distributions of Alu, L1 and SVA alternative allele counts normalized by total mapped reads and grouped by sample. **C)** distributions of germinal and somatic alternative alleles.

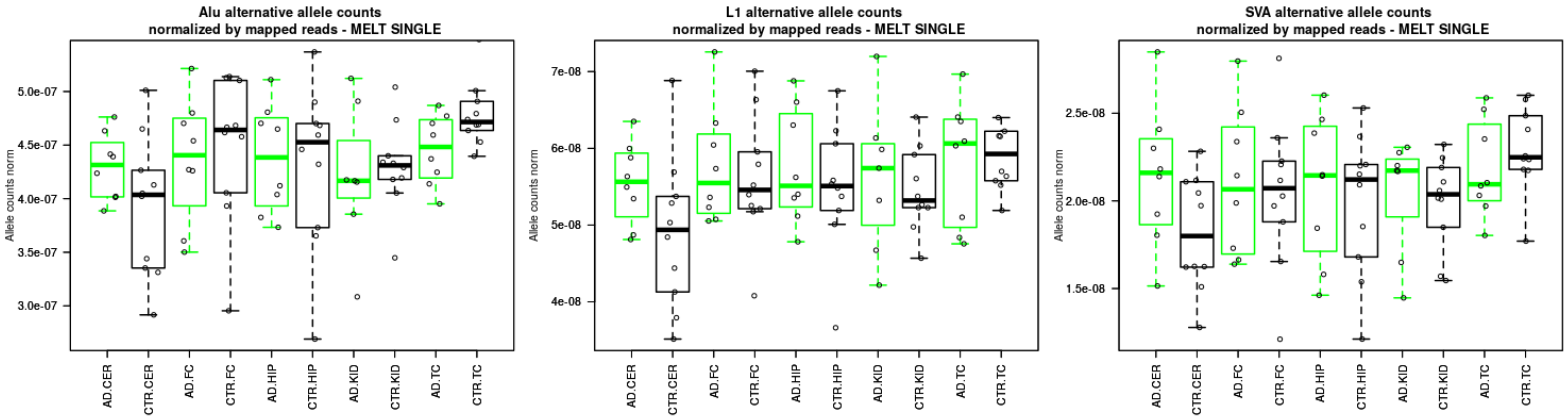
3.3.2.2.2 MELT Single analyses

I next focused on the analysis of retrotransposon integrations. While distributions grouped by cohort and tissue did not display any significant trend (figure 30-B), Alu alternative alleles were found to be higher in males CTRs with respect to males AD (p value 0.0043) and L1 alternative alleles were observed to be enriched in females AD samples (p value 0.029) (figure 30-A). To determine the impact of the genetic background on the results, I focused my attention to L1 data and classified alternative loci in *germinal* and *somatic*. I started by following a strategy similar to the one applied with the *MELT Deletion* results. Loci were called as "*shared*" when all 5 tissues supported an alternative locus while loci were defined as "*private*" when variants were found exclusively in one tissue. For both classes, the distributions of the reads that supported the evidences were displayed (see methods and figure 31), which lead to the identification of a threshold at 15 reads. According to the number of supporting reads, L1 data were then classified in *germinal* (with more than 15 reads) and *somatic* (with less than 15 reads) regardless of being "*private*" or "*shared*". Although no differences were observed when focusing on *somatic* loci, *germinal* sites showed a general trend shared by all tissues, in which AD presented higher alternative allele counts with respect to CTRs, which was in partial agreement with previous results (figure 30-C). Without grouping samples by gender and tissue, this evidence almost reached statistical significance suggesting that current sample size may be a limiting factor of our analyses (p value 0.059).

A)


Alu alternative allele counts
normalized by mapped reads - MELT SINGLE
Alternative allele counts / mapped reads
5.5e-07
5.0e-07
4.5e-07
4.0e-07
3.5e-07
3.0e-07
AD
CTR
Distribution of Alu allele counts
normalized by mapped reads - MELT SINGLE
Alternative lele counts / mapped reads
**
AD.F
CTR.F
AD.M
CTR.M
L1 alternative allele counts
normalized by mapped reads - MELT SINGLE
7.0e-08
6.5e-08
6.0e-08
5.5e-08
5.0e-08
4.5e-08
4.0e-08
Distribution of L1 allele counts
normalized by mapped reads - MELT SINGLE
*
7.5e-08
SVA alternative allele counts
normalized by mapped reads - MELT SINGLE
2.5e-08
2.0e-08
1.5e-08
Distribution of SVA allele counts
normalized by mapped reads - MELT SINGLE

B)



C)

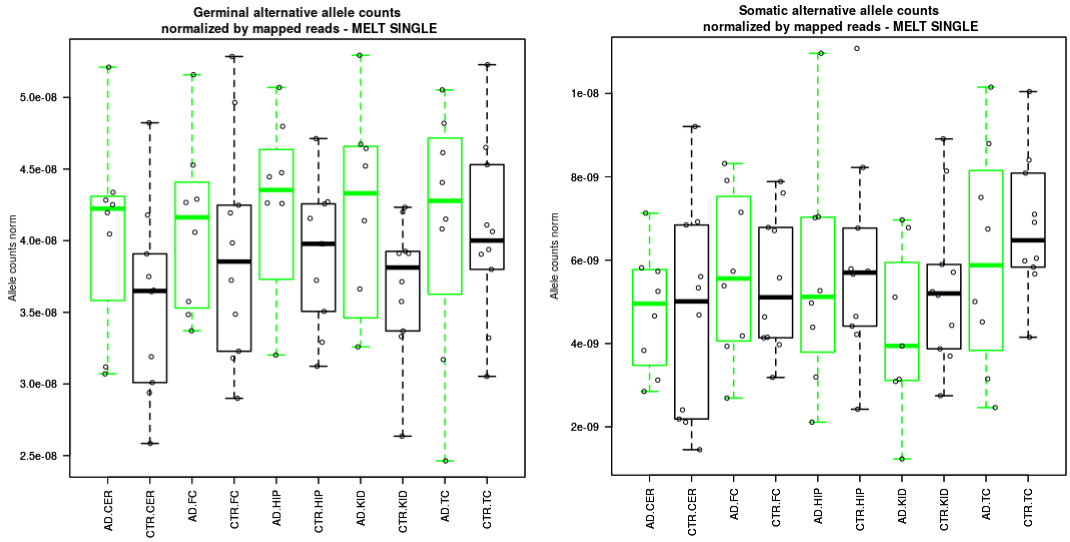


Figure 30: *MELT Single* results. **A)** Representation of Alu, L1 and SVA alternative allele counts normalized with total mapped reads. Results are pooled by cohort (left) and then stratified by cohort and gender (right). **B)** Distributions of Alu, L1 and SVA alternative allele counts normalized by total mapped reads grouped by sample. **C)** Distributions of germinal and somatic alternative alleles.

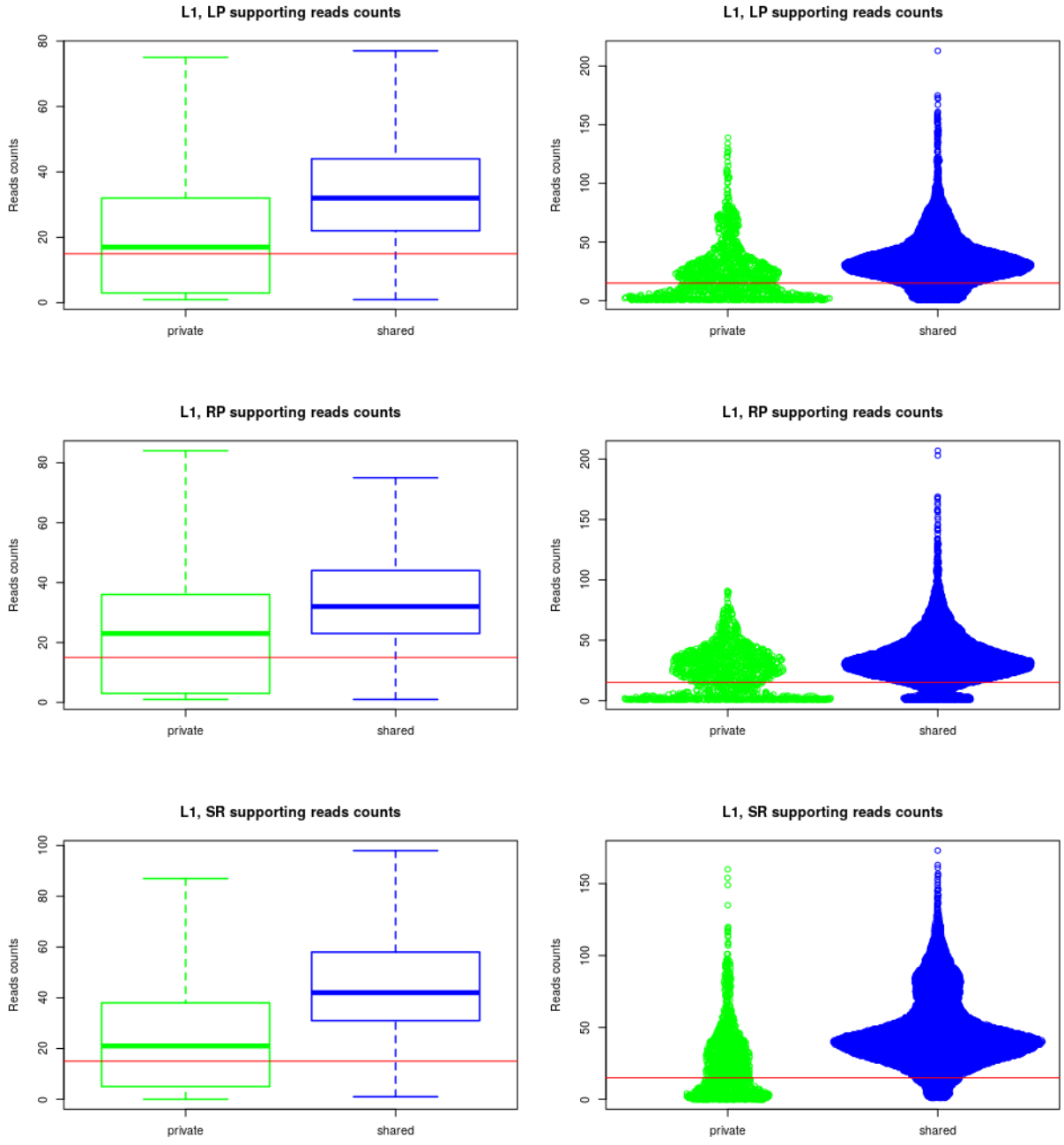


Figure 31: *MELT Single*, distributions of breakpoints supporting reads. LP: Discordant reads pairs that support Left breakpoints; RP: Discordant reads pairs that support Right breakpoints; SR: Split reads that support LINE-1 integrations.

3.3.2.2.3 MELT net alleles

Finally, both *MELT* analyses were integrated in order to estimate an overall retrotranposons CNVs. *MELT deletion* analyses, additionally resulted in the identification of loci in which the presence of mobile elements was not dissimilar from the condition observed in the reference genome. These reference loci were thus used to extract **reference allele counts**, which were next added to the **retrotransposon integrated alleles counts**, obtained with *MELT Single* analyses instead. The final count represented a good estimation of the overall retrotransposon CNVs.

From the analyses, no significant differences between the two cohorts were found, in agreement with the previous supporting reads analyses (figure 32).

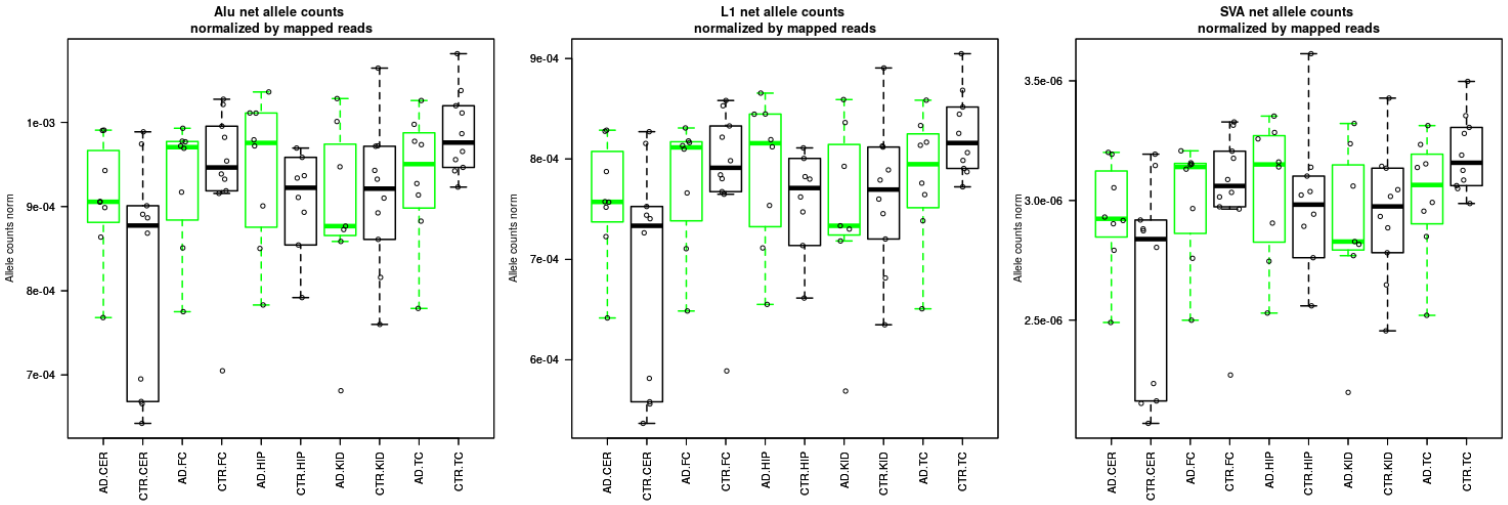


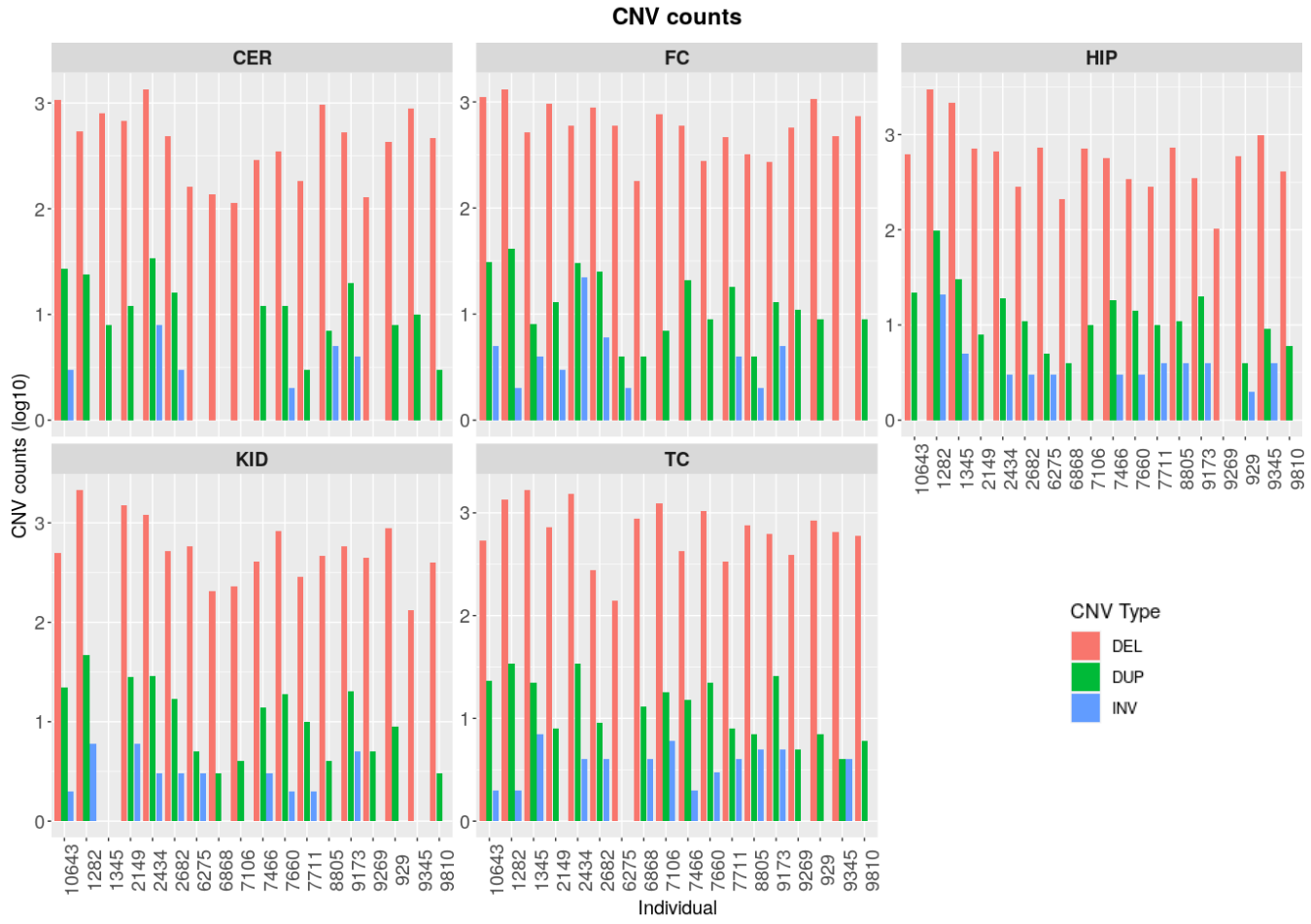
Figure 32: Distributions of Alu, L1 and SVA net allele counts normalized with total mapped reads per sample.

3.3.3 Exploratory analyses of genomic CNVs

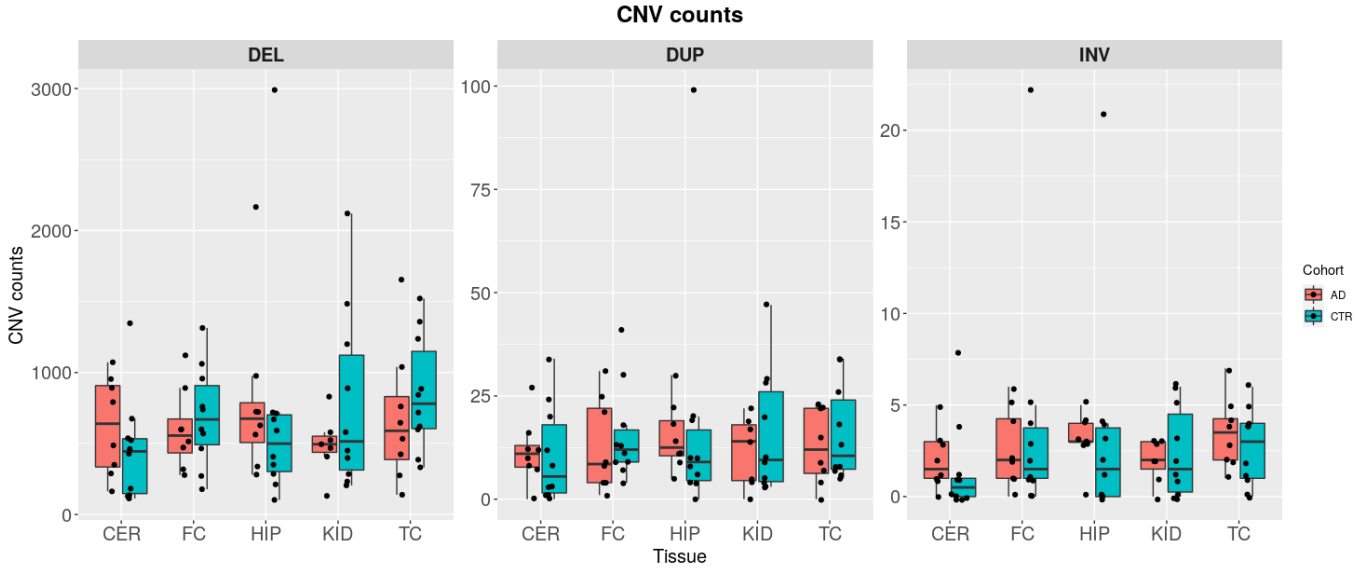
Lastly, with the ultimate goal of validating SNPs array results, I started investigating genomic CNVs from WGS data. Through the application of a robust pipeline (see methods), I was able to call 59,924 total deletions (DEL), 1,256 total duplications (DUP) and 244 total inversion (INV) (figure 33-A). Distribution of CNV counts showed no statistical significant differences when grouped by type, tissue and cohort (figure 33-B).

Subsequently, to validate SNPs arrays CNVs, I started performing positional overlaps using DEL and DUP called with both SNPs arrays and WGS data. From the results, it was found that 552 out of 1,122 chip CNVs were identified with the parallel technology. Moreover, counts of overlaps were not affected by the parameters used during the analyses, as changing the options did not increase the number of validated CNVs. First parallel validations were encouraging, having found about 50% of correspondence between SNPs array and WGS data. However, CNV counts from WGS were at least one order of magnitude higher than CNVs counts from SNPs arrays. A potential explanation could be represented by the fact that CNV calling from SNPs arrays requires signals from at least three consecutive probes. Therefore, to test this possibility, I investigated whether WGS CNVs were actually in overlap with the SNPs arrays probe panel. Data showed that 56,585 CNVs (55,920 DEL and 665 DUP) were in overlap with less than 3 chip probes, and thus, not being callable though SNPs arrays technology (figure 33-D).

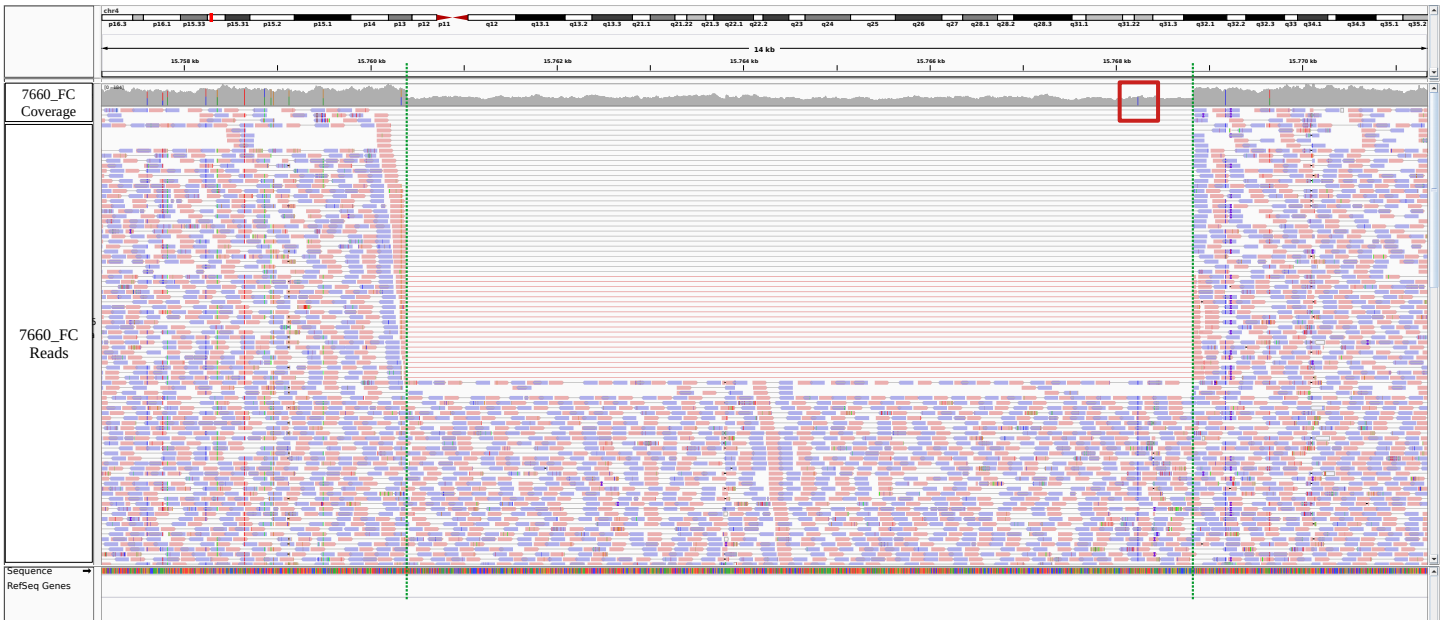
A)


CNV counts
CER
FC
HIP
KID
TC
CNV counts (log10)
Individual
10643
1282
1345
2149
2434
2682
6275
6868
7106
7466
7660
7711
8805
9173
9269
929
9345
9810
CNV Type
DEL
DUP
INV

B)


CNV counts
DEL
DUP
INV
CNV counts
Tissue
CER
FC
HIP
KID
TC
Cohort
AD
CTR

C)



D)

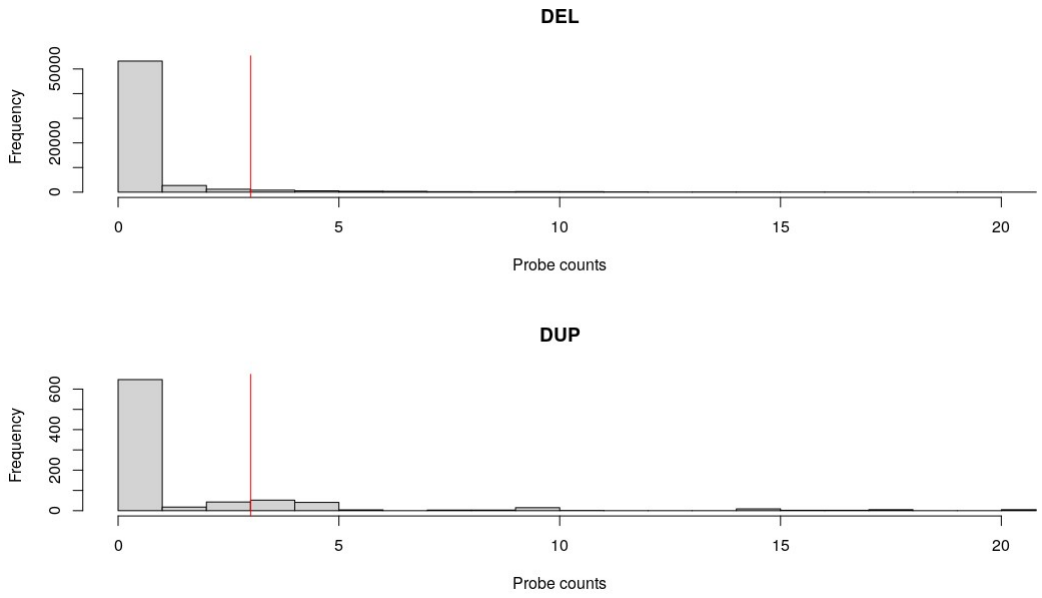


Figure 33: Genomic WGS CNVs exploratory analyses. **A)** Counts of CNVs per sample;**B)** Distribution of counts of CNVs grouped by tissues and cohorts. **C)** IGV screenshot showing a putative genomic heterozygous deletion identified with WGS from sample 7660_FC**.** Green dot lines mark the locus. From the coverage curve (top track) it can be appreciated the drop in coverage within the deletion. Mapping reads also supported the evidence as both split and discordant reads can be found at the breaking point levels. An additional supporting evidence is represented by the absence of heterozygous variants within the putative heterozygous deletion, which instead are abundant in the outside regions; **D)** Frequencies of WGS CNVs in overlap with the SNP array probe panel. Ref vertical lines mark the 3 probes cutoff.

3.4 Discussion and Conclusions

Here I provided a preliminary investigation of genome-wide SNVs and retrotransposons CNVs for a cohort of Alzheimer's disease individuals across 5 different tissues. WGS did not validate previous evidences of alteration in somatic SNVs quantity between AD and CTRs cohorts, as obtained with SNP-arrays. Instead, no quantitative differences in the numbers of early onset SNVs were found by comparing pair of tissues. I reconnected this discrepancy to intrinsic errors of the SNP array platforms that will require further experimental tests. However, it cannot be yet excluded that CNVs, in terms of genomic deletions and duplications, may have affected whole genome variant calling. In fact, it should be noted that whole genome variant calling approach was based on the *HaplotypeCaller* GATK 4 software, which is able to detect only germline variants by assuming a diploid state of the genome. Therefore, CNVs data from WGS, which also started to be explored, will be further used to investigate the possible bias inserted by genomic CNVs.

Although early onset SNVs were not quantitatively different within our cohorts, through mutational signature analyses we were able to track down the observed nucleotide substitutions to several mutagenics processes. Among them, SBS.1 is related to the endogenous and spontaneous deammination of 5-methylcytosine and it is linked to normal aging (*clock-like Signature*). It was expected since samples derived from old aged individuals. However, analyses showed that only about 25% of SNVs could be tracked to SBS1. In addition to SBS1, also SBS5, SBS6, SBS10b and SBS39 were detected in both AD and CTRs, although with no different frequencies ascribable to particular tissues or cohorts. SBS5, as SBS1, is a *clock-like* signature which to date is not associated with a known mutagen process. SBS6 instead, is related to defective DNA mismatch repair and was found to account for about 10% of the SNVs detected. Furthermore, also SBS10b was identified, which is associated to the effect of polymerase epsilon exonuclease domain. Overall, I did not found dissimilarities when focusing on particular tissues or cohorts, which, altogether to the identification of two *clock-like* signatures, could suggest that these may be a natural consequence of brain aging. With respect to SNPs array signatures, SBS5, SBS6 and SBS10b were a novelty. Moreover, with WGS we were not able to replicate and validate the discovery of SBS24 and the trend observed in SBS39 through SNPs arrays (although being able to call it with WGS data). These aspects could lead to several speculations. For instance, SNPs array analyses were focused on loci known to be polymorphic in the

human population, while NGS variants were the results of a genome-wide analyses that highly increased the number of tested loci. It could be hypothesized that this increase in variants identification may have improved signature assignation, and thus provided a different, and potentially more reliable, signature pattern. Nonetheless, both SNPs array and NGS signatures analyses evidenced how about 20% of variants cannot be ascribed to known signatures, underlying the current lack of knowledge regarding mutagens processes acting in non-cancer tissues, and potentially affecting the reliability to signature assignations.

With this study, I showed how the application of mutational signature analyses to other tissues and diseases than cancer, may further expand our understanding on the etiology of different pathologies. Importantly, the lack of non-cancer derived signature may have limited the ability to identify new AD-associated patterns of somatic variations which may remain hidden under the unknown signature results.

In this chapter I also investigated retrotransposons content in AD post-mortem tissues using WGS data, aiming to explore the significance of SNPs array results which suggested a loss of L1 elements due to genomic CNVs. By applying several bioinformatic analyses upon WGS data, I obtained two types of estimates of CNVs for L1, Alu and SVA. My approaches consisted in the analysis of *supporting reads* and in the *net retrotransposon alleles counts* (from the combination of *MELT* analyses). Both of them, did not show clear differences between the AD retrotransposons content and the CTRs ones. However, deeper investigation of the results provided by *MELT (both MELT-deletion and MELT-single)* pointed out that AD samples present higher SVA alternative alleles counts and, within non-reference annotations (*i.e.* integrations), higher L1 germinal alternative alleles. Although statistical significance was not always reached, concordance was found when comparing the trends from all the five tissues, suggesting that the extension of the analyses to a larger cohort of samples as well as to different than-L1 elements (*i.e.* SVA) could clarify whether the enrichments might truly be associated to the disease status. These exploratory analyses stresses the need of additional study of SNVs and retrotransposon CNVs in neurodegeneration diseases. To this purpose, I started to perform additional investigations upon the WGS dataset, focusing on genomic CNVs. Early exploratory analyses, which were mostly related on CNVs counts, were ineffective in showing significant differences between the two cohort. Nonetheless, I plan to further extend the analyses of CNVs by assessing L1 content in CNVs, trying to validate SNPs array results. Additionally, CNVs data might be also used to detect the existence of early

onset CNVs (*i.e.* CNVs present in the majority of the cells of the given tissue) which may provide further pieces of information.

In the next chapter, I will exhibit additional SNVs data, resulted from the application of a complete different algorithm, known as MUTECT2, which is specifically designed to unveil the presence of somatic SNVs, in particular late onset SNVs (*i.e.* variants present in a small numbers of cells from the given tissue), by comparing pairs of tissues collected from the same individual.

Chapter IV

Identification of Multi-nucleotide somatic variants from next-generation sequencing data

4.1 Introduction

Potential early onset somatic variants, discussed in the previous chapter, were identified by applying a germline variant caller upon different tissues of a single individual, followed by genotype comparisons. A limitation of such approach was the impossibility to unveil all the late onset somatic variants, that cannot be called as germinal events as their low allele frequencies would have impeded their detection. Therefore, I next applied a different variant caller algorithm, known as MUTECT2, upon the same Brazilian next-generation sequencing data. This tool is specifically designed to detect somatic SNVs with low allele frequencies from a “tumor” sample, by comparing it to a “normal” one from the same individual. I selected each individual’s KID as normal tissue and I called somatic SNVs for all the brain tissues available. By applying several stringent filtering steps, I was able to identify 15,646 total SNVs. Surprisingly, about 40% of them were in phase with another SNV locus at less than 10bps, which define them as multi-nucleotide variants (MNVs). To my knowledge this is one of the first identification of somatic MNVs, and the first one to came from brain tissues. Although MNVs were not tracked down to particular brain tissues and to the AD pathology, they were enriched in repetitive element annotations, and, more interesting, within Alu’s Pol III Box A and Box B. Finally, as candidate source for MNVs origin, I provided evidences that the observed pattern of nucleotide substitution is in agreement with a potential APOBEC involvement while excluding both polymerase zeta activity and CpG island modifications.

4.2 Materials and Methods

4.2.1 Somatic Variant calling

Somatic single nucleotide variants were called by comparing single brain WGS samples with respect to the relative kidney using the pair mode of *GATK 4 MUTECT2* (version 4.1.7.0) [Cibulskis et al., 2013]. The required panel of normals (PoN), used to filter our germinal contaminants, was previously generated according to GATK4 guidelines, by pooling together all the kidney samples available. Since sample 1345_KID was previously found to not be consistent with the other 1345 brain tissues, we discarded it from all the SNVs analyses. Raw VCF files with putative somatic SNVs were next filtered with the GATK 4 *FilterMutectCalls* command. Subsequently, resulting calls without the PASS value, calls with DP score lower than 100 and with reads showing to support the alternative alleles in the kidney were discarded. Pileup information used to assess reads support were retrieved with the *bam-readcounts* tool (version 0.8.0) [genome/bam-readcount, 2020] Final variants were called after a manual curation that was performed by directly observing mapping data, through the Integrative Genomic Viewer (IGV, version 2.8.0) [Robinson et al., 2011].

MAC (version 1.2) [Wei et al., 2015] was applied to identify the presence of mis-annotated MNVs from the final set of SNVs with a maximum search distance imposed to 10bps.

4.2.2 Variants characterization

4.2.2.1 Variants annotations

Annotations of somatic variants were performed using the ENSEMBL variant effect predictor (VEP, version 99.2) [McLaren et al., 2016], while pathogenic scores were estimated for both SNVs and MNVs with the Combined Annotation Dependent Depletion tool (*CADD)* (version 1.6) [Rentzsch et al., 2019]. Variants were next displayed with the *genVisR* (version 1.20.0) [Skidmore et al., 2016] R package. Genomic features were annotated using the annotatR R package (version 1.14.0) [Cavalcante

and Sartor, 2017] while repeats were annotated using the RepeatMasker UCSC track (version of 2020-02-20) [Smit et al., 2013-2015].

Statistical analyses on feature distributions were performed with Z scores methods. Random distribution was generated by randomly selecting the same amount of repeats from the repeat masker annotation 100 times, followed by the evaluation of variants overlaps counts.

4.2.2.2 Nucleotide substitutions analyses

The analysis of SNVs nucleotide substitutions were conducted with *the maftools* R package (version 2.4.05) [Mayakonda et al., 2018]. MNVs analyses were performed with a custom perl script that grouped variants according to the repeat annotation in overlap. Classes were: MNVs in Alu, MNVs in L1s, MNVs in other repeats and MNVs not in overlap with repeats. Assignation of 5' and 3' variants were made according to the forward strand of the reference genome, or when available, according on the plus strand orientation of the overlapping repeat element. Counts for MNVs in repeats were next normalized using the sum of all the possible combinations of random MNVs that can be generated within the observed elements, with a maximum of 10 bps distance.

4.2.2.3 Signature analyses

Mutational signature analysis was performed using the set of somatic SNVs. The *Helmsmann* software (version 1.4.2) [Carlson et al., 2018] was applied for matrix generation (parameters --length 3, –decomp nfm), while the *DesonstructSign* R package (version 1.9.0) [Rosenthal et al., 2016] was used for the signature assignation. COSMIC signatures database (version 3) was selected as reference repository of signatures (http://cancer.sanger.ac.uk/cosmic/signatures). Prior to assignation, matrix normalization was performed with the number of times that each trinucleotide context was observed in the genome (option tri.counts.method = 'genome').

4.2.2.4 Assessment of MNVs generation

Pileup analyses required for the assessment of alternative allele frequencies were performed with the *bam-readcount* software (version 0.8.0) [genome/bam-readcount, 2020]. Frequency of alternative alleles was then evaluated and ratios between pairs of variants within MNVs estimated. Statistical Z score method was used to evaluate ratios significance.

4.2.2.5 CpG island analyses

CpG islands coordinates were obtained from the UCSC track table (version of 2020-02-20) [Smit et al., 2013-2015] while L1 putative CpG island was identified with the *EMBOSS Cpgplot* tool (version 6.6.0.0) [Rice et al., 2000] from the L1.3 consensus sequence (GenBank accession: L19088.1).

4.2.2.6 MEME analyses

MNVs coordinates were extended in both directions by 10 nucleotides. Genomic sequences were next obtained with the sam*tools faidx* command (version 1.9). MEME from MEME Suit (version 5.1.1) [Bailey et al., 2009] was then applied (options: -dna, -nostatus, -mod zoops, -minw 4, -maxw 10 , -objfun classic, -revcomp, -markov_order 0). Finally, the most probable sequences of each meme was submitted to the *FootPrintDNA* online database (version of 23/01/2020) [Sebastian and Contreras-Moreira, 2014].

4.2.2.7 Alu box analyses

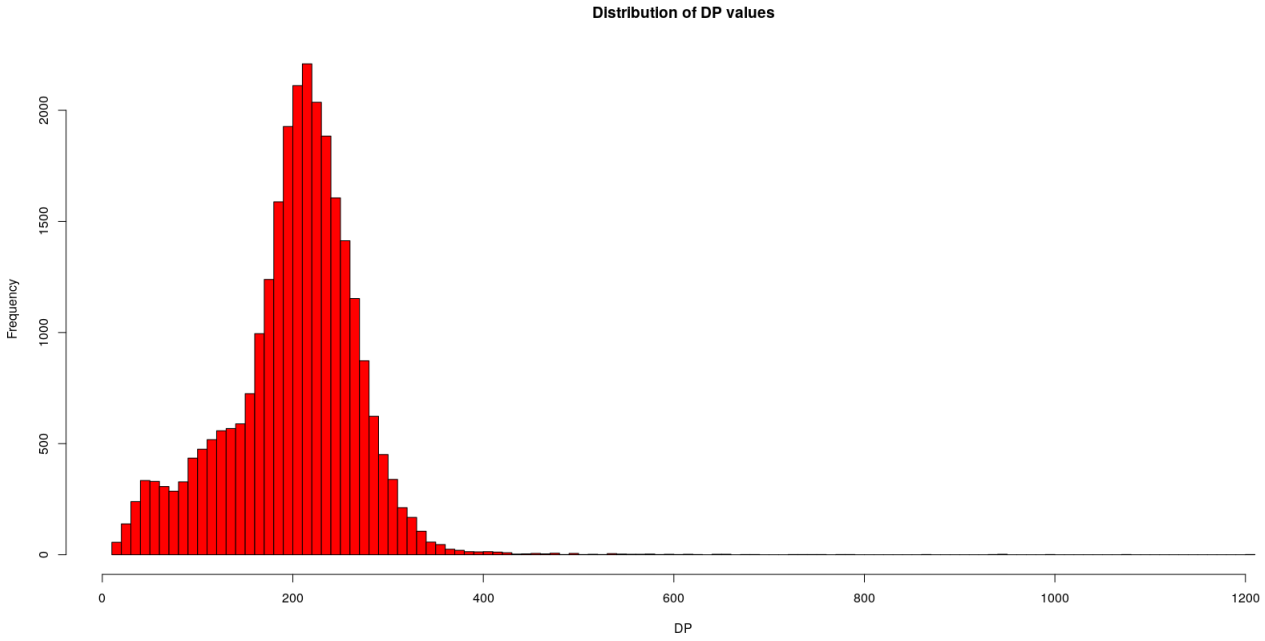
Consensus sequences of Box A (TGGCTCACGCC) and Box B (GWTCGAGAC) of Alu elements [Conti et al., 2015] were searched in the genome with the *EMBOSS fuzznuc* tool (6.6.0.0) [Rice et al., 2000]. Genomic occurrences in overlap with Alu annotations were intersected with the set of MNVs with the *bedtools intersect* command (version 2.28.0) [Quinlan and Hall, 2010]. Enrichment was statistically tested with Z score method. The random distribution was generated by shuffling 100x each box sequence followed by the evaluation of the overlaps with Alu's MNVs.

4.3 Results

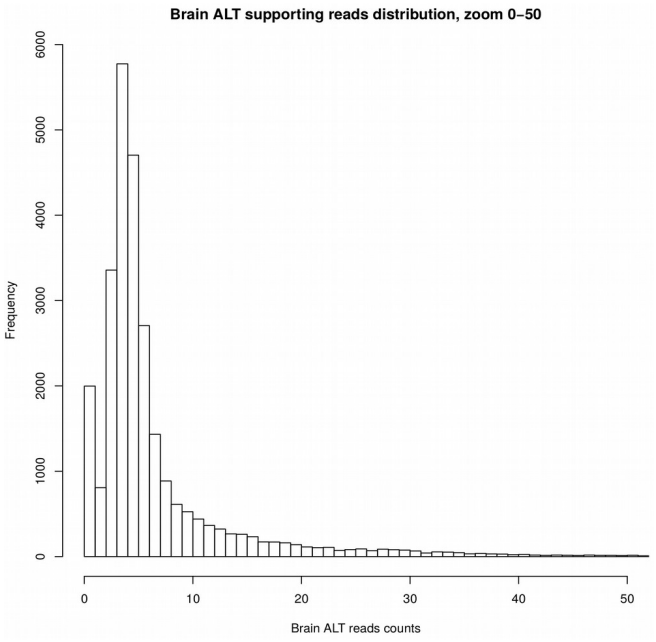
4.3.1 Identification of SNVs and MNVs from WGS data

To unveil the presence of low frequency somatic variants, defined as “late onset” SNVs, within the brain samples I performed a MUTECT2 analyses. By comparing single brain tissues (called also as “tumor” tissue) with respect to the kidney (called as “normal” tissue) of the same individual, I identified a total of 27,106 raw single nucleotides variants (SNVs). To test the correctness of the calls I started by observing the distribution of the reads depth (DP) scores and the raw number of reads that supported the alternative alleles, for both brain tissues and kidney (figure 34). From the distributions of the DP scores I found a peak centered at value ~200, which was coherent with the definition of DP score from MUTECT2 (*i.e* the sum of reads depth from “tumor” and “normal” tissues at a single locus) considering an average sample coverage of~100x (figure 34-A). Next, I focused on the distributions of the alternative alleles supporting reads. Within brain, I noted that alternative alleles were supported by a mean of 4 reads (figure 34-B), indicating an extremely low alternative allele frequency (~ 4%). In kidney instead, I observed that the vast majority of SNVs loci were supported by 0 reads, possibly hinting to true positive calls (figure 34-B). However, I also identified SNVs loci with alternative alleles supported by more than 1 reads which may represent instead false positive calls (figure 35-A). Therefore, to reduce the amount of potential false positive calls, conservative filters were imposed, and in particular only SNVs with DP >= 100 and without sign of support in the kidney were kept (16,139 / 27,106 SNVs). Furthermore, to remove potential germline contaminations and/or “early onset” SNVs, SNVs that were in overlap with the results of a GATK 4 Haplotypecaller analyses were also removed. As a result of these filters, I identified a total of 15,646 high-confidence SNVs. A manual inspection was then performed upon some random loci (figure 35-B), which often evidenced couples of SNVs in close proximity to each others (figure 35-C).

A)



B)



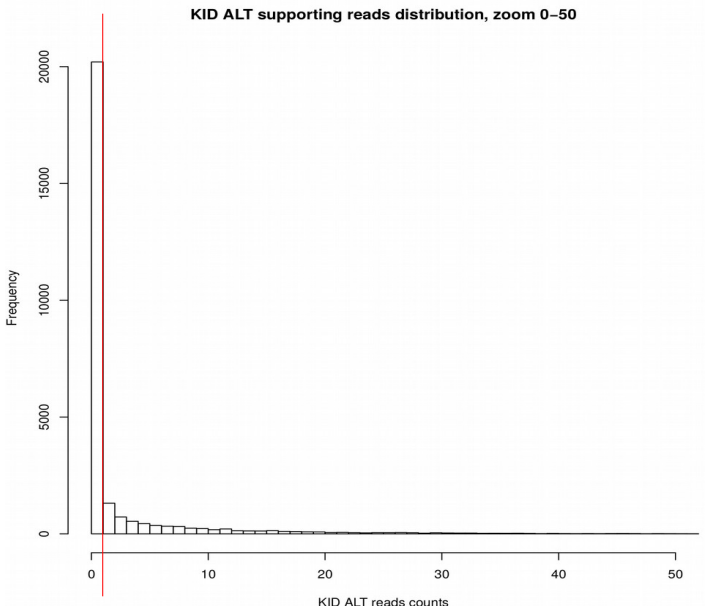
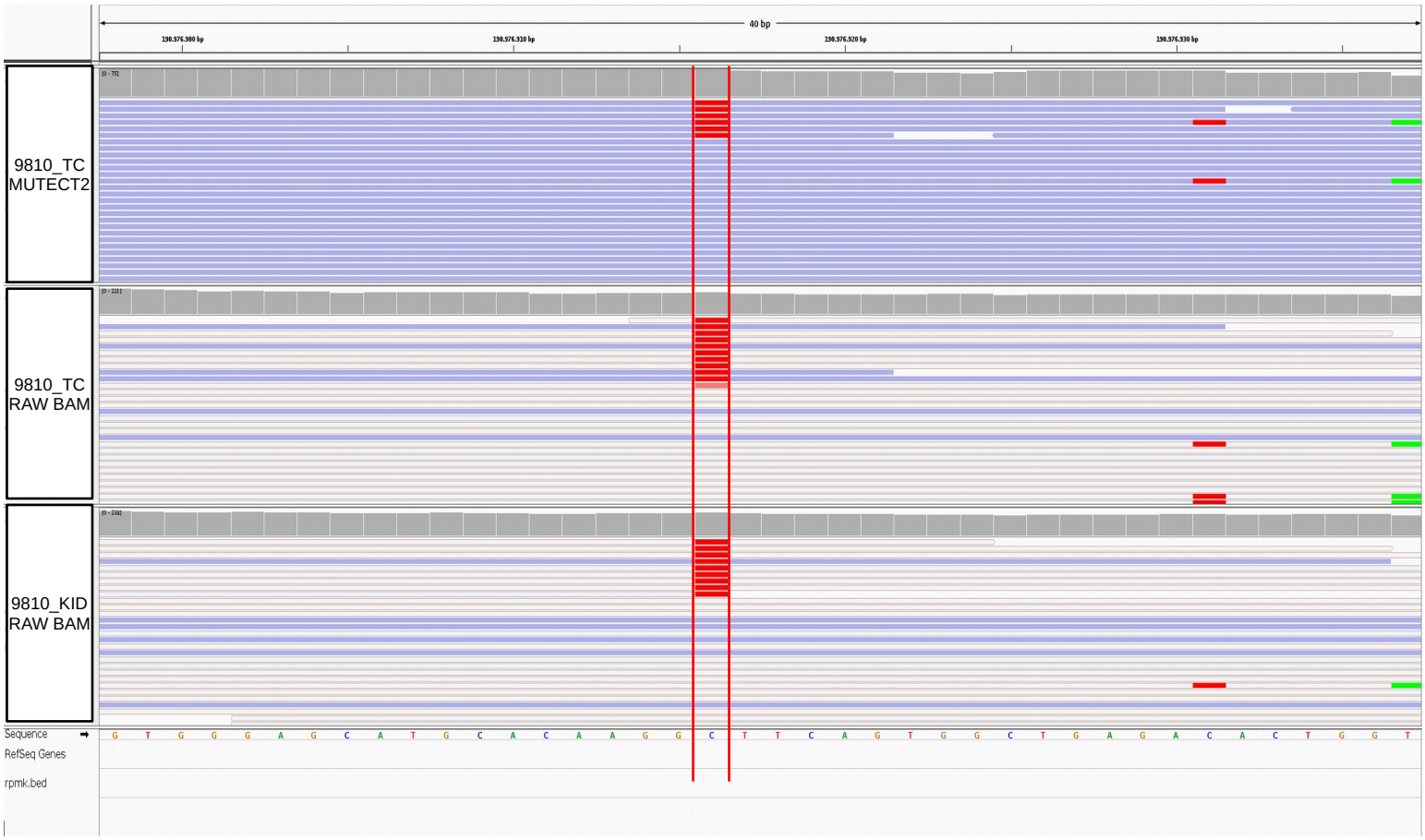
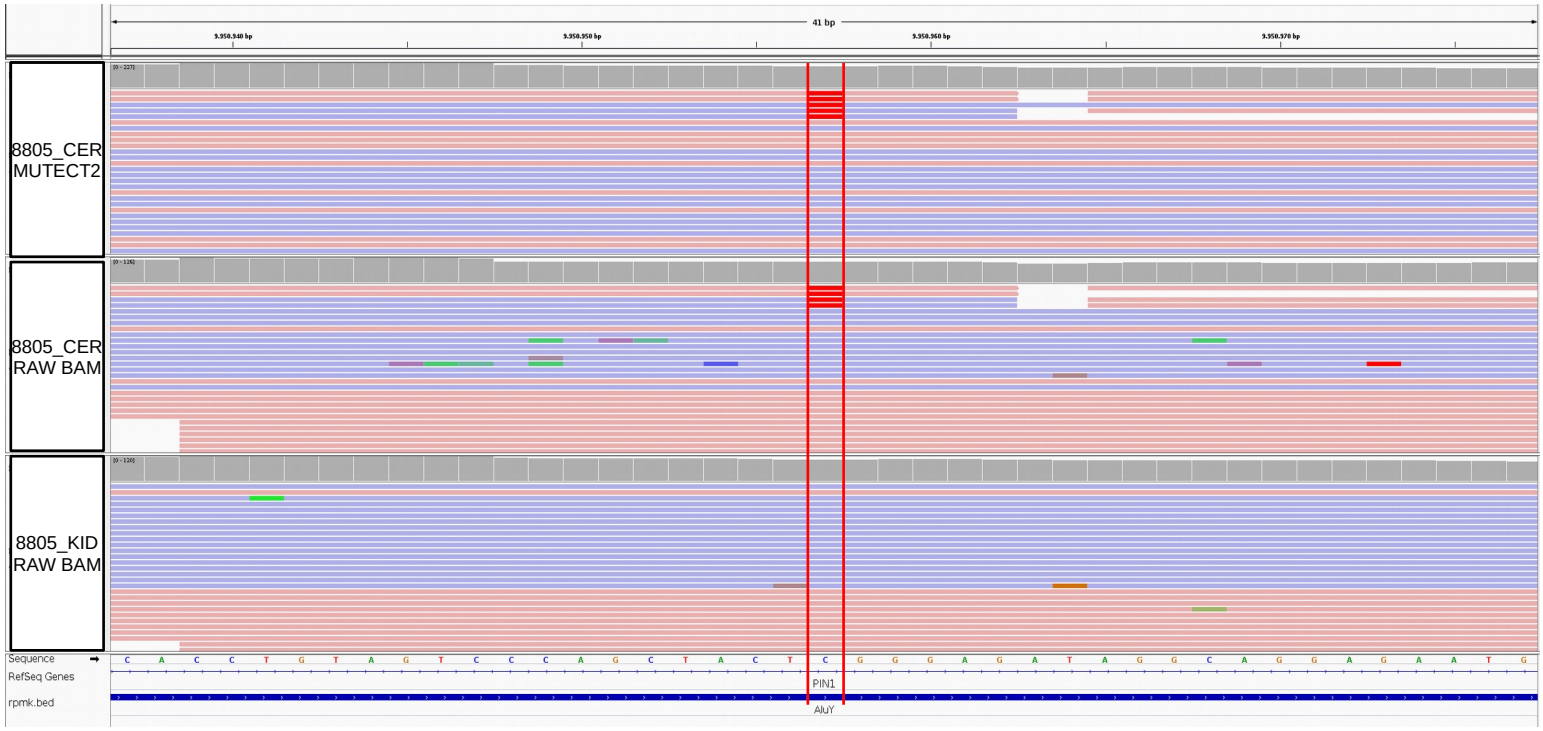


Figure 34: A) Distribution of SNVs DP scores. A peak at value ~200 was expected since DP is the result of the sum of both "tumor" tissue and "normal" tissue read depth; **B)** Distributions of Alternative (ALT) allele supporting reads for both brain tissues and kidney. The red vertical line indicates the reads threshold imposed.

A)


40 bp
190.576.900 bp
190.576.910 bp
190.576.920 bp
190.576.930 bp
9810_TC MUTECT2
9810_TC RAW BAM
9810_KID RAW BAM
Sequence
RefSeq Genes
rpmk.bed

B)


41 bp
9.950.940 bp
9.950.950 bp
9.950.960 bp
9.950.970 bp
8805_CER MUTECT2
8805_CER RAW BAM
8805_KID RAW BAM
Sequence
RefSeq Genes
rpmk.bed
PIN1
AluY

C)

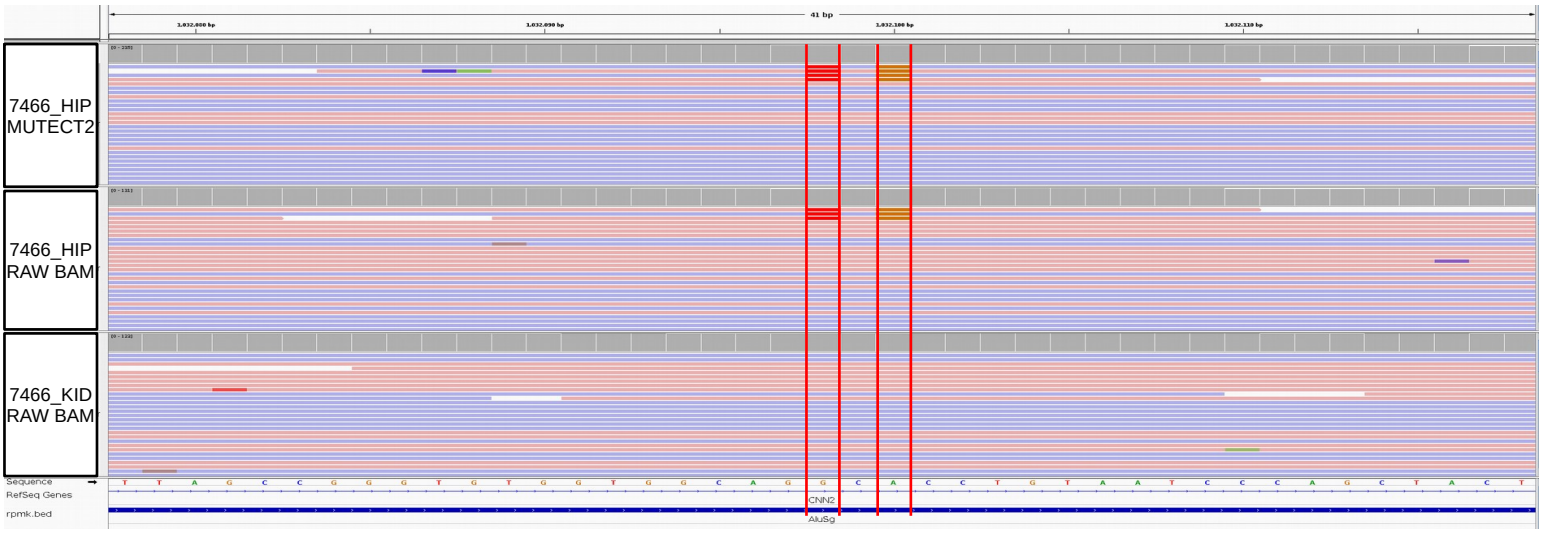
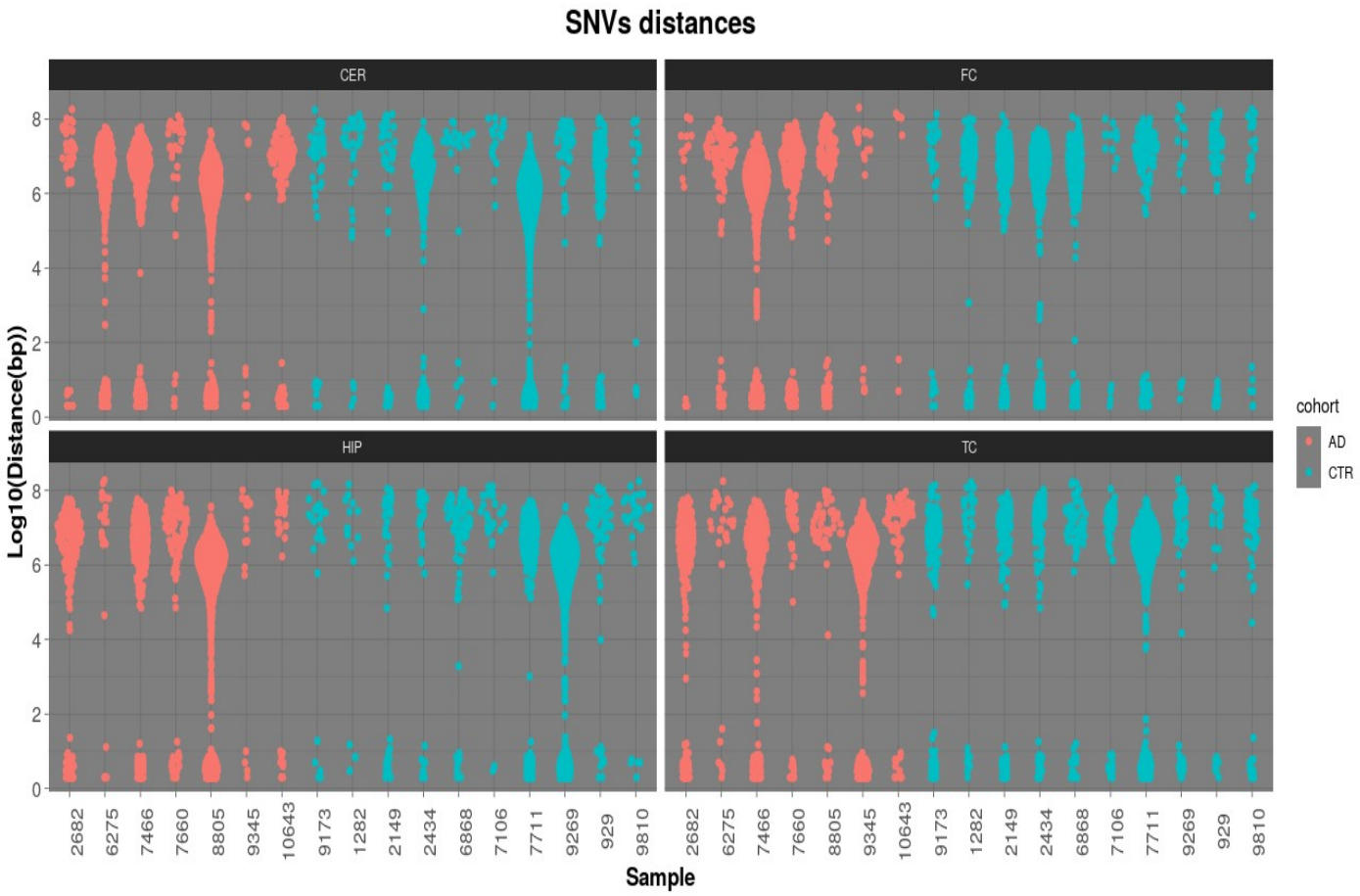


Figure 35: IGV screen-shots of MUTECT2 results. The screen-shots are composed by three tracks. From top to bottom: MUTECT2 realigned reads; raw alignment data of brain tissues; raw alignment data of kidney. **A)** IGV screen-shot of a potential false positive call from sample 9810_TC at chr4:190976916. Alternative allele was found to be supported also in KID; **B)** IGV screen-shot of a putative late onset somatic SNV found in sample 8805_CER at chr19:9950957; **C)** IGV screen-shot of putative late onset somatic SNVs found in sample 7466_HIP, at loci chr19:1032098 & chr19:1032100. SNVs were highlighted within red vertical lines.

Therefore, I proceed by inspecting the SNVs distances, observing a bimodal distribution (figure 36-A). A more precise investigation led me determinate that about 50% of SNVs per sample showed distances below 10 bps (figure 36-B). From the manual inspections, I also noted that close SNVs were also supported by the same read (figure 35-C), indicating that they were within the same haplotype phase. SNVs located within the same haplotype and in close proximity to each other, are defined as multi-nucleotide variants (MNVs).

It was then noticed that GATK is not able to distinguish between MNVs and SNVs. Therefore, I used the *MAC* software to correctly classify MNVs within the set of high-confidence SNVs. Notably, as a result, I observed 3,064 MNVs with a maximum distance of 10 bps and 9,518 SNVs.

A)



B)

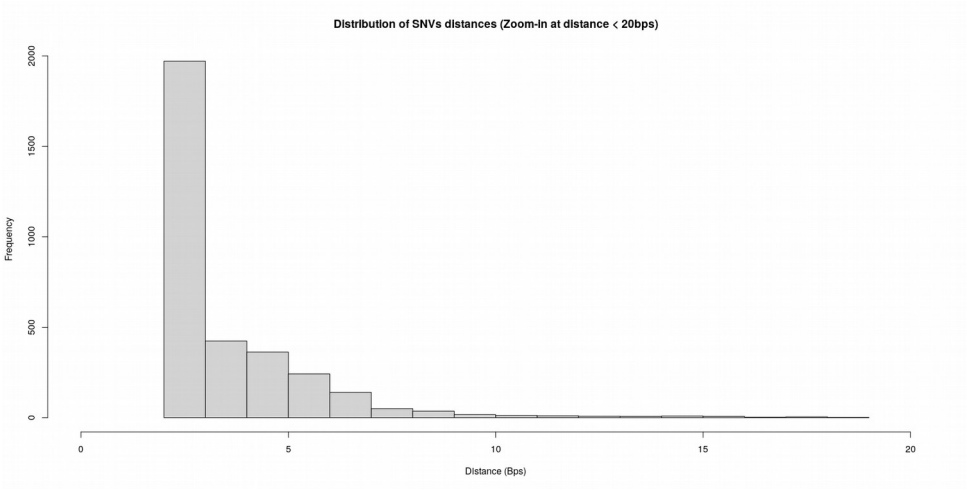


Figure 36: Late onset SNVs distances. **A)** Distribution of late onset SNVs distances for single samples; **B)** Zoom-in of SNVs distances distribution. A peak at distance equal to 2 bps is show.

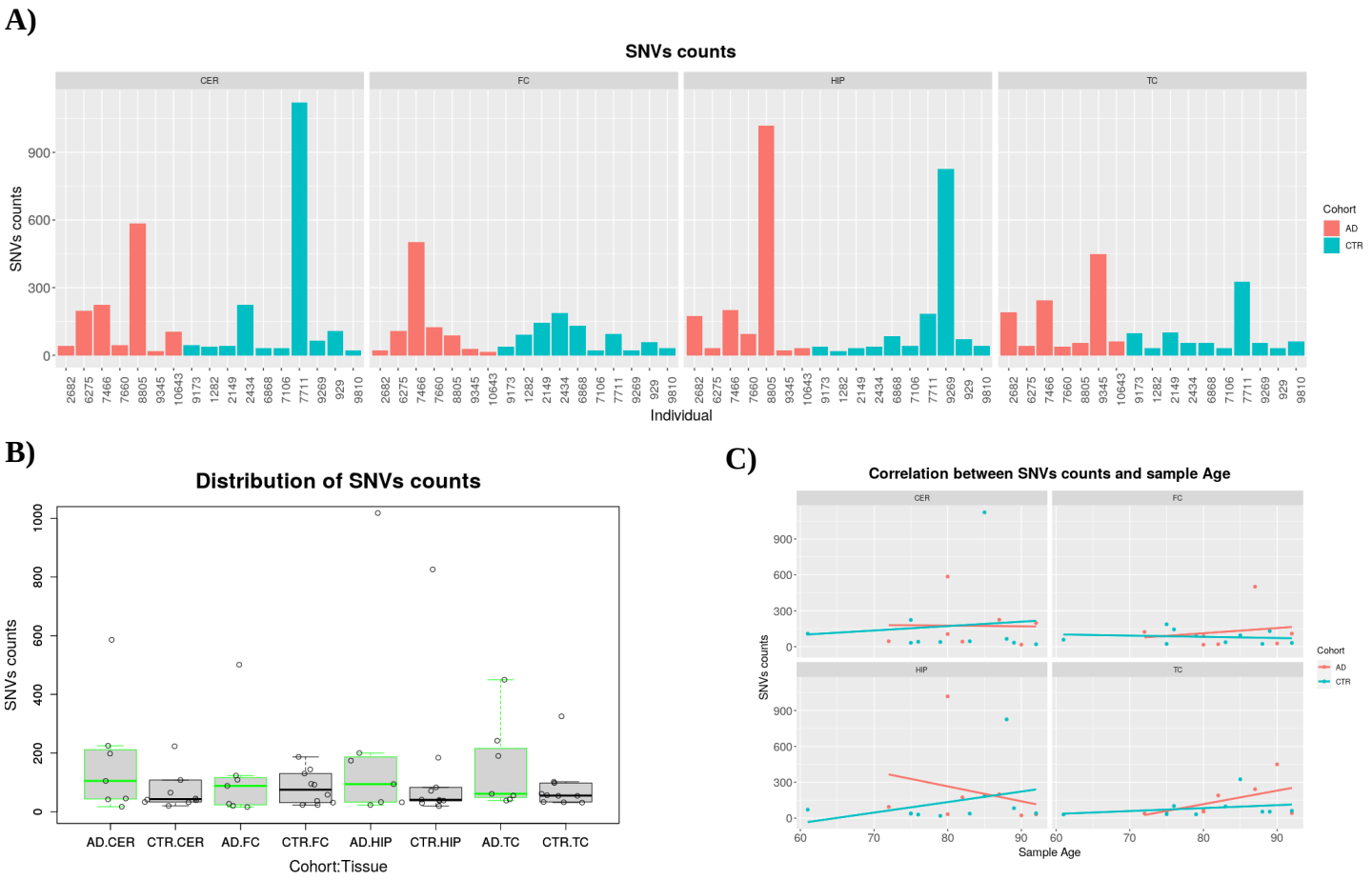
4.3.2 Single nucleotide somatic variants exploration

After the *MAC* analyses, I counted 2,947, 1,704, 2,943 and 1,924 late onset somatic SNVs, for a total of 9,518 SNVs, from CER, FC, HIP and TC tissues, respectively (figure 37-A). No differences were observed when sample counts were grouped according to their respective cohort (figure 37-B). However, it was noted that particular samples displayed extreme high counts of SNVs (> 300) that deviated from the mean value of ~140 SNVs per sample. It is known that the rate of somatic SNVs accumulation in brain strictly increase during age. Therefore, I investigated the correlation between SNVs counts and the age of our samples (figure 37-C). No significant correlations were found, however it cannot be excluded that this result was due to the limited sample size. Nonetheless, high-SNVs-counts outliers samples had generally more than 80 years old. I next determined that SNVs were mostly characterized by C>T (3,493/9,518) and T>C (2,888/9,518) nucleotide substitutions (figure 37-D). Considering all the possible types of nucleotide substitutions, transitions made up about 67% of the total set of SNVs while transversions represented only the 33%, on average.

I next proceeded by performing functional annotations, which evidenced that the vast majority of SNVs felt within intergenic (4,465 SNVs) and intronic regions (4,273 SNVs) (figure 38-A). Despite the low abundance of somatic SNVs within coding regions, the presence of potential pathogenic variants was tested by performing Combined Annotation-Dependent Depletion analyses (CADD analyses). From these, 12 potentially pathogenic SNVs (defined as having CADD score >= 20) were found, that, however, were not associated to the aetiology of the Alzheimer's disease and that were identified from both AD and CTR samples. Potential pathogenic SNVs were found to be related with transcription factors (VAX2, PHF3, NRF1), a transcriptional regulator (ZFAT), a host factor (HCFC1), enzymes (TDO2, PIP5K1C, GAL3ST3, ENDOU), a RNA binding proteins (NUFIP1) and protein coding genes (TTN, FRAS1) (figure 38-B).

I further annotated SNVs with repeats annotations, founding that 7,971 SNVs (83.75%) were in overlap with a mobile element sequence, with Alu and L1s as the most represented class of elements (5,474 and 1,190 SNVs, respectively).

Finally, I sought to decipher the molecular processes behind SNVs origin by performing signature analyses. To increase the strength of the analysis, and to possibly highlight signatures private of the AD cohort, SNVs were pooled together according to their cohort metadata (4,754 SNVs for AD and 4,764 SNVs for CTRs). From the analyses, I identified five signatures: SBS1, SBS6, SBS10b, SBS12 and SBS39 from both AD and CTRs, suggesting that no differences in molecular processes can be associated specifically to only one of the two cohorts (figure 39).



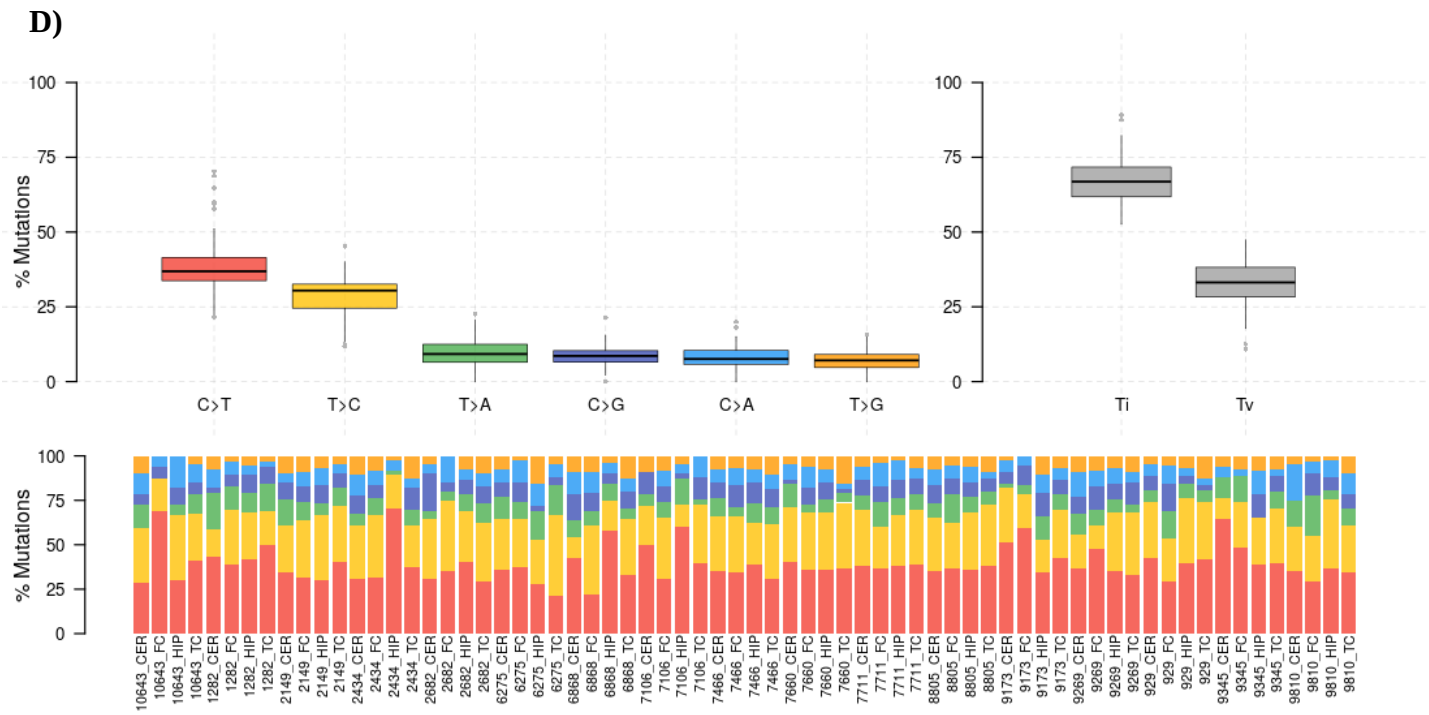
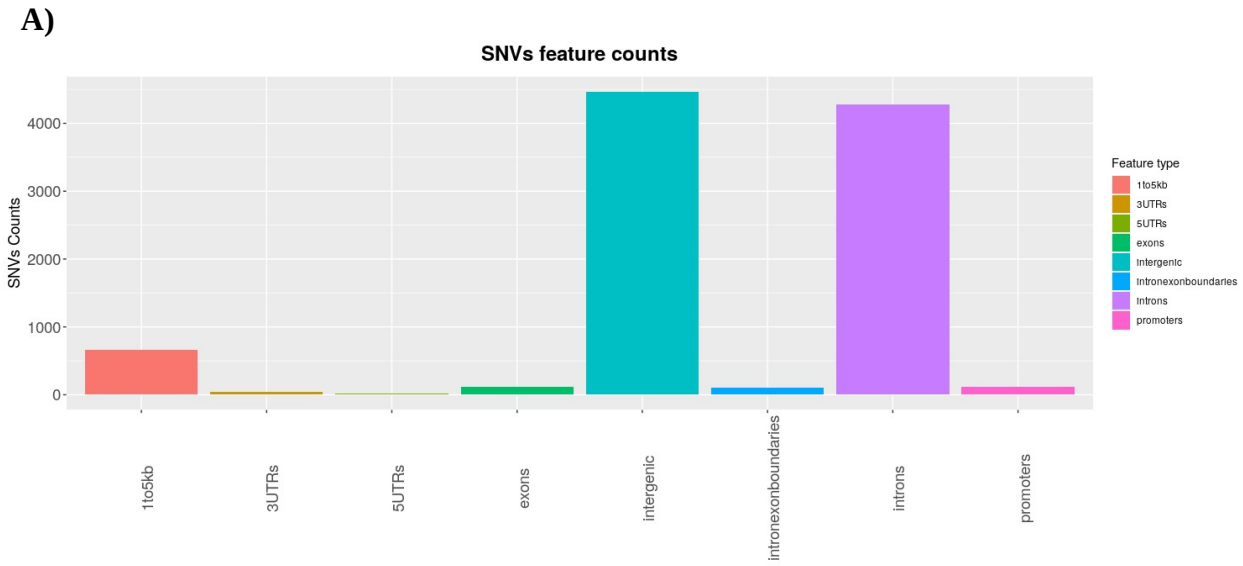


Figure 37: Late onset SNVs exploratory analyses. **A)** Late onset SNVs counts per sample; **B)** Distribution of late onset SNVs counts grouped by cohort and tissue. AD in green and CTRs in black; **C)** Correlation between sample's ages and late onset SNVs counts. Linear models are shown as colored lines; **D)** Late onset SNVs nucleotide substitutions analyses. Top left plot: Percentage of mutations per class of nucleotide substitution; Top right plot: Percentage of Transitions (Ti) and Tranversions (Ts); Bottom plot: Percentage of mutations per class of nucleotide substitutions displayed for each sample.



B)

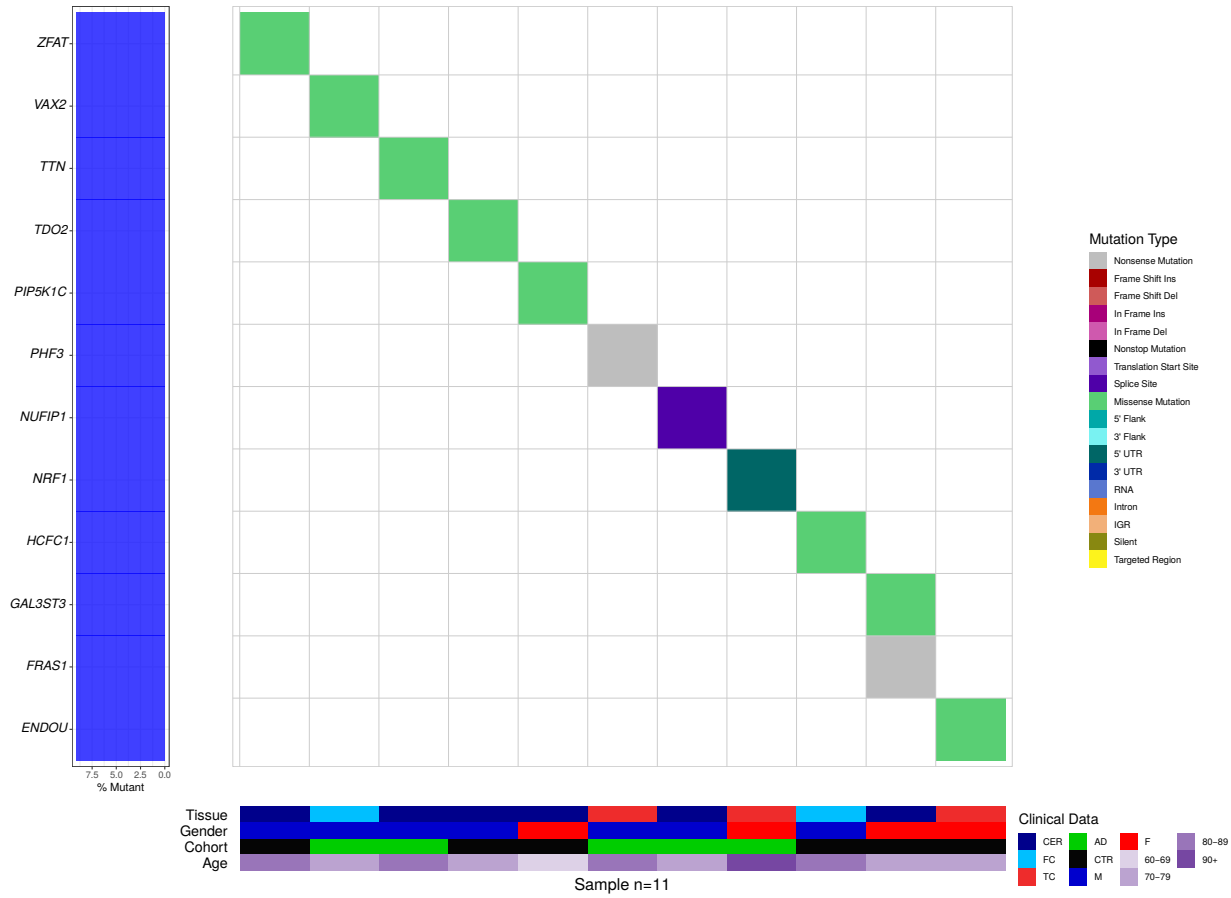


Figure 38: Features and CADD analyses. **A)** Features in overlap with late onset SNVs loci; **B)** Waterfall plot for potential pathogenic (CADD >= 20) late onset SNVs. Left plot: Gene feature and frequency of samples affected (calculated from a total of 11 samples); Center plot: Type of mutation; Bottom scheme: Clinical data. TDO2, PHF3, HCFC1 genes are of particular interest. TDO2 gene encodes for a heme enzyme that plays a critical role in tryptophan metabolism. Single nucleotide polymorphisms in this gene may be associated with autism [Nabi et al., 2004]. PHF3 gene is supposed to encode for a transcription factor and to be related with the development of glioblastoma [Fischer et al., 2001]. HCFC1 gene is a member of the host cell factor family. Interestingly, the protein encoded by this gene is involved in cell cycle control and transcriptional regulation during herpes simplex virus infection [Vogel and Kristie, 2013].

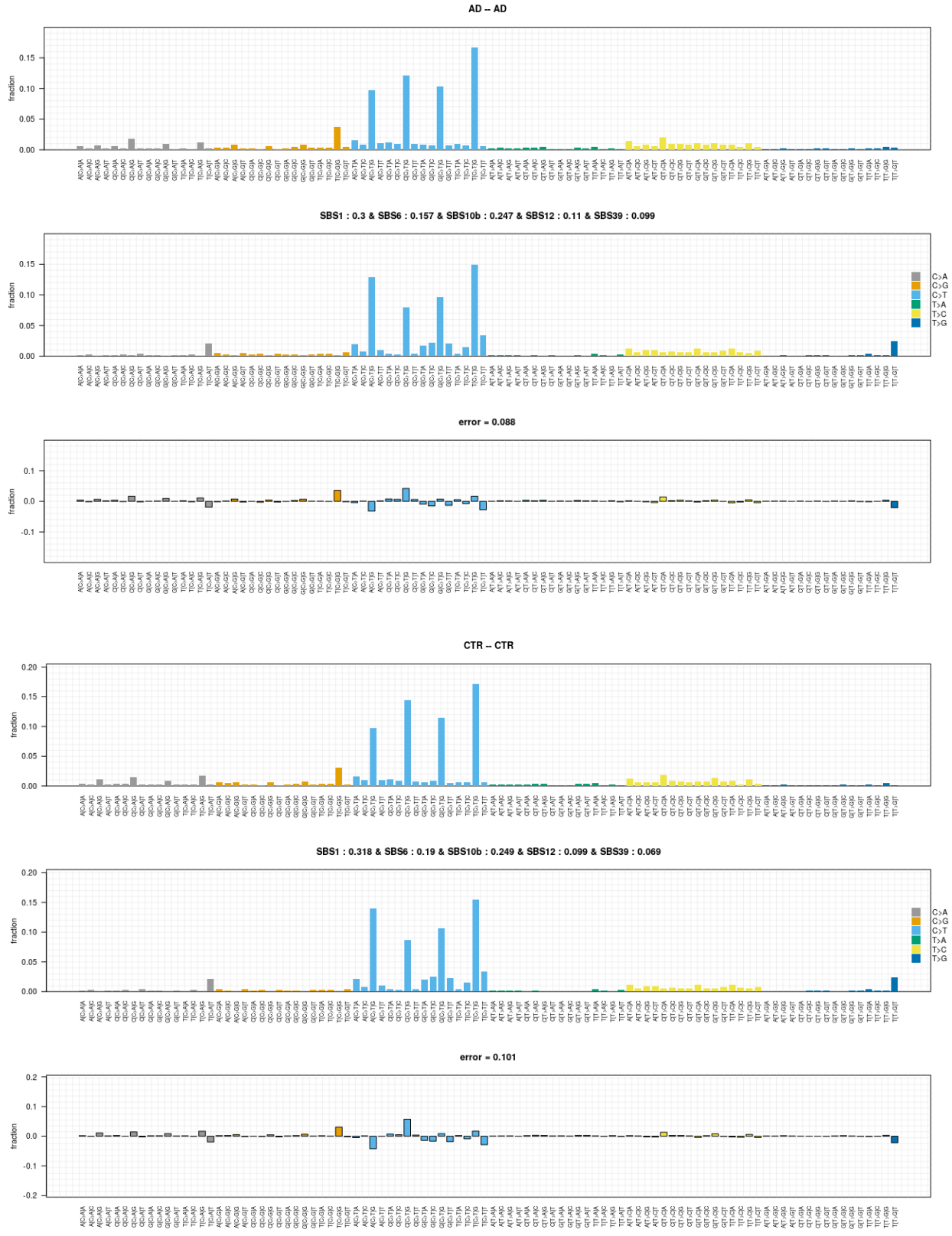


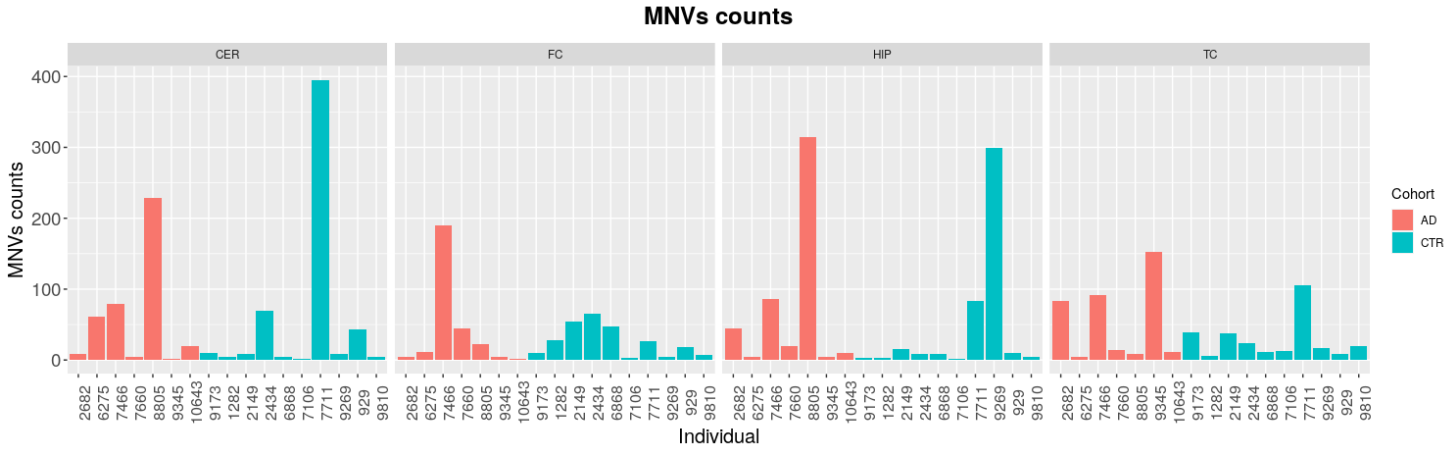
Figure 39: Signature analyses for AD and CTR late onset SNVs.

4.3.3 Multi-nucleotide somatic variants

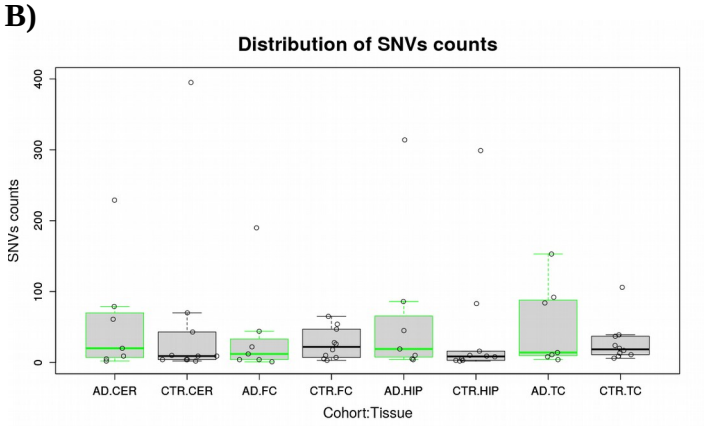
4.3.3.1 MNVs exploration

As a major result of the MAC annotations, I observed 6,128 SNVs forming 3,064 MNVs, which represented a consistent fraction of the overall somatic variants identified by MUTECT2 (6,128 out of 15,646 SNVs, ~39%). In particular, were annotated 955, 648, 921 and 540 somatic MNVs from CER, FC, HIP and TC, respectively (figure 40-A). As for SNVs, it was found that MNVs were not enriched in particular tissues nor cohorts (figure 40-B). Additionally, it was also observed that the same 7 samples that shown increased levels in SNVs, potentially due to their older age, presented also higher counts of MNVs (> 100) (figure 40-C). To understand whether variants in a MNVs were generated simultaneously or by multiple consecutive events, I next investigated the alternative allele frequencies of both variants within a MNVs. The results dictated that the vast majority of MNVs can be associated to a simultaneous origin, since a little number of MNVs (53 out 3064) showed significant differences in the alternative allele frequencies (figures 40-D and 40-E). Analyses of distances then, evidences a minimum value of 2 bps, indicating that no consecutive variants were present within the set of MNVs. Distance of 2 bps was also the most prevalent score (1,192 MNVs, 38.90%) (figure 40-F), which could lead to severe codon alteration that would have been undetected without a MNV-annotation step.

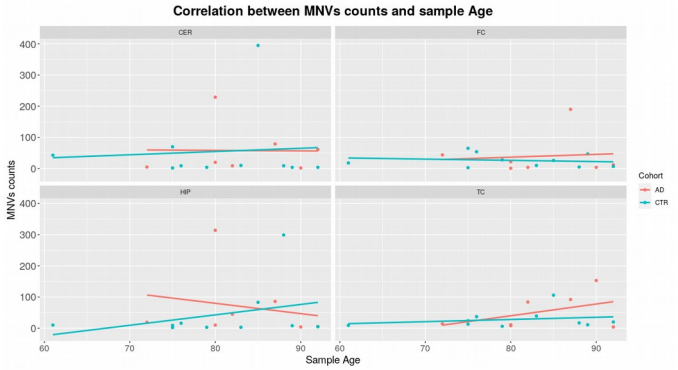
A)


MNVs counts
CER
FC
HIP
TC
MNVs counts
Individual
Cohort
AD
CTR

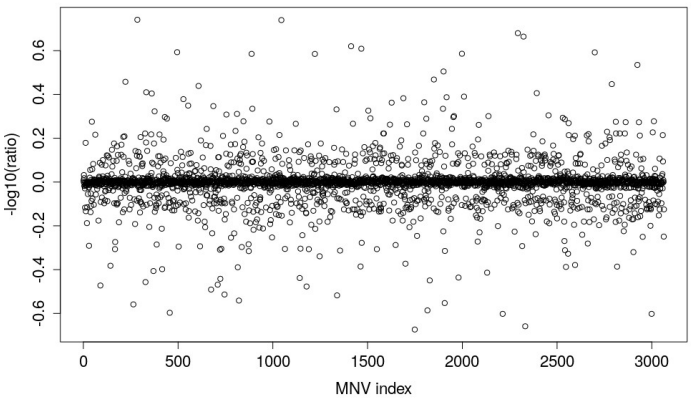
B)


Distribution of SNVs counts
SNVs counts
AD.CER
CTR.CER
AD.FC
CTR.FC
AD.HIP
CTR.HIP
AD.TC
CTR.TC
Cohort:Tissue

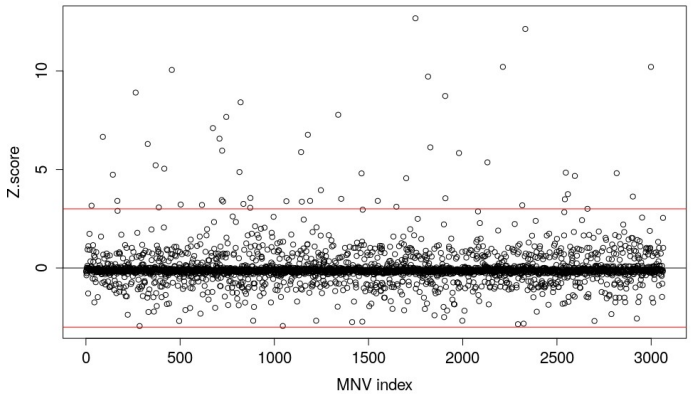
C)


Correlation between MNVs counts and sample Age
CER
FC
HIP
TC
MNVs counts
Sample Age
Cohort
AD
CTR

D)


ALT allele freq ratio (SNV1 / SNV2)
-log10(ratio)
MNV index

E)


Z.score of ratios
Z.score
MNV index

F)

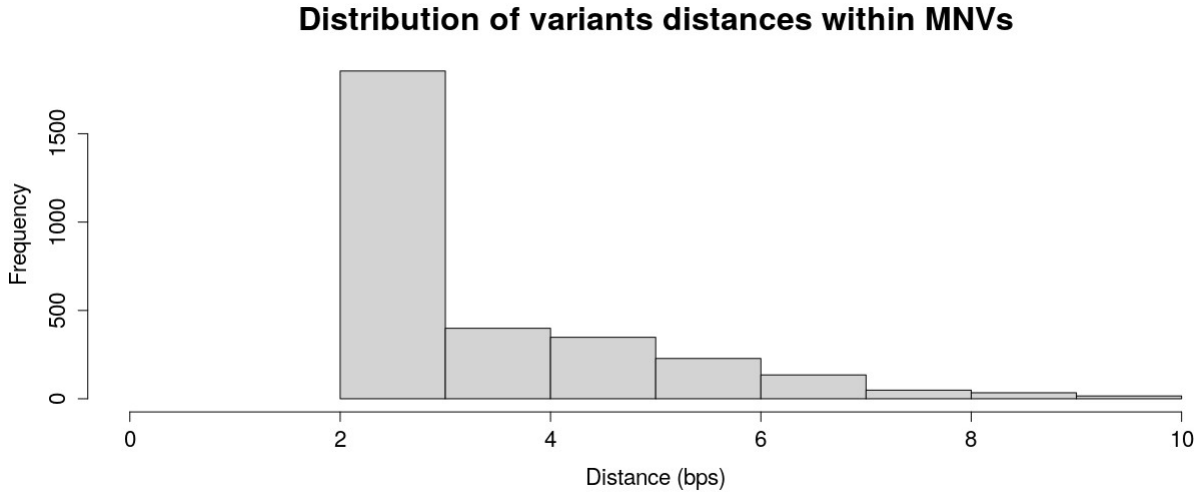
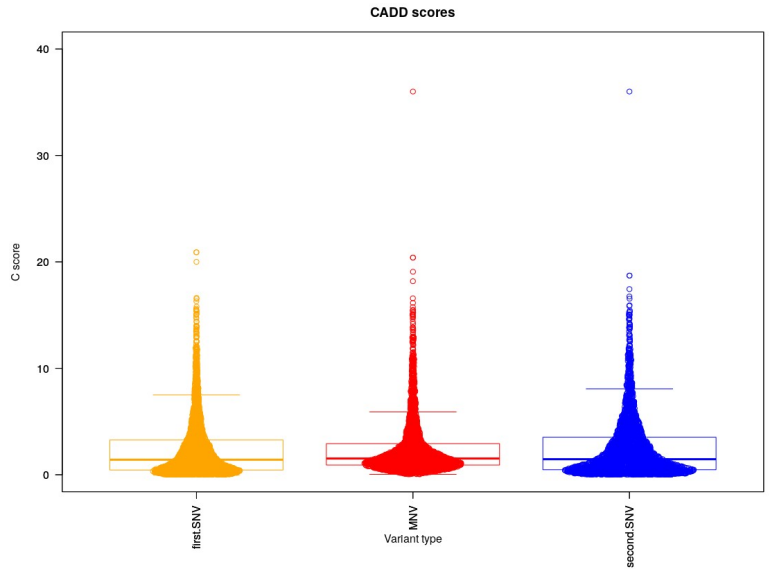


Figure 40: MNVs exploratory analyses. **A)** MNVs counts per sample; **B)** Distribution of MNVs counts grouped by cohort and tissue. AD in green and CTRs in black; **C)** Correlation between sample's ages and MNVs counts. Linear models are shown as colored lines; **D)** Alternative allele frequency ratios for SNVs within a MNVs call; **E)** Z scores for MNVs Alternative Allele ratios. Red horizontal lines were drawn at +3 and -3 standard deviation from the mean score to separate potential simultaneous from consecutive MNVs; **F)** Distribution of variants distances within a MNVs.

4.3.3.2 MNVs pathogenic scores and feature analyses

As several studies pointed out, the mis-annotation of MNVs may highly impact the functional annotation and, most important, the actual clinical effect of such class of variants. To appreciate whether the MNVs annotation step was valuable in providing potential pathogenic variants, possibly linked with the disease, I evaluated and compared the CADD scores for both MNVs and for their mis-annotated form (*i.e.* using the single SNVs that composed a MNVs). Although no appreciable increase in pathogenic variants (CADD > 20) was observed (figure 41-A), I speculated that this absence was an effect of the MNVs abundance in intronic and intergenic regions (only 21 MNVs in exonic regions) (figure 41). I next further annotated MNVs with repeats annotation, observing that 2,724 MNVs were found to be in overlap with a repeat element (88.90%). Alu was the most prevalent family of repeats (2,052 MNVs, 66.97% of MNVs), followed by L1s (422 MNVs, 13.78% of MNVs) (figure 42).

A)



B)

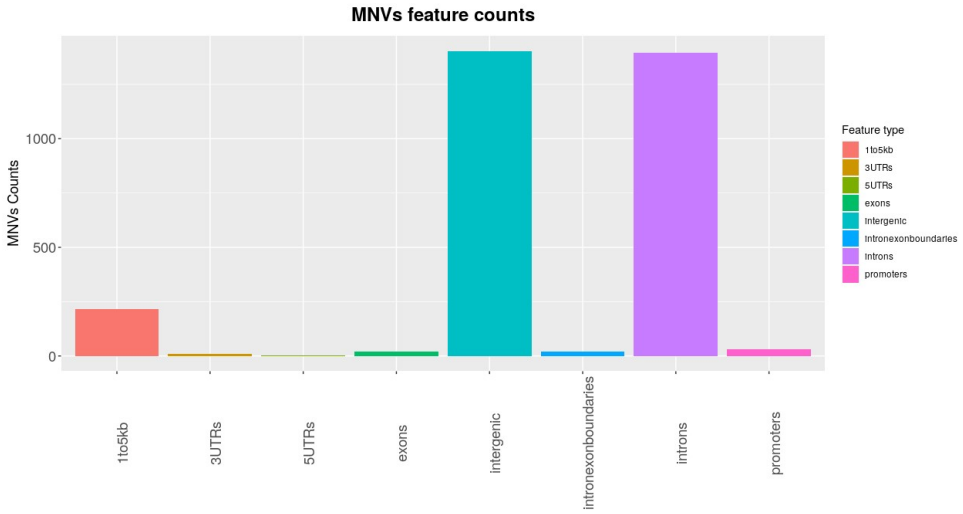
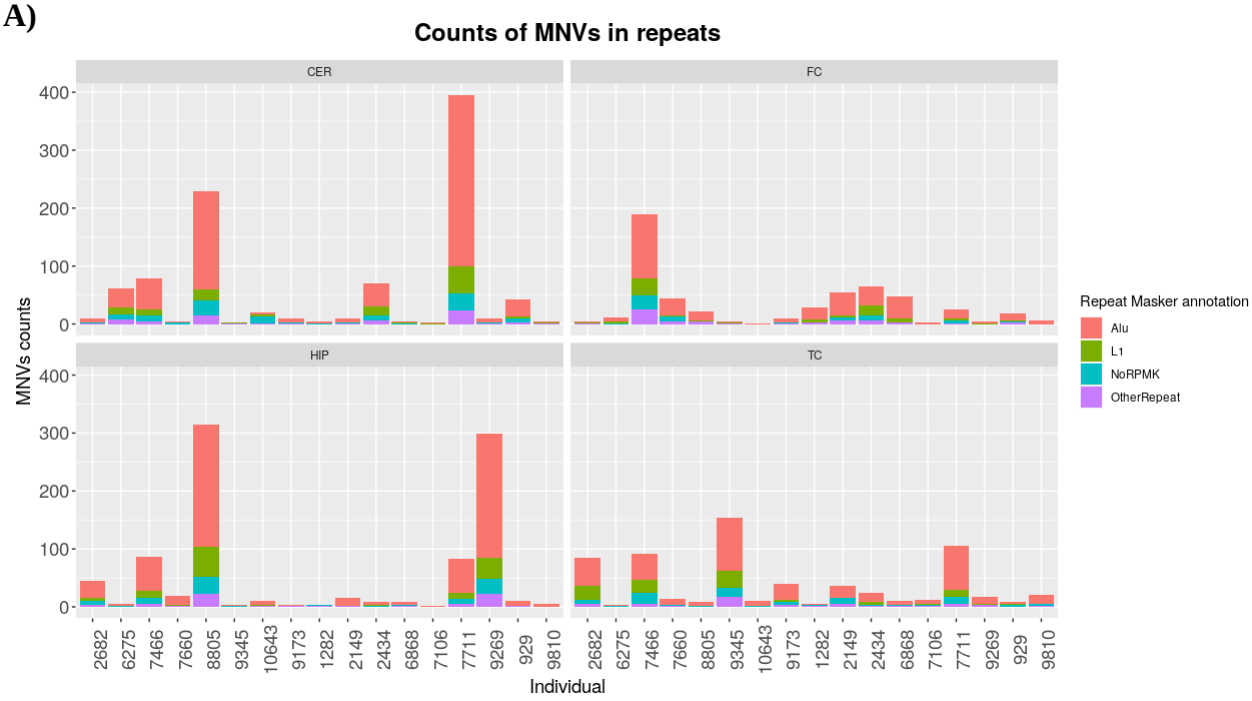


Figure 41: Features and CADD analyses. **A)** CADD scores distributions. CADD score was evaluated for both single variants from MNVs pairs and from the MNVs; Two potential pathogenic MNVs were identified. Chr18:30264933-30264938 GAGATC>AAGATT (CADD 20.4) was found from both 2682_HIP and 8805_HIP samples, within intron number 5 of the KLHL14 gene. The protein encoded by this gene is a member of the Kelch-like gene family, and associated with dystonia [Zhang et al., 2017]. Chr12:50475399-50475403 GGCCC>AGCCT (CADD 36) was found in sample 8805_CER, within exon 12 of the ASIC1 gene. This gene encodes a member of the acid-sensing ion channel (ASIC) family of proteins which functions in learning, pain transduction, touch sensation, development of memory and fear. **B)** Counts of features in overlap with MNVs**.**

A)



B)

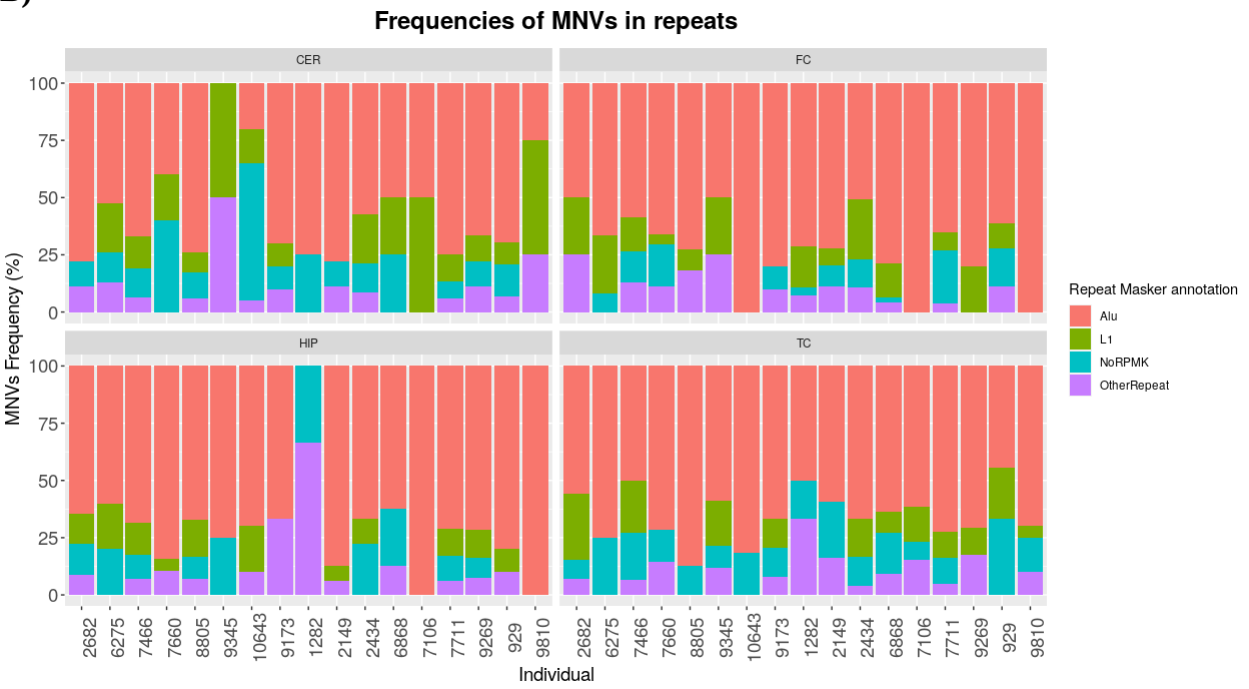


Figure 42: Analyses of repeat annotations. **A)** Counts of MNVs in overlap with repeats; **B)** Frequencies of MNVs in overlap with repeats. RPMK: Repeat Masker annotations; NoRPMK: MNVs not in overlap with RPMK annotations; OtherRepeats: MNVs in overlap with different than Alu and L1 repeats.

4.3.3.3 Investigation of MNVs origin

After having observed that MNVs were highly abundant in repeats, MNVs were classified according to a 4 group classification: MNVs in Alu, MNVs in L1s, MNVs in other repeats and MNVs not in repeats. Next, to gain information regarding MNVs origin, dedicated analyses of nucleotide substitutions were performed. MNVs nucleotides substitutions will be represented in the following form: AB>XY, in which A>X and B>Y represent the first and the second variation, respectively. For the analyses, the orientation of the paired variants (*i.e.* assigning which variant was at the 5` and which was at the 3`) was defined according to the transcriptional direction of the repeat (or to the forward strand of the reference genome). As a result, in Alu we observed several peaks at: AC>GT, CC>TT, CT>TC, TC>CT substitution classes (Figure 43-A). On the contrary, in L1s we identified a single major peak at CC>TT (figure 43-B). MNVs not in repeats displayed a pattern similar to L1s, presenting a major peak at CC>TT. However, also AC>GT classes showed to be strongly supported, as in Alus (figure 43-C). Interestingly, MNVs not in repeats did not evidenced similar peaks (figure 43-D).

CC>TT substitutions are known to be the result of both APOBEC activity and CpG islands mutations. I therefore investigated how many MNVs felt within annotated CpG islands, finding no overlaps. Additionally, I further extended CpG islands annotations by identifying a putative CpG island between bp 49 and 419 of the L1 consensus sequence. Although 14 MNVs were found to be in overlap with this putative CpG island, no appreciable changes in the nucleotide substitutions peaks were noted upon removal of these MNVs. Germinal MNVs are also speculated to be originated also by the effects of polymerase zeta. However, the absence of consecutive SNVs (MNVs with 0 bps of distance, figure 40-F) and the lack of TC>AA and GC>AA (and their reverse complement) substitution patterns, which are two distinctive markers of pol-zeta activity, are not supporting this hypothesis. This, in combination with the CC>TT peaks, led me to speculate that a direct involvement of APOBEC proteins is at the base of somatic MNVs origin in brain

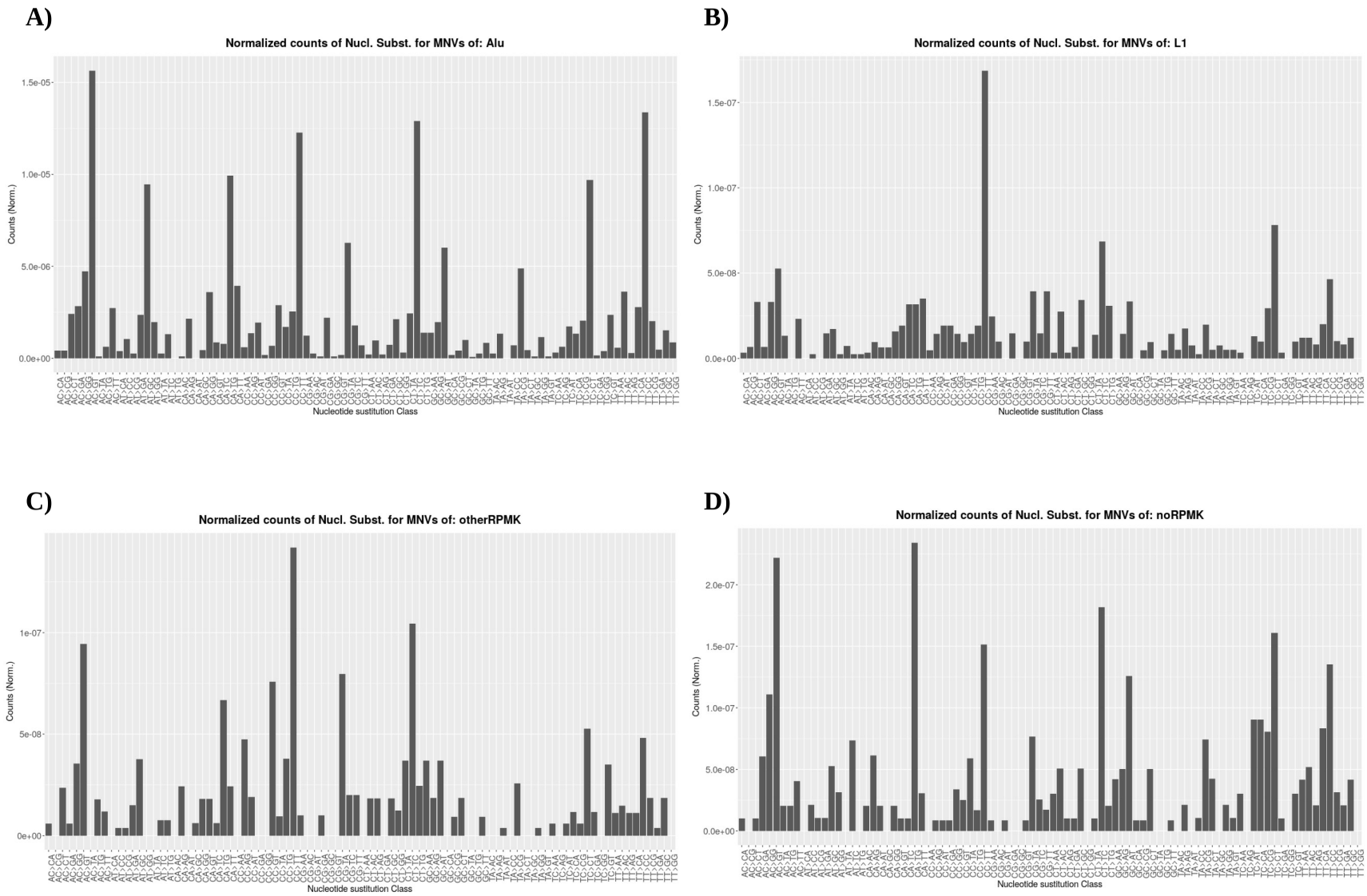
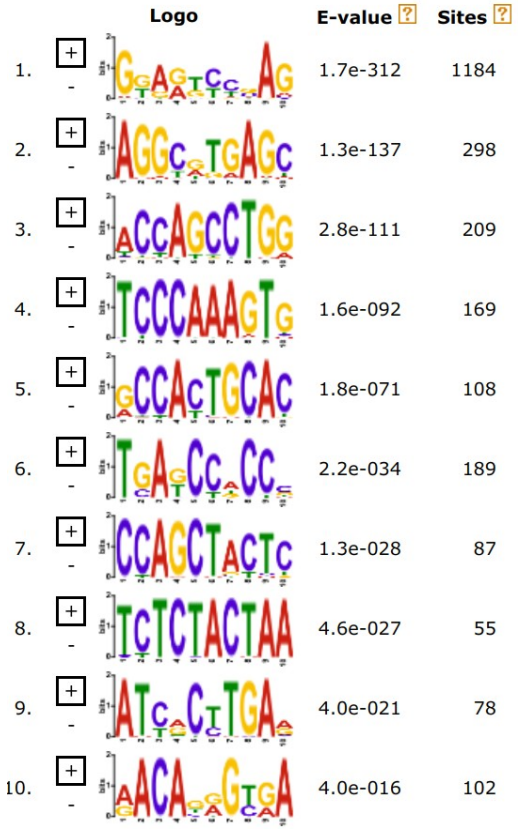


Figure 43: Nucleotide substitution analyses. Normalized counts of substitutions are displayed. **A)** MNVs in Alu; **B)** MNVs in L1; **C)** MNVs in other repeats Data; **D)** MNVs not in repeats.

Finally, I investigate the presence of consensus sequences within the genomic regions of MNVs (figure 44-A). The most supported consensus sequence identified, named as MEME-1, was further identified as a potential binding site for two Zinc fingers proteins (ZFN135 and ZFN460) linked to Herpes simplex virus type I infections (figure 44-B). This, in combination with the APOBEC-like signature observed, led me to speculate that a link between virus infections and somatic MNVs origin may be present, although further investigation are required.

A)



B)

footprintDB template	Source	Organisms	STAMP e-value	Motif similarity	footprinDB PWM / Consensus	Binding proteins	Interface sequences	Pfam domains
MA1587.1: ZNF135	JASPAR 2020	Homo sapiens	1.9e-06	7.35 / 10	---GGAGTCCGAG- yyrGGAGGTCGAGg	Show proteins	Show interfaces	Show domains
MA1596.1: ZNF460	JASPAR 2020	Homo sapiens	8.9e-06	7.23 / 10	----GGAGTCCGAG-- CyyGGGAGGCkGAGGy	Show proteins	Show interfaces	Show domains

Figure 44: MEME analyses results: **A)** List of discovered consensus sequences, coupled with numbers of supporting evidences; **B)** *FootPrintDNA* positive matches for the most probable MEME-1 sequence (GGAGTCCGAG).

4.3.3.4 MNVs enrichment in Alu boxes

From MNVs annotations, it was evidenced that a consistent fraction of variants felt within Alu repeats. I therefore tried to determinate whether MNVs cluster within particular Alu regions or were randomly dispersed throughout the consensus sequence. I identified two major peaks within the Left derived arm that were consistent with the position of Alu's Pol III Box A and Box B (figure 45). Genome-wide analysis was then performed, aiming at verifying whether MNVs were enriched in Box A/B consensus. After having derived Box A and Box B consensus sequences from literature, I identified 211,622 Box A and 251,572 Box B consensus with 0 mismatches within the reference genome. Next, I found that 56 and 103 MNVs were in overlap with the set of Box A and Box B consensus, respectively, mostly within AluY and AluS elements, which represent the youngest Alu families. Finally, by performing Z score analysis, I found that these overlaps were extremely significant (Z scores of 4.6 and 7.2 for Box A and Box B, respectively) (figure 46).

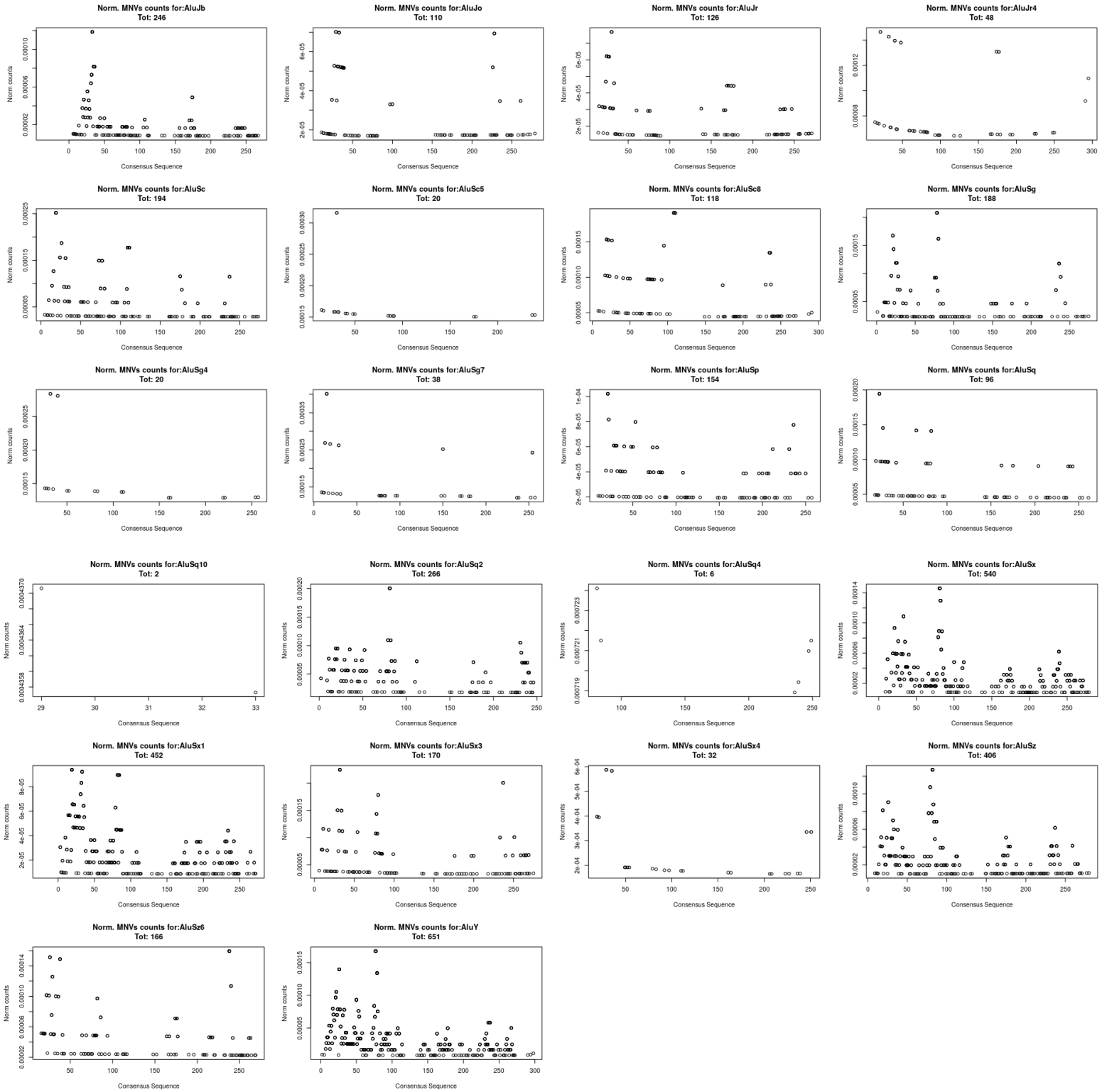


Figure 45: Normalized counts of MNVs within Alu's consensus sequence. Data are displayed according to Alu families. Total counts of MNVs per Alu family in reported in the title of each plot.

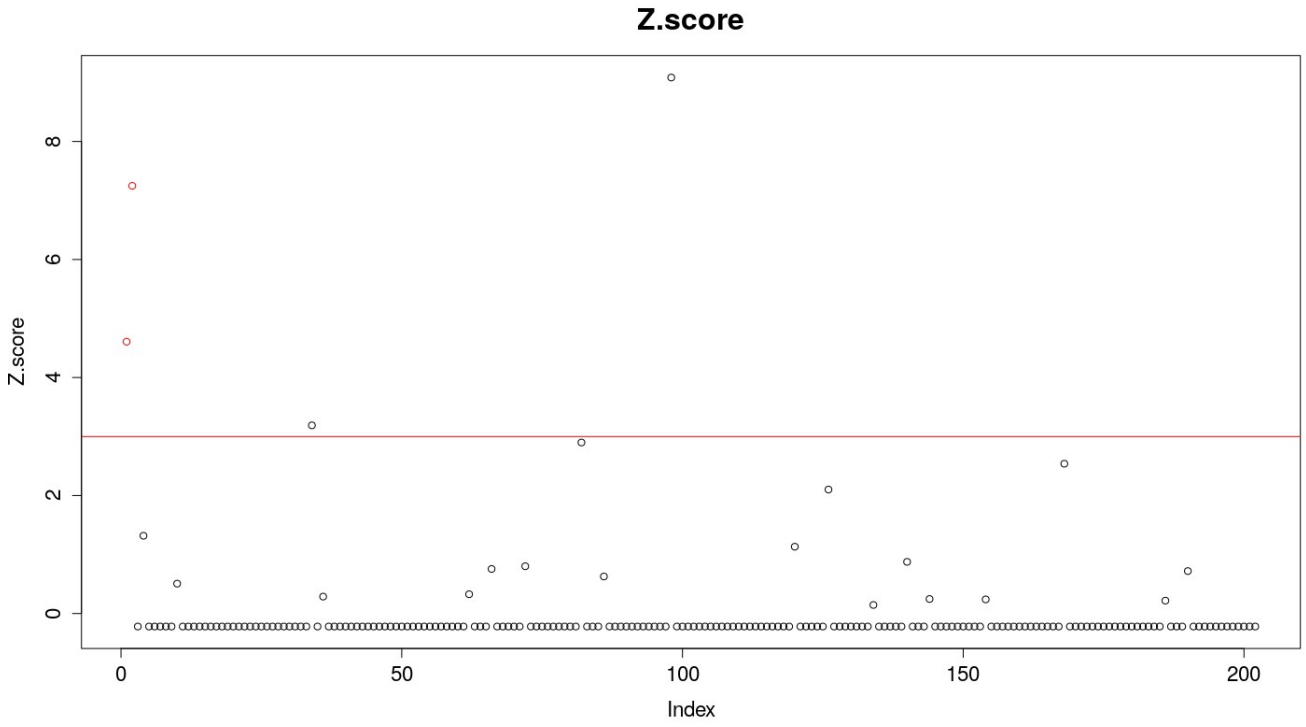


Figure 46: Z scores for 202 box sequences. The 2 original Box A and Box B are colored in red. Randomly shuffled regions (100 for Box A and 100 for Box B) are displayed in black. Z score threshold at 3 standard deviation is represented as a red horizontal line.

4.4 Discussion and Conclusion

Here I provided a deep investigation of somatic late-onset variants in brain tissues. From an initial set of 27,106 SNVs, by applying several filtering steps, a high-quality set of 15,646 variations was retained. During the variant calling step, I found software limitations that had the effect of mis-annotating 6,128 SNVs. By applying a correction step, I was able to fix the SNVs annotations. This led me to identify a total of 15,646 SNV of which 3,064 were MNVs and 9,518 were actual SNVs. SNVs were not observed to be enriched in particular tissues, nor to be prevalent in a particular cohort. Additionally, I evidenced that they did not provide greater pathogenic effects potentially due to their enrichment in non-coding regions. Finally, through signature analyses, SNVs origin was tracked down to SBS1, SBS6, SBS10b, SBS12 and SBS39. While SBS12 and SBS39 are currently without a proposed aetiology, the remaining 3 SBSs, which explained almost 80% of variants, are well known documented. Among them SBS.1 is related to the endogenous and spontaneous deammination of 5-methylcytosine and it is linked to the normal age progression (*clock-like Signature*). It was expected since samples derived from old aged individuals. SBS6 instead, is related to defective DNA mismatch repair while SBS10b is associated to the effect of polymerase epsilon exonuclease domain.

Most notably, the survey of WGS data led to the identification of somatic multi-nucleotide variants from brain. To my knowledge, this is the second documented observation of somatic MNVs (the first from samples not originating from tumors) and the first one from brain tissues. As for SNVs, I was unable to reconnect MNVs to particular tissues or to a single cohort. Moreover, I found that MNVs were not increasing the levels of pathogenicity, expressed in CADD scores, with respect to their mis-annotated forms. However, I speculate that this is due to the MNVs enrichment in non-coding regions. It was also noted that the vast majority of MNVs felt within repetitive elements, mostly Alu and L1s, and most notably, that MNVs were enriched within Alus Pol III boxes, which may interfere with the physiological levels of transcription.

MNVs were next demonstrated to be originated simultaneously through alternative allele frequency analyses. Additionally, with nucleotide substitution analyses, evidences of CC>TT variations enrichments were provided, **which I hypothesize being the result of APOBEC activity.** To support our hypothesis, I found no CpG islands involvement, and excluded the potential implication of Polymerase zeta as no consecutive MNVs were present in our set of variants. Interestingly, sequences

containing MNVs were found to contain consensus motifs for zinc fingers bindings. In particular, I observed that the most supported zinc fingers were linked to *Herpex simplex* virus type I (HSV-1) infections. This is interesting since HSV-1 infections were previously linked to Alzheimer's disease. In this regard, Moir and colleagues in 2018 proposed the antimicrobial protection hypothesis [Moir et al., 2018], for which Amyloid-β oligomerization is not intrinsically pathological as it emerges as an innate immune pathway. However, chronic infection would have led to neuroinflammation. Therefore, one could hypothesize a model in which HSV-1 brain chronical infections trigger APOBEC immune activity, resulting in MNVs accumulations in brain. Although intriguing, this model needs to be verified under several aspects, starting from the experimental validation of MNVs observations to the assessment of HSV-1 infections.

With these analyses I provided valuable data in support of somatic MNVs characterization stressing also the needs of integrating MNV annotations steps in current NGS pipelines.

Chapter V

Preliminary investigation of variants within AD-associated genes

5.1 Introduction

Despite LOAD arises from a less understood set of genetic, epigenetic, and environmental risk factors [Kingsbury et al., 2006; Gatz et al., 2006; Bertram et al., 2010], several studies have demonstrated a direct involvement of different genes and polymorphisms. Large-scale genome-wide association experiments, in this sense, provided valuable data regarding millions of polymorphisms from thousands of subjects [Bettens et al., 2013]. These collections can then be interrogated by researchers, potentially becoming fundamental in the understanding of the rise and progression of the disease. I therefore started to perform qualitative analyses, aimed to interrogate whether our variants collections, from both SNPs arrays and WGS experiments, may involve AD-associated genes and be therefore related with the etiology of the disease.

5.2 Materials and Methods

5.2.1 The AD-associated genes set

To test the presence of AD-associated variants within the available datasets, I selected genes and variants known to be correlated with AD. 10 genes associated with both EOAD and LOAD were selected. These are: ABCA7 [Vasquez et al., 2013], ADAM10 [Kim et al., 2009b], APOE [Karch et al., 2014], APP [Sleegers et al., 2006], BIN1 [Chapuis et al., 2013], PIN1 [Park et al., 2019], PSEN1 [Cruts and Broeckhoven, 1998], PSEN2 [Cai et al., 2015], SORL1 [Rogaeva et al., 2007] and TREM2 [Reitz and Mayeux, 2013]. Gene coordinates, were obtained form Gencode database (version 19) [Frankish et al., 2019] and extended by 10Kb both upstream and downstream (Table 6) with BEDtools (version 2.27.1, subcommand slop, parameter -b 10,000).

Then, I collected 11,180 single nucleotide variants associated to AD from the *gwasDB* version 2 [Li et al., 2016].

GENE SIMBOL	GENE STRAND	GENCODE ORIGINAL	GENCODE EXTENDED	LENGTH (bp)
ABCA7	+	chr19:1040102-1065571	chr19:1030102-1075571	45469
ADAM10	-	chr15:58887403-59042177	chr15:58877403-59052177	174774
APOE	+	chr19:45409011-45412650	chr19:45399011-45422650	23639
APP	-	chr21:27252861-27543446	chr21:27242861-27553446	310585
BIN1	-	chr2:127805603-127864931	chr2:127795603-127874931	79328
PIN1	+	chr19:9945933-9960358	chr19:9935933-9970358	34425
PSEN1	+	chr14:73603126-73690399	chr14:73593126-73700399	107273
PSEN2	+	chr1:227057885-227083806	chr1:227047885-227093806	45921
SORL1	+	chr11:121322912-121504402	chr11:121312912-121514402	201490
TREM2	-	chr6:41126244-41130924	chr6:41116244-41140924	24680

Table 6: Set of AD-associated genes selected for this study. Gencode original coordinates, reported in the GENCODE ORIGINAL column were extended in both direction by 10,000 nucleotides (GENCODE EXTENDED column).

5.2.2 Analyses of overlaps

Collections of AD-associated genes and variants in BED formats were intersected with SNPs arrays and WGS datasets by using *BEDtools* (version 2.27.1, subcommand *intersect*). Counts were obtained by parsing the intersection with R (version 3.6).

5.2.3 Feature analyses

VCF files were annotated with the Variant Effect Predictor (VEP, version 99.2) [McLaren et al., 2016] and then converted to MAF formats with the vcf2maf software (https://github.com/mskcc/vcf2maf). Feature analyses were performed with the *annotatR* R package (version 1.14.0) [Cavalcante and Sartor, 2017] with default genome libraries and waterfall plots were drawn with the *GenVisR* R package (version 1.20.0) [Skidmore et al., 2016].

5.3 Results

5.3.1 CNVs in overlap with AD associated genes

I started a preliminary analysis aimed to test whether germline CNVs called from SNPs array data were in overlap with a set of known AD-associated genes and variants (Table 6). I found 3 heterozygous deletions in overlap with APP, ABCA7, and APOE annotations (Table 7).

First, a high confidence (> 400) heterozygous deletion in the APP gene was identified from the CTR sample C_02_S001_C. The deletion, of 67,674 nucleotides in length, affected the genomic upstream region of the gene, without involving its protein coding region. Similarly, an heterozygous deletion was found in the ABCA7 locus. This CNV was observed in C02_S007_C CTR sample, and similarly to the APP deletion, it was in overlap with the upstream region of the gene. Finally. within the same C02_S007_C sample, I identified a small heterozygous deletion of 4,543 bps in length, located in the APOE region. This variation resulted in the loss of exon2, exon3 and exon4 (figure 47), therefore potentially resulting in an aberrant form of the protein product.

From the analysis, overlaps between CNVs and AD associated genes were observed in only two Spanish CTRs. Interestingly, this may indicate that CNVs do not represent a major factor for the aetiology of the disease in our samples.

GENCODE EXTENDED	GENE ID	CNV POS	CNV LENGTH	CNV_TYPE	CONFIDENCE	SUPPORTING SNPs	SAMPLE
Chr19:1030102-1075571	ABCA7	Chr19:988934-1032740	43807	Hetero. Del.	166.792	127	C02_S007_C
Chr19:45399011-45422650	APOE	Chr19:45409857-45414399	4543	Hetero. Del.	46.297	19	C02_S007_C
Chr21:27242861-27553446	APP	Chr21:27181483-27249156	67674	Hetero. Del.	404.81	130	C02_S001_C

Table 7: CNVs in overlap with AD-associated gene coordinates. CNV lengths are displayed in bps.

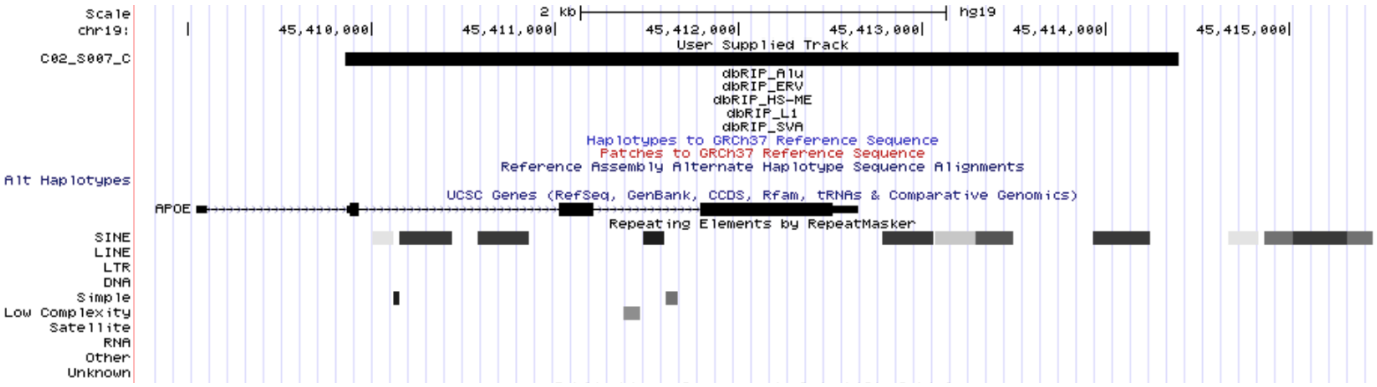


Figure 47: UCSC genome browser screen shot for the APOE heterozygous deletion. Deleted region is represented as a black rectangle at the C02_S007_C sample's track.

5.3.2 Single nucleotide variants from SNPs arrays

I next interrogated the SNPs array dataset of nucleotide variants. I observed the presence of 572 different SNPs loci in overlap with the selected AD-associated genes. 443 of these were found in intronic regions, while 70 were observed within exonic annotations. Additionally, 59 SNPs were called from the intergenic regions flanking the target genes. Among the selection of AD genes, APP showed the highest SNPs counts, with 188 variants. However, only 2 APP variants were observed in the exonic regions. Interestingly, 11 out of 572 SNPs were also identified as AD associated variants from GWAS studies (Table 8).

GENE SIMBOL	TOTAL SNPs	EXONIC SNPs	INTRONIC SNPs	INTERGENIC SNPs	SNPs in GWAS
ABCA7	36	10	22	4	1
ADAM10	78	6	66	6	0
APOE	12	2	4	6	4
APP	188	2	179	7	0
BIN1	65	6	41	18	4
PIN1	19	6	7	6	0
PSEN1	32	9	21	2	0
PSEN2	40	11	25	4	0
SORL1	96	18	77	1	2
TREM2	6	0	1	5	0
TOTAL	**572**	**70**	**443**	**59**	**11**

Table 8: Counts of SNPs in overlap with AD associated genes.

Next, to gain information on the potential impact of the variants and to provide a better overview of the features in overlap, I functionally annotated the set of variants in AD associated genes. The most damaging variants per gene and samples were displayed with *waterfall* form, from which we observed a high abundance for 3′ and 5′ UTRs and flanking regions. No gain in stop codons nor alteration in the protein sequences were noted (figure 48).

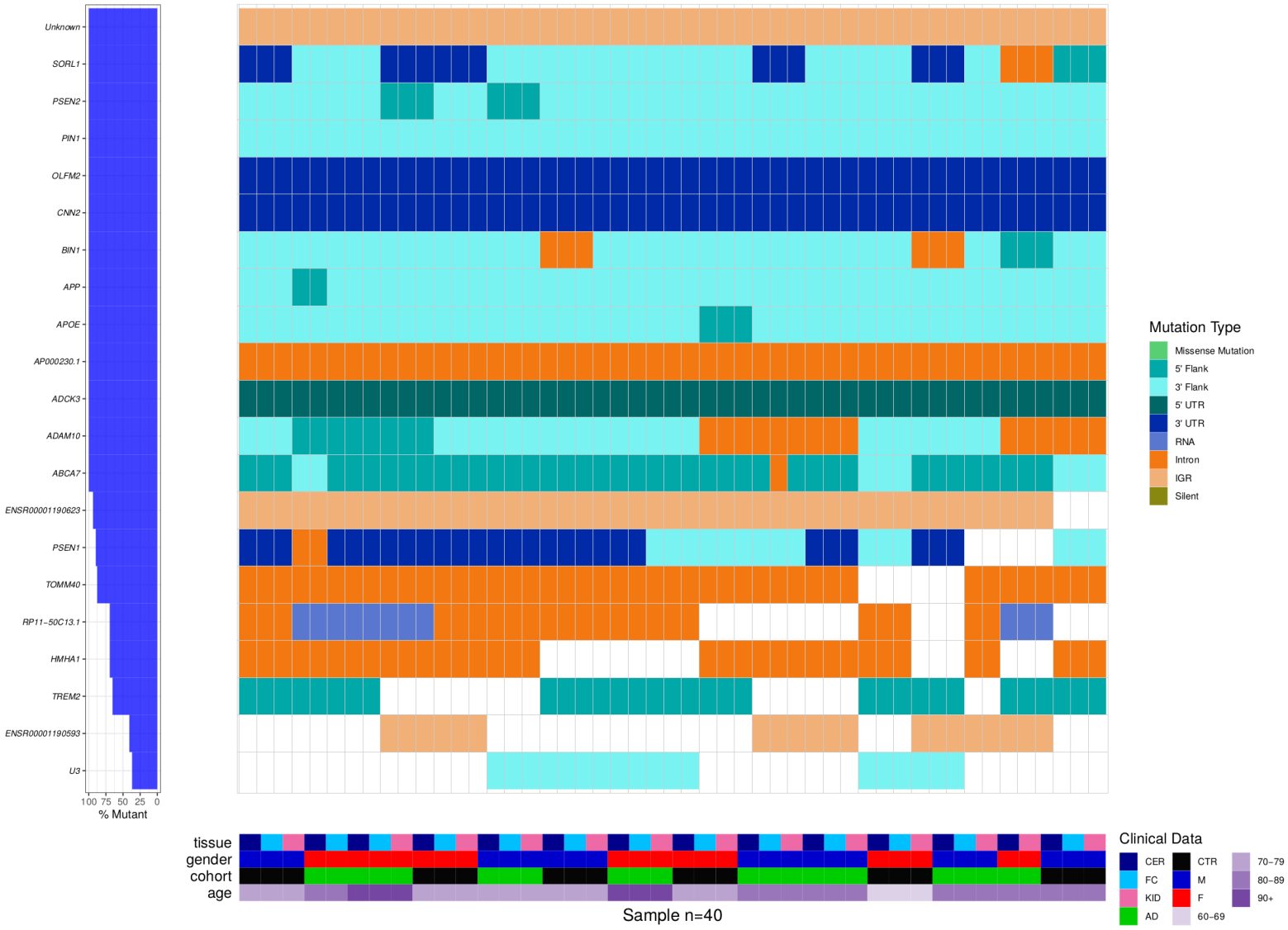


Figure 48: Waterfall plot of SNPs data. Center figure: Most damaging mutation type per sample and gene, left figure: gene names and frequency of samples with mutations; bottom figure: metadata information (Age, Tissue, Gender and Cohort) per sample.

Additionally, we seek to determinate the presence of SNPs in overlap with AD-associated variants. We found 11 positive results associated to the genomic loci of SORL1, ABCA7, APOE and BIN1. 2 SNPs associated to APOE locus were actually in overlap with the TOMM40 gene annotations. Out of 1,111 SNPs, 10 were observed in intragenic regions: 7 in introns and 3 in exons (Table 9) without altering the protein sequence or the splice sites.

SNP POS	SNP ID	REF	ALT	FEATURE	GWAS
chr11:121423552	rs11604897	G	A	intron	SORL1
chr11:121452354	rs7120354	G	A	intron	SORL1
chr19:1056492	rs3752246	G	C	exon	ABCA7
chr19:45401666	rs8106922	C	A	intron	TOMM40
chr19:45403412	rs1160985	C	A	intron	TOMM40
chr19:45407788	rs7259620	C	A	intergene	APOE
chr19:45410002	rs769449	G	A	exon	APOE
chr2:127826533	rs1060743	A	G	exon	BIN1
chr2:127839781	rs10194375	C	A	intron	BIN1
chr2:127841769	rs10200967	C	A	intron	BIN1
chr2:127859418	rs873270	A	G	intron	BIN1

Table 9: SNPs identified within AD-associated genes and GWAS studies. SNP POS: SNP annotation; REF: reference nucleotide; ALT: alternative nucleotide; GWAS: Overlapping gene from GWAS studies.

5.3.3 Germinal Single nucleotide variants from WGS

Then, I interrogated the WGS set of germinal variants. 4,405 loci were observed to be in overlap with the set of AD genes. In particular, 369, 3,522 and 514 variants were identified within exonic, intronic and intergenic regions, respectively. Additionally, similarly to what observed with the SNP array dataset, the APP gene arbored the highest amount of variants (1,411), while TREM2 the lowest (38). Interestingly, 31 sites were also found within the *gwasDB* set of AD-associated variants (Table 10).

GENE SIMBOL	TOTAL SNPs	EXONIC SNPs	INTRONIC SNPs	INTERGENIC SNPs	SNPs in GWAS
ABCA7	419	110	295	14	3
ADAM10	631	25	543	63	0
APOE	102	21	45	36	13
APP	1411	29	1321	61	0
BIN1	436	38	292	106	8
PIN1	121	27	55	39	0
PSEN1	451	34	343	74	0
PSEN2	189	29	129	31	0
SORL1	607	53	486	68	7
TREM2	38	3	13	22	0
TOTAL	**4405**	**369**	**3522**	**514**	**31**

Table 10: Counts of WGS SNPs in overlap with AD associated genes. Counts were evaluated for several features: exons, introns, intergenic regions and GWAS variants.

The set of 4,405 variants was next annotated with the variant effect predictor, from which no potentially pathogenic variants were evidenced in AD associated genes. However, the functional annotation led us extend the observation of variants within both 3′ and 5′ UTRs and gene flanking regions of AD associated genes (figure 49) with respect to the SNPs array.

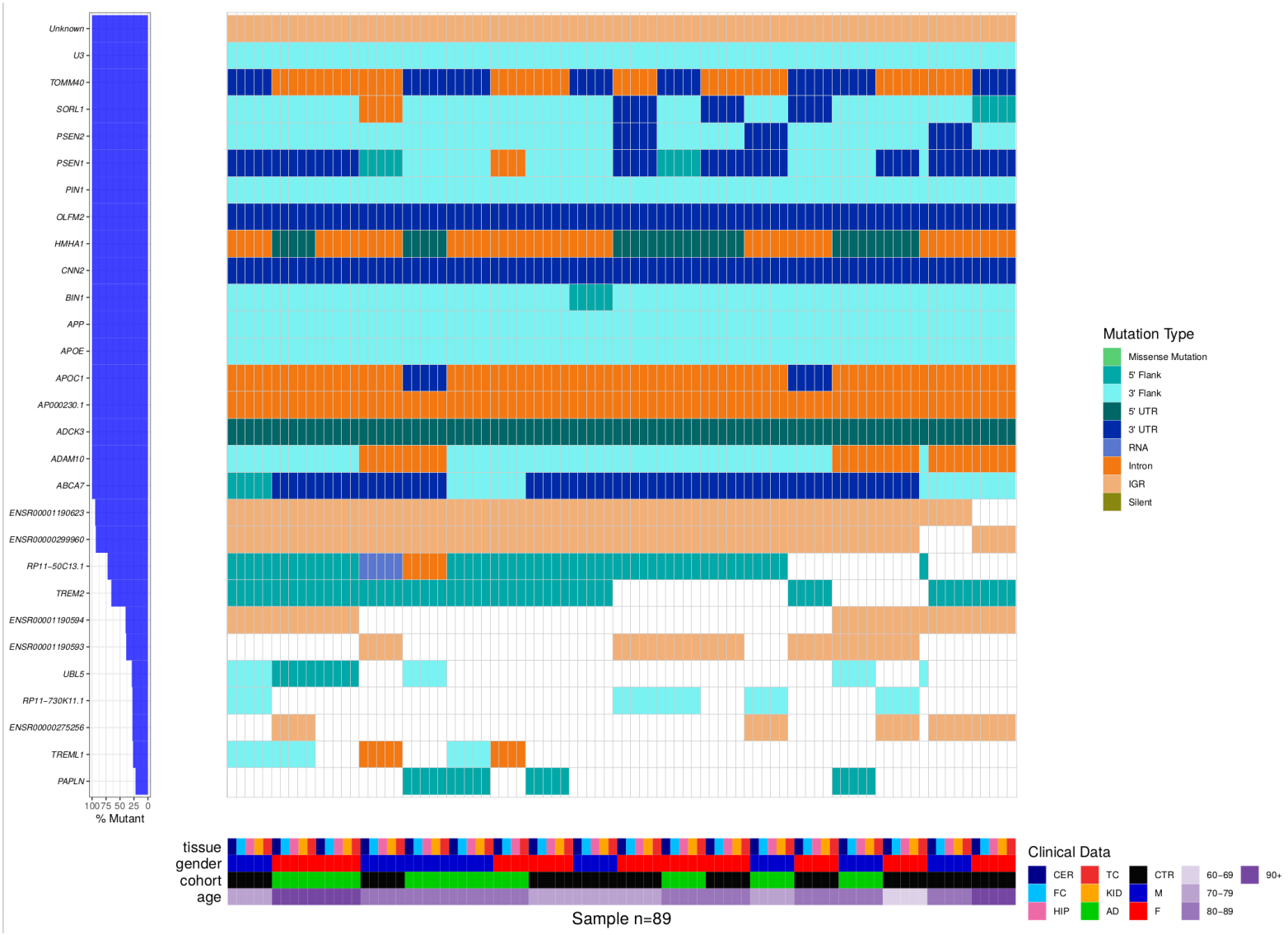


Figure 49: Waterfall plot of WGS germline variants data. Center figure: Most damaging mutation type per sample and gene, left figure: gene names and frequency of samples with mutations; bottom figure: metadata information (Age, Tissue, Gender and Cohort) per sample.

The set of 31 gwasDB variants contained all 11 loci detected through SNP arrays, which were thus validated. Interestingly, 8 out of 31 variants were discovered within exonic regions while only 5 where annotated as intergenic. Variants were observed within SORL1, ABCA7, TOMM40, APOE and BIN1 genes (Table 11).

SNP POS	SNP ID	REF	ALT	FEATURE	GWAS
chr11:121382172	rs6589884	A	T	intron	SORL1
chr11:121423552	rs11604897	C	T	intron	SORL1
chr11:121435587	rs11218343	T	C	intron	SORL1
chr11:121452354	rs7120354	G	A	intron	SORL1
chr11:121453779	rs58698151	A	T	intron	SORL1
chr11:121473391	rs11218360	T	C	intron	SORL1
chr11:121474025	rs12287339	T	C	intron	SORL1
chr19:1046520	rs3764650	T	G	intron	ABCA7
chr19:1056492	rs3752246	G	C	exon	ABCA7
chr19:1063443	rs4147929	A	G	intron	ABCA7
chr19:45401666	rs8106922	A	G	intron	TOMM40
chr19:45403412	rs1160985	C	T	intron	TOMM40
chr19:45403858	rs760136	A	G	intron	TOMM40
chr19:45404431	rs741780	T	C	intron	TOMM40
chr19:45404691	rs405697	A	G	exon	TOMM40
chr19:45404972	rs1038025	T	C	exon	TOMM40
chr19:45405062	rs1038026	A	G	exon	TOMM40
chr19:45406673	rs10119	G	A	exon	TOMM40
chr19:45407788	rs7259620	G	A	intergene	APOE
chr19:45408836	rs405509	T	G	intergene	APOE
chr19:45410002	rs769449	G	A	exon	APOE
chr19:45411941	rs429358	T	C	exon	APOE
chr19:45414451	rs439401	T	C	intergene	APOE
chr2:127826533	rs1060743	A	G	exon	BIN1
chr2:127839781	rs10194375	C	A	intron	BIN1
chr2:127841769	rs10200967	C	T	intron	BIN1
chr2:127852021	rs10207628	G	C	intron	BIN1
chr2:127859418	rs873270	T	C	intron	BIN1
chr2:127860830	rs754107	C	G	intron	BIN1
chr2:127872347	rs3856378	G	C	intergene	BIN1
chr2:127873035	rs4663098	T	C	intergene	BIN1

Table 11: SNPs identified within AD-associated genes and GWAS studies. SNP POS: SNP annotation; REF: reference nucleotide; ALT: alternative nucleotide; GWAS: Overlapping gene from GWAS studies. Loci in red were also observed with SNPs array data.

5.3.4 Somatic variants from WGS

I finally investigated the presence of somatic variants within the AD associated gene loci. 6 late onset somatic SNVs and 2 somatic MNVs (table 12) were identified. Variants were mostly localized in introns (6/8) with an additional SNV within the 5′ flanking region of the PSEN1 gene and one SNV in intergenic region (figure 50). No somatic variants were found in overlap with the set of AD associated variants.

CHR	START	END	REF	ALT	SAMPLE	GENE	VARIANT TYPE
chr14	73601106	73601116	GATTACAGGCA	AATTACAGGCG	929_HIP	PSEN1	MNV
chr14	73632210	73632211	T	C	9345_TC	PSEN1	SNV
chr15	58972581	58972582	T	C	8805_HIP	ADAM10	SNV
chr19	1032098	1032100	GCA	TCG	7466_HIP	ABCA7	MNV
chr19	1061194	1061195	G	A	7711_CER	ABCA7	SNV
chr19	9950957	9950958	C	T	8805_CER	PIN1	SNV
chr21	27244708	27244709	T	C	7711_TC	APP	SNV
chr21	27282525	27282526	C	T	1282_FC	APP	SNV

Table 12: Somatic nucleotide variants in AD associated genes

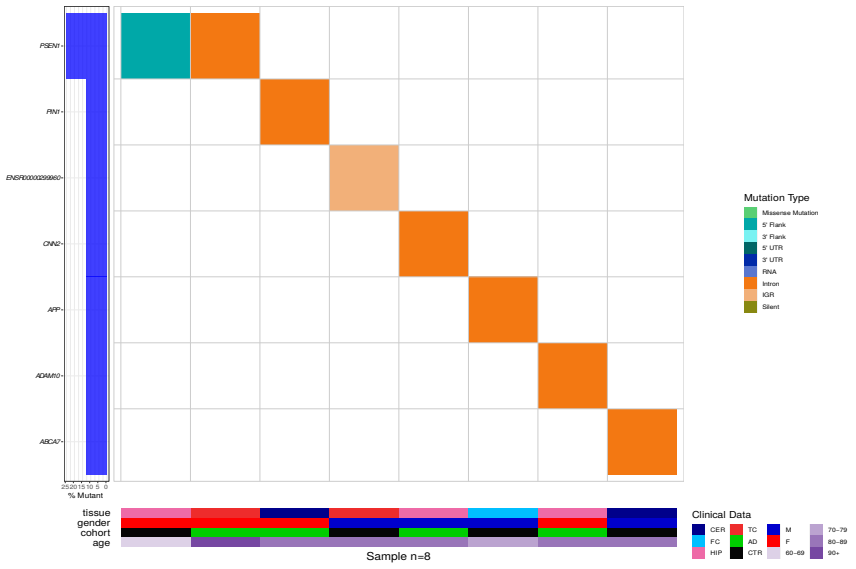


Figure 50: Waterfall plot of WGS somatic variants data. Center figure: Most damaging mutation type per sample and gene, left figure: gene names and frequency of samples with mutations; bottom figure: metadata information (Age, Tissue, Gender and Cohort) per sample.

5.4 Discussion and conclusion

Here I provided a preliminary qualitative analysis aimed to test a selection of 10 AD-associated genes for the presence of variants called with both SNPs arrays and WGS technologies. From the chip assay, three heterozygous deletions were detected. Two were in overlap with the genomic upstream regions of ABCA7 and APP and one with exons 2, 3 and 4 of APOE gene. Despite the potential pathological effect that deletions, in particular of exons, may confer, CNVs were found in only two CTRs. Therefore, I hypothesize that germline CNVs may not play a major role in the aetiology of AD in this particular cohort of samples. Nonetheless, further investigations based on WGS CNVs data would be still required to strengthen the results as well as to extend the analyses on smal CNVs. Additionally, 572 SNPs were found within our selection of AD related genes, 9 of which already annotated as AD associated variants with GWAS studies. Interestingly, three were observed in exonic regions. Then, by interrogating the WGS dataset, I further extend these observations. The number of SNPs in AD-related genes increased to 4,405, whereas 31 of them were also found to be already correlated to the disease from GWAS studies. Interestingly, a subset of 8 variants was also validated with SNPs experiments. Variants in overlap with GWAS studies were found from the ABCA7, APOE, BIN1, TOMM40 and SORL1 genes annotations and despite SORL1, at least one exonic mutation was observed for each gene. Finally, I also observed the presence of 6 single nucleotide somatic variants and 2 multi-nucleotide somatic variants in AD associated genes. To my knowledge, this is the first observation of multi-nucleotide somatic variants from hippocampal formations.

Overall, feature analyses showed that the vast majority of variants felt in intronic and intergenic regions. Nonetheless, I also observed high number of variants within the 3′ and 5′ UTRs and gene flanking regions.

While this preliminary analysis did not highlight the presence of variants that may have directly triggered the development of the pathology, UTRs and gene flanking regions are highly abundant in regulatory features. Therefore it cannot be excluded that potential modification of the gene function may contribute to the disease, requiring therefore further and deeper investigations. In this direction I plan to ***i)*** Extend the list of AD-associated genes; ***ii)*** Explore the effects of both germline and somatic CNVs from WGS data; ***iii)*** Interrogate variants in UTR and flanking regions to assess their effect of the gene function.

Chapter VI

Life-seq: a new targeted sequencing approach aimed at addressing current WGS limitations at LINE-1 regions

6.1 Introduction

As discussed in the main introduction, current whole genome short-reads sequencing methods have limitations that mostly affect structural variants discovery in repetitive regions. This is especially important when L1-dependent events lead to mosaicism that occur in a small number of cells in an heterogeneous tissue.

To confront these biological and technical complexities, Erwin and collaborators [Erwin et al., 2016] developed a machine learning-based, single-cell targeted sequencing approach, named somatic L1-associated variants (SLAV)-seq, to increase detection sensitivity in L1 regions. L1 specificity was conferred by a single L1-specific oligo able to pair within the 3′ of the retrotransposon element, that during sequencing generates L1-flanking genome split reads. Although providing a terrific level of resolution, even this state-of-art method has limitations conferred by its probe design. Since there is a single L1-specific oligo that pairs within the 3′ end of the L1 element sequence, SLAV-seq can indeed be used to detect new integration events, as deletions that involve downstream L1-proximal genomic regions. However, it lacks the ability to detect ***i)*** L1 polymorphic regions in the absence of the integrations; ***ii)*** L1-associated deletions that involve the whole repetitive sequence as these would not permit probe pairing; ***iii)*** structural variants that affect only the 5′ extremity as these would not be sequenced. To my knowledge, there are no reported targeted sequencing approaches able to detect all the aforementioned structural variants at once. Furthermore, after having performed SNPs array experiments, it was noticed that such technologies are generally depleted in markers within repetitive region, which was highly significant for L1 elements (figure 51). Following what stated in chapter II, this depletion may have limited CNV detection in retrotransposon regions and thus the initial study.

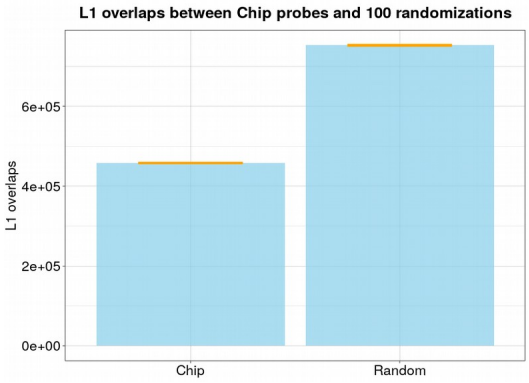


Figure 51: Overlaps between array markers and L1 annotations (from RPMK database) and 100 times randomly selected regions and L1 annotations. Z score = - 330. Counts of randomly selected regions match array markers numbers.

In the laboratories of Prof. Sanges and Gustincich, a new technology named LIne-1 Five prime End SEQuencing (LIFE-seq), was developed to address these complexities. LIFE-seq is a high-coverage, paired-end targeted-sequencing technology based on a capturing strategy. It consists in 2,783 probes of 80 bps in length, which are able to perfectly bind to the genomic region proximal to the 5′ end of just as many full-length L1 sequences. By designing probes on single sequences regions close to, but outside of the 5′ L1 annotations we aimed to assess the presence of L1 polymorphisms, L1 CNVs and L1-associated deletions, which would not be achievable with probes embedded within the element. Importantly, sequencing should be carried out at extremely high-coverage (ideally higher than 1000x per locus) to identify rare somatic events. In this context, I have developed the bioinformatic pipeline to process and analyze LIFE-seq data, from the raw sequencing data to the genotyping of the targeted sites.

6.2 Materials and Methods

6.2.1 The LIFE-seq techniques

Full-length L1 sequences were retrieved from L1base databases in genome build hg38 (L1 Full-length: FLI-L1, L1 with intact ORF2: ORF2-L1 and non-intact L1 longer than 4500 bp: FLnI-L1; Downloaded on April 2017, latest update: 2016-07-09) [Penzkofer et al., 2017]. I converted their coordinates by aligning the sequences to genome build GRCh37 (version GCA_000001405.1) using *MEGABLAST* (version 2.4.0, 2016-Aug-5, parameters: -perc_identity 100). From the results only coordinates that presented full coverage and 100% of sequence identity with the reference genome were kept. Sequences were next submitted to the Illumina DesignStudio™ tool to obtain custom probes (https://designstudio.illumina.com/). Design constrains were: 1) when aligned to the reference genome, probes sequences must not show multimapping properties; 2) probes must align in the upstream genomic region of target L1 elements, at a maximum distance of 100 bps. A set of 2,783 different custom probes was then obtained. Downstream checks consisted in further assessing the presence of potential multimapping events. These were conducted by mapping probes versus the reference human genome build GRCh37 (version GCA_000001405.1) with *BLAST+* (version 2.2.31+, blastn analysis using probes sequences as queries and reference genome build as database; parameters *-perc_identity* 100 *-qcov_hsp_perc* 100) [Camacho et al., 2009]. Mapping data were also combined with the targeted L1 annotations sites to obtain metrics on the distances between probes and related L1 sequences. Samples were selected from the Brazilian Alzheimer disease cohort and consisted in two male AD (ids: 2682, 7466) and two female controls (ids: 6868, 9269) (Table 4). LIFE-seq library preparation strictly followed the Nextera® Rapid Capture Enrichment Reference Guide (https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nexterarapidcapture/nextera-rapid-capture-enrichment-guide-15037436-01.pdf). Pilot phase library preparation was optimized by lowering the *tagment* enzyme quantity from 15 uL to 5 uL. Average insert length was observed to be equal to 460 bps by measurements carried out using the 2100 Bioanalyzer Instrument (Agilent Technologies, Inc) with Agilent DNA 1000 Kit. A second library was prepared with further protocol optimizations in order to achieve higher insert sizes. The process specifically involved the

modification of the post-tagmentation clean-up, with the retention of the supernatant. Final average insert size was observed to be 750 bps.

Libraries were next sequenced in paired-end strategy with the Hiseq1500 Illumina platform. Binary Base Calls (BCL) files were finally processed through the LIFE-seq bioinformatic pipeline. Library preparations and sequencing were carried out by Dr. Diego Vozzi (IIT).

6.2.2 Bioinformatic analyses

6.2.2.1 LIFE-seq pipeline

LIFE-seq sequencing protocol was developed in the research group of Prof. Sanges and Gustincich and I developed a specific bioinformatic pipeline for the analyses of the data. The analysis consist of several serials steps that, starting from the base call files (BCL) produced by the sequencer, can provide: 1) Demultiplexed fastq.gz files from BCL files; 2) Fastq quality assessment; 3) Removal of PCR-duplicated reads; 4) Mapping of the reads on the reference genome; 5) Distributions of inserts lengths; 6) Genotyping data; 7) Frequency of split and discordant reads within single loci (figure 52). To allow fast analyses of high coverage data and/or high number of samples, most of the steps were designed to support multi-threading options. LIFE-seq pipeline is written in Perl (version 5.22.1) and R languages (version 3.4.3), and it is meant to work in *unix* environments. The pipeline requires the presence of additional third-parts, freely available software and additional Perl libraries. The required software are: *bcl2fastq* [Illumina, https://emea.support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html?langsel=/it/], *fastQC* [Babraham Bioinformatics], *multiQC* [Ewels et al., 2016], *PRINSEQ* [Schmieder and Edwards, 2011], *BWA* [Li and Durbin, 2010], P*icard* [http://broadinstitute.github.io/picard], *java (*version >= *1.8)* [http://www.java.com], *SAMtools* [Li et al., 2009], *R* [R Core Team, 2012], while Perl libraries and modules are: IO::Tee [Neil B. https://metacpan.org/pod/IO::Tee], Statistics::PCA [Daniel S.T.H., https://metacpan.org/pod/Statistics::PCA], Cairo [Brian M., https://metacpan.org/release/Cairo], JSON [Kenichi I., https://metacpan.org/pod/JSON]. Additional software must be executable within the environmental *$PATH* variable, otherwise absolute paths must be inserted by the user at the very beginning of the pipeline. This last option was meant to give to the user the possibility to keep track of the softwares versions, providing a better reproducibility of the results. Pipeline was tested with a 64-bit Unix-system provided with Intel Xeon E5-2690 v3 2.6GHz chip set and 128 GB of memory. Here it follows a step-by-step description of the pipeline.

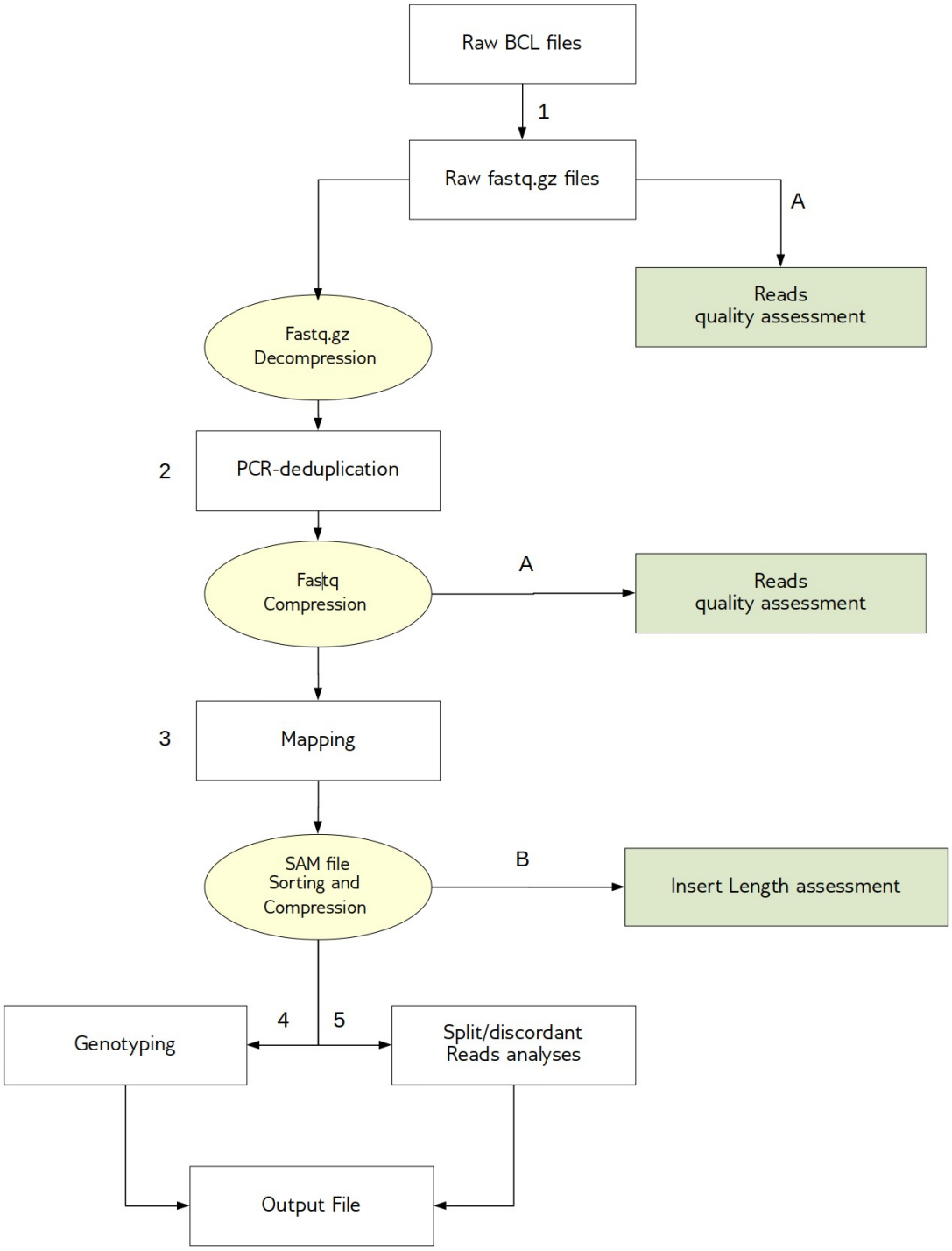


Figure 52: Representation of LIFE-seq pipeline workflow. 1- Conversion from BCL to Fastq.gz files; 2- *PRINSEQ* deduplication step; 3- Mapping step; 4- Genotyping and 5- Split/Discordant reads analyses.

BCL files are converted into fastq gzip compressed files (*fastq.gz*) using the *bcl2fastq* software. Reads quality is then determined with the *fastQC* tool. Both processes can be speed-up using multi-threading options. To facilitate sample's quality comparisons, *multiQC* software is used to generate a single run report.

I then focused my attention to select the best software to identify and remove PCR duplicates from raw fastq.gz files. PCR-duplicates are multiple and identical PCR products generated from the same template molecule. They can lead to false positive variant call and/or affect the coverage data by providing a distorted perception of the real quantity of the template molecule. Therefore, their removal is a crucial step for genotyping and mosaicism detection, where precise quantification is needed. I tested the following softwares: *PRINSEQ, FastUniq [Xu et al., 2012]* and *super deduper [Petersen et al., 2015]*. Percentages of duplicated reads were observed *prior* and after deduplication relying on the *fastqc* software. *PRINSEQ* showed two main disadvantages which primarily affected its speed: it does not support multi-threading options and it is unable to use paired-end fastq.gz files as inputs, thus requiring a decompressing preprocessing step. However, its deduplication performances were the best among all the tested software (data not shown) due to its duplicate read call strategy that relies on both pair-reads nucleotide sequences. We therefore selected *PRINSEQ* despite its slowness. Further enhancement in PCR-duplicates removal software may led us to reconsider this choice.

Deduplicated reads quality is assessed with the combination of *fastQC* and *multiQC* softwares, as previously reported. Deduplicated fastq.gz files are next mapped using *BWA* (version *0.7.17-r1194-dirty;* subcommand *mem;* default score parameters) versus the reference genome build GRCh37 (version GCA_000001405.1). *SAMtools* (version 1.9) is then applied to sort (subcommand *sort*) and to compress (subcommand *view*; parameters -S,-b,-h) the SAM files. Quality assessment, mapping, sorting and compression steps are optimized for multi-threading options. Inserts size analyses are next preformed with *Picard CollectInsertSizeMetrics* tool (version 2.18.20; default parameters). Although P*icard* doesn't support multi-threading options, Perl *threads* module is used to generate single forks for each sample. This allows the parallel analysis of single samples using different threads.

The next genotyping step is aimed at providing information about the presence or absence of the targeted L1 sequences at the level of single locus. It is performed through a separated R script, called by the main LIFE-seq pipeline, that is parallelized using the same strategy aforementioned for P*icard.* Genotyping is based on the assumption that each locus consists in two alleles and it relies on coverage

analyses performed at three different regions per locus. These are the 5′ genomic upstream region (named **5′ EXT**), the 5′ region (named **5′ INT**) and the 3′ genomic downstream region (named **3′ EXT**), all relatively to the targeted L1 element sequence (figure 53). Windows sizes (imposed to 500 bps as default setting) can be modified by users depending on their needs. Coverage information at nucleotides levels are extracted for all three regions with *samtools depth* (option -a). Distributions of coverages per region are produced, normalized using the total mapped reads per sample and then compared in the form of mapped reads per million. In the presence of both targeted L1 alleles (genotype 0/0), current LIFE-seq insert size library protocol does not produce reads relatively to the 3′ EXT region, thus its coverage results to be 0. On the contrary, in the total absence of the L1 alleles (genotype 1/1), we can obtain coverage data for the 3′ EXT without observing reads mapping on the 5′ INT region. Finally, in the case of a locus heterozygous for the presence of the L1 (genotype 0/1), coverage must be different from 0 in all the three regions (figure 53). To improve genotyping calls, poorly covered loci (< 400x) at the level of the probe regions are set as ungenotyped (genotype "./.") with a specific option.

Finally, to identify variants within L1 sequences, and in particular potential somatic SVs, I designed a novel split-reads/discordant-reads frequency analysis. Loci genotyped as 0/0 were selected and mapped read pairs were divided in 4 groups depending on their insert length calculated based on the mapping locations for each pair. The first class comprises pairs with insert length shorter than 1 Kb. The second cluster comprises pairs with insert length greater than 1 Kb but shorter than 15 Kbs, while the third group comprised pairs with insert length greater than 15 Kbs. Finally, pairs displaying reads mapped to different chromosomes, were clustered into a fourth group. By selecting loci genotyped as 0/0, in the absence of genotyping errors, different-than-class 1 reads should support somatic structural variants. In particular, class 2 reads were searched in order to detect potential somatic deletions. Filtered loci were manual checked with the Integrative Genomic Viewer (version 2.8.0).

Result obtained with the bioinformatic pipeline are stored within an organized hierarchy of directories rooted on the output folder, which is imposed by the user. Finally, as major output, the pipeline creates a probe-locus-oriented tabular-separated table in which genotyping and reads frequency data are associated with annotations regarding the targeted repeat element.

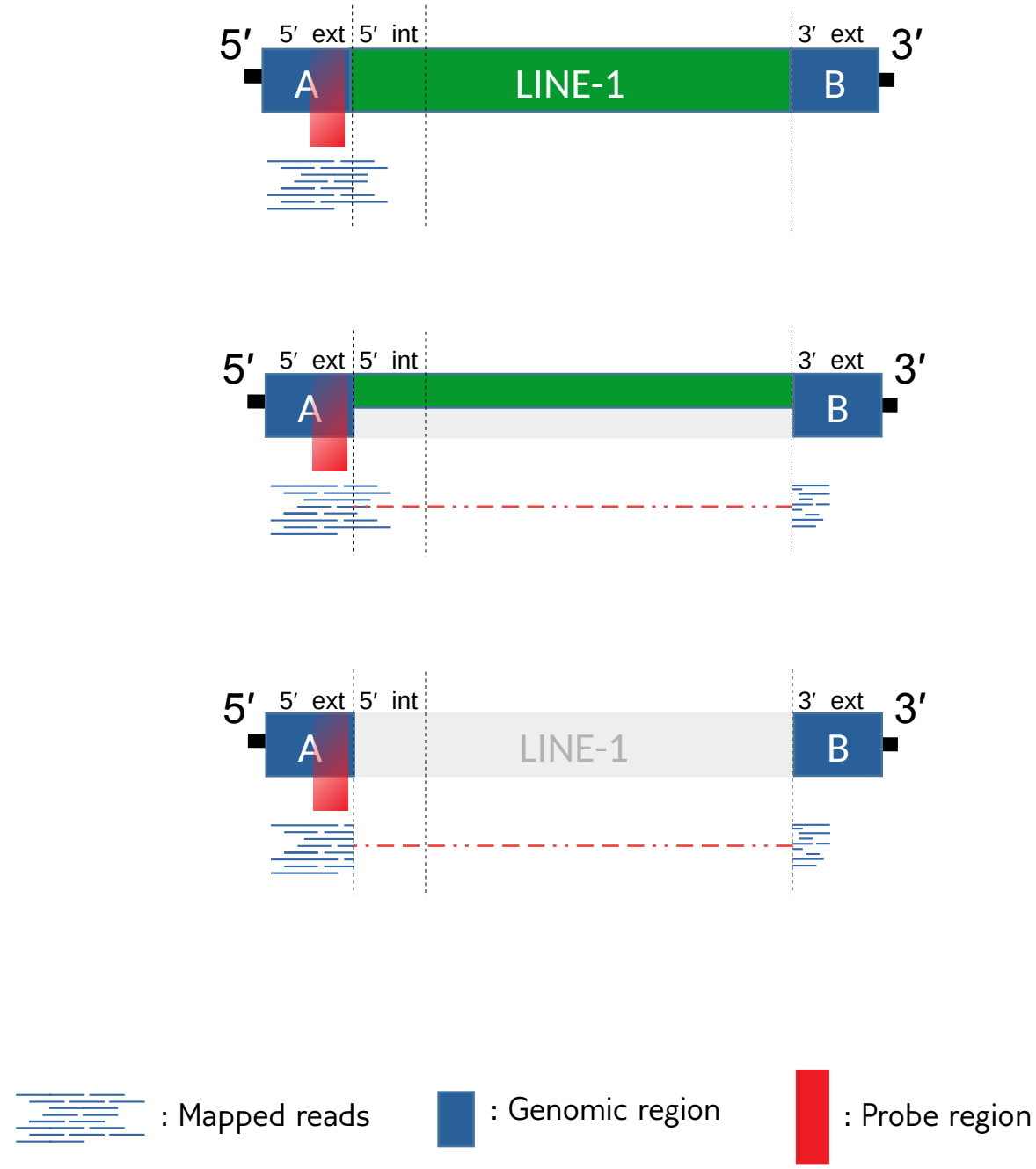


Figure 53: LIFE-seq genotyping rationale. 5′ EXT region: Genomic region flanking the L1 5′ UTR sequence; 5′ INT region: First portion of the L1 5′ UTR region; 3′ EXT : Genomic region flanking the L1 3′ UTR sequence. Depending on the presence (in green), or on the absence (in grey), of the targeted L1 sequence, reads can maps to only 5′ EXT and 5′ INT regions (genotype 0/0), to only 5′ EXT and 3′ EXT regions (genotype 1/1) or in all three regions (genotype 0/1).

6.2.2.2 Coverage analyses

TarSeqQC (version 1.8.0, november 2017) coverage analyses were performed with the pilot-phase data only. Coverage metrics were extracted for all samples following the *TarSeqQC* documentation (https://bioconductor.org/packages/release/bioc/vignettes/TarSeqQC/inst/doc/TarSeqQC-vignette.pdf). I modified coverage threshold, minimum base quality and minimum mapping quality parameters to 50, 15 and 60, respectively.

Putative CNVs were called relying on coverage analyses. Coverages of probe regions were obtained for all loci and samples with *SAMtools* (command depth -a). Values were normalized with total mapped reads per sample and reported with the per million annotation. Putative CNVs were next called from loci in which there was at least one individual with coverage between 150 and 250 and one with coverage outside this range.

Coordinates of LIFE-seq alternative loci were intersected with *BEDtools* (version 2.27.1, subcommand intersect, parameters -wa) with known L1 polymorphism annotations. These were retrieved from dbRIP database (version of 04/26/2018) [Wang et al., 2006] and from euL1 database (version 1.0) [Mir et al., 2015].

6.3 Results

6.3.1 Pilot phase

The correct functionalities of the LIFE-seq sequencing protocol and probe design were tested in a pilot phase experiment. This consisted in the application of the LIFE-seq technology to 4 cerebellum samples in technical triplicate (from AD 2682, AD 7466, CTR 6868 and CTR 9269 individuals, Table 4), for a total of 12 samples. With the pilot experiment, we aimed to: 1) confirm the robustness of probes design; 2) test LIFE-seq reproducibility and the need for technical replicates; 3) evaluate the ability to call L1 loci genotypes correctly.

Before sequencing, all 2783 probes designed with the Illumina DesignStudio tool were validated for their unique mapping position within the genome using a *Blastn* analysis. Furthermore, I confirmed that probes were all mapped outside of the target L1 element, at an average distance of 22 nucleotides. After sequencing, data were processed by applying a beta version of the LIFE-seq bioinformatic pipeline in combination with a *TarSeqQC* coverage analyses. Demultiplexing of BCL files allowed me to correctly assign 263.8 million reads to their respective samples (90.45% of the total reads produced). Reads quality was assessed prior and after a *PRINSEQ* PCR-duplicates removal step and was always higher than 35 on average, for each sample. Using the *PRINSEQ* PCR-duplicates removal tool, duplicates has been decreased by an 8% on average, while reads counts decreased by almost 50% (figure 54). Moreover, average reads length increased, from 212 bp to 222 bp. Outliers were tracked to the reads that were not assigned to specific samples (named as undetermined reads) during the demultiplexing process.

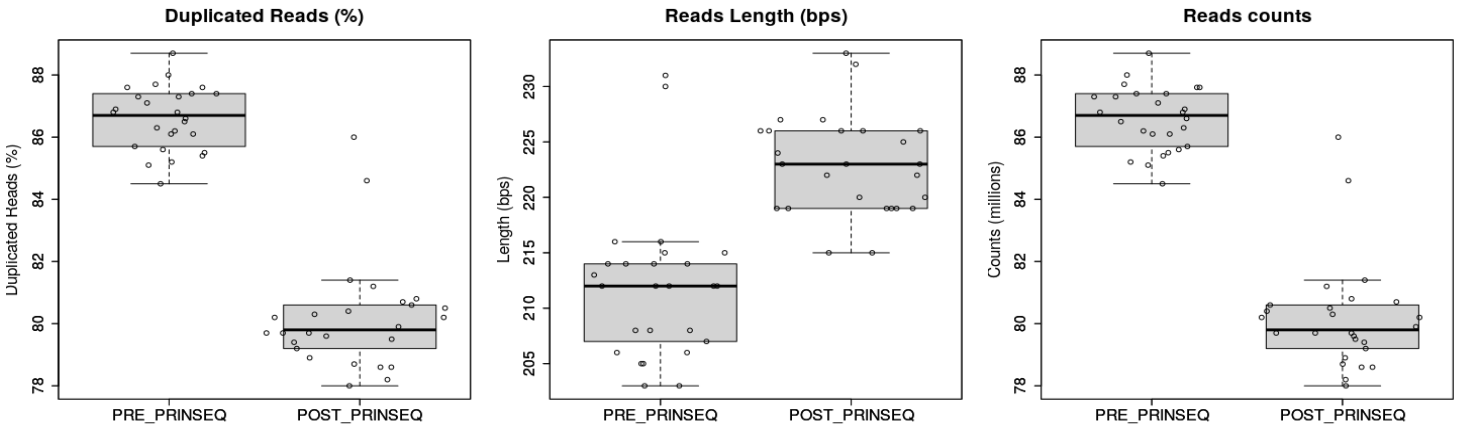
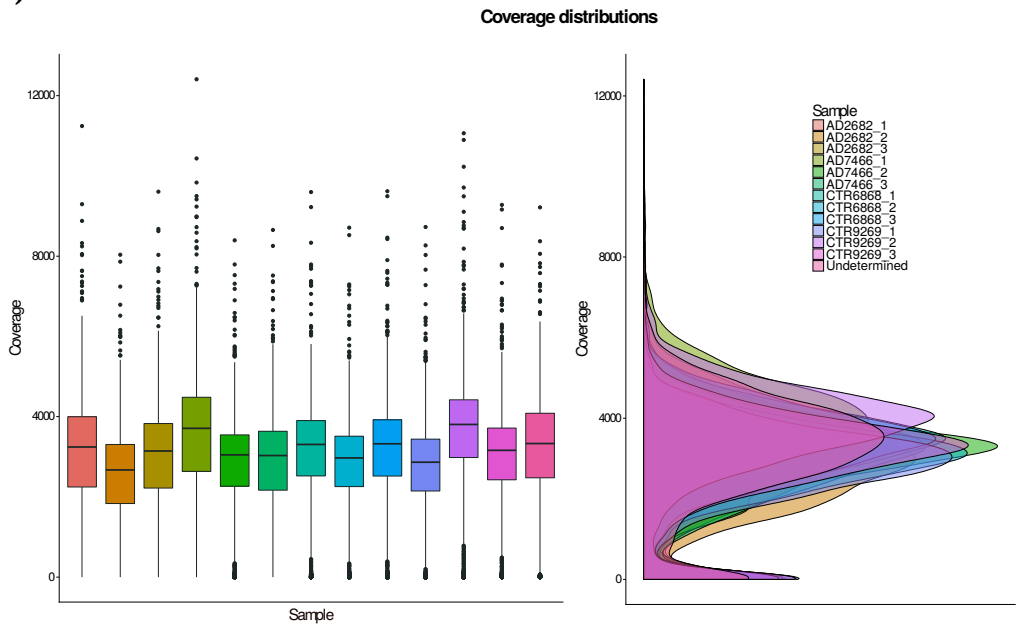


Figure 54: Comparison between pre-*PRINSEQ* and post-*PRINSEQ* data. From Left to Right: Distributions of duplicated reads percentages, average reads lengths and reads counts.

More than 99.9% of deduplicated reads were then mapped to the reference genome as described in the method section. Mapping data were further used to test whether there were nonfunctional probes. The *TarSeqQC* R package [Merino et al., 2017] was applied to assess probes coverages, revealing an average value higher than 3000x per locus, with peaks of more than 8000x. Moreover, *TarSeqQC* analysis showed that less than 5% of loci per sample had no coverage data, potentially indicating difficulties in probes pairing due to SNVs, or deletions of the genomic sequences required for their binding. Overall, the design was very robust since 75% of the targeted loci were sequenced with a coverage higher than 2000x and only 10% with coverage below 500x (figure 55).

A)



B)

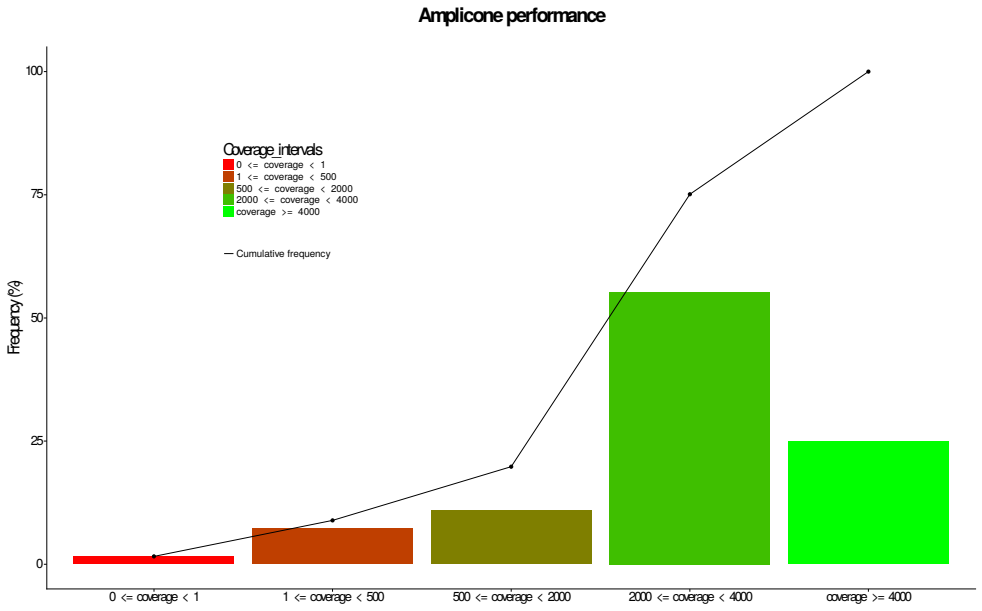
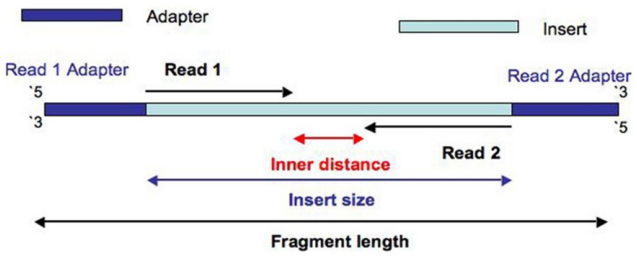


Figure 55: *TarSeqQC* coverage metrics. **A)** Coverage distributions for each sample and replicate. **B)** Amplicone performance. Percentages of probes were reported according to their coverage values.

Insert size (figure 56-A) was next estimated with the *Picard* tool, and found to be close to the reads lengths (~ 220 bps, figure 56-B). Further checks on the TLEN field of the SAM files (that reports the number of bases from the leftmost mapped base to the rightmost mapped base of a read pair) led me discovered that about 30% of reads showed sequence lengths equal to TLEN field values. This meant that about a third of the total paired reads consisted in the bidirectional sequencing of the same nucleotide sequence. This was most likely due to the small fragment length achieved with the pilot library preparation protocol, which on average was 460 bps. To obtain higher fragment sizes, the LIFE-seq library preparation protocol was modified in the next runs, as reported in the Methods section. Despite the small average insert size, and because the fragment size is predicted after the mapping on the reference genome, I also found a small peak centered at ~6,500 bps (figure 56-B), coherent with full-length L1 element lengths, supporting the existence of L1 polymorphisms visible as L1-specific deletions after mapping on the reference human genome and therefore resulting in a such long estimated length.

A)



B)

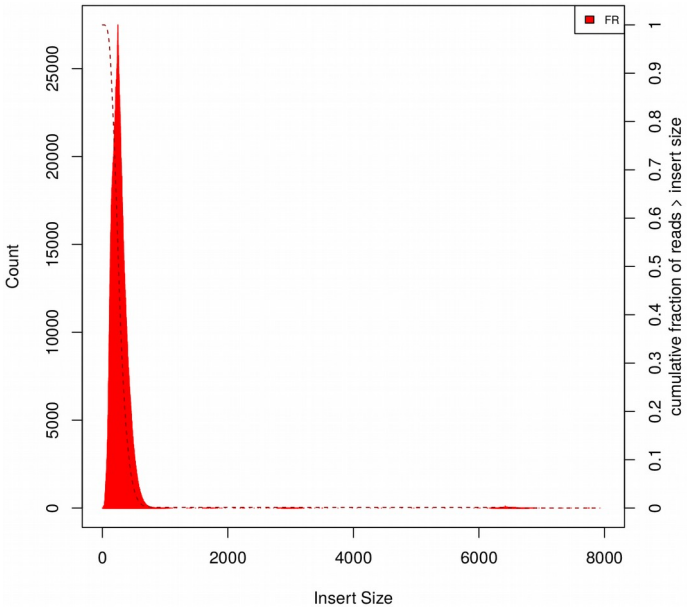


Figure 56: A) Overview on fragment nomenclature. Picture shows a single fragment composed by two read adapter flanking an insert; **B)** Insert size distribution. Counts of fragments (on y axis) are displayed according to their insert size (x axis) in base pairs. A very small peak is appreciable at insert size equal to ~6,500 (bps).

The analyses genotyped ~ 90% of the loci for which probes were designed. Genotyping concordance between replicates of the same sample was then tested (Table 13). 28 loci (15 unique) out of 11,172 (2,793 loci x 4 samples) loci were found to be discordant between replicates due to small coverages fluctuation across replicates. They had passed the imposed threshold by a very small number of reads. By checking these 15 loci, I observed a mean minimal coverage of 209 and a mean maximum coverage of 998 within all replicates. To remove possible errors due to low coverage, I set as “ungenotyped” these 28 loci for all the samples (Table 14).

Samples / Genotypes	**./.**	**0/0**	**0/1**	**1/1**
AD2682_1	248	2493	23	19
AD2682_2	250	2491	23	19
AD2682_3	249	2492	23	19
AD7466_1	249	2495	17	22
AD7466_2	250	2494	17	22
AD7466_3	248	2496	17	22
CTR6868_1	290	2456	19	18
CTR6868_2	289	2457	19	18
CTR6868_3	291	2455	19	18
CTR9269_1	291	2457	11	24
CTR9269_2	289	2459	11	24
CTR9269_3	290	2456	11	24

Table 13: Raw genotyping data from pilot phase. Data are displayed for all replicates. Genotypes as follows: “./.”: ungenotyped locus; “0/0”: loci homozygous for the presence of the targeted L1 element; “1/1”: loci homozygous for the absence of the targeted L1 element; “0/1”: heterozygous loci.

I next focused on ungenotyped regions (genotype "./."). These were found to be a) loci with coverage equal to 0; b) loci presenting coverages values below our internal quality threshold, fixed to 400x. To test whether ungenotyped loci were potentially due to genomic deletions, I took advantage of CNV data from WGS, finding only a single overlap. Moreover, I next compared the counts of ungenotyped regions between the 4 different samples. It was noted that there were ~40 more sites in CTRs (about 290 loci) with respect to the AD samples (about 250 loci) (Table 14). The discrepancy originates from the differences in genders between samples (CTRs are females while AD are males) and by the fact that both chromosome X and Y were included in the probe design. (with 66 probes in chrY). Importantly, this feature makes LIFE-seq able to discriminate between genders. Random manual checks were then performed with the *Integrative genomics viewer*, which confirmed the genotyping of all the alternative loci calls and 20 random reference calls per sample.

Genotyping, in combination with *TarSeqQC* coverage analyses, provided solid, reliable and reproducible data in support of the use of LIFE-seq protocol without the need of technical triplicates.

Samples / Genotypes	**./.**	**0/0**	**0/1**	**1/1**
AD2682	252	2489	23	19
AD7466	252	2492	17	22
CTR6868	291	2455	19	18
CTR9269	292	2456	11	24

Table 14: Refined genotyping data of pilot phase. Genotypes as follows: "./.": ungenotyped locus; "0/0": loci homozygous for the presence of the targeted L1 element; "1/1": loci homozygous for the absence of the targeted L1 element; "0/1": heterozygous loci.

6.3.2 Second test run

A second LIFE-seq experiment was performed after having improved the library preparation protocol and LIFE-seq analysis pipeline. With this new run I primarily seek to: 1) examine the ability to call CNVs associated to L1 elements which were not detected through SNPs array experiments; 2) further test the LINE-1 genotyping strategy; 3) explore a novel reads frequency approach aimed to detect L1-associated structural variants (in particular somatic deletions).

Sequencing was performed upon 5 different tissues (CER, FC, TC, HIP and KID) of the same pilot phase individuals. Sequencing raw data were then processed with the LIFE-seq analysis pipeline. 661.6 millions reads were generated, with an average of 15.5 millions reads per sample. Demultiplexing step resulted in the correct assignment of 95% of total reads produced to the relative samples, an improvement of 5% respect to the pilot phase. Demultiplexing step decreased duplicate reads fraction by a ~ 8%, increased average reads lengths and reduced total reads counts by a ~50%, in agreement with the observations made during the pilot phase (figure 57).

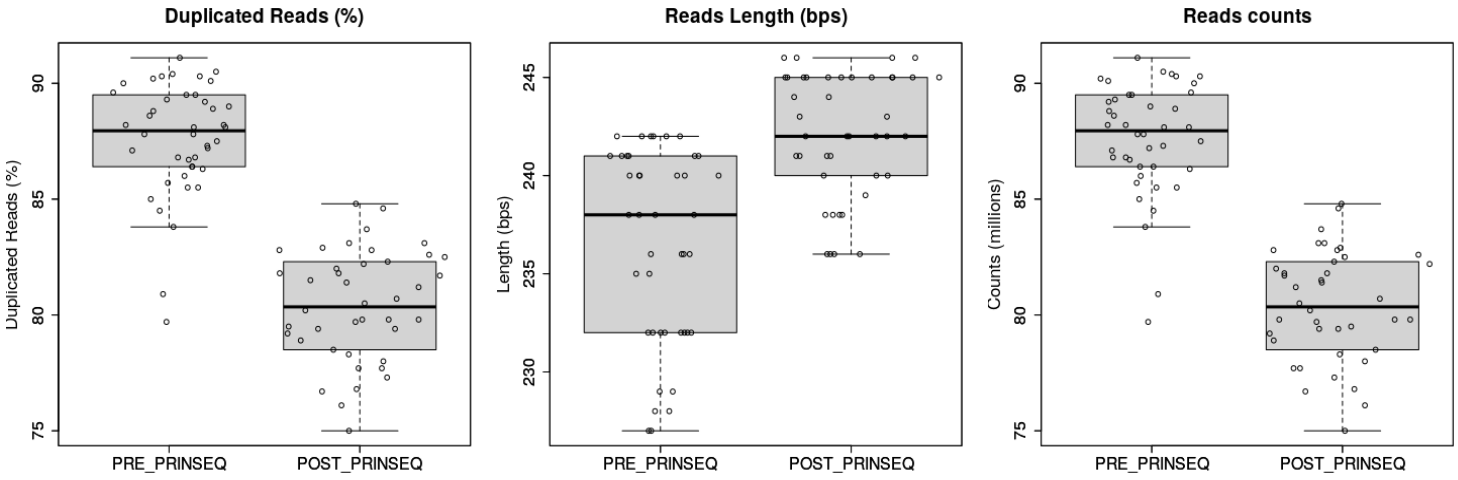


Figure 57: Comparison between pre-PRINSEQ and post-PRINSEQ data from the LIFE-seq second run test. From Left to Right: Distributions of duplicated reads percentages, average reads lengths and average reads counts.

Insert length distribution was observed to be, on average, of 750 bps. This increase of over 44%, with respect to the pilot phase, was coupled to a general increase in reads length between the two sequencing reactions. These data proved that modifications of the library preparation protocol were effective. As for the pilot phase, it was tested the ability to call individuals genders with LIFE-seq data. Instead of relying on genotyping results, I decided to rely on probe regions coverage, which cannot be biased by L1 polymorphisms. By assessing the probe coverage of sexual chromosomes loci, I found perfect agreement with the gender metadata (figure 58-A).

By applying the same coverage strategy at autosomal loci, I then seek to detect genomic CNVs. I focused on loci that had all five tissues of at least one individual with probe region coverage between ~150 and ~250 and all five tissues of at least one individual with coverage outside this range. 29 loci were thus identified (Table 15). They were supposed to be in overlap with CNVs but none of them resulted from our SNPs array CNV calling (figure 58-B).

Probe coordinates	Putative Copy Number			
	AD-2682	AD-7466	CTR-6868	CTR-9269
chr1:112703457-112703536	1	2	2	1
chr1:169226879-169226958	0	0	2	2
chr1:196756110-196756189	2	2	1	2
chr10:9561736-9561815	2	2	2	1
chr10:117347320-117347399	1	2	2	1
chr11:49746730-49746809	2	2	2	1
chr11:55453011-55453090	2	2	1	2
chr11:107243383-107243462	2	2	1	1
chr12:127585201-127585280	1	2	2	2
chr13:52849819-52849898	4	3	2	3
chr13:72845787-72845866	1	1	0	1
chr15:53976291-53976370	2	2	1	2
chr2:4205293-4205372	2	3	2	3
chr2:35991404-35991483	0	2	2	1
chr2:79338847-79338926	2	2	4	3
chr2:98141778-98141857	0	1	0	1
chr3:26432283-26432362	2	0	2	2
chr3:75538778-75538857	1	2	1	2
chr3:137379310-137379389	2	2	1	2
chr3:145645450-145645529	2	1	2	1
chr4:16944294-16944373	1	2	2	1
chr4:69381563-69381642	2	2	2	0
chr4:144586418-144586497	2	2	2	1
chr4:190058134-190058213	1	0	1	1
chr5:177199131-177199210	1	1	1	0
chr7:150302580-150302659	2	2	1	2
chr9:70199604-70199683	1	0	0	0
chrX:103287490-103287569	2	1	2	2
chrX:108356432-108356511	1	1	2	1

Table 15: Putative CNVs loci. Copy number at the probe regions are reported for single individuals since no inter-individual CNVs were detected. Coverage for the chr2:35991404-35991483 locus (highlighted in red) is displayed in figure 58-C

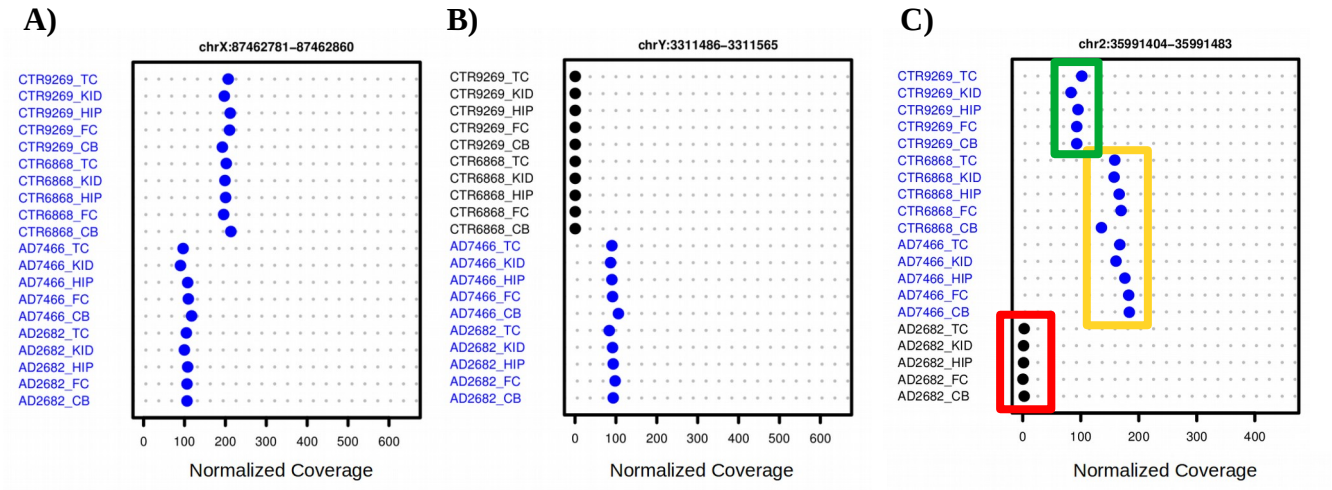


Figure 58: Coverages distributions. **A)** Coverage data for a chromosome X locus; **B)** Coverage data for a chromosome Y locus. **C)** Coverage data at a putative CNV locus. Differences in genders can be appreciated from A) and B). Females have about double of the normalized coverage with respect of Males in A), while they shown coverage equal to 0 in B). Copy number relative to two alleles was identified at coverage ~ 200 (yellow box). Heterozygous calls and Homozygous calls were proposed for coverage ~ 100 and 0 (green and red boxes), respectively. Samples in black displayed coverage equal to 0.

Next I focused my attention on genotyping. The method was improved with respect to the pilot phase by imposing fixed windows lengths for the coverage analyses of particular loci that presented consequent targeted L1s. Without this modification, I observed 3 loci in which 3′ EXT regions and 5′ EXT regions of consecutive L1 elements were in overlap, impeding the correct coverage evaluation and therefore resulting in ungenotyped calls. Consequently, genotyping step led to a refinement of the genotyping call as carried out in the pilot phase, decreasing the number of ungenotyped calls. Moreover, the improvements in sequencing library preparation led to longer fragments, that ultimately resulted in an increase of coverage at the 3′ EXT regions of alternative loci. This led us to improve sensitivity for alternative allele loci calls with respect to the pilot experiment. Alternative calls between tissues belonging to the same individuals were found to be equal, and further manually validated by displaying their coverage profile with the *IGV* software (Table 16).

Finally, I was able to validate some sites as L1 polymorphisms by scanning public databases of variants (DGVs, dbRIP). L1 polymorphic regions were loci for which there were no evidences of L1 annotations in the reference genome but known to arbor L1 insertions in the human population. I found that the vast majority of L1 polymorphisms were shared within different individuals, while only a small fraction of them was private and specific of single individuals (figure 59).

	AD-2682					AD-7466					CTR-6868					CTR-9269				
Genotypes / Tissues	**CER**	**FC**	**HIP**	**TC**	**KID**	**CER**	**FC**	**HIP**	**TC**	**KID**	**CER**	**FC**	**HIP**	**TC**	**KID**	**CER**	**FC**	**HIP**	**TC**	**KID**
./.	245	234	247	245	241	263	242	239	239	234	308	274	278	278	276	281	276	275	275	275
0/0	2493	2504	2491	2493	2497	2478	2499	2502	2502	2507	2436	2470	2466	2466	2468	2465	2470	2471	2471	2471
0/1	25	25	25	25	25	19	19	19	19	19	21	21	21	21	21	12	12	12	12	12
1/1	20	20	20	20	20	23	23	23	23	23	18	18	18	18	18	25	25	25	25	25

Table 16: Genotyping results of the LIFE-seq second run test. Counts of loci are reported according to their genotype data for each individual and tissue.

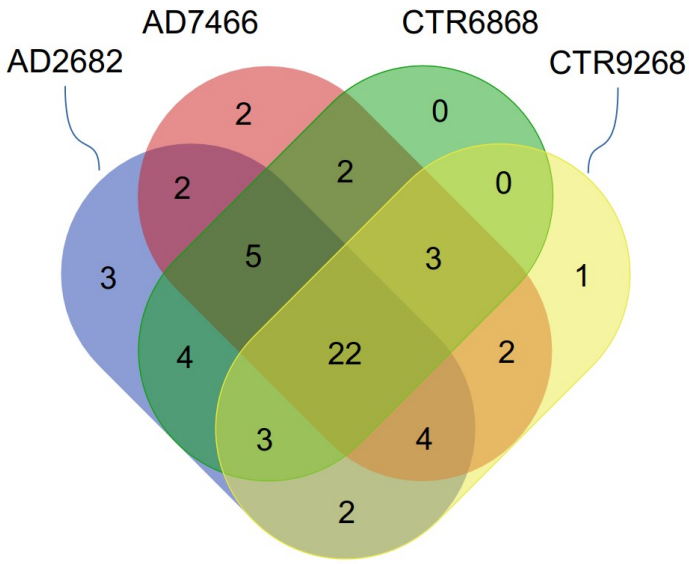


Figure 59: Overlaps between LIFE-seq L1 polymorphic loci from different samples. To note that the vast majority of L1 polymorphisms are shared between samples.

To further validate these results, I compared LIFE-seq genotyping data with the *MELT deletions* results from WGS of the very same samples. By focusing on LIFE-seq polymorphic loci, I found low genotyping correspondence with the MELT calls, which was less than 50% on average (Table 17).

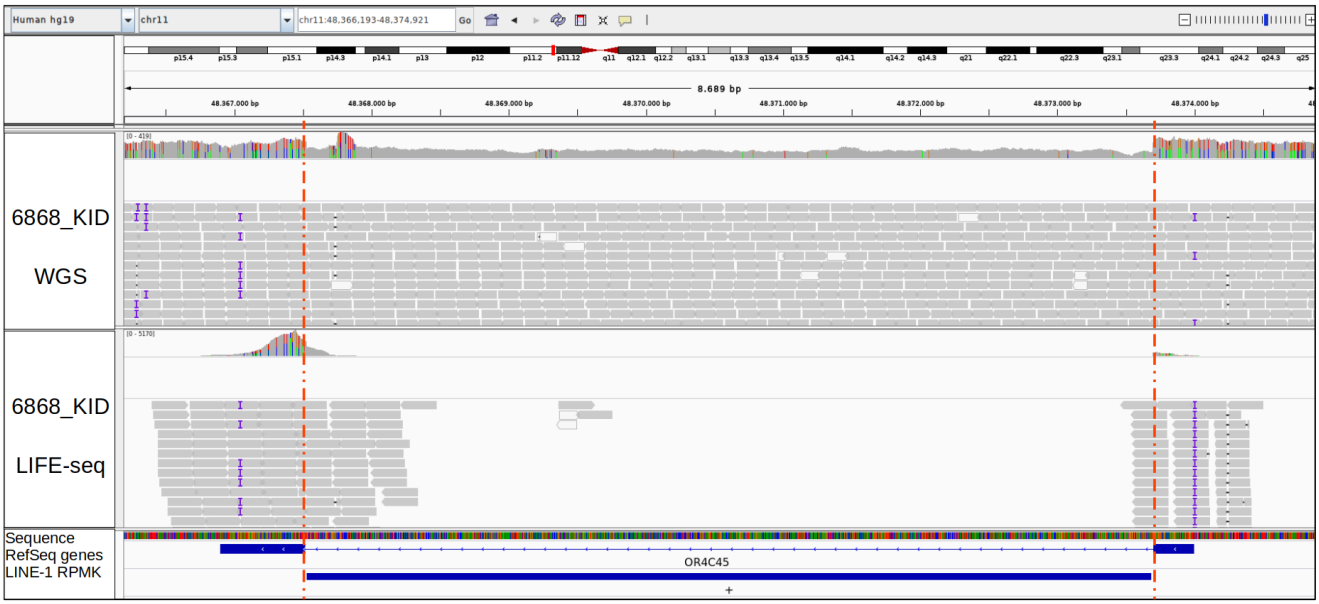
I then compared the coverage profiling of random concordant and discordant genotyping calls finding LIFE-seq results more robust with respect to MELT deletion, probably due to the higher mappings errors that affects WGS (figure 60).

I further focus on non-polymorphic LIFE-seq calls (loci genotyped as 0/0), observing that all but 1 were in agreement with *MELT deletion* results. The exception case was linked to a complex structural variant locus, which presented a deletion coupled with an inversion found to be in overlap with LIFE-seq probe region, thus affecting the mapping step of the bioinformatic pipeline.

Sample	**LIFE-seq polymorphic loci (counts)**	***MELT deletion* concordance (counts)**	**Frequency of concordance (%)**
2682	45	31	69
6868	39	14	36
7466	36	17	47
9269	37	11	30

Table 17: Genotyping concordance between LIFE-seq and MELT deletion calls. Counts of L1 polymorphic loci obtained with LIFE-seq approach are displayed in the second column. The same loci were genotyped with MELT-deletion. Counts of genotyping concordances between the two approaches is reported in the third column, while frequencies on the fourth.

A)



B)

Figure 60: LIFE-seq vs *MELT deletion* genotyping. On both *IGV* representations, first coverage track refers to WGS data while second coverage track refers to LIFE-seq data. The same sample (id: 6868_KID) is displayed on both tracks. Targeted L1 is showed with a blue bar on the bottom track, while L1 breakpoints are reported with two red vertical lines. **A)** Concordant call locus. It is genotyped as 0/1 with both *MELT Deletion* and LIFE-seq approaches; **B)** Discordant call locus. It is genotyped as 0/0 with *MELT Deletion* and 1/1 with LIFE-seq approaches.

To my knowledge there are no available tools to detect somatic variants embedded in L1 elements and only a few that can call germinal variants from targeted sequencing approaches. Therefore, to reach this goal, I took advantage of sequencing data to design and test a novel split/discordant-reads strategy based on genotyping calls. My rationale was based on the fact that loci genotyped as 0/0 would not present reads mapping to the 3′ EXT region unless L1-associated somatic variants are present (figure 61). Thus, my approach consisted in focusing on loci genotyped as 0/0 and in extracting split and discordant reads frequencies directly from the mapping data. Split and discordant reads frequencies were grouped depending on the mapping distance. For split reads, mapping distance is defined as the number of nucleotides that divide the reads from their split-fragments. For discordant reads instead, mapping distance consists in the inner distance of paired-reads. Groups were: *split1/discordant1*: with distance below 1 Kb; *split.1.15/discordant.1.15*: with distances between 1 and 15 Kb; *split.15/discordant.15*: with distances higher than 15 Kb; *split.chr/discordant.chr*: whit fragments mapping into two different chromosomes. I then discovered 12 loci presenting fragments classified as *split.1.15* and *discordant.1.15* and hand-checked them with the IGV tool. All 12 loci were in overlap with LIFE-seq polymorphic loci. Moreover, I observed the presence of SNPs that were not supported by other reads in the same sample. On the contrary, the same SNPs were found in samples from a different individual that harbored the L1 polymorphism suggesting potential inter-sample contaminations (figure 62).

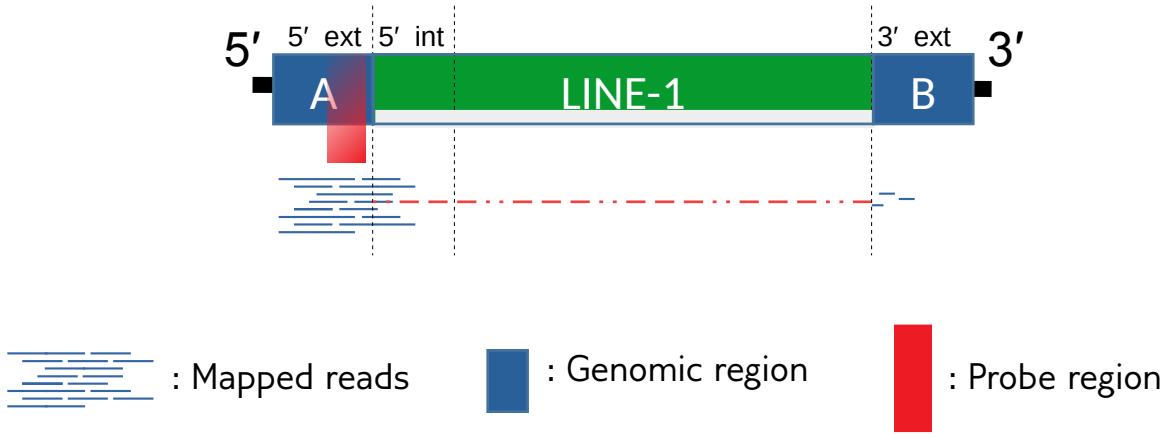
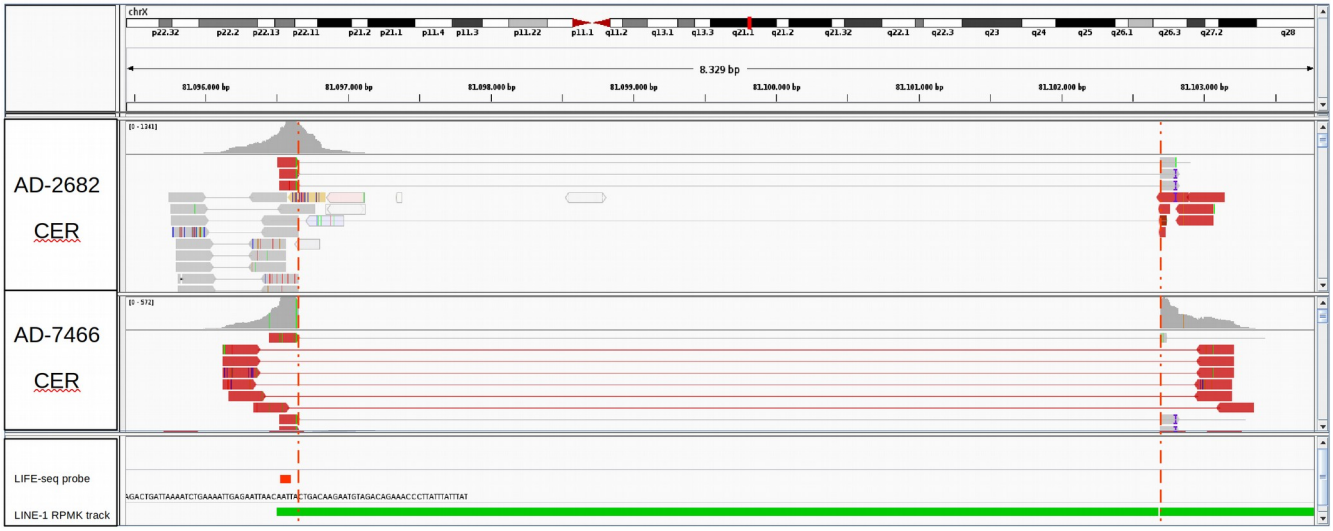


Figure 61: Rationale of split and discordant reads frequency approach. In loci genotyped as 0/0, somatic L1-associated structural variants would present a little amount of split and discordant reads mapped at the 3′ EXT region which are also anchored in the 5′ EXT region.

A)



B)

Figure 62: Candidate somatic locus represented with the IGV software. **A)** Overview on the locus; **B)** Zoom-in at the probe region. For both pictures, top track displays the sample in which the somatic call has been made (AD-2682 CER). Bottom track shows the same locus for a different sample (AD-7466 CER). In this last individual, locus was polymorphic for the presence of the targeted L1 element, thus no reads can be found mapping the targeted L1 sequence. Targeted L1 breaking points are shown with a vertical red dash line. Probe region and L1 annotations are displayed in red and green, respectively, on the bottom tracks part. Vertical colored bars within reads and coverage curves refers to substitutions with respect to the reference nucleotides. The absence of colored vertical bars within the coverage curves (see blue rectangle) strongly supports reference nucleotides.

6.4 Discussion and Conclusions

In confronting the needs for a technology able to asses structural variants within retrotransposon sequences, in particular L1 elements, the laboratories of Prof. Sanges and Gustincich developed a novel targeted sequencing technology called LIne 1 Five prime End sequencing (LIFE-seq). LIFE-seq was designed to sequence the genomic proximal 5′ region of about 3000 different full-length L1 loci at ultra-high depth of coverage. With a pilot phase, I demonstrated the robustness of the approach and probes design, being able to correctly sequence and genotype more than 90% of loci. This was possible after having designed and tested a comprehensive bioinformatic pipeline able to handle and to process raw targeted sequencing data till a genotyping step. Importantly, the technology was so robust that no technical triplicates were necessary to provide solid genotyping data. We then, successfully improved LIFE-seq library preparation and tested this new protocol in a second experiment were a 40% increase in insert dimension was observed. Coverage analyses on the probe regions discriminated samples according to their gender. Furthermore, when applied to CNVs discovery, coverage analyses pointed out the potential presence of 29 putative CNVs not detected through SNPs array technology. Although not yet validated with WGS CNVs data, I was successful in identifying about 30 CNVs out of ~ 3000 loci from the comparison of only 4 individuals. These numbers let me hypothesize that further increase in sample size would result in a much higher rate of CNV calls. However, the major strength of the bioinformatic pipeline was the capability to genotype loci for the presence of the targeted L1 sequence. In this regard, my strategy demonstrated to be reliable by using both different replicates and different tissues within the same individuals. LIFE-seq was effective in discovery previously known L1 polymorphisms. Finally, LIFE-seq genotyping was compared with WGS genotyping data, proving to be more robust and trustworthy. This can be explained by its extreme coverage and by the high quality of the probes design that make LIFE-seq less affected by mapping errors with respect to WGS. Altogether, coverage analyses and genotyping results proved that LIFE-seq is a valid technology to detect genomic CNVs as well as L1. Lastly, I started the design of a novel split/discordant reads frequency approach aimed to unveil the presence of structural variants embedded within L1 sequences, focusing so far on somatic L1 deletions. Although 12 putative somatic loci were identified with the approach, I observed that within some site, supporting reads were marked with SNPs that were tracked to belong to different samples, hinting to inter-sample contaminations. The level of inter-sample contaminations was

nonetheless found to be extremely low (less than 0.5% of the total mapped reads per putative somatic locus) and most importantly, have not affected the correctness of the coverage analyses and genotyping.

Preliminary data from WGS CNVs excluded the possibility that ungenotyped loci were due to genomic CNVs. Therefore, in the future work, I will clarify whether ungenotyped loci might be caused by errors in probes design which will result in the definition of a second more robust LIFE-seq probe panel. Additionally, I also plan to use WGS CNVs data to validate LIFE-seq CNV discoveries.

General conclusions and future directions

As a result of the recent technological advancements, mosaicism has been found to represent a major phenomenon acting in the human brain. Several observations opened to the possibility that somatic variations might be associated to the development of different brain-related disorders; nonetheless, current knowledge is not yet sufficient to define a clear correlation between them, requiring therefore additional studies. Motivated by this lack of knowledge, I decided to focus my attention on the characterization of two different sources of mosaicism, **somatic nucleotide variants** and **retrotransposon copy number variations**, in a particular form of neurodegeneration: the Alzheimer's disease. To approach mosaicism investigation, I analyzed brain data from different high-throughput technologies and contributed to the development of a new targeted sequencing approach. Additionally, to fulfill my goals, I tested the potential usage of uncommon strategies, as *signature analyses, supporting reads analyses* and multi-nucleotide variants annotations, highlighting their potential benefits and limitations. Through the exploitation of the different approaches described in this work, major results were obtained. Shortly, these can be summarized in ***i) t*he identification of known signatures from non-cancer brain data; *ii)* The first identification of somatic multi-nucleotide variants from brain tissues; *iii)* The development of a reliable targeted sequencing approach able to genotype ~3,000 L1 loci**.

However, despite the novelties reported in this dissertation, inconsistencies between SNPs array and WGS data were identified, which prevented the clarification of two main biological questions: ***i)*** Are early onset SNVs differentially represented in AD?; ***ii)*** Are retrotransposons contents altered within genomic CNVs in AD? To find the answers to these problems I here propose two additional analyses. First, to verify once and for all the correctness of the early onset SNVs calls, extended validations of the SNPs array data, with dedicated allele-specific PCRs analyses, should be performed. Second, to collect additional estimations of the retrotransposon contents in genomic CNVs, the already started WGS genomic CNVs exploratory analyses must be further expanded by taking into consideration mobile elements annotations. Indeed, by following the same workflow applied with SNPs arrays, it would be possible to observe whether genomic CNVs concur to retrotransposon CNVs alterations.

Moreover, several analyses are still ongoing to expand the results of my work. For instance, regarding the LIFE-seq technology, I'm testing the possibility of correctly genotyping L1 elements starting from medium coverages data (*i.e*. 30x per sample instead of 3000x). Whether true, this would highly increase the number of loadable samples within a single sequencing run (ideally 100 times more), thus allowing LIFE-seq to be extremely affordable for populational studies, where the analysis of hundreds of samples is required.

Further investigations could be conducted upon genomic WGS CNVs. On one hand, they might be used to validate LIFE-seq L1 genotyping data, since the comparison of LIFE-seq and *MELT deletion* results suggested that the latter was not strongly reliable. On the other hand, genomic CNVs from WGS data might also be used to validate the novel LIFE-seq CNV calling strategy, potentially providing an additional skill to this newly developed targeted sequencing approach.

Additionally, much of the future work will be focused on the observation of somatic MNVs, which experimental validation is currently undergoing in collaboration with Fabrizio Ecca from the IIT of Genova. Whether validated, MNVs discovery would strongly impact the current state of the art, demonstrating the presence of an additional type of somatic variants as well as stressing the need for further implementations to both available protocols for SNVs annotations and caller tools. The newly research field will include major biological open questions. Some examples of these questions are: ***i)*** Are somatic MNVs restricted to brain tissues? ***ii)*** Is APOBEC the actual source of brain-specific somatic MNVs? but most importantly, ***iii)*** Are somatic MNVs potentially related with the development of diseases, in particular neurodegenerative disorders?

Once again, the application of high-throughput technologies provided additional pieces of information in our understanding of the human biology, particularly regarding somaticism as in this case. Like in a mosaic, the data reported in this work represent just small tiles, which however will be extremely important to create the final picture.

References

Abyzov A, Mariani J, Palejev D, Zhang Y, Haney MS, Tomasini L, Ferrandino AF, Rosenberg Belmaker LA, Szekely A, Wilson M, Kocabas A, Calixto NE, Grigorenko EL, Huttner A, Chawarska K, Weissman S, Urban AE, Gerstein M, Vaccarino FM. 2012. Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature 492: 438–442.

Affymetrix Genome-Wide Human SNP Array 6.0 data sheet, 2007. http://tools.thermofisher.com/content/sfs/brochures/genomewide_snp6_datasheet.pdf

Alexandrov LB, Ju YS, Haase K, Loo PV, Martincorena I, Nik-Zainal S, Totoki Y, Fujimoto A, Nakagawa H, Shibata T, Campbell PJ, Vineis P, Phillips DH, Stratton MR. 2016. Mutational signatures associated with tobacco smoking in human cancer. Science 354: 618–622.

Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. 2013. Deciphering signatures of mutational processes operative in human cancer. Cell Rep. 3: 246–259.

Alsayyah A, ElMazoudy R, Al-Namshan M, Al-Jafary M, Alaqeel N. 2019. Chronic neurodegeneration by aflatoxin B1 depends on alterations of brain enzyme activity and immunoexpression of astrocyte in male rats. Ecotoxicol. Environ. Saf. 182: 109407.

Amir RE, Van den Veyver IB, Wan M, Tran CQ, Francke U, Zoghbi HY. 1999. Rett syndrome is caused by mutations in X-linked MECP2 , encoding methyl-CpG-binding protein 2. Nat. Genet. 23: 185–188.

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA, Nickerson DA, Schmidt JP, Sherry ST, Wang J, Wilson RK, Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, Lee S, Muzny D, Reid JG, Zhu Y, Wang J, Chang Y, Feng Q, Fang X, Guo X, Jian M, Jiang H, Jin X, Lan T, Li G, Li J, Li Y, Liu S, Liu X, Lu Y, Ma X, Tang M, Wang B, Wang G, Wu H, Wu R, Xu X, Yin Y, Zhang D, Zhang W, Zhao J, Zhao M, Zheng X, Lander ES, Altshuler DM, Gabriel SB, Gupta N, Gharani N, Toji LH, Gerry NP, Resch AM, Flicek P, Barker J, Clarke L, Gil L, Hunt SE, Kelman G, Kulesha E, Leinonen R, McLaren WM, Radhakrishnan R, Roa A, Smirnov D, Smith RE, Streeter I, Thormann A, Toneva I, Vaughan B, Zheng-Bradley X, Bentley DR, Grocock R, Humphray S, James T, Kingsbury Z, Lehrach H, et al. 2015. A global reference for human genetic variation. Nature 526: 68–74.

Azevedo FAC, Carvalho LRB, Grinberg LT, Farfel JM, Ferretti REL, Leite REP, Jacob Filho W, Lent R, Herculano-Houzel S. 2009. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. J. Comp. Neurol. 513: 532–541.

Babraham Bioinformatics. FastQC A Quality Control tool for High Throughput Sequence Data.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. 37: W202–W208.

Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, Talbot RT, Gustincich S, Freeman TC, Mattick JS, Hume DA, Heutink P, Carninci P, Jeddeloh JA, Faulkner GJ. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. Nature 479: 534–537.

Ballif BC, Rorem EA, Sundin K, Lincicum M, Gaskin S, Coppinger J, Kashork CD, Shaffer LG, Bejjani BA. 2006. Detection of low-level mosaicism by array CGH in routine diagnostic specimens. Am. J. Med. Genet. A. 140: 2757–2767.

Bar-Shira A, Rashi-Elkeles S, Zlochover L, Moyal L, Smorodinsky NI, Seger R, Shiloh Y. 2002. ATM-dependent activation of the gene encoding MAP kinase phosphatase 5 by radiomimetic DNA damage. Oncogene 21: 849–855.

Beck CR, Garcia-Perez JL, Badge RM, Moran JV. 2011. LINE-1 Elements in Structural Variation and Disease. Annu. Rev. Genomics Hum. Genet. 12: 187–215.

Beck JA, Poulter M, Campbell TA, Uphill JB, Adamson G, Geddes JF, Revesz T, Davis MB, Wood NW, Collinge J, Tabrizi SJ. 2004. Somatic and germline mosaicism in sporadic early-onset Alzheimer's disease. Hum. Mol. Genet. 13: 1219–1224.

Becker KG, Swergold GD, Ozato K, Thayer RE. 1993. Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element. Hum. Mol. Genet. 2: 1697–1702.

Behjati S, Huch M, van Boxtel R, Karthaus W, Wedge DC, Tamuri AU, Martincorena I, Petljak M, Alexandrov LB, Gundem G, Tarpey PS, Roerink S, Blokker J, Maddison M, Mudie L, Robinson B, Nik-Zainal S, Campbell P, Goldman N, van de Wetering M, Cuppen E, Clevers H, Stratton MR. 2014. Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature 513: 422–425.

Behm M, Öhman M. 2016. RNA Editing: A Contributor to Neuronal Dynamics in the Mammalian Brain. Trends Genet. TIG 32: 165–175.

Belancio VP, Hedges DJ, Deininger P. 2008. Mammalian non-LTR retrotransposons: For better or worse, in sickness and in health. Genome Res. 18: 343–358.

Belgnaoui SM, Gosden RG, Semmes OJ, Haoudi A. 2006. Human LINE-1 retrotransposon induces DNA damage and apoptosis in cancer cells. Cancer Cell Int. 6: 13.

Belshaw R, Dawson ALA, Woolven-Allen J, Redding J, Burt A, Tristem M. 2005. Genomewide Screening Reveals High Levels of Insertional Polymorphism in the Human Endogenous Retrovirus Family HERV-K(HML2): Implications for Present-Day Activity. J. Virol. 79: 12507–12514.

Bertram L, Lill CM, Tanzi RE. 2010. The genetics of Alzheimer disease: back to the future. Neuron 68: 270–281.

Bettens K, Sleegers K, Van Broeckhoven C. 2013. Genetic insights in Alzheimer's disease. Lancet Neurol. 12: 92–104.

Bianconi E, Piovesan A, Facchin F, Beraudi A, Casadei R, Frabetti F, Vitale L, Pelleri MC, Tassani S, Piva F, Perez-Amodio S, Strippoli P, Canaider S. 2013. An estimation of the number of cells in the human body. Ann. Hum. Biol. 40: 463–471.

Biesecker LG, Spinner NB. 2013. A genomic view of mosaicism and human disease. Nat. Rev. Genet. 14: 307–320.

Bitar M, Barry G. 2018. Multiple Innovations in Genetic and Epigenetic Mechanisms Cooperate to Underpin Human Brain Evolution. Mol. Biol. Evol. 35: 263–268.

Bogerd HP, Wiegand HL, Hulme AE, Garcia-Perez JL, O'Shea KS, Moran JV, Cullen BR. 2006. Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. Proc. Natl. Acad. Sci. U. S. A. 103: 8780–8785.

Bras J, Guerreiro R, Hardy J. 2012. Use of next-generation sequencing and other whole-genome strategies to dissect neurological disease. Nat. Rev. Neurosci. 13: 453–464.

Breckpot J, Thienpont B, Gewillig M, Allegaert K, Vermeesch JR, Devriendt K. 2012. Differences in Copy Number Variation between Discordant Monozygotic Twins as a Model for Exploring Chromosomal Mosaicism in Congenital Heart Defects. Mol. Syndromol. 2: 81–87.

Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH. 2003. Hot L1s account for the bulk of retrotransposition in the human population. Proc. Natl. Acad. Sci. U. S. A. 100: 5280–5285.

Bruder CEG, Piotrowski A, Gijsbers AACJ, Andersson R, Erickson S, Diaz de Ståhl T, Menzel U, Sandgren J, von Tell D, Poplawski A, Crowley M, Crasto C, Partridge EC, Tiwari H, Allison DB, Komorowski J, van Ommen G-JB, Boomsma DI, Pedersen NL, den Dunnen JT, Wirdefeldt K, Dumanski JP. 2008. Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am. J. Hum. Genet. 82: 763–771.

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.

Burke WD, Malik HS, Rich SM, Eickbush TH. 2002. Ancient lineages of non-LTR retrotransposons in the primitive eukaryote, Giardia lamblia. Mol. Biol. Evol. 19: 619–630.

Bushman DM, Chun J. 2013. The genomically mosaic brain: aneuploidy and more in neural diversity and disease. Semin. Cell Dev. Biol. 24: 357–369.

Bushman DM, Kaeser GE, Siddoway B, Westra JW, Rivera RR, Rehen SK, Yung YC, Chun J. 2015. Genomic mosaicism with increased amyloid precursor protein (APP) gene copy number in single neurons from sporadic Alzheimer's disease brains. eLife 4: e05116.

Cai X, Evrony GD, Lehmann HS, Elhosary PC, Mehta BK, Poduri A, Walsh CA. 2015. Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. Cell Rep. 10: 645.

Cai X, Evrony GD, Lehmann HS, Elhosary PC, Mehta BK, Poduri A, Walsh CA. 2014. Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. Cell Rep. 8: 1280–1289.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

Cancer Genome Atlas Research Network. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455: 1061–1068.

Carlson J, Li JZ, Zöllner S. 2018. Helmsman: fast and efficient mutation signature analysis for massive sequencing datasets. BMC Genomics 19: 845.

del Carmen Seleme M, Disson O, Robin S, Brun C, Teninges D, Bucheton A. 2005. In vivo RNA localization of I factor, a non-LTR retrotransposon, requires a cis-acting signal in ORF2 and ORF1 protein. Nucleic Acids Res. 33: 776–785.

Caspersson T, Zech L, Johansson C. 1970. Differential binding of alkylating fluorochromes in human chromosomes. Exp. Cell Res. 60: 315–319.

Castelnuovo M, Massone S, Tasso R, Fiorino G, Gatti M, Robello M, Gatta E, Berger A, Strub K, Florio T, Dieci G, Cancedda R, Pagano A. 2010. An Alu-like RNA promotes cell differentiation and reduces malignancy of human neuroblastoma cells. FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. 24: 4033–4046.

Cavalcante RG, Sartor MA. 2017. annotatr: genomic regions in context. Bioinforma. Oxf. Engl. 33: 2381–2383.

Cavallucci V, D'Amelio M, Cecconi F. 2012. Aβ Toxicity in Alzheimer's Disease. Mol. Neurobiol. 45: 366–378.

Chapuis J, Hansmannel F, Gistelinck M, Mounier A, Van Cauwenberghe C, Kolen KV, Geller F, Sottejeau Y, Harold D, Dourlen P, Grenier-Boley B, Kamatani Y, Delepine B, Demiautte F, Zelenika D, Zommer N, Hamdane M, Bellenguez C, Dartigues J-F, Hauw J-J, Letronne F, Ayral A-M, Sleegers K, Schellens A, Broeck LV, Engelborghs S, De Deyn PP, Vandenberghe R, O'Donovan M, Owen M, Epelbaum J, Mercken M, Karran E, Bantscheff M, Drewes G, Joberty G, Campion D, Octave J-N, Berr C, Lathrop M, Callaerts P, Mann D, Williams J, Buée L, Dewachter I, Van Broeckhoven C, Amouyel P,

Moechars D, Dermaut B, Lambert J-C, GERAD consortium. 2013. Increased expression of BIN1 mediates Alzheimer genetic risk by modulating tau pathology. Mol. Psychiatry 18: 1225–1234.

Chen L-L, Yang L. 2017. ALUternative Regulation for Gene Expression. Trends Cell Biol. 27: 480–490.

Cheung SW, Shaw CA, Scott DA, Patel A, Sahoo T, Bacino CA, Pursley A, Li J, Erickson R, Gropman AL, Miller DT, Seashore MR, Summers AM, Stankiewicz P, Chinault AC, Lupski JR, Beaudet AL, Sutton VR. 2007. Microarray-based CGH detects chromosomal mosaicism not revealed by conventional cytogenetics. Am. J. Med. Genet. A. 143A: 1679–1686.

Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. 2015. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat. Methods 12: 966–968.

Chiu Y-L, Witkowska HE, Hall SC, Santiago M, Soros VB, Esnault C, Heidmann T, Greene WC. 2006. High-molecular-mass APOBEC3G complexes restrict Alu retrotransposition. Proc. Natl. Acad. Sci. U. S. A. 103: 15588–15593.

Chow JC, Ciaudo C, Fazzari MJ, Mise N, Servant N, Glass JL, Attreed M, Avner P, Wutz A, Barillot E, Greally JM, Voinnet O, Heard E. 2010. LINE-1 Activity in Facultative Heterochromatin Formation during X Chromosome Inactivation. Cell 141: 956–969.

Christensen T. 2016. Human endogenous retroviruses in neurologic disease. APMIS Acta Pathol. Microbiol. Immunol. Scand. 124: 116–126.

Chu WM, Liu WM, Schmid CW. 1995. RNA polymerase III promoter and terminator elements affect Alu RNA expression. Nucleic Acids Res. 23: 1750–1757.

Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31: 213–219.

Conlin LK, Thiel BD, Bonnemann CG, Medne L, Ernst LM, Zackai EH, Deardorff MA, Krantz ID, Hakonarson H, Spinner NB. 2010. Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. Hum. Mol. Genet. 19: 1263–1275.

Conti A, Carnevali D, Bollati V, Fustinoni S, Pellegrini M, Dieci G. 2015. Identification of RNA polymerase III-transcribed Alu loci by computational screening of RNA-Seq data. Nucleic Acids Res. 43: 817–835.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. 10: 691–703.

Cost GJ, Feng Q, Jacquier A, Boeke JD. 2002. Human L1 element target-primed reverse transcription in vitro. EMBO J. 21: 5899–5910.

Coufal NG, Garcia-Perez JL, Peng GE, Marchetto MCN, Muotri AR, Mu Y, Carson CT, Macia A, Moran JV, Gage FH. 2011. Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. Proc. Natl. Acad. Sci. 108: 20382–20387.

Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, Morell M, O'Shea KS, Moran JV, Gage FH. 2009. L1 retrotransposition in human neural progenitor cells. Nature 460: 1127–1131.

Cruchaga C, Karch CM, Jin SC, Benitez BA, Cai Y, Guerreiro R, Harari O, Norton J, Budde J, Bertelsen S, Jeng AT, Cooper B, Skorupa T, Carrell D, Levitch D, Hsu S, Choi J, Ryten M, Sassi C, Bras J, Gibbs JR, Hernandez DG, Lupton MK, Powell J, Forabosco P, Ridge PG, Corcoran CD, Tschanz JT, Norton MC, Munger RG, Schmutz C, Leary M, Demirci FY, Bamne MN, Wang X, Lopez OL, Ganguli M, Medway C, Turton J, Lord J, Braae A, Barber I, Brown K, Pastor P, Lorenzo-Betancor O, Brkanac Z, Scott E, Topol E, Morgan K, Rogaeva E, Singleton AB, Hardy J, Kamboh MI, George-Hyslop PS, Cairns N, Morris JC, Kauwe JSK, Goate AM. 2014. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. Nature 505: 550–554.

Cruts M, Broeckhoven CV. 1998. Presenilin mutations in Alzheimer's disease. Hum. Mutat. 11: 183–190.

Cui C, Shu W, Li P. 2016. Fluorescence In situ Hybridization: Cell-Based Genetic Diagnostic and Research Applications. Front. Cell Dev. Biol. 4.

Dai L, Taylor MS, O'Donnell KA, Boeke JD. 2012. Poly(A) Binding Protein C1 Is Essential for Efficient L1 Retrotransposition and Affects L1 RNP Formation. Mol. Cell. Biol. 32: 4323–4336.

Damert A. 2015. Composite non-LTR retrotransposons in hominoid primates. Mob. Genet. Elem. 5: 67–71.

Damert A, Raiz J, Horn AV, Löwer J, Wang H, Xing J, Batzer MA, Löwer R, Schumann GG. 2009. 5'-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. Genome Res. 19: 1992–2008.

Dawson A, Hartswood E, Paterson T, Finnegan DJ. 1997. A LINE-like transposable element in Drosophila, the I factor, encodes a protein with properties similar to those of retroviral nucleocapsids. EMBO J. 16: 4448–4455.

Deininger PL, Moran JV, Batzer MA, Kazazian HH. 2003. Mobile elements and mammalian genome evolution. Curr. Opin. Genet. Dev. 13: 651–658.

DeMattos RB, Cirrito JR, Parsadanian M, May PC, O'Dell MA, Taylor JW, Harmony JAK, Aronow BJ, Bales KR, Paul SM, Holtzman DM. 2004. ApoE and Clusterin Cooperatively Suppress Aβ Levels and Deposition: Evidence that ApoE Regulates Extracellular Aβ Metabolism In Vivo. Neuron 41: 193–202.

Dewannieux M, Heidmann T. 2005. Role of poly(A) tail length in Alu retrotransposition. Genomics 86: 378–381.

Ding W, Lin L, Chen B, Dai J. 2006. L1 elements, processed pseudogenes and retrogenes in mammalian genomes. IUBMB Life 58: 677–685.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485: 376–380.

Erickson RP. 2010. Somatic gene mutation and human disease other than cancer: an update. Mutat. Res. 705: 96–106.

Erwin JA, Paquola ACM, Singer T, Gallina I, Novotny M, Quayle C, Bedrosian TA, Alves FIA, Butcher CR, Herdy JR, Sarkar A, Lasken RS, Muotri AR, Gage FH. 2016. L1-associated genomic regions are deleted in somatic cells of the healthy human brain. Nat. Neurosci. 19: 1583–1591.

Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, Lehmann HS, Parker JJ, Atabay KD, Gilmore EC, Poduri A, Park PJ, Walsh CA. 2012. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell 151: 483–496.

Evrony GD, Lee E, Park PJ, Walsh CA. 2016. Resolving rates of mutation in the brain using single-neuron genomics. eLife 5.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32: 3047–3048.

Ewing AD. 2015. Transposable element detection from whole genome sequence data. Mob. DNA 6: 24.

Fanning T, Singer M. 1987. The LINE-1 DNA sequences in four mammalian orders predict proteins that conserve homologies to retrovirus proteins. Nucleic Acids Res. 15: 2251–2260.

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, Waki K, Hornig N, Arakawa T, Takahashi H, Kawai J, Forrest ARR, Suzuki H, Hayashizaki Y, Hume DA, Orlando V, Grimmond SM, Carninci P. 2009. The regulated retrotransposon transcriptome of mammalian cells. Nat. Genet. 41: 563–571.

Feng Q, Moran JV, Kazazian HH, Boeke JD. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87: 905–916.

Fischer U, Struss A-K, Hemmer D, Michel A, Henn W, Steudel W-I, Meese E. 2001. PHF3 expression is frequently reduced in glioma. Cytogenet. Genome Res. 94: 131–136.

Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, Barnes I, Berry A, Bignell A, Carbonell Sala S, Chrast J, Cunningham F, Di Domenico T, Donaldson S, Fiddes IT, García Girón C, Gonzalez JM, Grego T, Hardy M, Hourlier T, Hunt T, Izuogu OG, Lagarde J, Martin FJ, Martínez L, Mohanan S, Muir P, Navarro FCP, Parker A, Pei B, Pozo F, Ruffier M, Schmitt BM, Stapleton E, Suner M-M, Sycheva I, Uszczynska-Ratajczak B, Xu J, Yates A, Zerbino D, Zhang Y, Aken B, Choudhary JS, Gerstein M, Guigó R, Hubbard TJP, Kellis M, Paten B, Reymond A, Tress ML, Flicek P. 2019. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 47: D766–D773.

Frumkin D, Wasserstrom A, Kaplan S, Feige U, Shapiro E. 2005. Genomic Variability within an Organism Exposes Its Cell Lineage Tree. PLOS Comput. Biol. 1: e50.

Gao Y, Sun Y, Frank KM, Dikkes P, Fujiwara Y, Seidl KJ, Sekiguchi JM, Rathbun GA, Swat W, Wang J, Bronson RT, Malynn BA, Bryans M, Zhu C, Chaudhuri J, Davidson L, Ferrini R, Stamato T, Orkin SH, Greenberg ME, Alt FW. 1998. A Critical Role for DNA End-Joining Proteins in Both Lymphogenesis and Neurogenesis. Cell 95: 891–902.

Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, Mills RE, 1000 Genomes Project Consortium, Devine SE. 2017. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. 27: 1916–1929.

Gasior SL, Wakeman TP, Xu B, Deininger PL. 2006. The Human LINE-1 Retrotransposon Creates DNA Double-strand Breaks. J. Mol. Biol. 357: 1383–1393.

Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, Fiske A, Pedersen NL. 2006. Role of genes and environments for explaining Alzheimer disease. Arch. Gen. Psychiatry 63: 168–174.

genome/bam-readcount. 2020. The McDonnell Genome Institute.

Gilbert N, Lutz S, Morrish TA, Moran JV. 2005. Multiple fates of L1 retrotransposition intermediates in cultured human cells. Mol. Cell. Biol. 25: 7780–7795.

Gilbert N, Lutz-Prigge S, Moran JV. 2002. Genomic deletions created upon LINE-1 retrotransposition. Cell 110: 315–325.

Gire V, Roux P, Wynford-Thomas D, Brondello J-M, Dulic V. 2004. DNA damage checkpoint kinase Chk2 triggers replicative senescence. EMBO J. 23: 2554–2563.

Gleeson JG, Minnerath S, Kuzniecky RI, Dobyns WB, Young ID, Ross ME, Walsh CA. 2000. Somatic and Germline Mosaic Mutations in the doublecortin Gene Are Associated with Variable Phenotypes. Am. J. Hum. Genet. 67: 574–581.

Glinsky GV. 2018. Contribution of transposable elements and distal enhancers to evolution of human-specific features of interphase chromatin architecture in embryonic stem cells. Chromosome Res. 26: 61–84.

Goodier JL. 2016. Restricting retrotransposons: a review. Mob. DNA 7.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17: 333–351.

Goriely A, Wilkie AOM. 2012. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. Am. J. Hum. Genet. 90: 175–200.

Gottlieb B, Beitel LK, Alvarado C, Trifiro MA. 2010. Selection and mutation in the “new” genetics: an emerging hypothesis. Hum. Genet. 127: 491–501.

Griciuc A, Serrano-Pozo A, Parrado AR, Lesinski AN, Asselin CN, Mullin K, Hooli B, Choi SH, Hyman BT, Tanzi RE. 2013. Alzheimer’s disease risk gene CD33 inhibits microglial uptake of amyloid beta. Neuron 78: 631–643.

Groesser L, Herschberger E, Ruetten A, Ruivenkamp C, Lopriore E, Zutt M, Langmann T, Singer S, Klingseisen L, Schneider-Brachert W, Toll A, Real FX, Landthaler M, Hafner C. 2012. Postzygotic HRAS and KRAS mutations cause nevus sebaceous and Schimmelpenning syndrome. Nat. Genet. 44: 783–787.

Guo C, Jeong H-H, Hsieh Y-C, Klein H-U, Bennett DA, De Jager PL, Liu Z, Shulman JM. 2018. Tau Activates Transposable Elements in Alzheimer’s Disease. Cell Rep. 23: 2874–2880.

Hafner C, López-Knowles E, Luis NM, Toll A, Baselga E, Fernández-Casado A, Hernández S, Ribé A, Mentzel T, Stoehr R, Hofstaedter F, Landthaler M, Vogt T, Pujol RM, Hartmann A, Real FX. 2007. Oncogenic PIK3CA mutations occur in epidermal nevi and seborrheic keratoses with a characteristic mutation pattern. Proc. Natl. Acad. Sci. U. S. A. 104: 13450–13454.

Hafner C, van Oers JMM, Vogt T, Landthaler M, Stoehr R, Blaszyk H, Hofstaedter F, Zwarthoff EC, Hartmann A. 2006. Mosaicism of activating FGFR3 mutations in human skin causes epidermal nevi. J. Clin. Invest. 116: 2201–2207.

Hafner C, Toll A, Gantner S, Mauerer A, Lurkin I, Acquadro F, Fernández-Casado A, Zwarthoff EC, Dietmaier W, Baselga E, Parera E, Vicente A, Casanova A, Cigudosa J, Mentzel T, Pujol RM, Landthaler M, Real FX. 2012. Keratinocytic epidermal nevi are associated with mosaic RAS mutations. J. Med. Genet. 49: 249–253.

Han JS. 2010. Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. Mob. DNA 1: 15.

Han K, Konkel MK, Xing J, Wang H, Lee J, Meyer TJ, Huang CT, Sandifer E, Hebert K, Barnes EW, Hubley R, Miller W, Smit AFA, Ullmer B, Batzer MA. 2007. Mobile DNA in Old World monkeys: a glimpse through the rhesus macaque genome. Science 316: 238–240.

Han K, Sen SK, Wang J, Callinan PA, Lee J, Cordaux R, Liang P, Batzer MA. 2005. Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. Nucleic Acids Res. 33: 4040–4052.

Hancks DC, Goodier JL, Mandal PK, Cheung LE, Kazazian HH. 2011. Retrotransposition of marked SVA elements by human L1s in cultured cells. Hum. Mol. Genet. 20: 3386–3400.

Hancks DC, Kazazian HH. 2016. Roles for retrotransposon insertions in human disease. Mob. DNA 7: 9.

Hancks DC, Kazazian HH. 2010. SVA retrotransposons: Evolution and genetic instability. Semin. Cancer Biol. 20: 234–245.

Haraksingh RR, Abyzov A, Urban AE. 2017. Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans. BMC Genomics 18: 321.

Havecker ER, Gao X, Voytas DF. 2004. The diversity of LTR retrotransposons. Genome Biol. 5: 225.

Heckmann JM, Low W-C, de Villiers C, Rutherfoord S, Vorster A, Rao H, Morris CM, Ramesar RS, Kalaria RN. 2004. Novel presenilin 1 mutation with profound neurofibrillary pathology in an indigenous Southern African family with early onset Alzheimer's disease. Brain 127: 133–142.

Hirschhorn K, Decker WH, Cooper HL. 1960. Human intersex with chromosome mosaicism of type XY/XO. Report of a case. N. Engl. J. Med. 263: 1044–1048.

Hohjoh H, Singer MF. 1996. Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. EMBO J. 15: 630–639.

Hohjoh H, Singer MF. 1997. Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. EMBO J. 16: 6034–6043.

Holtzman DM, Morris JC, Goate AM. 2011. Alzheimer's Disease: The Challenge of the Second Century. Sci. Transl. Med. 3: 77sr1-77sr1.

Hook EB. 1977. Exclusion of chromosomal mosaicism: tables of 90%, 95% and 99% confidence limits and comments on use. Am. J. Hum. Genet. 29: 94–97.

Huda A, Mariño-Ramírez L, Landsman D, Jordan IK. 2009. Repetitive DNA elements, nucleosome binding and human gene expression. Gene 436: 12–22.

International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, Hemsley FM, Jennings JL, Kerr D, Klatt P, Kolar P, Kusada J, Lane DP, Laplace F, Youyong L, Nettekoven G, Ozenberger B, Peterson J, Rao TS, Remacle J, Schafer AJ, Shibata T, Stratton MR, Vockley JG, Watanabe K, Yang H, Yuen MMF, Knoppers BM, Bobrow M, Cambon-Thomsen A, Dressler LG, Dyke SOM, Joly Y, Kato K, Kennedy KL, Nicolás P, Parker MJ, Rial-Sebbag E, Romeo-Casabona CM, Shaw KM, Wallace S, Wiesner GL, Zeps N, Lichter P, Biankin AV, Chabannon C, Chin L, Clément B, de Alava E, Degos F, Ferguson ML, Geary P, Hayes DN, Hudson TJ, Johns AL, Kasprzyk A, Nakagawa H, Penny R, Piris MA, Sarin R, Scarpa A, Shibata T, van de Vijver M, Futreal PA, Aburatani H, Bayés M, Botwell DDL, Campbell PJ, Estivill X, Gerhard DS, Grimmond SM, Gut I, Hirst M, López-Otín C, Majumder P, Marra M, McPherson JD, Nakagawa H, Ning Z, Puente XS, Ruan Y, Shibata T, Stratton MR, Stunnenberg HG, Swerdlow H, Velculescu VE, Wilson RK, Xue HH, Yang L, Spellman PT, Bader GD, et al. 2010. International network of cancer genome projects. Nature 464: 993–998.

Jacobs KB, Yeager M, Zhou W, Wacholder S, Wang Z, Rodriguez-Santiago B, Hutchinson A, Deng X, Liu C, Horner M-J, Cullen M, Epstein CG, Burdett L, Dean MC, Chatterjee N, Sampson J, Chung CC, Kovaks J, Gapstur SM, Stevens VL, Teras LT, Gaudet MM, Albanes D, Weinstein SJ, Virtamo J, Taylor PR, Freedman ND, Abnet CC, Goldstein AM, Hu N, Yu K, Yuan J-M, Liao L, Ding T, Qiao Y-L, Gao Y-T, Koh W-P, Xiang Y-B, Tang Z-Z, Fan J-H, Aldrich MC, Amos C, Blot WJ, Bock CH, Gillanders EM, Harris CC, Haiman CA, Henderson BE, Kolonel LN, Le Marchand L, McNeill LH, Rybicki BA, Schwartz AG, Signorello LB, Spitz MR, Wiencke JK, Wrensch M, Wu X, Zanetti KA, Ziegler RG, Figueroa JD, Garcia-Closas M, Malats N, Marenne G, Prokunina-Olsson L, Baris D, Schwenn M, Johnson A, Landi MT, Goldin L, Consonni D, Bertazzi PA, Rotunno M, Rajaraman P, Andersson U, Beane Freeman LE, Berg CD, Buring JE, Butler MA, Carreon T, Feychting M, Ahlbom A, Gaziano JM, Giles GG, Hallmans G, Hankinson SE, Hartge P, Henriksson R, Inskip PD, Johansen C, Landgren A, McKean-Cowdin R, Michaud DS, Melin BS, Peters U, Ruder AM, Sesso HD, Severi G, et al. 2012. Detectable clonal mosaicism and its relationship to aging and cancer. Nat. Genet. 44: 651–658.

Jamuar SS, Lam A-TN, Kircher M, D'Gama AM, Wang J, Barry BJ, Zhang X, Hill RS, Partlow JN, Rozzo A, Servattalab S, Mehta BK, Topcu M, Amrom D, Andermann E, Dan B, Parrini E, Guerrini R, Scheffer IE, Berkovic SF, Leventer RJ, Shen Y, Wu BL, Barkovich AJ, Sahin M, Chang BS, Bamshad M, Nickerson DA, Shendure J, Poduri A, Yu TW, Walsh CA. 2014. Somatic Mutations in Cerebral Cortical Malformations. N. Engl. J. Med. 371: 733–743.

Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, Marzluff WF, Sharpless NE. 2013. Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA 19: 141–157.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258: 818–821.

Kaneko-Ishino T, Irie M, Ishino F. 2017. Mammalian-Specific Traits Generated by LTR Retrotransposon-Derived SIRH Genes. In: Pontarotti P, editor. Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts. Cham: Springer International Publishing, p 129–145.

Kano H, Godoy I, Courtney C, Vetter MR, Gerton GL, Ostertag EM, Kazazian HH. 2009. L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. Genes Dev. 23: 1303–1312.

Kaplanis J, Akawi N, Gallone G, McRae JF, Prigmore E, Wright CF, Fitzpatrick DR, Firth HV, Barrett JC, Hurles ME, on behalf of the Deciphering Developmental Disorders study. 2019. Exome-wide assessment of the functional impact and pathogenicity of multinucleotide mutations. Genome Res. 29: 1047–1056.

Karch CM, Cruchaga C, Goate AM. 2014. Alzheimer's Disease Genetics: From the Bench to the Clinic. Neuron 83: 11–26.

Karch CM, Goate AM. 2015. Alzheimer's Disease Risk Genes and Mechanisms of Disease Pathogenesis. Biol. Psychiatry 77: 43–51.

Karch CM, Jeng AT, Nowotny P, Cady J, Cruchaga C, Goate AM. 2012. Expression of novel Alzheimer's disease risk genes in control and Alzheimer's disease brains. PloS One 7: e50976.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32: D493–D496.

Kazazian HH. 2011. Mobile DNA transposition in somatic cells. BMC Biol. 9: 62.

Kazazian HH, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332: 164–166.

Khan A, Mathelier A. 2017. Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. BMC Bioinformatics 18: 287.

Kim J, Basak JM, Holtzman DM. 2009a. The Role of Apolipoprotein E in Alzheimer's Disease. Neuron 63: 287–303.

Kim J, Castellano JM, Jiang H, Basak JM, Parsadanian M, Pham V, Mason SM, Paul SM, Holtzman DM. 2009b. Overexpression of low-density lipoprotein receptor in the brain markedly inhibits amyloid deposition and increases extracellular A beta clearance. Neuron 64: 632–644.

Kim WS, Guillemin GJ, Glaros EN, Lim CK, Garner B. 2006. Quantitation of ATP-binding cassette subfamily-A transporter gene expression in primary human brain cells. NeuroReport 17: 891–896.

Kim WS, Li H, Ruberu K, Chan S, Elliott DA, Low JK, Cheng D, Karl T, Garner B. 2013. Deletion of Abca7 Increases Cerebral Amyloid-β Accumulation in the J20 Mouse Model of Alzheimer's Disease. J. Neurosci. 33: 4387–4394.

Kim WS, Weickert CS, Garner B. 2008. Role of ATP-binding cassette transporters in brain lipid transport and neurological disease. J. Neurochem. 104: 1145–1166.

Kingsbury MA, Yung YC, Peterson SE, Westra JW, Chun J. 2006. Aneuploidy in the normal and diseased brain. Cell. Mol. Life Sci. CMLS 63: 2626–2641.

Kinsey JA. 1990. Tad, a LINE-like transposable element of Neurospora, can transpose between nuclei in heterokaryons. Genetics 126: 317–323.

Kobayashi K, Nakahori Y, Miyake M, Matsumura K, Kondo-Iida E, Nomura Y, Segawa M, Yoshioka M, Saito K, Osawa M, Hamano K, Sakakihara Y, Nonaka I, Nakagome Y, Kanazawa I, Nakamura Y, Tokunaga K, Toda T. 1998. An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. Nature 394: 388–392.

Koito A, Ikeda T. 2013. Intrinsic immunity against retrotransposons by APOBEC cytidine deaminases. Front. Microbiol. 4.

Kolosha VO, Martin SL. 2003. High-affinity, Non-sequence-specific RNA Binding by the Open Reading Frame 1 (ORF1) Protein from Long Interspersed Nuclear Element 1 (LINE-1). J. Biol. Chem. 278: 8112–8117.

Kramerov DA, Vassetzky NS. 2011. Origin and evolution of SINEs in eukaryotic genomes. Heredity 107: 487–495.

Krug L, Chatterjee N, Borges-Monroy R, Hearn S, Liao W-W, Morrill K, Prazak L, Rozhkov N, Theodorou D, Hammell M, Dubnau J. 2017. Retrotransposon activation contributes to neurodegeneration in a Drosophila TDP-43 model of ALS. PLoS Genet. 13: e1006635.

Krych-Goldberg M, Moulds JM, Atkinson JP. 2002. Human complement receptor type 1 (CR1) binds to a major malarial adhesin. Trends Mol. Med. 8: 531–537.

Kubo S, Seleme MDC, Soifer HS, Perez JLG, Moran JV, Kazazian HH, Kasahara N. 2006. L1 retrotransposition in nondividing and primary human somatic cells. Proc. Natl. Acad. Sci. U. S. A. 103: 8036–8041.

Kurek KC, Luks VL, Ayturk UM, Alomari AI, Fishman SJ, Spencer SA, Mulliken JB, Bowen ME, Yamamoto GL, Kozakewich HPW, Warman ML. 2012. Somatic mosaic activating mutations in PIK3CA cause CLOVES syndrome. Am. J. Hum. Genet. 90: 1108–1115.

Kurnosov AA, Ustyugova SV, Nazarov VI, Minervina AA, Komkov AY, Shugay M, Pogorelyy MV, Khodosevich KV, Mamedov IZ, Lebedev YB. 2015. The Evidence for Increased L1 Activity in the Site of Human Adult Brain Neurogenesis. PLOS ONE 10: e0117854.

Kuwabara T, Hsieh J, Muotri A, Yeo G, Warashina M, Lie DC, Moore L, Nakashima K, Asashima M, Gage FH. 2009. Wnt-mediated activation of NeuroD1 and retro-elements during adult neurogenesis. Nat. Neurosci. 12: 1097–1105.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, 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 JP, 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 JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Larsen PA, Lutz MW, Hunnicutt KE, Mihovilovic M, Saunders AM, Yoder AD, Roses AD. 2017. The Alu neurodegeneration hypothesis: A primate-specific mechanism for neuronal transcription noise,

mitochondrial dysfunction, and manifestation of neurodegenerative disease. Alzheimers Dement. J. Alzheimers Assoc. 13: 828–838.

Laurie CC, Laurie CA, Rice K, Doheny KF, Zelnick LR, McHugh CP, Ling H, Hetrick KN, Pugh EW, Amos C, Wei Q, Wang L, Lee JE, Barnes KC, Hansel NN, Mathias R, Daley D, Beaty TH, Scott AF, Ruczinski I, Scharpf RB, Bierut LJ, Hartz SM, Landi MT, Freedman ND, Goldin LR, Ginsburg D, Li J, Desch KC, Strom SS, Blot WJ, Signorello LB, Ingles SA, Chanock SJ, Berndt SI, Le Marchand L, Henderson BE, Monroe KR, Heit JA, de Andrade M, Armasu SM, Regnier C, Lowe WL, Hayes MG, Marazita ML, Feingold E, Murray JC, Melbye M, Feenstra B, Kang JH, Wiggs JL, Jarvik GP, McDavid AN, Seshan VE, Mirel DB, Crenshaw A, Sharopova N, Wise A, Shen J, Crosslin DR, Levine DM, Zheng X, Udren JI, Bennett S, Nelson SC, Gogarten SM, Conomos MP, Heagerty P, Manolio T, Pasquale LR, Haiman CA, Caporaso N, Weir BS. 2012. Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat. Genet. 44: 642–650.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 15: R84.

Lee JH, Huynh M, Silhavy JL, Kim S, Dixon-Salazar T, Heiberg A, Scott E, Bafna V, Hill KJ, Collazo A, Funari V, Russ C, Gabriel SB, Mathern GW, Gleeson JG. 2012. De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. Nat. Genet. 44: 941–945.

Lee T-H, Guo H, Wang X, Kim C, Paterson AH. 2014. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics 15: 162.

Lehnert S, Loo PV, Thilakarathne PJ, Marynen P, Verbeke G, Schuit FC. 2009. Evidence for Co-Evolution between Human MicroRNAs and Alu-Repeats. PLOS ONE 4: e4456.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won H-H, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation Consortium. 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536: 285–291.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinforma. Oxf. Engl. 27: 2987–2993.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. 26: 589–595.

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.

Li H, Xing L, Zhang M, Wang J, Zheng N. 2018. The Toxic Effects of Aflatoxin B1 and Aflatoxin M1 on Kidney through Regulating L-Proline and Downstream Apoptosis. BioMed Res. Int. 2018: e9074861.

Li JB, Church GM. 2013. Deciphering the functions and regulation of brain-enriched A-to-I RNA editing. Nat. Neurosci. 16: 1518–1522.

Li MJ, Liu Z, Wang P, Wong MP, Nelson MR, Kocher J-PA, Yeager M, Sham PC, Chanock SJ, Xia Z, Wang J. 2016. GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. Nucleic Acids Res. 44: D869-876.

Lindhurst MJ, Parker VER, Payne F, Sapp JC, Rudge S, Harris J, Witkowski AM, Zhang Q, Groeneveld MP, Scott CE, Daly A, Huson SM, Tosi LL, Cunningham ML, Darling TN, Geer J, Gucev Z, Sutton VR, Tziotzios C, Dixon AK, Helliwell T, O'Rahilly S, Savage DB, Wakelam MJO, Barroso I, Biesecker LG, Semple RK. 2012. Mosaic overgrowth with fibroadipose hyperplasia is caused by somatic activating mutations in PIK3CA. Nat. Genet. 44: 928–933.

Lindhurst MJ, Sapp JC, Teer JK, Johnston JJ, Finn EM, Peters K, Turner J, Cannons JL, Bick D, Blakemore L, Blumhorst C, Brockmann K, Calder P, Cherman N, Deardorff MA, Everman DB, Golas G, Greenstein RM, Kato BM, Keppler-Noreuil KM, Kuznetsov SA, Miyamoto RT, Newman K, Ng D, O'Brien K, Rothenberg S, Schwartzentruber DJ, Singhal V, Tirabosco R, Upton J, Wientroub S, Zackai EH, Hoag K, Whitewood-Neal T, Robey PG, Schwartzberg PL, Darling TN, Tosi LL, Mullikin JC, Biesecker LG. 2011. A Mosaic Activating Mutation in AKT1 Associated with the Proteus Syndrome. N. Engl. J. Med. 365: 611–619.

Liu D, Niu Z-X. 2009. The structure, genetic polymorphisms, expression and biological functions of complement receptor type 1 (CR1/CD35). Immunopharmacol. Immunotoxicol. 31: 524–535.

Lodato MA, Rodin RE, Bohrson CL, Coulter ME, Barton AR, Kwon M, Sherman MA, Vitzthum CM, Luquette LJ, Yandava CN, Yang P, Chittenden TW, Hatem NE, Ryu SC, Woodworth MB, Park PJ, Walsh CA. 2018. Aging and neurodegeneration are associated with increased mutations in single human neurons. Science 359: 555–559.

Lovšin N, Peterlin BM. 2009. APOBEC3 Proteins Inhibit LINE-1 Retrotransposition in the Absence of ORF1p Binding. Ann. N. Y. Acad. Sci. 1178: 268–275.

Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell 72: 595–605.

Lupski JR. 2015. Structural variation mutagenesis of the human genome: Impact on disease and evolution. Environ. Mol. Mutagen. 56: 419–436.

Lynch M. 2010. Rate, molecular spectrum, and consequences of human mutation. Proc. Natl. Acad. Sci. 107: 961–968.

Lyon MF. 1998. X-Chromosome inactivation: a repeat hypothesis. Cytogenet. Genome Res. 80: 133–137.

Maden M. 2007. Retinoic acid in the development, regeneration and maintenance of the nervous system. Nat. Rev. Neurosci. 8: 755–765.

Malik M, Simpson JF, Parikh I, Wilfred BR, Fardo DW, Nelson PT, Estus S. 2013. CD33 Alzheimer's risk-altering polymorphism, CD33 expression, and exon 2 splicing. J. Neurosci. Off. J. Soc. Neurosci. 33: 13320–13325.

Mangeat B, Turelli P, Caron G, Friedli M, Perrin L, Trono D. 2003. Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. Nature 424: 99–103.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. Trends Genet. TIG 24: 133–141.

Martin SL. 1991. Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells. Mol. Cell. Biol. 11: 4804–4807.

Martin SL, Bushman FD. 2001. Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. Mol. Cell. Biol. 21: 467–475.

Martínez A, Otal R, Sieber B-A, Ibáñez C, Soriano E. 2005. Disruption of ephrin-A/EphA binding alters synaptogenesis and neural connectivity in the hippocampus. Neuroscience 135: 451–461.

Mathias SL, Scott AF, Kazazian HH, Boeke JD, Gabriel A. 1991. Reverse transcriptase encoded by a human transposable element. Science 254: 1808–1810.

Mayakonda A, Lin D-C, Assenov Y, Plass C, Koeffler HP. 2018. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 28: 1747–1756.

McCombie WR, McPherson JD, Mardis ER. 2019. Next-Generation Sequencing Technologies. Cold Spring Harb. Perspect. Med. 9: a036798.

McConnell MJ, Moran JV, Abyzov A, Akbarian S, Bae T, Cortes-Ciriano I, Erwin JA, Fasching L, Flasch DA, Freed D, Ganz J, Jaffe AE, Kwan KY, Kwon M, Lodato MA, Mills RE, Paquola ACM, Rodin RE, Rosenbluh C, Sestan N, Sherman MA, Shin JH, Song S, Straub RE, Thorpe J, Weinberger DR, Urban AE, Zhou B, Gage FH, Lehner T, Senthil G, Walsh CA, Chess A, Courchesne E, Gleeson JG, Kidd JM, Park PJ, Pevsner J, Vaccarino FM, Brain Somatic Mosaicism Network. 2017. Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. Science 356.

McCulloch SD, Kunkel TA. 2008. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. Cell Res. 18: 148–161.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. Genome Biol. 17: 122.

Merino GA, Murua YA, Fresno C, Sendoya JM, Golubicki M, Iseas S, Coraglio M, Podhajcer OL, Llera AS, Fernández EA. 2017. TarSeqQC: Quality control on targeted sequencing experiments in R. Hum. Mutat. 38: 494–502.

Mir AA, Philippe C, Cristofari G. 2015. euL1db: the European database of L1HS retrotransposon insertions in humans. Nucleic Acids Res. 43: D43-47.

Moir RD, Lathe R, Tanzi RE. 2018. The antimicrobial protection hypothesis of Alzheimer's disease. Alzheimers Dement. 14: 1602–1614.

Moran JV, DeBerardinis RJ, Kazazian HH. 1999. Exon shuffling by L1 retrotransposition. Science 283: 1530–1534.

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH. 1996. High frequency retrotransposition in cultured mammalian cells. Cell 87: 917–927.

Morris JC, Roe CM, Xiong C, Fagan AM, Goate AM, Holtzman DM, Mintun MA. 2010. APOE predicts amyloid-beta but not tau Alzheimer pathology in cognitively normal aging. Ann. Neurol. 67: 122–131.

Muotri AR, Chu VT, Marchetto MCN, Deng W, Moran JV, Gage FH. 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435: 903–910.

Muotri AR, Marchetto MCN, Coufal NG, Oefner R, Yeo G, Nakashima K, Gage FH. 2010. L1 retrotransposition in neurons is modulated by MeCP2. Nature 468: 443–446.

Myers JS, Vincent BJ, Udall H, Watkins WS, Morrish TA, Kilroy GE, Swergold GD, Henke J, Henke L, Moran JV, Jorde LB, Batzer MA. 2002. A Comprehensive Analysis of Recently Integrated Human Ta L1 Elements. Am. J. Hum. Genet. 71: 312–326.

Nabi R, Serajee FJ, Chugani DC, Zhong H, Huq AHMM. 2004. Association of tryptophan 2,3 dioxygenase gene polymorphism with autism. Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet. 125B: 63–68.

Nigumann P, Redik K, Mätlik K, Speek M. 2002. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. Genomics 79: 628–634.

Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, Martincorena I, Alexandrov LB, Martin S, Wedge DC, Van Loo P, Ju YS, Smid M, Brinkman AB, Morganella S, Aure MR, Lingjærde OC, Langerød A, Ringnér M, Ahn S-M, Boyault S, Brock JE, Broeks A, Butler A, Desmedt C, Dirix L, Dronov S, Fatima A, Foekens JA, Gerstung M, Hooijer GKJ, Jang SJ, Jones DR, Kim H-Y, King TA, Krishnamurthy S, Lee HJ, Lee J-Y, Li Y, McLaren S, Menzies A, Mustonen V, O'Meara S, Pauporté I, Pivot X, Purdie CA, Raine K, Ramakrishnan K, Rodríguez-González FG, Romieu G, Sieuwerts AM,

Simpson PT, Shepherd R, Stebbings L, Stefansson OA, Teague J, Tommasi S, Treilleux I, Van den Eynden GG, Vermeulen P, Vincent-Salomon A, Yates L, Caldas C, van't Veer L, Tutt A, Knappskog S, Tan BKT, Jonkers J, Borg Å, Ueno NT, Sotiriou C, Viari A, Futreal PA, Campbell PJ, Span PN, Van Laere S, Lakhani SR, Eyfjord JE, Thompson AM, Birney E, Stunnenberg HG, van de Vijver MJ, Martens JWM, Børresen-Dale A-L, Richardson AL, Kong G, Thomas G, Stratton MR. 2016. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534: 47–54.

Oliver KR, Greene WK. 2011. Mobile DNA and the TE-Thrust hypothesis: supporting evidence from the primates. Mob. DNA 2: 8.

Ostertag EM, Goodier JL, Zhang Y, Kazazian HH. 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. Am. J. Hum. Genet. 73: 1444–1451.

Ostertag EM, Kazazian Jr HH. 2001. Biology of Mammalian L1 Retrotransposons. Annu. Rev. Genet. 35: 501–538.

Pagnamenta AT, Lise S, Harrison V, Stewart H, Jayawant S, Quaghebeur G, Deng AT, Murphy VE, Sadighi Akha E, Rimmer A, Mathieson I, Knight SJL, Kini U, Taylor JC, Keays DA. 2012. Exome sequencing can detect pathogenic mosaic mutations present at low allele frequencies. J. Hum. Genet. 57: 70–72.

Parcerisas A, Rubio SE, Muhaisen A, Gómez-Ramos A, Pujadas L, Puiggros M, Rossi D, Ureña J, Burgaya F, Pascual M, Torrents D, Rábano A, Ávila J, Soriano E. 2014. Somatic Signature of Brain-Specific Single Nucleotide Variations in Sporadic Alzheimer's Disease. J. Alzheimers Dis. 42: 1357–1382.

Park JS, Lee J, Jung ES, Kim M-H, Kim IB, Son H, Kim S, Kim S, Park YM, Mook-Jung I, Yu SJ, Lee JH. 2019. Brain somatic mutations observed in Alzheimer's disease associated with aging and dysregulation of tau phosphorylation. Nat. Commun. 10: 1–12.

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

Petersen KR, Streett DA, Gerritsen AT, Hunter SS, Settles ML. 2015. Super deduper, fast PCR duplicate detection in fastq files. In: Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics BCB '15. Atlanta, Georgia: Association for Computing Machinery, p 491–492.

Pinkel D, Gray JW, Trask B, van den Engh G, Fuscoe J, van Dekken H. 1986. Cytogenetic analysis by in situ hybridization with fluorescently labeled nucleic acid probes. Cold Spring Harb. Symp. Quant. Biol. 51 Pt 1: 151–157.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat. Genet. 20: 207–211.

Piotrowski A, Bruder CEG, Andersson R, Diaz de Ståhl T, Menzel U, Sandgren J, Poplawski A, von Tell D, Crasto C, Bogdan A, Bartoszewski R, Bebok Z, Krzyzanowski M, Jankowski Z, Partridge EC, Komorowski J, Dumanski JP. 2008. Somatic mosaicism for copy number variation in differentiated human tissues. Hum. Mutat. 29: 1118–1124.

Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin M-L, Ordóñez GR, Bignell GR, Ye K, Alipaz J, Bauer MJ, Beare D, Butler A, Carter RJ, Chen L, Cox AJ, Edkins S, Kokko-Gonzales PI, Gormley NA, Grocock RJ, Haudenschild CD, Hims MM, James T, Jia M, Kingsbury Z, Leroy C, Marshall J, Menzies A, Mudie LJ, Ning Z, Royce T, Schulz-Trieglaff OB, Spiridou A, Stebbings LA, Szajkowski L, Teague J, Williamson D, Chin L, Ross MT, Campbell PJ, Bentley DR, Futreal PA, Stratton MR. 2010. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463: 191–196.

Poduri A, Evrony GD, Cai X, Elhosary PC, Beroukhim R, Lehtinen MK, Hills LB, Heinzen EL, Hill A, Hill RS, Barry BJ, Bourgeois BFD, Riviello JJ, Barkovich AJ, Black PM, Ligon KL, Walsh CA. 2012. Somatic activation of AKT3 causes hemispheric developmental brain malformations. Neuron 74: 41–48.

Prak ET, Kazazian HH. 2000. Mobile elements and the human genome. Nat. Rev. Genet. 1: 134–144.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–575.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

Qureshi H, Hamid SS, Ali SS, Anwar J, Siddiqui AA, Khan NA. 2015. Cytotoxic effects of aflatoxin B1 on human brain microvascular endothelial cells of the blood-brain barrier. Med. Mycol. 53: 409–416.

R Core Team. 2012. R: A Language and Environment for Statistical Computing.

Rademakers R, Cruts M, Sleegers K, Dermaut B, Theuns J, Aulchenko Y, Weckx S, De Pooter T, Van den Broeck M, Corsmit E, De Rijk P, Del-Favero J, van Swieten J, van Duijn CM, Van Broeckhoven C. 2005. Linkage and Association Studies Identify a Novel Locus for Alzheimer Disease at 7q36 in a Dutch Population-Based Sample. Am. J. Hum. Genet. 77: 643–652.

Rademakers R, Cruts M, Van Broeckhoven C. 3. Genetics of Early-Onset Alzheimer Dementia. ScientificWorldJournal 3: etsw.2003.39.

Raiz J, Damert A, Chira S, Held U, Klawitter S, Hamdorf M, Löwer J, Strätling WH, Löwer R, Schumann GG. 2012. The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. Nucleic Acids Res. 40: 1666–1683.

Raux G, Guyant-Maréchal L, Martin C, Bou J, Penet C, Brice A, Hannequin D, Frebourg T, Campion D. 2005. Molecular diagnosis of autosomal dominant early onset Alzheimer's disease: an update. J. Med. Genet. 42: 793–795.

Reitz C, Mayeux R. 2013. TREM2 and Neurodegenerative Disease. N. Engl. J. Med. 369: 1564–1565.

Ren G, Vajjhala P, Lee JS, Winsor B, Munn AL. 2006. The BAR domain proteins: molding membranes in fission, fusion, and phagy. Microbiol. Mol. Biol. Rev. MMBR 70: 37–120.

Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. 2019. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 47: D886–D894.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16: 276–277.

Richards RI, Robertson SA, O'Keefe LV, Fornarino D, Scott A, Lardelli M, Baune BT. 2016. The Enemy within: Innate Surveillance-Mediated Cell Death, the Common Mechanism of Neurodegenerative Disease. Front. Neurosci. 10: 193.

Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Perez JL, Moran JV. 2015. The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. Microbiol. Spectr. 3: MDNA3-0061–2014.

Ridge PG, Ebbert MTW, Kauwe JSK. 2013. Genetics of Alzheimer's Disease. BioMed Res. Int. 2013: e254954.

Rivière J-B, Mirzaa GM, O'Roak BJ, Beddaoui M, Alcantara D, Conway RL, St-Onge J, Schwartzentruber JA, Gripp KW, Nikkel SM, Worthylake T, Sullivan CT, Ward TR, Butler HE, Kramer NA, Albrecht B, Armour CM, Armstrong L, Caluseriu O, Cytrynbaum C, Drolet BA, Innes AM, Lauzon JL, Lin AE, Mancini GMS, Meschino WS, Reggin JD, Saggar AK, Lerman-Sagie T, Uyanik G, Weksberg R, Zirn B, Beaulieu CL, Finding of Rare Disease Genes (FORGE) Canada Consortium, Majewski J, Bulman DE, O'Driscoll M, Shendure J, Graham JM, Boycott KM, Dobyns WB. 2012. De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. Nat. Genet. 44: 934–940.

Rizzi F, Caccamo AE, Belloni L, Bettuzzi S. 2009. Clusterin is a short half-life, poly-ubiquitinated protein, which controls the fate of prostate cancer cells. J. Cell. Physiol. 219: 314–323.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative Genomics Viewer. Nat. Biotechnol. 29: 24–26.

Rodríguez-Santiago B, Malats N, Rothman N, Armengol L, Garcia-Closas M, Kogevinas M, Villa O, Hutchinson A, Earl J, Marenne G, Jacobs K, Rico D, Tardón A, Carrato A, Thomas G, Valencia A, Silverman D, Real FX, Chanock SJ, Pérez-Jurado LA. 2010. Mosaic Uniparental Disomies and Aneuploidies as Large Structural Variants of the Human Genome. Am. J. Hum. Genet. 87: 129–138.

Rogaeva E, Meng Y, Lee JH, Gu Y, Kawarai T, Zou F, Katayama T, Baldwin CT, Cheng R, Hasegawa H, Chen F, Shibata N, Lunetta KL, Pardossi-Piquard R, Bohm C, Wakutani Y, Cupples LA, Cuenco KT, Green RC, Pinessi L, Rainero I, Sorbi S, Bruni A, Duara R, Friedland RP, Inzelberg R, Hampe W, Bujo H, Song Y-Q, Andersen OM, Willnow TE, Graff-Radford N, Petersen RC, Dickson D, Der SD, Fraser PE, Schmitt-Ulms G, Younkin S, Mayeux R, Farrer LA, St George-Hyslop P. 2007. The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. Nat. Genet. 39: 168–177.

Rogers J, Li R, Mastroeni D, Grover A, Leonard B, Ahern G, Cao P, Kolody H, Vedders L, Kolb WP, Sabbagh M. 2006. Peripheral clearance of amyloid β peptide by complement C3-dependent adherence to erythrocytes. Neurobiol. Aging 27: 1733–1739.

Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. 2016. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. 17: 31.

Rousseau F, Bonaventure J, Legeai-Mallet L, Pelet A, Rozet JM, Maroteaux P, Le Merrer M, Munnich A. 1994. Mutations in the gene encoding fibroblast growth factor receptor-3 in achondroplasia. Nature 371: 252–254.

Roy-Engel AM, Salem A-H, Oyeniran OO, Deininger L, Hedges DJ, Kilroy GE, Batzer MA, Deininger PL. 2002. Active Alu element “A-tails”: size does matter. Genome Res. 12: 1333–1344.

Ryan NJ, Hogan GR, Hayes AW, Unger PD, Siraj MY. 1979. Aflatoxin B1; its role in the etiology of Reye’s syndrome. Pediatrics 64: 71–75.

Ryan NS, Rossor MN. 2010. Correlating familial Alzheimer’s disease gene mutations with clinical phenotype. Biomark. Med. 4: 99–112.

Sala Frigerio* C, Piscopo* P, Calabrese E, Crestini A, MalvezziCampeggi L, Civita di Fava R, Fogliarino S, Albani D, Marcon G, Cherchi R, Piras R, Forloni G, Confaloni A. 2005. PEN–2 gene mutation in a familialAlzheimer’s disease case. J. Neurol. 252: 1033–1036.

Saleh A, Macia A, Muotri AR. 2019. Transposable Elements, Inflammation, and Neurological Disease. Front. Neurol. 10: 894.

Sandmeyer SB, Aye M, Menees T. 2002. Ty3, a Position-Specific, Gypsy-Like Element in Saccharomyces cerevisiae. Mob. DNA II: 663–683.

Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, DeBerardinis RJ, Gabriel A, Swergold GD, Kazazian HH. 1997. Many human L1 elements are capable of retrotransposition. Nat. Genet. 16: 37–43.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinforma. Oxf. Engl. 27: 863–864.

Schorn AJ, Gutbrod MJ, LeBlanc C, Martienssen R. 2017. LTR-Retrotransposon Control by tRNA-Derived Small RNAs. Cell 170: 61-71.e11.

Schrijvers EMC, Koudstaal PJ, Hofman A, Breteler MMB. 2011. Plasma Clusterin and the Risk of Alzheimer Disease. JAMA 305: 1322–1326.

Sebastian A, Contreras-Moreira B. 2014. footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. Bioinforma. Oxf. Engl. 30: 258–265.

Shaikh TH. 2017. Copy Number Variation Disorders. Curr. Genet. Med. Rep. 5: 183–190.

Shen L, Wu LC, Sanlioglu S, Chen R, Mendoza AR, Dangel AW, Carroll MC, Zipf WB, Yu CY. 1994. Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. J. Biol. Chem. 269: 8466–8476.

Shiang R, Thompson LM, Zhu YZ, Church DM, Fielder TJ, Bocian M, Winokur ST, Wasmuth JJ. 1994. Mutations in the transmembrane domain of FGFR3 cause the most common genetic form of dwarfism, achondroplasia. Cell 78: 335–342.

Skene PJ, Illingworth RS, Webb S, Kerr ARW, James KD, Turner DJ, Andrews R, Bird AP. 2010. Neuronal MeCP2 Is Expressed at Near Histone-Octamer Levels and Globally Alters the Chromatin State. Mol. Cell 37: 457–468.

Skidmore ZL, Wagner AH, Lesurf R, Campbell KM, Kunisaki J, Griffith OL, Griffith M. 2016. GenVisR: Genomic Visualizations in R. Bioinforma. Oxf. Engl. 32: 3012–3014.

Sleegers K, Brouwers N, Gijselinck I, Theuns J, Goossens D, Wauters J, Del-Favero J, Cruts M, Duijn CM van, Broeckhoven CV. 2006. APP duplication is sufficient to cause early onset Alzheimer's dementia with cerebral amyloid angiopathy. Brain 129: 2977–2983.

Smalheiser NR. 2014. The RNA-centred view of the synapse: non-coding RNAs and synaptic plasticity. Philos. Trans. R. Soc. B Biol. Sci. 369: 20130504.

Speek M. 2001. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol. Cell. Biol. 21: 1973–1985.

Srinivasan S, Kalinava N, Aldana R, Li Z, Hagen S van, Rodenburg SYA, Wind-Rotolo M, Sasson AS, Tang H, Qian X, Kirov S. 2020. Mis-annotated multi nucleotide variants in public cancer genomics datasets can lead to inaccurate mutation calls with significant implications. bioRxiv: 2020.06.05.136549.

Stenglein MD, Harris RS. 2006. APOBEC3B and APOBEC3F inhibit L1 retrotransposition by a DNA deamination-independent mechanism. J. Biol. Chem. 281: 16837–16841.

Strachan T, Read A. 2018. Human Molecular Genetics, 5th Edition. Garland Science. 761 p.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y, Konkel MK, Malhotra A, Stütz AM, Shi X, Casale FP, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, Mu XJ, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer E-W, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, 1000 Genomes Project Consortium, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO. 2015. An integrated map of structural variation in 2,504 human genomes. Nature 526: 75–81.

Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD. 2002. Molecular archeology of L1 insertions in the human genome. Genome Biol. 3: research0052.

Tan H, Wu C, Jin L. 2018. A Possible Role for Long Interspersed Nuclear Elements-1 (LINE-1) in Huntington's Disease Progression. Med. Sci. Monit. Int. Med. J. Exp. Clin. Res. 24: 3644–3652.

Taniguchi-Ikeda M, Kobayashi K, Kanagawa M, Yu C, Mori K, Oda T, Kuga A, Kurahashi H, Akman HO, DiMauro S, Kaji R, Yokota T, Takeda S, Toda T. 2011. Pathogenic exon-trapping by SVA retrotransposon and rescue in Fukuyama muscular dystrophy. Nature 478: 127–131.

Tanzi RE. 2012. The genetics of Alzheimer disease. Cold Spring Harb. Perspect. Med. 2.

Tariq A, Jantsch MF. 2012. Transcript diversification in the nervous system: A to I RNA-editing in CNS function and disease development. Front. Neurosci. 6.

The Deciphering Developmental Disorders Study, Fitzgerald TW, Gerety SS, Jones WD, van Kogelenberg M, King DA, McRae J, Morley KI, Parthiban V, Al-Turki S, Ambridge K, Barrett DM, Bayzetinova T, Clayton S, Coomber EL, Gribble S, Jones P, Krishnappa N, Mason LE, Middleton A, Miller R, Prigmore E, Rajan D, Sifrim A, Tivey AR, Ahmed M, Akawi N, Andrews R, Anjum U, Archer H, Armstrong R, Balasubramanian M, Banerjee R, Baralle D, Batstone P, Baty D, Bennett C, Berg J, Bernhard B, Bevan AP, Blair E, Blyth M, Bohanna D, Bourdon L, Bourn D, Brady A, Bragin E, Brewer C, Brueton L, Brunstrom K, Bumpstead SJ, Bunyan DJ, Burn J, Burton J, Canham N, Castle B, Chandler K, Clasper S, Clayton-Smith J, Cole T, Collins A, Collinson MN, Connell F, Cooper N, Cox H, Cresswell L, Cross G, Crow Y, D'Alessandro M, Dabir T, Davidson R, Davies S, Dean J, Deshpande C, Devlin G, Dixit A, Dominiczak A, Donnelly C, Donnelly D, Douglas A, Duncan A, Eason J, Edkins S, Ellard S, Ellis P, Elmslie F, Evans K, Everest S, Fendick T, Fisher R, Flinter F, Foulds N, Fryer A, Fu B, Gardiner C, Gaunt L, Ghali N, Gibbons R, et al. 2015. Large-scale discovery of novel genetic causes of developmental disorders. Nature 519: 223–228.

Treon SP, Xu L, Yang G, Zhou Y, Liu X, Cao Y, Sheehy P, Manning RJ, Patterson CJ, Tripsas C, Arcaini L, Pinkus GS, Rodig SJ, Sohani AR, Harris NL, Laramie JM, Skifter DA, Lincoln SE, Hunter ZR. 2012. MYD88 L265P somatic mutation in Waldenström's macroglobulinemia. N. Engl. J. Med. 367: 826–833.

Uesaka M, Nishimura O, Go Y, Nakashima K, Agata K, Imamura T. 2014. Bidirectional promoters are the major source of gene activation-associated non-coding RNAs in mammals. BMC Genomics 15: 35.

Ullu E, Tschudi C. 1984. Alu sequences are processed 7SL RNA genes. Nature 312: 171–172.

Upton KR, Gerhardt DJ, Jesuadian JS, Richardson SR, Sánchez-Luque FJ, Bodea GO, Ewing AD, Salvador-Palomeque C, van der Knaap MS, Brennan PM, Vanderver A, Faulkner GJ. 2015. Ubiquitous L1 mosaicism in hippocampal neurons. Cell 161: 228–239.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinforma. 43: 11.10.1-11.10.33.

Vansant G, Reynolds WF. 1995. The consensus sequence of a major Alu subfamily contains a functional retinoic acid response element. Proc. Natl. Acad. Sci. U. S. A. 92: 8229–8233.

Vartanian J-P, Guétard D, Henry M, Wain-Hobson S. 2008. Evidence for Editing of Human Papillomavirus DNA by APOBEC3 in Benign and Precancerous Lesions. Science 320: 230–233.

Vasquez JB, Fardo DW, Estus S. 2013. ABCA7 expression is associated with Alzheimer's disease polymorphism and disease status. Neurosci. Lett. 556: 58–62.

Vázquez-Manrique RP, Farina F, Cambon K, Dolores Sequedo M, Parker AJ, Millán JM, Weiss A, Déglon N, Neri C. 2016. AMPK activation protects from neuronal dysfunction and vulnerability across nematode, cellular and mouse models of Huntington's disease. Hum. Mol. Genet. 25: 1043–1058.

Veltman JA, Brunner HG. 2012. De novo mutations in human genetic disease. Nat. Rev. Genet. 13: 565–575.

Verghese PB, Castellano JM, Garai K, Wang Y, Jiang H, Shah A, Bu G, Frieden C, Holtzman DM. 2013. ApoE influences amyloid-β (Aβ) clearance despite minimal apoE/Aβ association in physiological conditions. Proc. Natl. Acad. Sci. 110: E1807–E1816.

Viollet S, Monot C, Cristofari G. 2014. L1 retrotransposition. Mob. Genet. Elem. 4: e28907.

Voelkerding KV, Dames SA, Durtschi JD. 2009. Next-Generation Sequencing: From Basic Research to Diagnostics. Clin. Chem. 55: 641–658.

Vogel JL, Kristie TM. 2013. The Dynamics of HCF-1 Modulation of Herpes Simplex Virus Chromatin during Initiation of Infection. Viruses 5: 1272–1291.

Voytas DF, Boeke JD. 2002. Ty1 and Ty5 of Saccharomyces cerevisiae. Mob. DNA II: 631–662.

Wallace NA, Belancio VP, Deininger PL. 2008. L1 mobile element expression causes multiple types of toxicity. Gene 419: 75–81.

Walter C, Clemens LE, Müller AJ, Fallier-Becker P, Proikas-Cezanne T, Riess O, Metzger S, Nguyen HP. 2016. Activation of AMPK-induced autophagy ameliorates Huntington disease pathology in vitro. Neuropharmacology 108: 24–38.

Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, Batzer MA. 2005. SVA elements: a hominid-specific retroposon family. J. Mol. Biol. 354: 994–1007.

Wang J, Song L, Grover D, Azrak S, Batzer MA, Liang P. 2006. dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Hum. Mutat. 27: 323–329.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA, 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.

Wang Q, Pierce-Hoffman E, Cummings BB, Alföldi J, Francioli LC, Gauthier LD, Hill AJ, O'Donnell-Luria AH, Karczewski KJ, MacArthur DG. 2020. Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. Nat. Commun. 11: 2539.

Wei L, Liu LT, Conroy JR, Hu Q, Conroy JM, Morrison CD, Johnson CS, Wang J, Liu S. 2015. MAC: identifying and correcting annotation for multi-nucleotide variations. BMC Genomics 16: 569.

Wildschutte JH, Baron A, Diroff NM, Kidd JM. 2015. Discovery and characterization of Alu repeat sequences via precise local read assembly. Nucleic Acids Res. 43: 10292–10307.

Wildschutte JH, Williams ZH, Montesion M, Subramanian RP, Kidd JM, Coffin JM. 2016. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proc. Natl. Acad. Sci. U. S. A. 113: E2326-2334.

Wong LH, Choo KHA. 2004. Evolutionary dynamics of transposable elements at the centromere. Trends Genet. TIG 20: 611–616.

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.

Zemojtel T, Penzkofer T, Schultz J, Dandekar T, Badge R, Vingron M. 2007. Exonization of active mouse L1s: a driver of transcriptome evolution? BMC Genomics 8: 392.

Zhang J, Weinrich JAP, Russ JB, Comer JD, Bommareddy PK, DiCasoli RJ, Wright CVE, Li Y, van Roessel PJ, Kaltschmidt JA. 2017. A Role for Dystonia-Associated Genes in Spinal GABAergic Interneuron Circuitry. Cell Rep. 21: 666–678.

Zhang Y, Maksakova IA, Gagnier L, Lagemaat LN van de, Mager DL. 2008. Genome-Wide Assessments Reveal Extremely High Levels of Polymorphism of Two Active Families of Mouse Endogenous Retroviral Elements. PLOS Genet. 4: e1000007.

Zhao X, Li C, Paez JG, Chin K, Jänne PA, Chen T-H, Girard L, Minna J, Christiani D, Leo C, Gray JW, Sellers WR, Meyerson M. 2004. An Integrated View of Copy Number and Allelic Alterations in the Cancer Genome Using Single Nucleotide Polymorphism Arrays. Cancer Res. 64: 3060–3071.