Scuola Internazionale Superiore di Studi Avanzati
Ph.D. course in Functional and Structural Genomics

Aberrant transcription regulation in a subset of individuals affected by Autism Spectrum Disorders

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

Candidate:
Giovanni Spirito

Supervisor:
Prof. Remo Sanges

Co-Supervisor:
Prof. Stefano Gustincich

Academic year 2019/2020

Ma tu,

mentalmente,

come stai?

Table of Contents

Abstract

Autism Spectrum Disorder (ASD) is a set of heterogeneous neurodevelopmental conditions mainly involving impaired communication and repetitive behaviors. It is known that genetics have a fundamental but yet not completely understood role in its etiology. Indeed, several hundreds genes have been implicated into ASD pathogenesis and susceptibility. Interestingly, not all these genes present a direct role in neuronal functions. Indeed ASD-related genes may be mainly divided into two categories: genes with a critical role in synaptic functions and genes involved in chromatin remodeling or regulation of transcription. The link between the latter and the neurological manifestations of ASD has not been uncovered yet.

Starting from this insight I sought to stratify ASD cases into two experimental groups on the basis of the presence or absence of potentially deleterious mutations within ASD-related genes involved in transcriptional regulation and/or chromatin remodeling.

I then explored the putatively different transcriptional landscape unique to each one of the two subset of patients in terms of gene and transposable elements expression. I detected a pervasive up-regulation of LINE transposable elements and down-regulation of genes important for synaptic functions only in the experimental group including ASD individuals carrying a putatively deleterious mutation within a regulatory gene. Interestingly, up-regulated LINEs are enriched within down-regulated genes. Furthermore I characterized the genomic location of DE non-coding genomic elements in order to try to reconstruct the regulatory alterations at the basis of the differential expression observed. Finally I replicated my experiments on independent datasets in order to test their reproducibility.

Abbreviations

ASD: Autism Spectrum Disorder

LINE: Long Interspersed Nuclear Element

SINE: Short Interspersed Nuclear Element

LTR: Long Terminal Repeat

WES: Whole-Exome Sequencing

ACC: Anterior Cingulate Cortex

PFC: Pre-Frontal Cortex

iPSC: induced Pluripotent Stem Cells

NPC: Neural Precursor Cell

CNS: Central Nervous System

DE: Differentially Expressed

TE: Transposable Element

1. Introduction

1.1 The complexity of the human genome

At the completion of the Human Genome Project the whole sequence of the human genome was available to researchers[1]. Remarkably the result of the project revealed that only ~2% of the whole genome is occupied by coding sequences[1]. Despite the greater relative importance of coding sequences, research conducted over the past two decades gave increasing importance to the non-coding portion of the genome, especially in regulating core physiological functions. Moreover several lines of evidence suggest that the source for many human-specific characteristics may lie in the complex relationship between coding and non-coding genomic sequences.

In this chapter I will describe how the complexity of the human genome may be defined, and how intricate regulatory mechanisms may allow the development of complex organisms. Additionally, I will illustrate how the alteration of these mechanisms can impact human health and how the scientific community exploits the latest technologies to provide the means necessary for their study.

1.1.1 The non-coding genome: a possible driver of biological complexity

Although science is supposed to reach precise terms able to synthesize elaborated concepts, quite often definitions can be vague and interpretable. It may be therefore not surprising that a universally accepted interpretation of the concept of genome complexity is lacking. However, biological complexity may be considered as the range of sub-cellular structures, number of cell types, functional repertoire of regulatory capabilities, level of sophistication in neural functions and the intricate developmental processes necessary for the generation of these characteristics[2]. Therefore we may regard genome complexity as the multiple and sophisticated layers of information embedded within the genome which allow for the development of biologically complex organisms.

The genome has been the object of intense study since in the mid-late 18th century, when Gregor Mendel proposed the law of independent assortment, explaining why characters are linked to the genetic asset[3]. From thereafter, biologists have leveraged increasingly advanced technologies to explore the human genome and its link with phenotype. Since 1960 molecular geneticists have started to demonstrate that changes in the base-pair composition of DNA are translated into changes in protein structure or developmental plans[4]. It is therefore possible that the complexity of an organism is a

reflection of the physical complexity of its genome, defined as the amount of information stored in its sequence.

It is now established that the human genome includes about 20,000 protein coding genes, accounting for less than 2% of its total length. Sequences encoding for coding genes show clear signs of evolutionary conservation[5] and allow for the synthesis of proteins, which are the main effector molecules of living organisms. However, since the number of human genes does not exceed, or is even inferior to the one in some simpler organisms[2], it is hardly believable that the number of coding genes alone correlates with organism complexity.

A fairly straightforward work proposed a compelling perspective by calculating the ratio between coding and non-coding DNA within the genome of several prokaryote and eucaryote organisms. The results show that the ratio of noncoding DNA to total genomic DNA increases as biological complexity of the analyzed organisms increments[2] (figure 1.1). Interestingly humans, which at least as far as neural capacity, may be considered as the most refined organisms in the biosphere, hold the highest ratio between non-coding and coding DNA[2]. It is therefore conceivable that increased organism complexity is primarily associated with an expansion in regulatory control systems, rather than an increased number of functional components[2,6]. As a result, the non-coding portion of the genome, previously regarded primarily as “junk”, may hold an outstanding amount of information essential for the development and function of multiple human-specific traits. This speculation may be strengthened by the established notion according to which the vast majority (> 95%) of the non-coding genome is actively transcribed[7].

The higher level of complexity distinguishing the human species is mainly due to the staggering level of intricacy within structures of the central nervous system (CNS). It is therefore likely that the regulatory potential of the non-coding genome is exploited to the fullest during the development and life of the brain. The study of this topic may therefore be of keen interest to understand the development and the basis for several diseases of the human CNS.

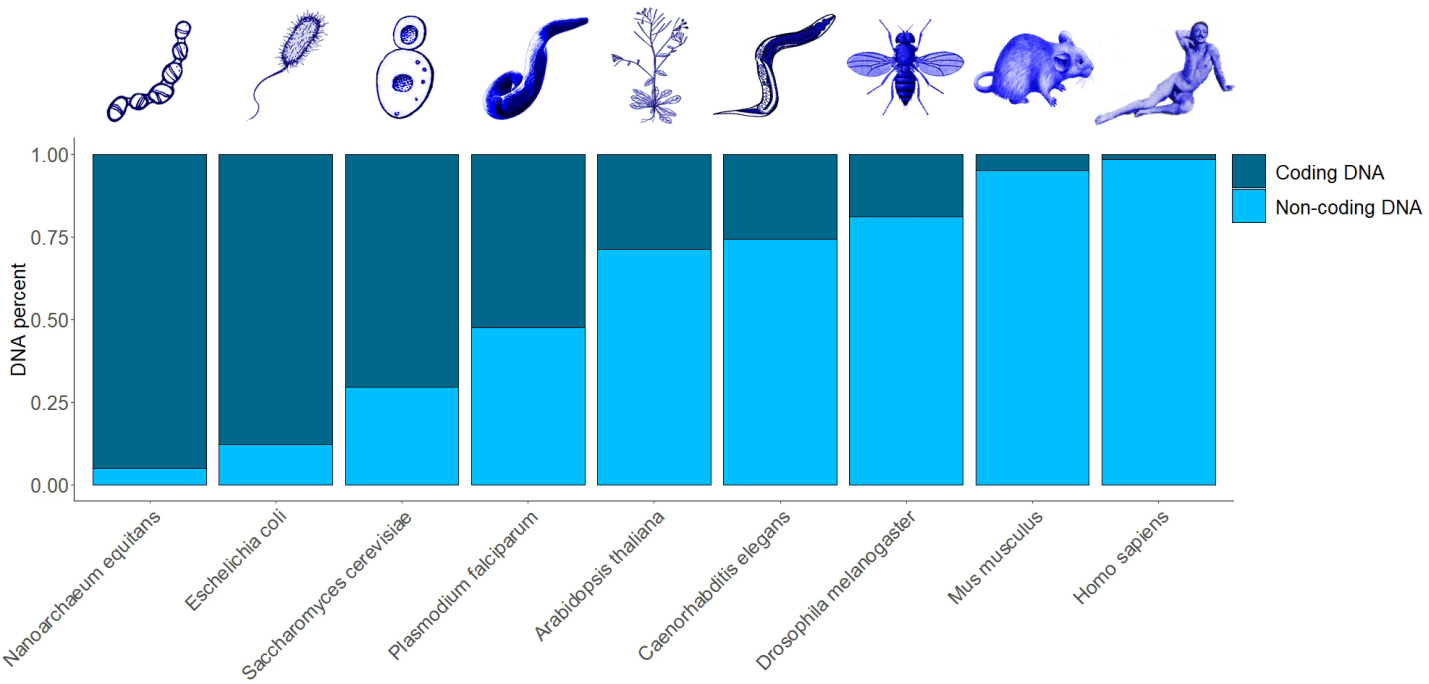

Figure 1.1 – Percentage of genome occupied by coding and non-coding DNA for some of the most studied organisms. Species are sorted in order of crescent biological complexity. Data relative to the coding/non-coding DNA ratio was retrieved from Ryan J. Taft and John S. Mattick, 2003

1.1.2 Epigenetics: the functional regulator between coding and non-coding DNA

The refined mechanisms that allow for the development of complex organisms need to be precisely regulated in a highly dynamic fashion. However the sequence of the genome is considered to be overall fixed and almost equal in all the cells of a given individual. Nonetheless genomes of all species evolved systems able to influence gene expression, and therefore the phenotype, without changing the nucleotide composition of the genome. These mechanisms act by modifying the structure of the chromatin, and are collectively referred to as epigenetic regulation[8].

The chromatin is finely organized at the structural level, and its architecture has a fundamental role in its function. The core of the topological organization of eukariotic genomes is the nucleosome[8]. The nucleosome itself is a fairly simple structure made up by octamer of proteins called histones encircled by ~147 bp of DNA[8]. The level of tightness to which DNA is wrapped around histones at a given moment is correlated to the level of transcriptional activity of DNA[8]. The core function of epigenetic mechanisms is in fact to regulate chromatin accessibility, which directly impacts gene expression. To give an idea of the scope of this mechanism, it may suffice to say that by determining which genes and

how much they are expressed at a given time, it is possible to produce every kind of human cell given the same genome. Furthermore, this organization makes gene expression regulation, and to a certain extent, cell identity, extremely dynamic. Indeed most of the known epigenetic changes are reversible[8].

The main kinds of epigenetic modifications are DNA methylation and histone modifications (mainly methylation and acetylation)[8]. These mechanisms are put in place by adding and removing specific chemical moieties to either the DNA or histone proteins[8]. The molecular effectors responsible for this may be mainly divided into three roles: writer and eraser proteins (which respectively add and remove chemical groups from their targets) and reader proteins, which bind to specific epigenetic marks thus influencing chromatin structure and therefore gene expression[8]. Hence the activity of these proteins allows epigenetic regulation to take place constantly during the life of the cell.

The study of epigenetic marks revealed reproducible patterns in terms of the effect of specific marks on gene expression[9]. It was revealed that the effects of a specific modification were due to the moiety added to the target and to the overall chromatin context in which the modification takes place[9]. This suggests an interplay among different kinds of epigenetic mechanisms. For example, DNA methylation can result in an increased or decreased gene expression according to whether the methyl group is added within the gene body or in proximity to the promoter[8,9]. Similarly, histonic markers such as H3K27Ac, H3K9Ac, H3K36me3, H3K4me3 and H3K4me1 are associated with an increased chromatin accessibility; while H3K9me3 and H3K27me3 imply a more closed and less accessible chromatin state[8,9]. Interestingly, it is known that the combination of different histone markers may produce unique chromatin states. For example, genomic loci characterized by the co-presence of H3K4me3 and H3K27me3 are considered ‘bivalent’ regions[10]. These regions are characterized by an inactive but poised state, meaning that the level of chromatin accessibility, and therefore transcription within these loci can be finely and quickly altered[10]. These classes of genomic loci are indeed considered especially important during delicate cellular processes such as development and differentiation[10].

Overall, the study of epigenetic mechanisms led to a functional profiling of the non-coding portion of the genome, defining an impressive amount of regulatory genomic segments. Consequently, the role of regulatory non-coding sequences became a primary topic in the study of human diseases characterized by a genetic etiology.

1.1.3 Next generation sequencing technologies allow for the exploration of genomes

The human genome is an intricate structure comprising a certain number of segments encoding for effector molecules and a comparatively much larger amount of sequences whose main purpose is to regulate the former. Decades of studies revealed that the patterns within nucleotide sequences are strongly linked to the function of each segment.

With the conclusion of the Human Genome Project in the early 2000s, it has been possible to read the almost totality of the human genome sequence[1]. This opened the possibility to understand the role of each segment of DNA. However, the discovery of the sequence of the human genome only allowed scientists to access it like some kind of instruction manual without precisely knowing its language. Consequently, in the past decades, a large portion of biologists' work has been directed to understand the underlying grammar and vocabulary underlying the sequence of the human genome.

The most common approach to link a gene to its functions in an organism is to use animal models by producing KO models and thus selectively mutating genes of interest. This and similar 'reverse-engineering' approaches are however not applicable, for evident ethical implications, to human individuals. Geneticists have therefore focused their studies on the genomic assets of individuals born with pathological phenotypes (usually referred to as genetic diseases). Genomic loci which differ from the reference sequence in an individual are usually called genomic variants. Genomes belonging to individuals affected by genetic diseases can be compared to the human reference genome in order to pinpoint variants which may be at the basis of the pathological phenotype. Yet, due to the large number of variants characterizing each individual, this is often a rather demanding task. Indeed, an average human genome differs from all other human genomes for about 0.01% of its sequence. Since a human genome is about 3 billion base-pair long, each individual is characterized by at least 3.5 million variants[11]. In order to discriminate between common (and usually harmless) and rare deleterious variation it is necessary to analyze a large number of human genomes.

Only in the past few years due to exponential technological advances in sequencing DNA and in the development of bioinformatic analytical tools, large-scale projects similar to the Human Genome Project have begun to unravel the secrets of the human genome. The rapid decrease in sequencing costs allowed for the analysis of an unprecedented amount of genomic data, to the point where it may be conceivable to think that the amount of genomic material analyzed in a single issue of *Nature Genetics* possibly exceeds the sum of the nucleotide sequences analyzed from the discovery of Mendelian laws to the start of the Human Genome Project. Indeed, in the past decade several research consortia were

formed with the aim of exploring the human genome by sequencing thousands of different individuals. Each consortium chose specific strategies in order to address separate issues regarding our genome. To note, part of the knowledge and data produced by these works has been made freely available to the community.

For example, the gnomAD[12] and EXaC[13] projects aggregate and make publicly available both genome and exome sequencing data from a wide variety of large-scale sequencing projects. These consortia leveraged genomic data from several thousand individuals from various disease-specific and population genetic studies, allowing the classification of millions of genomic variants. This kind of data mostly proves its usefulness when researchers need to classify the variants found in a particular subject. Variants annotation, in fact, allows for the classification of a variant as common (and probably harmless) or rare (and possibly deleterious). Due to the large number of variants characterizing each individual, this utility is essential in large-scale studies on genetic diseases. Moreover, the massive amount of annotated variants can be exploited to perform analyses called Genome-Wide Association Studies (GWAS), which allow to link specific common variants to complex traits.

Other project such as ENCODE[14], Fantom[15] and Roadmap Epigenomics[16] aimed at assessing which parts of the non-coding genome are functional; that is, painting the landscape of non-coding genes and gene-regulatory sequences of the genome. These resources are crucial to understand the interplay between coding and non-coding genome, and to explore how the alteration of specific epigenetic mechanisms may be related to pathological phenotypes.

Other consortia focused on assessing the link between genotype and phenotype by sequencing both genome and transcriptiome of the same individuals. One of the first and most important of such is the 1000 Genome Project[11]. At the completion of the final phase, 2,500 individuals from 26 populations were sequenced. The final release of the project (phase III) described over 88 million variants: 84.7 million SNPs, 3.6 million indels and over 60,000 structural variants (which in turn are divided in deletions, inversions, duplications and transposable elements insertions)[17]. Alongside the 1000 Genome Project, the Geuvadis consortium performed both RNA and small RNA sequencing on lymphoblastoid cell lines (LCL) of 445 individuals who participated to the 1000 Genome Project[11,18]. It is therefore possible to integrate these individual genomic and transcriptomic datasets in order to assess the impact of genomic variants on gene expression in this cell line. These kinds of analyses are performed by regressing gene expression levels for individual samples against locus-specific genotypes for matched samples[19], and they are usually referred to as expression Quantitative Trait Loci (eQTL). Overall, the

integrative analysis of genotype and transcriptomic data from multiple human individuals demonstrated that differences in the genomic asset can manifest in changes in overall gene expression as well as the relative abundance of transcripts from the same gene. Indeed, the majority of analyzed genes showed substantial variation in expression levels linked to genomic differences among individuals[18].

Collectively, the knowledge resulting from multiple consortia born to explore the intricacies of the human genome has been, and still is, of fundamental importance to unravel the function of the variety of sequences found in the genome; thus laying the foundations for the understanding of the molecular mechanisms behind genetic diseases.

1.1.4 Transcriptional regulation is altered in specific diseases

In the past decade several consortia were formed with the aim of exploring the extent of human genomic variation and its impact on phenotype through massive use of NGS technologies. The results of these large-scale projects overall converge in outlining a surprisingly high level of variation among human genome sequences[11,17,18]. This is a result of the process of variant accumulation which characterized human evolution. Genomic variants (or mutations) mainly arise from errors occurring during the process of DNA replication, and they are essential for evolution, since they allow for species adaptation to a specific environment. While the most common kinds of alterations manifest as substitutions of single nucleotides, mutational phenomena involving larger genomic sequence occur as well. This is mainly the case of copy-number-variants (CNVs) and especially duplications and insertions[20]. These kinds of variation imply the addition of genomic material to a species genome, and likely allowed for a fast rate in the evolution of non-coding regulatory sequences, hence allowing the existence of increasingly refined organisms[20].

Due to the random nature of mutations, some variants may have a detrimental impact on the organism. Several human diseases can in fact be directly traced back to specific mutations which result in an aberrant activity of one or multiple genes[21]. These are generally called genetic diseases.

The vast majority of known disease-causing variants reside within the coding genome[20]. However, the specific cause for more than 50% of these diseases is still unknown. Specifically, for some diseases, while it is known that there is a strong genetic etiological component, it is hard to pinpoint the actual cause. That is, it is not trivial to trace the observed phenotype to the dysfunction of a single gene. These diseases are usually referred to as complex diseases. Complex diseases are often caused by the

combination of rare and common variants with moderate impact[20]. Additionally, for many of these, the environment is also considered to influence the disease manifestation[20].

Broadly speaking, possibly the most common kinds of phenotypic manifestations arising from a genetic deleterious mutation include neurodevelopmental defects[22]. This may be due to the fact that the CNS is the most sophisticated structure in humans. Therefore a high number of genes is involved in its development and function. Furthermore, its complex development requires a fine regulation, thus even small changes may lead to macroscopic aberrant phenotypes. Some of the most common neurodevelopmental conditions are considered to be complex diseases. Indeed, in several cases the link between genetic mutations and an aberrant neurological phenotype is not clear. Neurodevelopmental disorders are often characterized by aberrant and delayed early-life development of the brain[23]. As previously mentioned, epigenetic control is crucial in determining cell identity during development. Therefore, it is perhaps not surprising that disruptions of genes involved in epigenetic functions are known to be causative for several mental retardation/intellectual disability syndromes[24,25]. Recent work has highlighted genes with epigenetic functions as being implicated in neurodevelopmental disorders, primarily in autism spectrum disorders (ASD) and schizophrenia (SCZ)[24]. Indeed, an aberrant epigenetic landscape is often observed in such conditions[26]. The control of epigenetic regulation is up to the activity of non-coding regulatory sequences. These findings suggest a link between genome regulation and neurological disease and thus put a spotlight on the non-coding genome in the context of neurodevelopmental diseases study and possible treatments.

Exploring the epigenetic role in neurodevelopmental diseases may be particularly attractive due both to the lack of extensive knowledge on the subject and to the possibility for correcting the causative epigenetic perturbation granted by the reversible nature of epigenetic regulation[27].

1.2 Transposable elements: the genomes functional junk DNA

Transposable elements (TEs) are genomic sequences able to move from one location to another within the genome[28]. Despite being widely considered junk DNA at first, increasing evidence is showing how they may be involved in several regulatory functions[29,30]. As a consequence, the interest of the scientific community is progressively shifting towards the understanding of the functional roles of TEs.

1.2.1 Transposable elements classification

As a result of their ancient and deep evolutionary origins, TEs come in a plethora of forms and sizes. They can be primarily divided into classes based on their transposition mechanism; in turn each class can be divided into families based on their nucleotide structure. Namely, class 2 TEs are called DNA transposons; and they mobilize through a 'cut-and-paste' mechanism[31]. Instead class 1 TEs are called retrotransposons; because they mobilize through a 'copy-and-paste' mechanism thanks to an RNA intermediate[31]. Retrotransposons can be in turn divided into two main sub-classes: LTR and non-LTR. The former integrates thanks to cleavage and strand-transfer reaction catalyzed by an integrase[31], in a similar fashion to retroviruses; the latter through a process called target-primed reverse transcription, and it is coupled with reversed transcription[32].

Retrotransposons include the only classes of TEs believed to be still active in the human genome: long interspersed nuclear elements (LINE), short interspersed nuclear elements (SINE) Alu and SINE-VNTR-Alu (SVA)[33]. Full length LINE elements are about 6 Kbp long, they make up about 17% of the human genome[34] and are the only known autonomous TEs currently active in the human genome[33]. L1 elements sequence include two open reading frames: ORF1, encoding for a protein with RNA-binding activity and ORF2, encoding for a protein with endonuclease and reverse-transcriptase activity[34]. In addition to having an impact on the genome through their own retrotransposition, L1s protein machinery can cause the mobilization of non-autonomous transcribed TEs such as Alu and SVA[35] and may also cause the birth of pseudogenes[36]. Alu elements are only a few hundreds nucleotides long[37], they are unique to primates and, being the most common class of SINE in the human genome, they represent a major factor of genetic diversity[38]. Similarly, SINE-VNTR-Alu (SVA) elements are the evolutionarily youngest class of human TEs. Due to their multi-domain composition, these elements have been described as composite non-coding retrotransposons[39]. Each domain is believed to derive from either a retrotransposon or a simple repeat sequence[39]. Canonically, starting from the 5′ end an

SVA element is composed by: an hexameric CCCTCT repeat, followed by sequence sharing homology to two antisense Alu fragments, a variable number of GC-rich tandem repeats (VNTR), a sequence likely derived partially from the ENV gene and partially from the 3′ end of a repeated sequence of an ancient endogenous retrovirus (HERV-K10)[40], and a polyadenylation signal.

The mechanism through which retrotransposons mobilize, allowed them to multiply their sequences within the host genome. In fact, retrotransposons activity throughout eukaryotes evolution increased the number of their copies within the genome to the point where approximately 45% of the human genome is constituted by retrotransposons-derived sequences[33]. However, their contribution is likely to be even larger since the identification of ancestral TE-derived sequences is made harder by the accumulation of mutations over time. As a result, a great portion of the human genome non-coding landscape is made of retrotransposons.

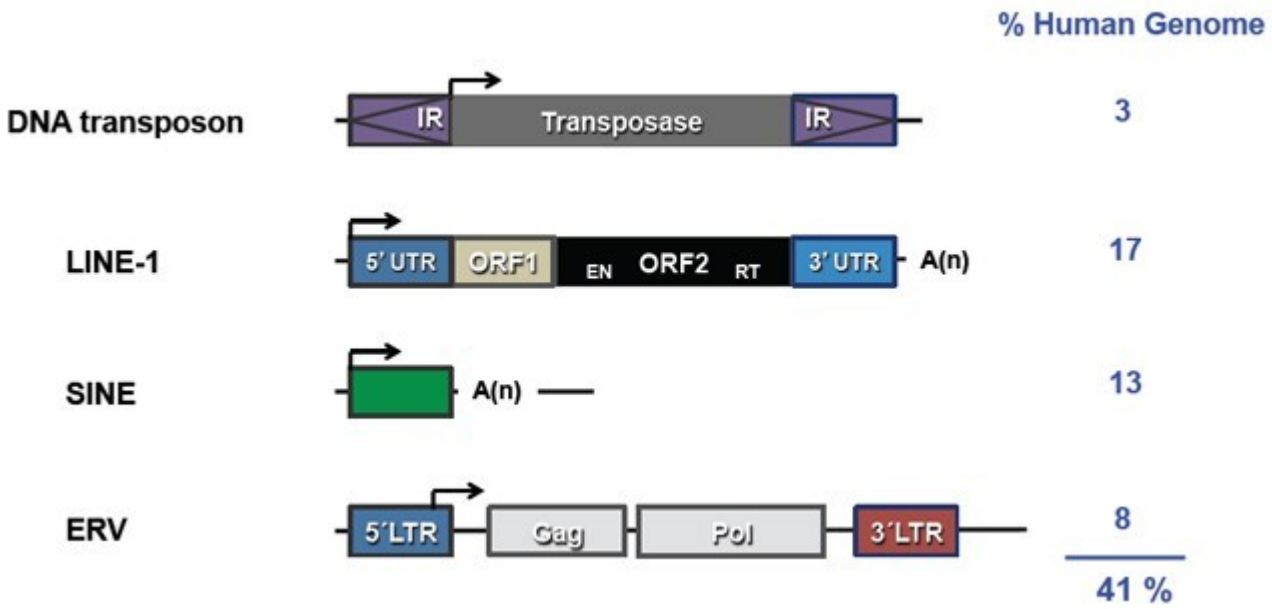

Figure 1.2 – Different classes of TEs: DNA transposons, LTR, LINE, SINE. The fractions of human genome occupied by each TE class is shown on the right. IR (inversed repeat), RT (reverse transcriptase), ORF (open reading frame), UTR (untranslated region), EN (endonuclease); (Garcia-Perez, Widmann and Adams, 2016)

1.2.2 Transposable elements act as gene-regulatory sequences

Only about 2% of the human genome is composed of genes translated into proteins, the main effector molecules of the cell. A crucial feature essential for the development and the physiology of complex organisms, lies in the capability to shape intricate networks of gene expression regulation. It is now known that most of eukariotic gene regulation is mediated by *cis*-regulatory elements, which are embedded within the largely understudied non-coding portion of the genome[41]. TEs, in their ancestral form, carry their own regulatory sequences[41], which are essential to exploit the host genomic asset to their own benefit. Multiple lines of evidence point to a co-option of these regulatory sequences by the host organisms, which along evolution, may have 'domesticated' TEs to their own benefit[41–43]. Indeed, TEs have been often shown to contribute to the regulation of genes by supplying *cis*-regulatory sequences[31,44,45]. In human, TE-derived sequences provide functional elements to promoters[46,47], enhancers[48,49], transcription terminators[50], transcription factors binding sites[51] and splice-sites[52,53]. These kinds of regulatory features are akin for their ability to bind specific proteins, thus regulating gene expression. Moreover, a single human TE can contribute to the regulation of multiple genes by influencing chromatin structure[47].

Transcription factors (TF) are proteins that bind to specific nucleotide motifs called binding sites thus regulating genes expression. Genome-wide analysis of TF binding sites revealed that up to 40% of human binding motifs are derived by TEs[54]. Furthermore it has been established that TE-derived binding motifs are non-randomly distributed among TE families[54]. Generally evolutionarily young TEs are more likely to contain TF binding sites since they accumulated a lower number of mutations compared to older elements[55]. This is in line with the notion according to which TEs are responsible for several species-specific regulatory networks[55]. Indeed TE-derived regulatory sequences are enriched nearby genes involved in immunity, response to external stimuli, neurodevelopmental functions[48] and overall lineage- or species-specific genes[44], consistently with the fact that single TEs are mostly lineage-specific genomic components, and are therefore prone to impact fast-evolving species-specific phenotypic characteristics. Conversely, TEs are depleted nearby conserved genes[48]. An additional level of regulatory complexity lies in the fact that TE-derived binding sites may be more or less accessible to TFs due to local chromatin epigenetic changes[56]. Epigenetic regulation of specific nucloetide sequences may therefore impact gene regulation in a genome-wide fashion, as TEs belonging to the same families are spread all over the genome and share almost identical sequences. Moreover, recent evidence shows a striking association between transposable elements classes and genes with different functions[57]. In

particular, SINEs are enriched within genes with ‘housekeeping’ functions, such as translation, nucleic acids binding and proliferation. In contrast, LINE enriched genes are strongly enriched in specialized functions, such as immunoglobulin function, retinol metabolism and olfactory receptors, typical of terminally differentiated cells[57]. This ‘barcoding’ capability of transposable elements may be crucial in order to correctly orchestrate the regulation of gene expression during embryogenesis. Of particular interest is the case of LINE elements. In human embryonic cells, genes enriched for LINE elements are transcriptionally repressed. This is likely a result of the binding between RNA derived from the expression of LINE elements embedded within LINE-enriched genes and LINE DNA sequences[57]. Through this non completely understood mechanism, epigenetic regulation of nucleotide sequences characteristic of LINEs may promote the silencing and re-activation of specific sets of genes during development.

Overall TEs likely gave rise and spread a plethora of regulatory modules within the genome throughout human evolution, thus giving rise to novel regulatory networks.

1.2.3 Transposable elements transcription is influenced by epigenetic regulation

Transposable elements physiologically might contribute to gene regulation in humans by providing binding sites for transcription factors or other proteins involved in chromatin regulation. On the other hand, dysregulation of TEs may lead to aberrant gene expression and therefore be harmful for the host. Indeed the regulatory potential of most transposons is usually silenced by epigenetic mechanisms[58,59].

Nevertheless, TEs are known to escape silencing at embryonic stages[60], affecting early human development by regulating nearby protein-coding genes. Waves of hypomethylation during embryogenesis are linked with higher rates of TEs transcription and retrotransposition[61]. The precise role of TEs in early mammal development is however not yet well understood. TEs activation during development and cell differentiation may be especially important during neural development[62,63]. In particular, recent evidence suggests a role of selective LINE activation during neurogenesis. This may manifest as a change in the epigenetic patterns at the level of LINE promoters[62,63].

Moreover, altered DNA methylation levels and patterns of CpG residues within LINE sequences have been reported in multiple neurodevelopmental diseases[64]; and disruption of specific recently evolved non-coding regulatory elements have been found to be disrupted in autism[65]. Taken together, this data points to an alteration of epigenetic patterns influencing the activity of specific transposable elements.

Interestingly, chromatin remodeling patterns observed in neurons during learning resemble those found in neurodevelopmental conditions[66]. This outcome suggests a possible involvement of transposable elements activity in physiological mechanisms characteristic of the CNS. Indeed, aside from developing cells, many TEs are actively transcribed in adult tissues as well. Recent evidence shows that approximately 2% of TEs are transcriptionally active[67]. Interestingly, TEs expression is very often tissue-specific[67], suggesting that the activation of different TE subfamilies may differentially impact specific cell types.

Although the regulation of TEs transcription and retrotransposition in somatic and developing tissues remains unclear, accumulating evidence suggests that epigenetic mechanisms, including DNA methylation and histone modifications, are involved in TEs activity and may impact the expression of target genes[68]. It is especially unclear whether TEs transcription may be either a mere marker of specific developmental and somatic cellular mechanism or may be essential for such processes to occur. Furthermore, being lineage-specific genomic elements, TEs regulation may contribute to species-specific molecular functions, which in humans may be especially important within the CNS.

1.2.4 Altered transposable elements expression is found in neurodevelopmental disorders

Recent studies suggested that specific TEs are transcriptionally activated during neuronal differentiation[69]. This activation may impact the regulation of specific genes essential for neurogenesis. Consequently, it may be speculated that an aberrant TEs transcriptional activity may bear a negative and somewhat “specific” impact on genes involved in neuronal functions. Indeed, an altered TEs (especially of LINE) expression has been reported in numerous neurodegenerative and neurological conditions[69].

Schizophrenia is a neurological disease characterized by behavioral deficits mainly developed throughout adolescence and adulthood. Nonetheless several evidences point to an at least partially genetic etiology. Increased LINE[70] activity and decreased LINE methylation have been reported in brain tissue deriving from schizophrenic patients compared to controls[71].

Similarly, increased rates of LINE expression and retrotransposition have been observed in the brain of RETT syndrome patients and mouse models[72]. RETT syndrome is a rare neurodevelopmental condition which leads to intellectual disability, seizures, and autistic behaviors[72]. It can be caused by a loss of

function of the gene MECP2, which normally represses LINE 5’ regions[72]. It is therefore likely that the lack of this methylation-dependent LINE repression is involved in the etiology of the disease.

Activity of LINE elements has been implicated in many Autism Spectrum Disorders (ASD) phenotypes as well[73,74]. Recent works underlined a reduction of methylation and an increase in LINE-1 expression in ASD post-mortem brains[74,75]. Interestingly, trimethylation of histone H3K9 (H3K9me3), which is responsible for the formation of condensed heterochromatin and prevents LINE activation, was significantly reduced at LINE ORF1 and ORF2 sequences but not at the 5′-UTR in the autism samples[74].

However, despite all the efforts, researchers did not assess a clear causal link between LINE activation and neurodevelopmental disorders. Moreover, it could be speculated that LINE aberrant activity may be an effect of an upstream causative pathological event rather than a driver of the aberrant phenotype. Nevertheless, it would not be too reckless to speculate that the consequences of an aberrant LINE activation may be more likely to be detrimental than beneficial.

1.3 Autism Spectrum Disorders: the most common neurological complex disease

Autism Spectrum Disorders (ASD) are a set of heterogeneous neurodevelopmental conditions mainly involving impaired communication and repetitive behaviors[76]. These symptoms are frequently co-morbid with other neuropsychiatric symptoms, including intellectual disability, developmental delay, seizures and attention-deficit disorder[76]. ASD represents the most common neurodevelopmental condition in human: the median worldwide prevalence of autism is around 1% according to the latest world-wide surveys[76], and it features a strong bias against males[76].

Most of the mechanisms underlying ASD etiology still remain a mystery. In particular scientists have not yet been able to solve the puzzle at the basis of ASD genetic heterogeneity. The main reason behind this intricacy may depend of the fact that hundreds of genes are thought to be involved in the pathology. ASD therefore has all the rights to be considered a complex disease.

To understand etiologies at the basis of ASD development, clarification of the substantial heterogeneity by subgroup is essential. Moreover, it is of keen importance to understand how alteration of specific molecular mechanisms act during early ASD development, as early recognition and therapeutic intervention rely on it.

1.3.1 The genetics of ASD: an incomplete puzzle

The molecular mechanisms at the basis of the development of ASD have not been thoroughly described yet. However, it is by now clear that there is a strong genetic component to the etiology to autism: twin and sibling studies have consistently shown that ASD is one of the most highly heritable complex disorders in humans[77]. Nonetheless, from the beginning it was clear that ASD phenotypes were not due to deleterious mutation within a single gene.

In the past decade, next-generation sequencing technologies had a transformative role in accelerating the process of gene discovery in ASD[78]. The most common approach for assessing genomic variants and genes related to ASD is whole-exome sequencing (WES) followed by bioinformatic analyses[78]. Exome capture uses thousands of DNA probes to bind and extract DNA in the ~2% of the genome that encodes for proteins, called the exome. Currently, the low overall costs of this technology enable the study of large numbers of individuals. Several early studies highlighted the potential of WES in the identification of risk-mediating variants in individual families or small cohorts with developmental delay or ASD[78]. Furthermore, by comparing the rate of deleterious mutations within genes in cases versus controls it has been possible to pinpoint a large number of genes contributing to the risk of ASD.

The results of these kinds of analysis have been corroborated and replicated in large independent ASD cohorts over the past few years[78]. The increasing scope of WES studies and the ever growing cohort sizes allowed to expand the discovery rate of ASD genes, to the point where almost 1000 genes show a degree of involvement in ASD etiology[79].

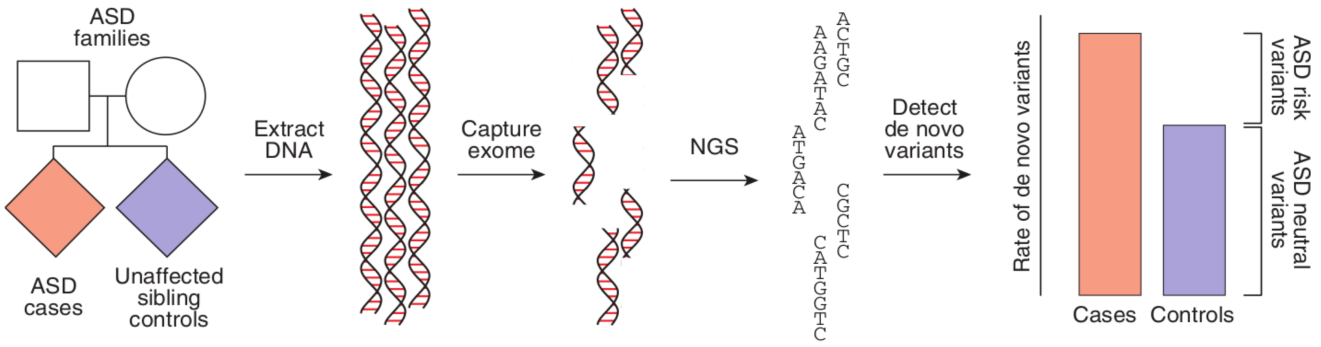

Figure 1.3 - Exome sequencing in autism spectrum disorder (ASD). Blood-derived DNA is collected from individuals with ASD (cases), both their parents, and, ideally, an unaffected sibling control. The ~2% of the genome that encodes proteins is captured and undergoes next-generation sequencing (NGS). *De novo* variants are identified by comparing the childs sequence to the parents' and the rate of *de novo* variants between cases and controls is assessed. If other factors, such as parental age and sequencing quality metrics, can be excluded, the excess of *de novo* variants represents risk-mediating variants identified in the cases only (Stephan J. Sanders, 2020).

Overall, next-generation sequencing has had a revolutionary impact in identifying genes that contribute to ASD. This progress in gene discovery, alongside NGS-derived functional data, is already beginning to provide insights into the neurobiology of ASD. However, despite this progress, a coherent model for ASD pathogenesis and effective therapies remains elusive, prompting the need for further research to follow up on the leads NGS methods have been critical in finding. In the near future WES may become the standard for ASD diagnosis. Furthermore, as the cost of NGS has continued to decrease, short-read whole-genome sequencing (WGS) may replace WES as a method to identify genetic factors associated with ASD. Notably, WGS would allow to explore and possibly asses the importance of the non-coding genome in the pathogenesis of ASD.

Figure 1.4 – Word cloud representative of SFARI genes (SFARI version 2020 Q2). The sizes of the gene names are proportional to the number of publications regarding ASD and involving each gene.

1.3.2 The two faces of ASD-related genes

The application of NGS technologies to the study of ASD allowed researchers to draw up a list of genes involved in the etiology of ASD[79]. At the present time the most reliable and publicly available database of ASD-related genes is the SFARI database[79]. The SFARI ASD list is a manually curated set of candidate genes implicated by common variant association, candidate gene studies, genes within ASD-associated CNV, in ASD and, to a lesser extent, syndromic forms of ASD.

Interestingly, mutations within none of the hundreds of genes present in the SFARI database account for more than 1% of ASD cases[80]. Undoubtedly, the understanding of how this large number of genes may converge to affect human brain development is critical to correctly interpret ASD etiology[81]. Most of the genes involved is ASD, similarly to other neuropsychiatric disorders, are genes coding for neuronal components crucial for brain function. The majority of the genes preset in the SFARI database play an important role in synaptic function[79].

Recent genome-wide and unbiased analyses applied to large ASD cohorts, have highlighted another class of genes that do not present a direct role in neuronal functions. In particular a substantial number of these genes results involved in transcription regulation and/or chromatin remodeling[24,78]. The most recurrent classes of these regulatory genes are KDM and KTM. These are enzymes whose main role is

to respectively add and remove methyl residues to specific lysines located on the tail of histone proteins. Lysine methylation is one of the main kinds of histone modifications in the human genome. As previously mentioned, histone methylation can have different effects on chromatin accessibility according to which specific lysine is chemically modified[8,9]. Consequently, it is not surprising that the activity of members of the KTM (histone lysine methyltransferase) and KDM (histone lysine demethylase) family are crucial for regulating a plethora of events including gene expression, cell cycle, and differentiation. Importantly, regulation of lysine methylation emerged as critical for neurological function and disease[82].

Genome regulation can also occur through DNA methylation. Several DNA-demethylases have been involved in ASD and other neurological conditions[58,83,84]. The DNMT family member most strongly linked to ASD is DNMT3A. Recent evidence showed how DNMT3A deposits a unique form of non-CG DNA methylation across the genome of neurons during development[83]. Therefore the alteration of these mechanisms may very well be related to the development of neurological conditions. Both common and rare variants associated with ASD, are within genes coding for proteins which regulate chromatin structure by binding methylated DNA, such as most members of the MBD family (MBD1, MBD3, MBD4, MBD5 and MBD6)[85]. Furthermore, mouse models of MBD1 KO show clear deficits frequently associated with autism[85]. Rare mutations within methylation-dependent transcriptional regulator MeCP2 are also linked to autism[72]. Finally, among ASD-related regulatory genes, there are genes that may lead to ASD phenotypes by directly influencing the expression of neural genes. These are for example the helicase CHD8 and transcription factors such as FOXP1 and SOX5[86].

A recent work by Ramaswami et al[87] integrated data from gene expression, DNA methylation and histone acetylation from ASD and healthy individuals, proposing the existence of two major subgroups of ASD. The first one is indistinguishable from control samples in terms of transcriptional and epigenetic alterations. On the other hand, the other subgroup recapitulates all known molecular changes typical of ASD[87]. A clear molecular cause at the basis of either one of the two groups is however lacking. It would be therefore compelling to think that deleterious genomic variants within the two classes of ASD-related genes would be at the basis of the two distinct ASD individual subgroups.

The significant advancement in ASD-related gene discovery recently laid a foundation from which to understand the molecular neurobiology of ASD. Reasoning on all these observations led me to the crucial consideration, followed throughout my thesis, that ASD candidate genes might be divided in two major groups: synaptic genes and regulatory genes. The link between deleterious mutations

affecting a regulatory gene and those triggering primarily a neurological phenotype has not been uncovered yet. The plainest explanation would be for the regulatory genes to somehow specifically control the expression of the genes involved in synaptic functions. However the link between these two groups of genes remains unclear.

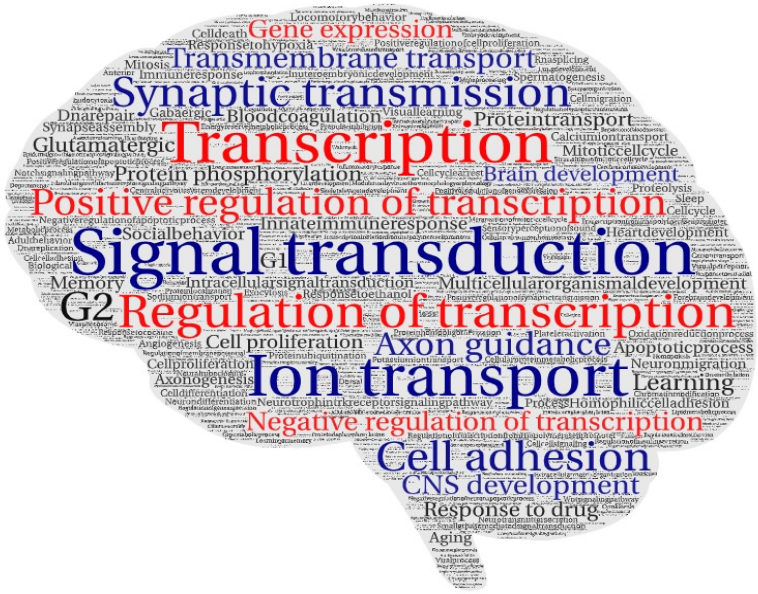

Figure 1.5 – Word cloud representative of the functional enrichments relative to SFARI genes. Enrichment analysis was performed on the complete list of SFARI genes with EnrichR. The size of the words are proportional to the inverse log10 p-value relative to each enrichment. I colored in red GO relative to transcription regulation/chromatin remodeling and in blue GO terms involved in synaptic functions.

1.3.3 ASD phenotypes converge at the transcriptomic level

Genome-wide analyses allowed to establish a great number of genes involved in ASD. At the same time transcriptomic analyses began to explore how the transcriptional landscape changes in individuals affected by ASD and healthy controls. The answer to why the alteration of such a large number of genes may converge on a heterogeneous but coherent spectrum of neurodevelopmental defects, may lie in the genes featuring an aberrant expression in the ASD-affected brain. Recently NGS technologies prompted a significant leap forward in the study of gene expression, very much the same way it impacted the study of genomes and the identification of disease-causative mutations. RNA sequencing, with its ever-decreasing costs, has indeed become the standard for high-throughput transcriptomic studies on large cohorts[88].

Overall, two main strategies have been used to explore the possible convergence of genetically heterogeneous ASD cases at the gene expression level. On one hand it is possible to compare the transcriptional landscape of *post-mortem* tissues of ASD patients to controls[88]. In this regard, *post-mortem* brain tissues are considered the most valuable resource, as they are ought to be most representative for a neurodevelopmental condition. On the other hand, several authors focused their analyses on how the expression of ASD-related genes (i.e.: SFARI genes) changes among separate brain regions and within different stages of normal development[88]. These studies helped to assess which brain regions and which points during development were mostly affected in ASD. In both cases rather than focusing on the expression of single genes, scientists employed gene co-expression modules analyses, as an approach aimed at identifying functionally related genes that co-vary across samples.

Studies of *post-mortem* brain tissues are usually stirred by data availability and are therefore limited in terms of cohort size. However, several recent works were able to find shared abnormalities in gene expression in a large subset of autism cases, despite genetic variability. Namely, two main kinds of genes co-expression modules are consistently dysregulated: 1) modules of down-regulated genes involved in synaptic function, for neuronal markers and enriched for ASD-associated genes; and 2) modules of up-regulated genes involved in immune and inflammatory responses, enriched for markers of microglia and astrocytes, but not for genes associated with ASD[89–91].

The alternative approach includes the assumption according to which the spatial and temporal expression patterns of ASD-related genes could indicate when and where the role of these genes is mostly relevant. Works as, for example, Parikshak et al.[90] and Willsey et al.[92], employed this strategy to map ASD-associated genes onto co-expression networks representing developmental trajectories and different cortical layers. In short, these studies found that ASD-related genes were especially relevant within superficial cortical layers and glutamatergic projection neurons during the mid-fetal phase, suggesting that specific times and places may be more impacted in ASD development.

Overall, transcriptomic analyses have strongly contributed to better understand the molecular causes of ASD, especially assessing a degree of convergence downstream the largely heterogeneous genomic landscape characteristic of ASD. Furthermore, the most recent studies point to a potential critical involvement of the non-coding genome in the pathogenesis of autism. Consequently, the investigation of epigenetic regulatory networks associated with autism may be a crucial step to bring all the pieces of the ASD puzzle together.

1.3.4 Specific epigenetics alterations characterize ASD

The identification of a broad array of genes as risk factors for ASD made the detection of a common thread rather troublesome. However, the integration of genomic and transcriptomic data highlighted gene modules that are similarly dysregulated across different cohorts of ASD patients. Therefore, it is possible that a specific genome dysregulation may be distinctive of autism.

Indeed there has been a growing interest in the epigenetics of autism[26]. Recently, multiple independent studies, observed epigenetic processes perturbations in *post-mortem* brain tissue of individuals with ASD compared to typically developing controls[87,93,94].

Both Wong et al[87] and Nardone et al[93] reported an altered DNA methylation landscape among multiple brain regions of ASD individuals. Interestingly, they observed an inverse correlation between gene expression and DNA methylation within the individuals. In particular, genomic areas responsible for immune functions were enriched for hypomethylated CpGs, whereas genes related to synaptic membrane were enriched among hypermethylated CpGs. These observations are in line with the overexpression of gene modules relative to immune functions and decreased expression of gene modules specific for neuronal functions detected in ASD. Interestingly, several genes crucial for the regulation of DNA methylation are included in the SFARI database. For example, both rare and common variants within DNA methyltransferases DNMT1 and DNMT3A have been identified as potentially involved in ASD pathogenesis[95]. It may therefore be conceivable that deleterious variants within these genes may be at the basis of the alterations in DNA methylation patterns specific for ASD.

Additionally, several regulatory genes associated to ASD act on the regulation of specific histone marks. Interestingly, most of these marks are typical of bivalent chromatin[9] (namely H3K27me3, and H3K4me3). For example, four lysine methyltransferases (KMT2A, KMT2C, KMT2D, KMT2F), four lysine demethylases (KDM1A, KDM5A, KDM5B, KDM5C), and two reader proteins (PHF21A, PHF8) are specific for H3K4me and are mutated in multiple neurodevelopmental disorders, including autism[24,25]. Instead SFARI genes such as KDM6A are specific for the histone marker H3K27me3[96].

Several of these genes are also considered crucial regulators of the genome during development. This is consistent with the notion according to which bivalent chromatin regions are both tissue-specific and developmentally regulated. Since transcriptomic analyses pointed to middle-fetal developmental phase as the most important phase in ASD development, the loss of function of specific regulatory genes might strongly impact brain development. Furthermore, bivalent chromatin regions are highly enriched

within genes involved in neuronal development, immune responses, and synaptic transmission. Since transcriptomic analyses revealed the modules of genes dysregulated in ASD were mainly involved in immune response and synaptic transmission, it could be speculated that the alteration of bivalent chromatin regions may be the key underlying the convergence in terms of transcriptional de-regulation distinctive of ASD.

Overall, typical epigenetic changes are found in ASD, and they are directly related to observed changes in gene expression. Deleterious mutations within specific chromatin regulating genes may be at the basis for these phenomena.

1.3.5 ASD-related epigenetic changes are concentrated within specific regulatory regions

A specific epigenetic asset likely linked with damaging phenotypic manifestations has been observed in ASD. Recently, multiple authors focused their efforts on the identification and functional characterization of the specific genomic loci modified in ASD.

A work by Corley et at compared the genome-wide methylome in the *post-mortem* brain tissue of ASD individuals with that from healthy fetal brain at different neurodevelopmental stages[26]. The epigenetic alterations detected in ASD were preferentially directed at intragenic and bivalently modified chromatin domains of genes predominately involved in neurodevelopment. Many of these genes were associated with aberrant precursor messenger RNA splicing events in ASD. Interestingly, the methylation landscape in adult neurons affected by ASD closely resemble that characteristic of earlier time points in fetal brain development[26]. These findings may implicate that a delay in the epigenetic program specific for ASD may lead to deleterious transcriptomic events.

In line with these observations, multiple works such as Wang et al.[97] and Marchetto et al.[98] detected an increased proliferation in neuronal precursor cells (NPCs) derived from de-differentiation of fibroblasts isolated from ASD individuals, compared to their healthy counterpart. In particular, Wang et al reported an altered DNA replication program and increased DNA damage in ASD-derived NPCs, especially in replication stress-susceptible genes[97], several of which are associated with ASD pathogenesis. On the other hand, Marchetto et al. highlighted abnormal neurogenesis and reduced synaptogenesis in neurons differenciated from ASD-derived NPCs[98]. Furthermore, a study[99] compared intra-individual differences in gene expression between cerebellum and pre-frontal cortex both within ASD and within healthy individuals. Interestingly, the authors detected a considerably higher number of differentially expressed genes between the two brain regions within healthy compared to ASD individuals[99]. Together this data

may imply that the ASD-specific genome alterations would result in a dysregulation of regulatory regions normally operating in developing neurons, leading to neuronal defects.

Several neurodevelopmental conditions are suspected to be linked to aberrant activity of regulatory regions that recently evolved in human[100]. A recent publication identified a set of human regulatory regions emerged after the separation from old world monkeys[65] and highlighted how these are enriched within genomic regulatory regions altered in ASD[65]. Indeed, genomic loci relative to hominoid-specific regulatory gains show a significant overlap with regions that lose the H3K27ac histone mark, typical of active enhancers in ASD patients[65]. This evidence suggests that ASD-related epigenetic defects may be caused by an abnormal activity of specific evolutionarily young regulatory regions.

The transcription of regulatory regions such as enhancers may be considered evidence for their active state. Specific transposable elements are known to have largely contributed to the formation of regulatory regions, especially those involved in neuronal functions[42,48].

I previously discussed the increased LINE expression in ASD outlined in some recent publications[74]. This evidence may point to an ASD-specific alteration in regulatory genomic elements including LINE transposons resulting from an aberrant epigenetic landscape and leading to transcriptional dysregulations typical of ASD. Intriguingly, Tangsuwansri et al.[75] demonstrated that, in lymphoblastoid cell lines derived from a subset of ASD patients with severe language impairment, the overall methylation level of LINE elements was decreased compared to controls, and this was inversely correlated with the level of expression of LINE-containing genes[75]. That is, only down-regulated genes were enriched in less methylated LINE content. This data may point to an involvement of LINE expression in the dysregulation of genes involved in ASD development. However, it is not clear whether the LINE expression would be a cause or an effect of the underlying molecular alterations.

Final considerations

The reconstruction of the genomic alterations underlying ASD is a rather challenging task. However, accumulating evidence is revealing a fascinating outlook were the existence of a remarkably intricate system of transcriptional regulation has on one hand allowed for the existence of our complex brain, while, on the other hand, increased the susceptibility to neural diseases. Indeed the alteration of specific networks of genomic non-coding regulatory regions may represent a major causative factor in the development of ASD. The phenotypic heterogeneity of ASD implies that such mechanisms may not be shared among all affected individuals, and may be instead distinctive of specific subsets of ASD patients. To this regard, the recently developed technologies may prompt for a proper stratification of ASD into more homogeneous subsets.

1.4 Aim of the project

Genes involved in the genetic etiology of ASD may be divided into two main functional categories. My main objective has been to understand whether ASD individuals carrying deleterious mutations within each of the two classes of genes may present different molecular alterations compared to healthy controls. In particular I wanted to assess whether genes and transposable elements transcription was altered in either of the two groups of patients. I also aimed at functionally characterize the genomic regions potentially involved in the transcriptional dysregulation. Finally, I wanted to assess whether my results may be reproduced and therefore validated from the analysis of additional, independent experiments on ASD-derived samples.

2. Methods

2.1 Datasets used

RNA-seq and WES datasets

I took advantage of public data provided by Velmeshelv et al.[101] (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA434002). This dataset comprises *post-mortem* whole RNA-seq data from anterior cingulate cortex (ACC) and pre-frontal cortex (PFC) and blood-derived WES data of 15 ASD cases and 16 matched controls. Since RNA-seq was performed for both tissues for only 7 out of 15 ASD individuals, while for the remaining individuals only sequencing of either ACC or PFC was performed, the two datasets are only partially overlapping. Furthermore whole exome sequencing (WES) data were available for the 15 ASD samples.

Additionally I retrieved whole RNA-seq data from neurons resulting from the de-differentiation of fibroblasts extracted from 7 ASD patients featuring macrocephaly and 4 matched healthy controls provided by Marchetto et al[98] (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67528).

Chromatin marks

I retrieved already processed chip-seq data from the NIH Roadmap epigenomics mapping consortium[16] (https://egg2.wustl.edu/roadmap/web_portal/processed_data.html). In particular I retrieved bed format data relative to chip-seq experiments for six major histone modification (H3K4me1, H3K4me3, H3K9me3, H3K9ac, H3K27me3 and H3K27ac) from *post-mortem* mid frontal lobe of an healthy individual. I used this data to calculate the enrichment for different histonic markers within differentially expressed (DE) LINEs. Enrichments were computed by overlapping the coordinates of the peaks of each histone mark with the coordinates of DE LINEs. The number of overlaps was then compared with the average and standard deviation resulting from 100 randomizations. For each random permutation, the number of overlaps between the histone marker peaks and a set of LINEs randomly selected among non-DE LINEs and equal in size compared to each set of DE LINEs. Results featuring a Z score < -2 or > 2 (p-value < 0.05) were considered respectively significantly enriched or depleted.

Active enhancer regions

Eight brain regions enhancer-gene networks were used as representative for active enhancers. The study used data from 127 human samples from ENCODE and Roadmap Epigenomics[16] and 808 human samples from FANTOM5[15]. Briefly, the average H3K4me1, H3K27ac, H3K27me3, and Dnase-seq signals were computed for each of the 127 samples; subsequently, this data was integrated with cap analysis of gene expression (CAGE) signals and predicted active enhancers of 808 samples were collected from FANTOM5. Detailed information about this approach can be found in the published study[102]. I retrieved coordinates of active enhancers of eight brain regions, including angular gyrus, substantia nigra, hippocampus middle, germinal matrix, inferior temporal lobe, anterior caudate, dorsolateral prefrontal cortex, and cingulate gyrus.

I used enhancers genomic coordinates to calculate the enrichment DE LINEs and genes. Enrichments were computed by overlapping the coordinates of enhancers with the coordinates of DE LINEs. The number of overlaps was then compared with the average and standard deviation resulting from 100 randomizations computed with 100 sets of random non-DE LINEs, the same as the histone marker enrichment analysis. Results featuring a $Z\ score < -2$ or > 2 (p-value < 0.05) were considered respectively significantly enriched or depleted.

Gene expression ratio between neurons and iPSC

Gene expression ratios between neurons and iPSC cells were retrieved from a recent publication (Lu et al 2020[103]) (https://ars.els-cdn.com/content/image/1-s2.0-S1097276520303920-mmc6.xlsx). The authors calculated expression levels of all genes in iPSC and mature neurons obtained by de-differentiation of *ex-vivo* collected fibroblasts from healthy human individuals. Next the log of the gene expression ratio of each gene between mature neurons and iPSC was calculated. I used this data to assess the expression of specific sets of DE genes between neurons and undifferentiated cells in healthy individuals.

2.2 Samples stratification

Variant discovery was performed on whole exome sequencing (WES) data relative to ASD samples. The exome sequencing raw data were used as input for an automatized pipeline with the purpose to identify and annotate all loci differing from the human reference genome, that I define as variants. The pipeline workflow could be summarized in three steps: 1) alignment, where reads from the sequencing

are mapped to the reference genome; 2) variant discovery, where the alignments are examined to detect variations with respect to the reference human genome; 3) annotation, where a series of functional information are associated with each detected variation. In general, the GATK4 best practices (https://gatk.broadinstitute.org/hc/en-us/articles/360035894731-Somatic-short-variant-discovery-SNVs-Indels-) for the discovery of germinal short variants from trios were followed. Alignment of raw paired-end reads to the reference genome (version hg19) was performed with bwa mem[104] (version 0.7.17) with standard parameters. Subsequently, duplicated reads were marked with picard (version 2.18.20). Variant discovery was then performed with the GATK4 utility HaplotypeCaller, using the appropriate file containing the coordinates of the sequences targeted by the exome sequencing. At the end of this phase a single vcf file was produced. Finally all variants were annotated using annovar[105] (databases updated to 27/05/2019).

ASD samples were assigned to the ASD_reg group according to the presence or of at least one variant which included all the following characteristics: 1) Unannotated or annotated as pathogenic in the database of pathogenic variants clinvar[106], 2) Minimum of 15 supporting reads, 3) Non-synonymous or splicing effect on the transcript, 4) CADD phred score > 20, 5) SIFT pred. = D, 6) Within a SFARI gene (score 1, 2, 3 or S) related chromatin remodeling/transcription regulation. The last point was assessed through a manual review of the SFARI genes bearing the putatively deleterious mutations. Information such as gene ontology, cellular localization and gene function inferred through literature search was used to assign the gene to the 'ASD_reg' group. The SFARI gene database release 2020 Q2 was used. All ASD samples which did not feature at least one genomic variant comprising all the characteristics described above was assigned to the ASD_other group.

2.3 Transposable elements and gene quantification and differential expression

Raw RNA-seq reads were aligned to the reference genome (version hg19) with STAR[107] (version 2.6, standard parameters). Gene expression was quantified with HTseq-count (version 0.11.3; parameters: -f bam -r pos -t exon -i gene_name -a 10) on the BAM files obtained with STAR. The number of reads mapping to each genomic feature defined as 'exon' in the GTF file retrieved from genecode[108] version v32lift37 were quantified. The number of total reads mapping outside exons and the total number of reads displaying a quality too low to be counted were assessed with HTseq as well.

In order to quantify transposable elements expression I used two different software: SQuIRE[109] and TEspeX. The two tools employ distinct strategies to assess TEs transcriptional activity.

TEspeX quantifies TEs transcription levels by counting the number of reads mapping to a reference transcriptome built merging the RepBase[110] human TEs fasta sequences and the Ensembl[111] transcript sequences containing all the human coding and non-coding annotated transcripts using STAR (v2.6.0c). Only reads flagged as primary (−F 0 × 100 parameter) and not mapping to either coding or non coding transcripts are counted using Python scripts and Picard FilterSamReads (v2.18.4). The amount of reads mapping to each TE subfamily fasta sequence is therefore obtained.

On the other hand SQuIRE (software for quantifying interspersed repeat expansion) allows for the quantification of TE upon spliced alignment of RNA-seq data to a reference genome. SQuIRE is a set of computational tools which use different strategies to combine counts from uniquely mapping and multi-mapping reads and then generate counts for individual TE, as well as TEs by class and subfamily. Therefore using SQuIRE it has been possible to analyze transposable element expression on both subfamily and locus-specific (single copies of TEs annotated on RepeatMasker) level. SQuIRE (v0.9.9.92) was used.

Both in the case of genes and of transposable elements, DEseq2[112] (release 3.12) was used for differential expression analysis. DEseq2 analysis was performed with standard parameters on the raw gene counts assessed by HTseq, SQuIRE and TEspeX. Both genes and transposable elements shorter than 200 bp and featuring an average of less than 15 mapped reads among the all the samples were not considerate suitable for differential expression analysis, and were therefore filtered out. Genes and TEs with a FDR < 0.05 and LogFoldChange > 0.5 or < 0.5 were considered as differentially expressed.

2.3 Linear regression experiments

Genomic coordinates of expressed LINE elements were overlapped with those of expressed protein-coding genes with bedtools intersect[113], obtaining couples each made up by a gene and an overlapping LINE. For each couple, a linear regression was performed with the lineregress function of SciPy between the normalized expression of the TE and the gene. Linear regressions characterized by an FDR below 0.05 were considered significant. FDR calculation was performed with Python statsmodels fdrcorrection function. All genes featuring a significant FDR and a coefficient of correlation below 0 were considered as negatively correlated with respect to the expression of the overlapping LINE element.

2.4 Functional enrichments

The tool GREAT[114] was used to perform all enrichment analyses of differentially expressed genes and transposable elements. GREAT assigns biological meaning to a set of non-coding genomic regions by analyzing the annotations of the nearby genes. I assessed functional enrichments for biological process, as most representative of the overall functions of the genes, independently for DE genes and DE LINEs. Gene ontology terms with an FDR < 0.05 were considered significant.

2.5 PCA analyses

PCA experiments were performed with sing the fast.prcomp() function of the gmodels R library upon normalization of genes and transposable elements raw counts performed with DEseq2 (using both estimateSizeFactors and estimateDispersions functions).

2.6 Transposable elements gene coverage

Transposable element coverage relative to each gene was assessed as the percentage of each gene genomic locus occupied by expressed TEs. Genomic coordinates of human genes were overlapped with those of expressed TEs using bedtoos intersect. Subsequently, the percentage of the length of each gene occupied by each of the three main classes of human TEs (LINE, SINE and LTR) was calculated with a custom Python script.

2.7 Transcription factor motifs enrichment

Transcription factor binding sites enrichment analyses were performed with the online tool AME, part of the MEME suite[115]. AME allows for the identification of known motifs that are relatively enriched in a set of query sequences compared with user-defined control sequences. AME supports several types of sequence scoring functions, and it treats motif occurrences the same, regardless of their locations within the sequences. In this case I used the database HOCOMOCO human v11 to search for enrichments. In the case, I used a fasta file containing the sequences of the genomic loci in correspondance to the coordinates of up-regulated LINEs as query and a fasta file containing the sequences of the same number of random LINEs as background for the enrichment. The fasta sequences were extracted from the human reference genome hg19 with bedtools getfasta[113].

3. Results

3.1 Stratification of ASD cases based on the mutational landscape

ASD is a complex neurodevelopmental disease with a strong genetic etiology. Bioinformatic analysis of WES data allows to detect germinal genomic variants within exons; that is, genomic loci which differs from the human reference genome within coding regions. Moreover, variant annotation grants the possibility to functionally characterize each mutation. Filtering of annotated variants can therefore be exploited to separate harmless mutations from potentially deleterious ones. Furthermore, the SFARI gene database provides an extensive list of genes involved in ASD[79]. I took advantage of this knowledge and tools to pinpoint potentially disease-causative variants for each ASD sample. Consequently, I proposed a stratification of ASD cases into two main groups on the basis of the function of mutated genes. All analyses have been performed separately for the two tissues available.

3.1.1 Samples stratification

I retrieved public WES data of *post-mortem* ASD blood samples from Velmeshev et al.[101]. The dataset consisted of a total of 31 samples (from 16 controls and 15 ASD donors). Upon aligning raw fastq reads to the reference genome (version hg19), I used the GATK4 variant discovery pipeline in order to detect all non-reference short variants, following GATK4 best practices as described in methods. The variant discovery phase resulted in the detection of about 40,000 total non-reference variants for each sample. Subsequently, I employed a custom filtering strategy (described in methods) in order to pinpoint only putatively deleterious variants within SFARI genes (SFARI score 1, 2, 3 or S). Variants filtering revealed a total of 16 potentially deleterious variants within SFARI genes in a total of 10 samples, which are reported in table 1. I manually evaluated the molecular function of these genes, highlighting 7 regulatory genes, corresponding to 6 individuals. Furthermore I detected 9 mutations in as many SFARI non-regulatory genes. I was not able to find a suggestive candidate variant for all ASD samples. Table 2 provides detailed information about the SFARI genes involved in deleterious mutations and about the manual curation performed in order to to divide them in the two categories representative of the experimental groups.

In short, I divided ASD samples into two experimental groups: the first (ASD_reg) comprising 6 samples featuring at least one suggestive deleterious mutation affecting a SFARI gene involved in

transcriptional regulation/chromatin remodeling, the second (ASD_other) including the remaining 9 ASD cases. All further analyses take into account this stratification.

RNA-seq data derived from anterior cingulate cortex (ACC) and pre-frontal cortex (PFC) was retrieved for both ASD individuals and healthy controls. However, data from both tissues was available for only 10 individuals, for the remaining 21 individuals only expression data derived from either ACC or PFC was produced. Therefore, the following analyses were performed on a total of 41 RNA-seq derived samples, 18 for ACC (4 ASD_reg, 5 ASD_other and 10 controls) and 21 for PFC (5 ASD_reg, 7 ASD_other and 10 controls).

Chr	Start	End	Ref	Alt	Gene.refGene	ExonicFunc.refGene	SIFT_pred	CADD_phred	SFARI score	Genotype	Sample	Gene class
chr7	107638840	107638840	G	A	LAMB1	nonsynonymous SNV	D	25.8	2	0/1	5841	Non-regulatory
chr11	106558300	106558300	A	C	GUCY1A2	nonsynonymous SNV	D	26.3	3	0/1	5565	Non-regulatory
chr2	166183366	166183366	C	A	SCN2A	nonsynonymous SNV	D	24.2	1	0/1	4849	Non-regulatory
chr6	72952094	72952094	T	A	RIMS1	nonsynonymous SNV	D	26.9	1	0/1	5864	Non-regulatory
chr7	91714220	91714220	A	C	AKAP9	nonsynonymous SNV	D	24.4	2	0/1	5945	Non-regulatory
chr6	43008316	43008316	C	A	CUL7	nonsynonymous SNV	D	24.8	2	0/1	5841	Non-regulatory
chrX	38525554	38525554	G	T	TSPAN7	nonsynonymous SNV	D	28	3	0/1	4899	Non-regulatory
chr20	62124639	62124639	A	T	EEF1A2	nonsynonymous SNV	D	26.5	S	0/1	5531	Non-regulatory
chr17	7096434	7096434	T	G	DLG4	nonsynonymous SNV	D	28.3	1	0/1	5403	Non-regulatory
chrX	53271087	53271087	A	C	IQSEC2	nonsynonymous SNV	D	28	1	0/1	5565	Non-regulatory
chrX	44938475	44938475	C	G	KDM6A	nonsynonymous SNV	D	24	2	1/1	5945	Regulatory
chr18	47800229	47800229	C	T	MBD1	nonsynonymous SNV	D	27	3	0/1	6033	Regulatory
chr14	21899786	21899786	A	G	CHD8	nonsynonymous SNV	D	21	1	0/1	5864	Regulatory
chr12	42854164	42854164	C	T	PRICKLE1	nonsynonymous SNV	D	31	2	0/1	4849	Regulatory
chr5	88100610	88100610	A	T	MEF2C	nonsynonymous SNV	D	26.9	3	0/1	5939	Regulatory
chr5	37010221	37010221	T	C	NIPBL	nonsynonymous SNV	D	21.3	1	0/1	5945	Regulatory
chrX	147924957	147924957	C	T	AFF2	nonsynonymous SNV	D	22.2	1	0/1	5278	Regulatory

Table 1 – putatively deleterious variants within SFARI genes (score 1, 2, 3 or S) found in all WES samples

Gene	Description	Cellular localization	Comment	Experimental group
AFF2	Putative transcriptional activator that is a member of the AF4FMR2 gene family	Nucleus	AFF2 influences splicing and gene expression (PMID: 21330300), it is strongly associated with ASD (PMID: 22065534).	ASD_reg
CHD8	DNA helicase that functions as a transcription repressor by remodeling chromatin structure	Nucleus	CHD8 is a DNA elicase which regulate gene expression Batsukh et al. (2010), multiple studies associated rare CHD8 mutations to ASD (PMID 22495309, PMID 23160995, PMID 25363768, PMID 22521361, PMID 25363760)	ASD_reg
KDM6A	Catalyzes the demethylation of tri/dimethylated histone H3	Nucleus	This gene demethylates specific histone markers thus influencing gene expression (PMCID: PMC4811158), rare LoF mutations suggest that it is an ASD related gene	ASD_reg
MBD1	Binds specifically to methylated DNA and can repress transcription	Nucleus	MBD1 repressed transcription by binding methylated DNA (PMID: 12711603), Genetic association and rare variants have been found in the MBD1 gene associated with autism in a Caucasian and African-American cohort (PMID: 23055267)	ASD_reg
MEF2C	Transcription factor involved in diverse developmental processes including hematopoiesis, cardiogenesis and neurogenesis	Nucleus	This transcription factor is involved in multiple neurological conditions incuding ASD (PMID: 20513142; PMID: 20412115)	ASD_reg
PRICKLE1	Nuclear receptor that may be a negative regulator of the Wnt/beta-catenin signaling pathway and is involved in the planar cell polarity pathway that controls convergent extension during gastrulation and neural tube closure	Nucleus	PRICKLE1 is expressed in distinct neuron populations and contributes to the development of CNS (PMID: 23420014), mouse models Prickle1 +/- showed ASD-like traits (PMID: 24312498)	ASD_reg
NIPBL	Bipartite nuclear targeting sequence and a putative HEAT repeat	Nucleus	This gene is involved in chromatin structure and likely important during development (PMID: 32433956), it has been associated with ASD and other neurodevelopmental conditions (PMID: 15146186)	ASD_reg
GUCY1A2	Soluble guanylate cyclases are heterodimeric proteins that catalyze the conversion of GTP to 3',5'-cyclic GMP and pyrophosphate	Cytosol	The impact of this gene on neuron function is not clear, however it is mostly expressed in the brai,, it shows weak association with ASD (PMID: 30184081).	ASD_other
LAMB1	Extracellular matrix glycoproteins	Extracellular	LAMB1 probably has a crucial role in neuron development, and it is associated to ASD (PMID: 15128462)	ASD_other
AKAP9	Binds to type II regulatory subunits of protein kinase A, specifically found in the neuromuscular junction	Cytosol	This gene encodes for a scaffold protein likely important in the regulation of the integrity between membrane receptors and the inside of the cell, it is therefore likely important for the organization of postsynaptic specialization. LoF variants have been assciated to ASD (PMID: 26402605).	ASD_other
ANK2	Links the integral membrane proteins to the underlying spectrin-actin cytoskeleton	Cytosol	ANK2 links membrane proteins to the cytoscheleton (PMID: 17242276) and it is related to multiple neurological phenotypes including ASD (MIM:600919, PMID 2536376)	ASD_other
DLG4	Membrane-associated guanylate kinase, recruited into NMDA receptor and potassium channel clusters	Cytosol	DLG4 is important for the regulation of postsynaptic density in neurons (PMID: 24778134). It is associated to ASD (PMID: 28191889)	ASD_other
IQSEC2	Guanine nucleotide exchange factor for the ARF family of small GTP-binding proteins, component of the postsynaptic density at excitatory synapses	Cytosol	This gene is involved in is involved in cytoskeletal organization, dendritic spine morphology, and excitatory synaptic organization (PMID: 30328660). It is associated with intellectual disabilities including ASD (PMID: 28815955)	ASD_other
RIMS1	Rab effector involved in exocytosis. May act as scaffold protein that regulates neurotransmitter release at the active zone	Cytosol	This gene is believed to be important for neurotrasmitter release during short-term synaptic plasticity (Etsuko Takao-Rikitsu, 2004). De novo frameshift variants in this gene have been identified in unrelated ASD cases from the Simons Simplex Collection (Iossifov et al., 2012; Dong et al., 2014).	ASD_other
SCN2A	Voltage-gated ion channel essential for the generation and propagation of action potentials	Plasma membrane	SCN2A is a fundamental component of voltage-gated sodium channels (PMID: 28379373), it is therefore likely important for neuronal function. It is strongly associated to ASD (PMID 28379373).	ASD_other
CUL7	Component of an E3 ubiquitin-protein ligase complex that interacts with TP53, CUL9, and FBXW8 proteins	Cytosol	This gene may be an important mediator of proteasomal and lysosomal degradations of potassium channels (PMID: 28098200), its mutation may therefore impact neuronal function, rare mutations have been found in ASD (PMID: 25961944)	ASD_other

Table 2 – summary of the manual curation used to assign each gene to either of the two experimental groups

3.1.2 The transcriptional landscape is altered in ASD_reg samples

ASD_reg samples are characterized by the presence of putative deleterious mutations within genes involved in transcription regulation/chromatin remodeling. I therefore sought to explore whether I could highlight any difference in the transcriptional landscape of ASD_reg samples compared to the other experimental groups. First I quantified gene expression with HTseq, as described in methods. Interestingly, PCA analyses based on the whole expression data show ASD_reg samples to be slightly separated from the other groups in both the tissues analyzed (figure 3.1.2a). This suggests the presence of more differences in gene expression between ASD_reg and control samples with respect to the differences between controls and the rest of ASD samples.

In order to examine the overall gene expression at a genomic structural level, I plotted the normalized number of reads displaying low mapping quality and extra-exonic mapping, as reported by HTseq (figure 3.1.2b). Interestingly, while the number of low-quality reads does not significantly change among experimental groups, the number of reads showing extra-exonic mapping is significantly higher for ASD_reg samples compared to ASD_other and control samples in the ACC (p-value = 0.04), while a similar non-significant trend is present for the PFC (p-value = 0.06) (figure 3.1.2b). This results hints at an overall increase in extra-exonic expression in ASD_reg samples.

A substantial portion of mammal non-coding genome is made up by transposable elements[33,42,48]. I therefore speculated that the increased expression of extra-exonic sequences might involve the transcription of transposable element fragments. I therefore calculated the amount reads mapping on transposons. I used a new method developed in my laboratory, called TEspeX, which assesses TE expression whilst not counting reads possibly generated from TE fragments embedded within annotated transcripts. Details about the transposable elements quantification performed with TEspeX are reported in methods. TEspeX is able to cumulatively quantify the expression of the multiple human TE classes. Since an increased LINE expression in ASD is reported in literature[74], I decided to report the expression level of LINEs separately from the expression level of all other TE classes. Indeed, LINE expression-related PCA analyses show a separation of ASD_reg samples from ASD_other and control samples (figure 3.1.2c). In line with this observation, ASD_reg samples are characterized by a significantly increased expression of only LINE elements in ACC (p-value = 0.01) and PFC (p-value = 0.02) (Figure 3.1.2d). This suggests that the observed increase in the expression of extra-exonic genomic regions in ASD_reg samples may be specifically related to an increased expression of LINE elements.

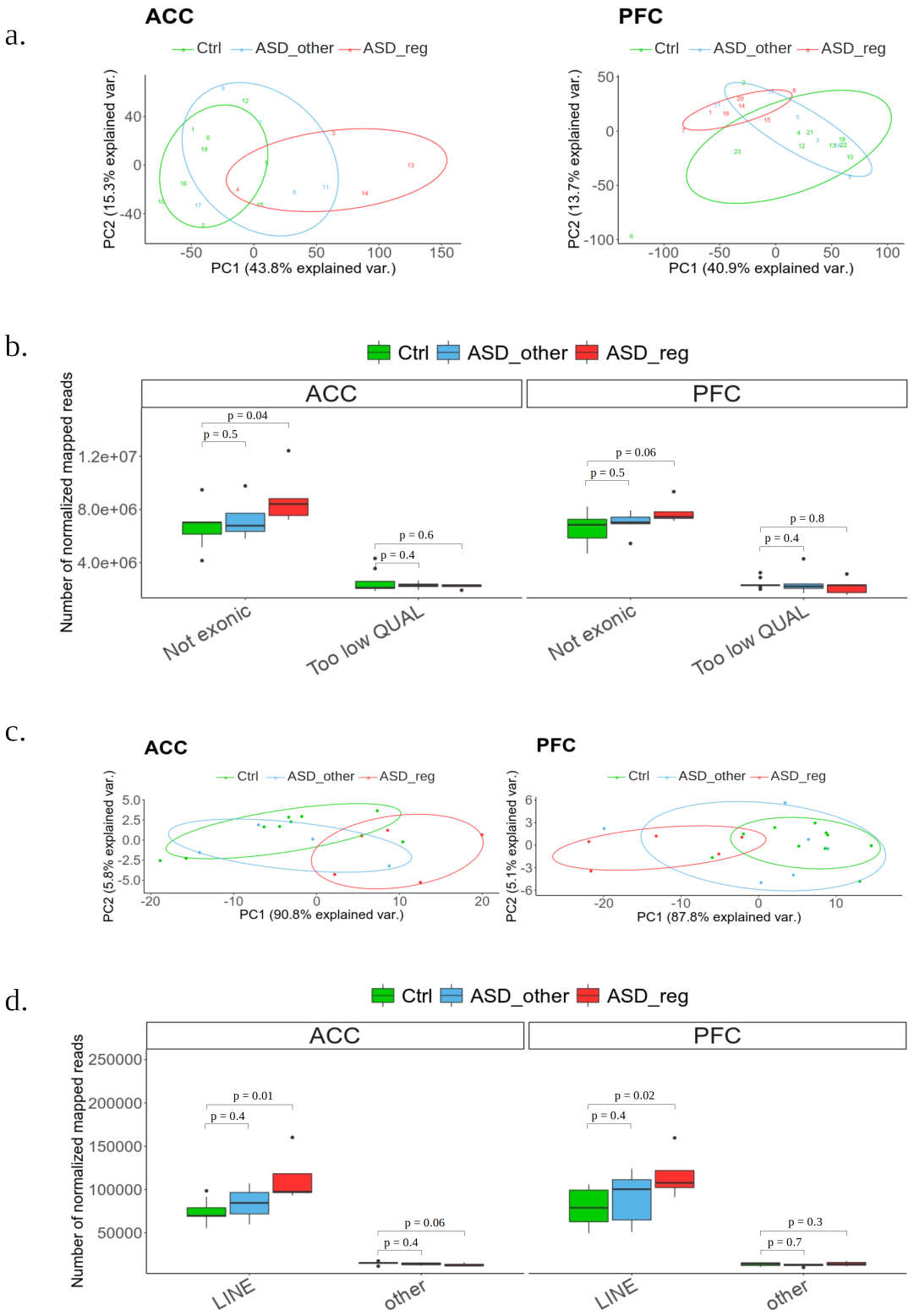

Figure 3.1.2 – a. PCA analysis showing the distribution of all samples according to normalized gene expression levels, samples had been previously divided into three experimental groups as described in methods; b. Global amount of normalized reads mapping outside exons and total amount of normalized reads with a mapping quality score too low to be quantified for the three experimental groups in ACC and PFC; c. PCA analysis showing the distribution of all samples according to normalized LINE expression levels quantified with TEspeX; d. Global amount of normalized reads mapping within LINE elements and within other classes of TEs quantified with TEspeX.

3.2 Specific TEs and genes are differentially expressed in ASD_reg samples

Stratification of ASD samples revealed than one of the two ASD patients subset, ASD_reg, may be characterized by transcriptional differences in gene and transposable elements expression compared to controls, while the other does not. I therefore sought to explore the transcriptional landscape of the experimental groups more in detail. I performed differential expression (DE) analyses relative to gene and transposable elements expression, by comparing each experimental group with controls separately. Consequently I focused on the functional characterization of differentially expressed genes exploring the possible relationship between DE TEs and DE genes.

3.2.1 Transposable elements are differentially expressed only in ASD_reg samples

Several SFARI genes are involved in chromatin remodeling[24]. It is known that aberrant activity of genome regulatory genes may lead to an increased TEs expression[117,118]. Additionally, an increased expression of TEs (especially LINEs) has been observed in autism[74].

Cumulatively, I observed higher LINE expression only in ASD_reg samples. I performed TE-related differential expression analysis comparing independently ASD_reg samples *versus* controls and ASD_other samples *versus* controls in order to confirm these observations and specifically detect which transposable elements drive this increased expression. All differential expression analyses have been performed with DEseq2, as described in methods.

I used two different software to quantify transposable elements expression: SQuIRE and TEspeX. Each of the four main classes of human transposable elements (LINE, SINE, LTR and DNA) is grouped into a number of subfamilies according to the specific sequence divergence from the common ancestor sequence of each class. A total of 1022 human TE subfamilies are annotated in human, each comprising multiple TE fragments spread throughout the genome. TEspeX cumulatively quantifies the expression of TE subfamilies by counting the sequencing reads mapped to the nucleotide sequence of each TE subfamily. On the other hand, SQuIRE cumulatively assesses the amount of reads mapping to TEs at the subfamily level by counting reads mapped to the genomic loci where annotated TEs are found, and by aggregating the counts for all TEs relative to each subfamily. Furthermore, SQuIRE takes into account also reads generated from TE-fragments that are embedded in coding and non-coding annotated transcripts. Additionally SQuIRE can be used to quantify the expression of

transposable elements with a locus-specific resolution, that is, it is able to quantify the expression of each transposable element annotated within the human genome.

Overall subfamily-level differential expression analyses confirm the results observed in the exploratory analysis. Indeed, while no significantly differentially expressed TE subfamily was detected in all ASD_other vs controls experiments, a substantial number of significantly DE TE subfamilies (FDR < 0.05) was detected in all ASD_reg vs controls comparisons, in both tissues analyzed (Figure 3.2.1a,b). On a total of 1022 TE subfamilies analyzed, TEspeX detected a total of 49 differentially expressed TE subfamilies in the ACC (29 up-regulated and 20 down-regulated) and 37 in the PFC (16 up-regulated and 21 down-regulated). As expected, the majority of up-regulated subfamilies are LINEs (25 on 29 in ACC and 9 on 16 in PFC), while only a no LINE subfamilies are down-regulated in both tissues (table 3a) . On the other hand, non-LINE TE subfamilies are mostly down-regulated (table 3a). The total number of up- and down-regulated TEs in ACC and PFC detected with TEspeX is reported in table 3a.

Differential expression analysis was performed on TE subfamily-level expression assessed with SquIRE as well. The results were concordant with the previous analysis. Indeed, on a total of 1022 TE subfamilies analyzed, SQuIRE detected 375 total differentially expressed TE subfamilies in the ACC and 402 in the PFC (FDR < 0.05) (figure 3.2.1c,d). Also in this case LINE subfamilies are mostly up-regulated (102 on 106 in ACC and 75 on 86 in PFC), while other classes subfamilies are mostly down-regulated (figure 3.2.1c,d). The detailed list of DE subfamilies is reported in table 3b.

In order to obtain the specific dysregulated transposable elements differentially expressed in ASD_reg samples, I performed differential expression analysis on locus-specific TE expression levels computed with SQuIRE. In total Squire revealed ~70,000 expressed locus-specific TE fragments (~25,000 LINEs, ~32,000 SINEs, ~8,000 LTR and ~5,000 DNA). Each TE fragment needed to feature an average of at least 15 reads mapped among all samples in order to be considered expressed.

Similarly to previous results, differentially expressed TEs were found only in the ASD_reg *versus* controls comparisons (figure 3.2.1e,f). In total 12658 DE TE fragments were detected in the ACC and 12336 in the PFC (FDR < 0.05). Also in this case the number of up-regulated TEs is higher than the number of down-regulated TEs only for the LINE class. Indeed 4709 out of 6129 LINEs are up-regulated in the ACC and 3034 out of 5342 LINEs are up-regulated in the PFC. On the other hand, all other TE classes are characterized by an higher number of down-regulated fragments compared to up-regulated ones. The detailed list of DE fragments detected with SQuIRE is reported in table 3c.

a.

Number of differentially expressed transposable elements subfamilies (TEspeX)

ACC				PFC			
TE	Up-regulated	Down-regulated	Total	TE	Up-regulated	Down-regulated	Total
LINE	25	0	25	LINE	9	0	9
Non-LINE	4	20	24	Non-LINE	7	21	28
Total	29	20	49	Total	16	21	37

b.

Number of differentially expressed transposable elements subfamilies (SQuIRE)

ACC				PFC			
TE	Up-regulated	Down-regulated	Total	TE	Up-regulated	Down-regulated	Total
LINE	102	4	106	LINE	75	11	86
SINE	2	7	9	SINE	9	9	18
DNA	21	67	88	DNA	26	83	109
LTR	64	104	168	LTR	67	122	189
other	0	4	4	other	0	0	0
Total	189	186	375	Total	177	225	402

c.

Number of differentially expressed transposable elements fragments (SQuIRE)

ACC				PFC			
TE	Up-regulated	Down-regulated	Total	TE	Up-regulated	Down-regulated	Total
LINE	4706	1423	6129	LINE	3034	2308	5342
SINE	1078	2746	3824	SINE	1388	2916	4304
DNA	183	834	1017	DNA	220	1142	1362
LTR	253	1329	1582	LTR	186	1109	1295
other	1	105	106	other	20	13	33
Total	6221	6437	12658	Total	4848	7488	12336

Table 3 – a. number of different classes of TE subfamilies detected with TEspeX; b. number of different classes of TE subfamilies detected with SQuIRE; c. number of different classes of TE fragments detected with SQuIRE.

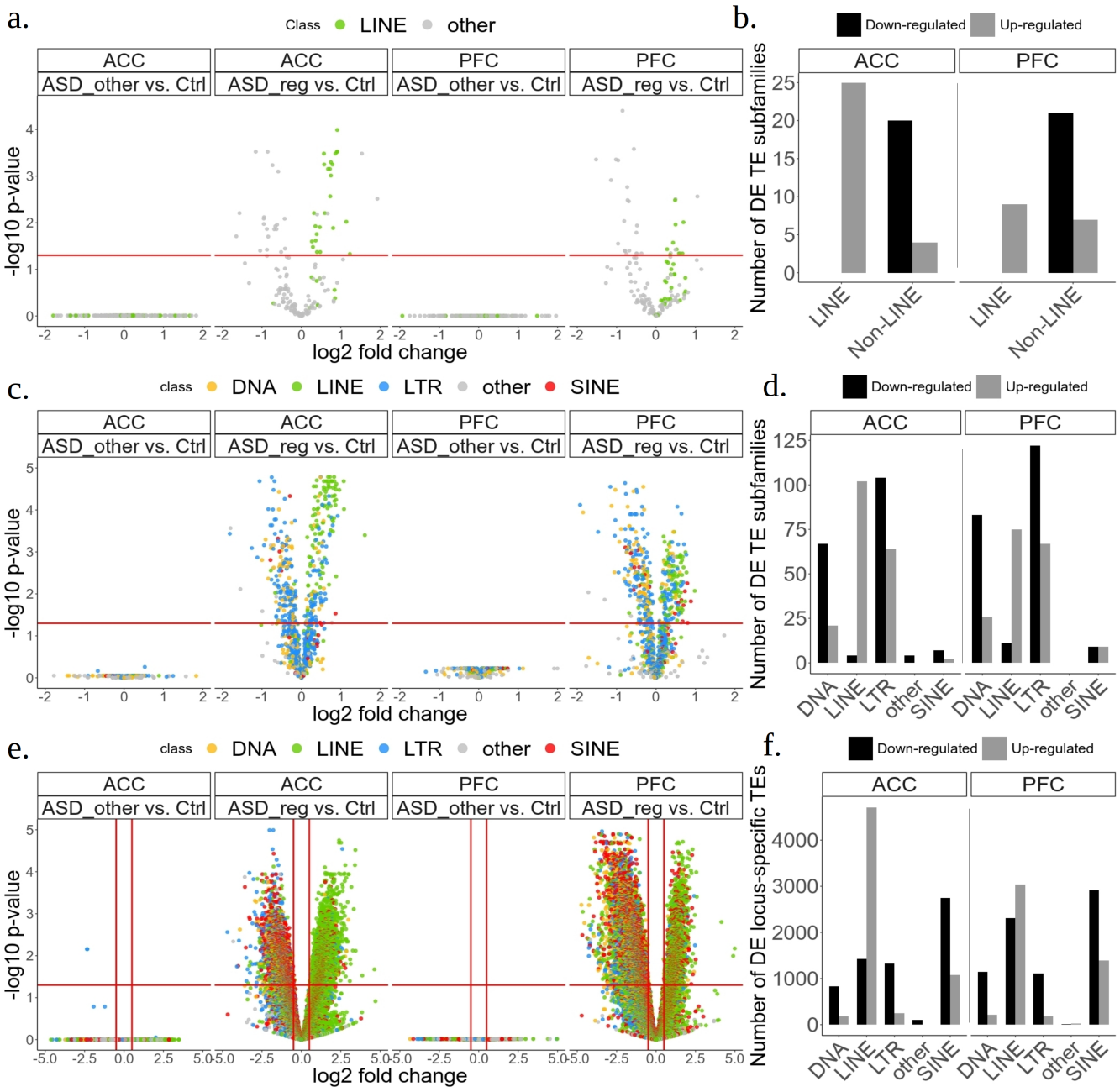

Figure 3.2.1 – a. Volcano plot representing differentially expressed TE LINE and non-LINE subfamilies quantified with TEspeX between ASD_other and controls samples and between ASD_reg and controls samples for ACC and PFC, each dot represents a TE subfamily plotted with respect to the inverse logarithm of its FDR value and the logarithm of its fold change expression with respect to controls, significance line: FDR = 0.05, |fold-change| > 0.5; b. Number of DE TE LINE and non-LINE subfamilies quantified with TEspeX in ACC and PFC, DE elements are divided into up-regulated (fold-chance > 0) and down-regulated (fold-change < 0); c. Volcano plot representing differentially expressed TE subfamilies quantified with SQuIRE between ASD_other and controls samples and between ASD_reg and controls samples for ACC and PFC, each dot represents a TE subfamily plotted with respect to the inverse logarithm of its FDR value and the logarithm of its fold change expression with respect to controls, significance line: FDR = 0.05, |fold-change| > 0.5; d. Number of DE TE subfamilies quantified with SQuIRE in ACC and PFC, DE elements are divided into up-regulated (fold-chance > 0) and down-regulated (fold-change < 0); e. Volcano plot representing differentially expressed TE fragments quantified with SQuIRE between ASD_other and controls samples and between ASD_reg and controls samples for ACC and PFC, each dot represents a TE fragment plotted with respect to the inverse logarithm of its FDR value and the logarithm of its fold change expression with respect to controls, significance line: FDR = 0.05, |fold-change| > 0.5; f. Number of DE TE fragments for each TE class quantified with SQuIRE in ACC and PFC, DE fragments are divided into up-regulated (fold-chance > 0) and down-regulated (fold-change < 0).

3.2.2 LINE up-regulation is concentrated within intronic LINEs of neuronal genes

An increased expression of extra-exonic sequences and LINE elements has been observed in ASD_reg samples. In order to assess the average genomic localization of DE LINEs in further detail, I overlapped the genomic coordinates of locus-specific DE LINEs with the genomic coordinates of exons, and introns retrieved from Ensembl[111], as described in methods. All LINEs not overlapping with either introns or exons have been considered as extra-genic. The analysis was performed separately for up-and down-regulated LINEs. Interestingly, the distribution of up- and down-regulated LINEs among the three kinds of genomic localizations is remarkably different. Indeed, in ACC and PFC ~50-70% of down-regulated LINEs are exonic, with intronic and extra-genic LINEs fairly equally distributed among the remaining percentage. On the other hand, ~90% of up-regulated LINEs are intronic in both ACC and PFC, while the percentage of both exonic and extra-genic LINEs is below 10% (figure 3.2.2a). The percentage of intronic up-regulated LINEs in both ACC and PFC is significantly higher compared to the fraction of total annotated human LINEs in the genome (fisher test, p-value < 10e-05). Several transposable elements fragments can be transcribed independently from the host gene, as they carry promoter sequences[33,48,119]. This is especially evident when the strand of transcription of the expressed TE is the opposite compared to the host transcript. I wanted to test whether the expression of up-regulated LINEs, which are mostly intronic, was more likely independent or linked to the transcription of the host transcript. I therefore calculated the percentage of up-regulated LINEs overlapping expressed genes transcribed from the same strand as the overlapping gene. Interestingly, 99% of up-regulated LINEs is transcribed from the same strand as the overlapping host gene in both ACC and PFC (Figure 3.2.2b). On the other hand, a lower percentage, about 83% of down-regulated LINEs and about 52% of total annotated LINEs are oriented on the same strand as the overlapping gene.

In order to evaluate the potential functional impact specific for up- and down-regulated LINEs, I performed functional enrichments analyses focused on the genomic loci of DE LINEs with the tool GREAT[114]. GREAT allows to assign biological meaning to a set of non-coding genomic regions by analyzing the annotations of nearby protein coding genes. Interestingly, the top 5 most significant (FDR < 0.05) functional biological process enrichments are strongly different between up- and down-regulated LINEs (figure 3.2.2e). Down-regulated LINEs are enriched in proximity to genes involved in metabolic processes such as response to fluid shear stress and erythrocyte maturation in the ACC and

mitochondrial ATP synthesis and purine ribonucleoside triphosphate biosynthetic process in the PFC. On the other hand, up-regulated LINEs are enriched within genes important for regulation of GTPase activity, double strand breaks repair and neuronal maturation in the ACC and several biological processes important for neuronal functions, such as regulation of axon guidance and action potential in the PFC. Furthermore intriguingly, increased LINE transcription, especially in the PFC, is concentrated in proximity to genes important for neuronal functions. The dysregulation of gene module involved in neuronal functions reported in ASD may suggest that increased LINE expression may contribute to the dysregulation of these genes.

Taken together, these results support to the idea that only the subgroup of ASD patients I named ASD_reg is characterized by pervasive up-regulation of LINE elements. Specifically, intronic LINE fragments overexpression might drive the observed alterations. Furthermore, LINE up-regulation occurs within genomic regions involved in crucial neuronal functions (especially in the PFC), and might therefore be involved in the development of the neurodevelopmental phenotype.

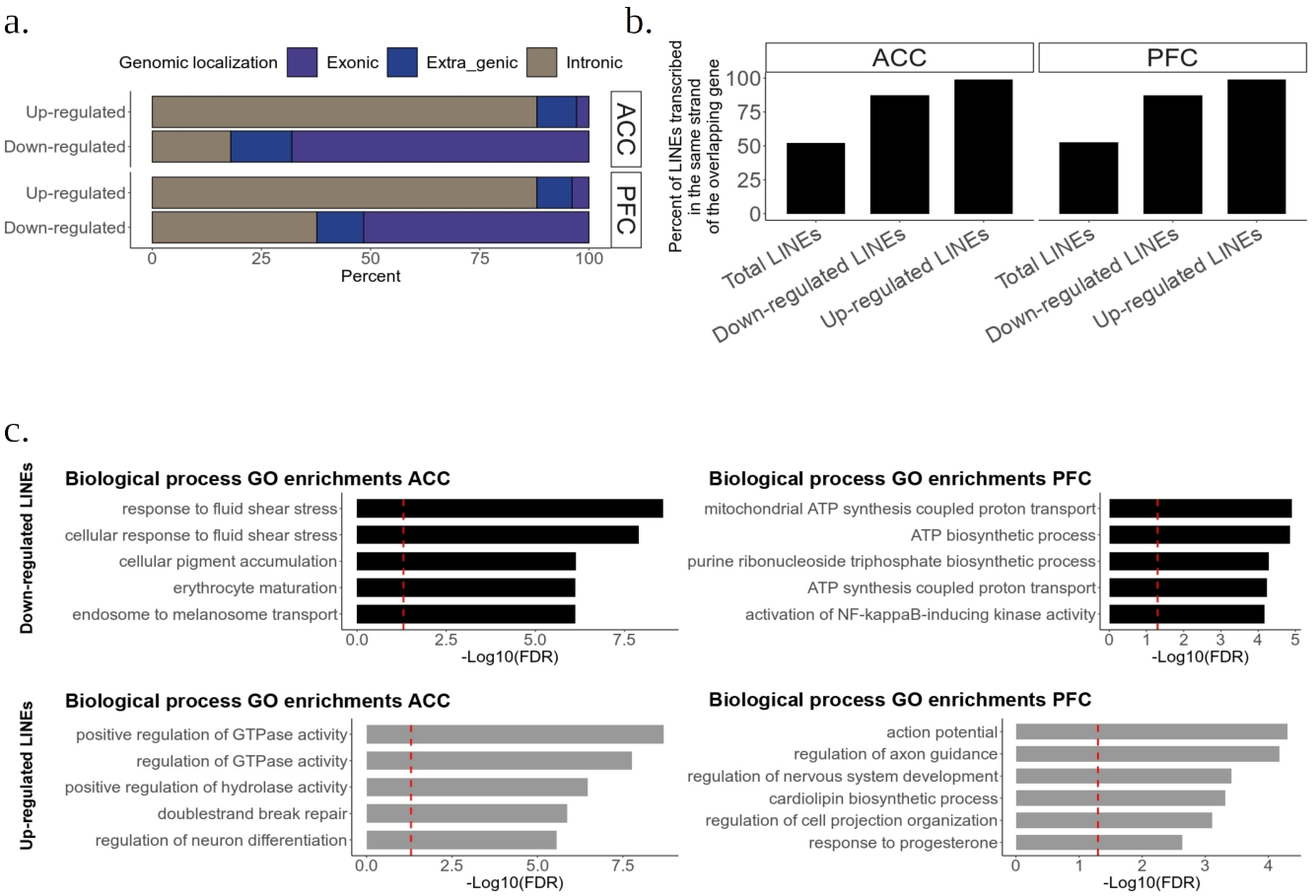

Figure 3.2.2 – a. percentage of up- and down- regulated LINEs overlapping intron, exons and extra-genic regions for ACC and PFC; b. percentage of up-regulated, down-regulated and total annotated LINEs whose sense of transcription is the same as the overlapping gene (*|Z-score| > 2, p-value < 0.05); c. Functional enrichments relative to DE LINEs for ACC and PFC, the significance line is relative to an FDR = 0.05.

3.2.3 Genes are differentially expressed only in ASD_reg samples

Differential expression analyses highlighted the presence of a consistent amount of differentially expressed TEs in ASD_reg samples compared to controls. I performed differential expression analysis with DEseq2 also on canonical gene expression. Gene quantification analysis resulted in ~16,000 expressed protein coding genes, that is, genes identified by at least an average of 15 mapped reads among all samples. Similarly to DE TEs results, gene-level DE analyses resulted in no significantly DE genes for all ASD_other vs controls comparisons (figure 3.2.3a,b). On the other hand, a substantial number of both significantly up- and down- regulated genes resulted from all ASD_reg vs controls comparisons (FDR < 0.05) (figure 3.2.3a,b). Specifically, I detected 389 down-regulated and 631 up-regulated genes in ACC and 833 down-regulated and 952 up-regulated genes in PFC. Overall, all differential expression experiments showed pervasive genome-wide dysregulation of genes and TEs expression only in ASD_reg samples.

Overall I observed pervasive alterations in both genes and transposable elements expression only in ASD_reg samples, which are characterized by a putatively deleterious mutation within a gene involved in chromatin organization and transcriptional regulation.

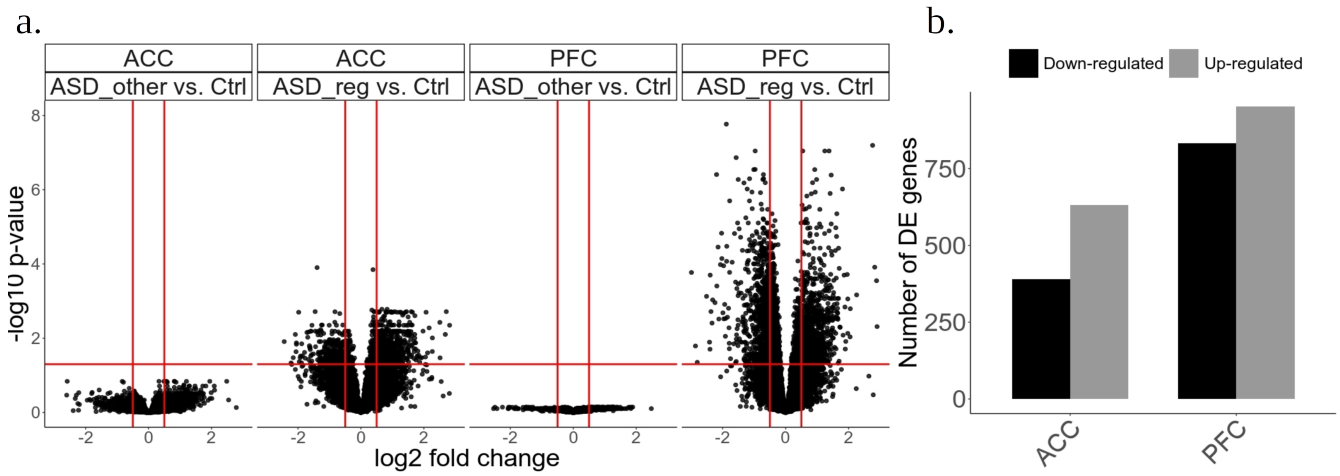

Figure 3.2.3 – a. Volcano plot representing differentially expressed genes quantified with HTseq between ASD_other and controls samples and between ASD_reg and controls samples for ACC and PFC, each dot represents a gene plotted with respect to the inverse logarithm of its FDR value and the logarithm of its fold change expression with respect to controls, significance line: FDR = 0.05, |fold-change| > 0.5; b. Number of genes quantified with HTseq in ACC and PFC, DE genes are divided into up-regulated (fold-chance > 0) and down-regulated (fold-change < 0).

3.2.4 Down-regulated genes are mostly related to neuronal functions

ASD_reg samples are characterized by the presence of differentially expressed genes and transposable elements, while ASD_other samples are not. Several articles describe transcriptional alteration in ASD[88–90]. I therefore explored DE genes more in detail in order to assess whether the aberrant transcriptional landscape is consistent with literature.

In order to functionally characterize DE genes, I performed GO biological process enrichment analyses with GREAT on both up-regulated and down-regulated genes separately. Interestingly, down-regulated genes in ASD_reg samples are enriched for GO classes relative to functions potentially related to ASD, such as chemical synaptic transmission in ACC and potassium ion transport in PFC (FDR < 0.05) (Figure 3.2.4a). On the other hand, up-regulated genes did not show any significant enrichment in biological process for both ACC and PFC. This result is consistent with multiple works focused on ASD transcriptomics which highlight the presence of gene co-expression modules involved in synaptic functions only within down-regulated genes[88–91].

The expression of neural-specific genes usually increases during the differentiation of neural cells. Consequently, I speculated that down-regulated genes would, on average, feature an higher expression level in mature neurons compared to undifferentiated cells. To test this hypothesis I leveraged data from Lu et al. 2020[103], in particular I retrieved the ratio of expression between neurons and iPSC for each gene. Genes resulting down-regulated in the comparison of ASD_reg *versus* controls show a significantly higher average expression ratio between mature neurons and iPSCs compared to up-regulated genes in both ACC and PFC (figure 3.2.4b). Suggesting that genes down-regulated in ASD_reg samples may have an higher importance in neuronal functions compared to up-regulated genes. Furthermore, while ~75% of down-regulated genes are characterized by a positive expression ratio between neuron and iPSCs, more than 70% of up-regulated genes feature a negative ratio (figure 3.2.4b).

Finally, only down-regulated genes are significantly enriched for genes present in the SFARI database (Z-score > 3, p-value < 0.01), highlighting the fact that most ASD-related DE genes are down-regulated (figure 3.2.4c) in ASD_reg patients.

Taken together these results point to a critical relevance of down-regulated genes in the pathogenesis of ASD_reg patients.

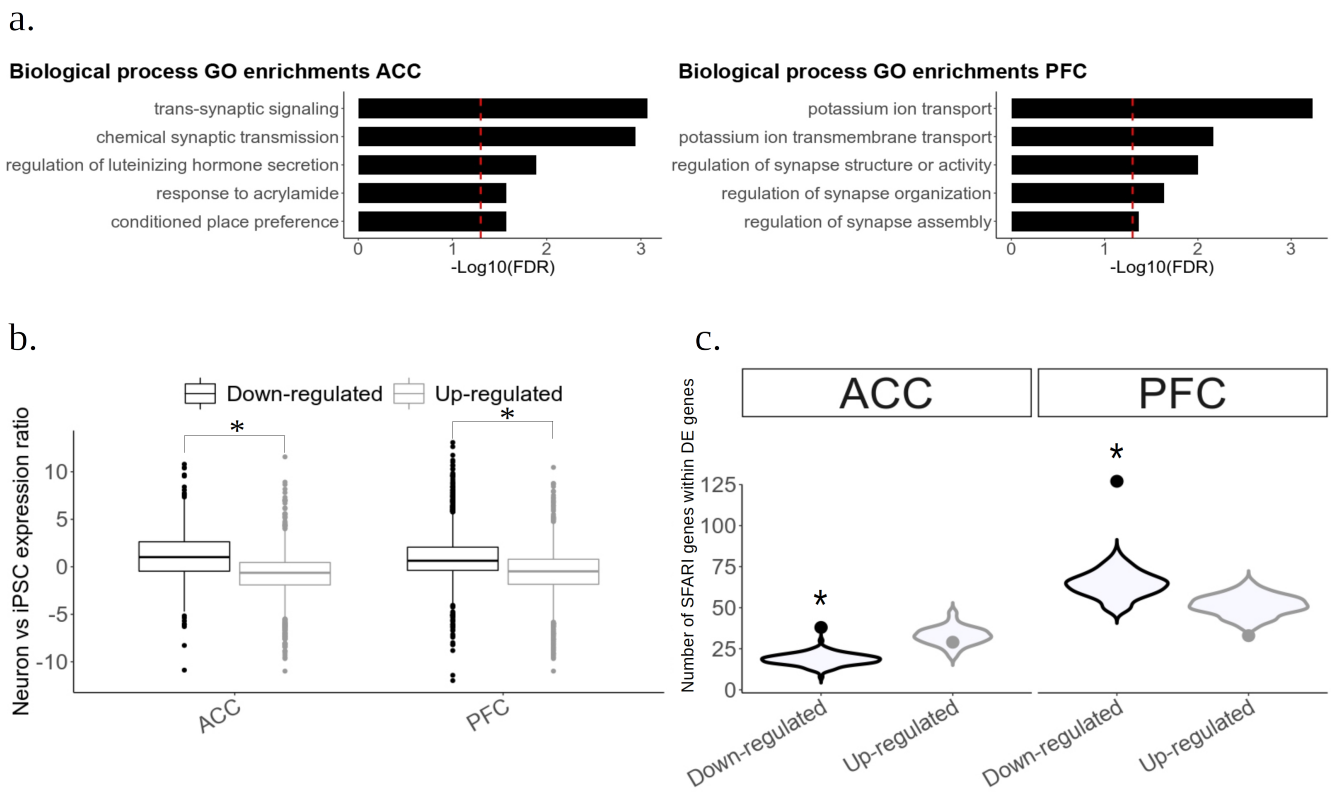

Figure 3.2.4 – a. Functional enrichments related to biological processes for genes down-regulated in ASD_reg *versus* control samples comparisons in ACC and PFC, the significance line corresponds to an FDR = 0.05; b. Normalized gene expression ratio between neurons and iPSC (data from Lu et al. 2020) relative to each gene resulted up- and down-regulated in the ASD_reg *versus* control samples comparisons in ACC and PFC, *p-value < 0.05; c. SFARI genes enrichments within up- and down-regulated genes in the ASD_reg versus control samples comparisons in ACC and PFC, the dot represents the number of SFARI genes (score 1, 2, 3 or S) present within each set of DE genes, while the violin represents a random distribution obtained by counting the number of SFARI genes present in 100 sets of randomly selected expressed genes, *Z-score > 2, p-value < 0.05.

3.2.5 Up-regulated LINEs may be involved in the down-regulation of neuronal genes

ASD_reg samples feature an overall different transcriptional landscape compared to other ASD samples and controls. Most importantly, I observed an increased LINE expression and down-regulation of ASD-related genes. Given that functional enrichments relative to both up-regulated LINEs and down-regulated genes mainly include biological processes crucial for neuronal activity, I wanted to assess whether LINE up-regulation may be related to the dysregulation of genes related to neuronal functions.

SFARI genes are genes whose alteration is strongly related to the development of an ASD phenotype[79]. I reasoned that an higher presence of LINE elements expressed in the brain within SFARI genes genomic loci may make SFARI genes specifically susceptible to alterations in LINE elements expression. I therefore calculated the percentage of each SFARI gene occupied by each of the three main classes of human transposable elements (LINE, SINE and LTR) expressed in my samples, and defined it as TE coverage. SFARI genes TE coverage was compared with a random distributions of coverage computed with the same number of random genes. Interestingly, expressed LINEs occupy a significantly higher fraction of the gene length for SFARI genes compared to random genes (Z-score > 3, p-value < 0.05) (figure 3.2.5a); while expressed SINEs and LTRs do not (figure 3.2.5a). This evidence may suggest that the observed alteration in the expression of LINE elements could be related to the etiology of ASD.

It is generally expected that transposable element fragments embedded within coding genes would vary their expression levels together with the gene transcript. However, since ASD-related genes are enriched in active LINE elements, I sought to test whether up-regulated LINEs may be found within down-regulated genes in ASD_reg samples. Interestingly, I observed the presence of up-regulated LINE fragments overlapping down-regulated genes. Strikingly up- regulated LINEs are significantly enriched within down-regulated genes in both ACC (Z-score = 2.6, p-value < 0.01) and PFC (Z-score = 7, p-value < 10e-05) (figure 3.2.5b). This result may suggest an impact of LINE up-regulation on down-regulation of specific genes. Furthermore, in the PFC, up-regulated genes are significantly depleted of both up-regulated and down-regulated LINEs (Z-score < 2, p-value < 0.01). Previous results showed that up-regulated genes are more expressed in undifferentiated cells compared to mature neurons. This result may imply that genes generally devoid of LINE elements and highly expressed in developing cells are up-regualted in ASD_reg samples.

The presence of up-regulated LINEs within down-regulated genes begged the question whether this kinds of phenomena would be unique to ASD or present in healthy tissues as well. Namely I wondered whether a subset of genes would change their expression level in the opposite direction with respect to the overlapping LINE. Therefore, I overlapped the genomic coordinates of all expressed genes with the genomic coordinates of all expressed intronic LINEs, obtaining a list of 13976 gene-LINE couples. Then, for each couple, I computed a linear regression between the normalized expression level of the gene and the LINE element among healthy control samples. On the total gene-LINE couples correlations, 10465 couples, relative to 2990 genes, featured a significant FDR (< 0.05), and among these, a fraction of 706 genes was characterized by a negative coefficient of correlation ($r < 0$). This set of genes I called 'NEG genes' putatively change their expression level discordantly with respect to an overlapping intronic LINE.

Interestingly this gene set is significantly enriched for genes resulting down-regulated in the ACC of ASD_reg samples (Z-score > 2, p-value < 0.05) and in the PFC of ASD_reg samples (Z-score > 3, p-value < 0.01) (figure 3.2.5c) compared to 100 random distributions computed with 100 sets of the same number of random non-DE genes. Furthermore, functional enrichments performed with GREAT revealed that NEG genes are significantly enriched for regulation of dendritic spine morphogenesis biological process (FDR = 4e-04). These results suggest that in normal healthy brain tissues a subset of transcribed LINE fragments vary their expression discordantly compared to overlapping genes involved in neuronal functions, and this mechanism may be enhanced in a subset of ASD_reg patients.
Interestingly, in both tissues analyzed, there is more convergence in DE LINEs than DE genes. Indeed, while 50-60% of total DE LINEs are shared between ACC and PFC, only <15% of total DE genes are shared between ACC and PFC (figure 3.2.5d). This result suggests that the impact of deleterious mutations within SFARI regulatory genes may have differential impact in different tissues, and that such differences may be more marked at gene level than at LINE expression level.
Taken together, these results highlight a potential impact of LINE dysregulation on the aberrant expression of genes crucial for neuronal functions. The increased LINE expression observed in ASD_reg samples may therefore be a causative factor in the development of the observed neurological phenotype.

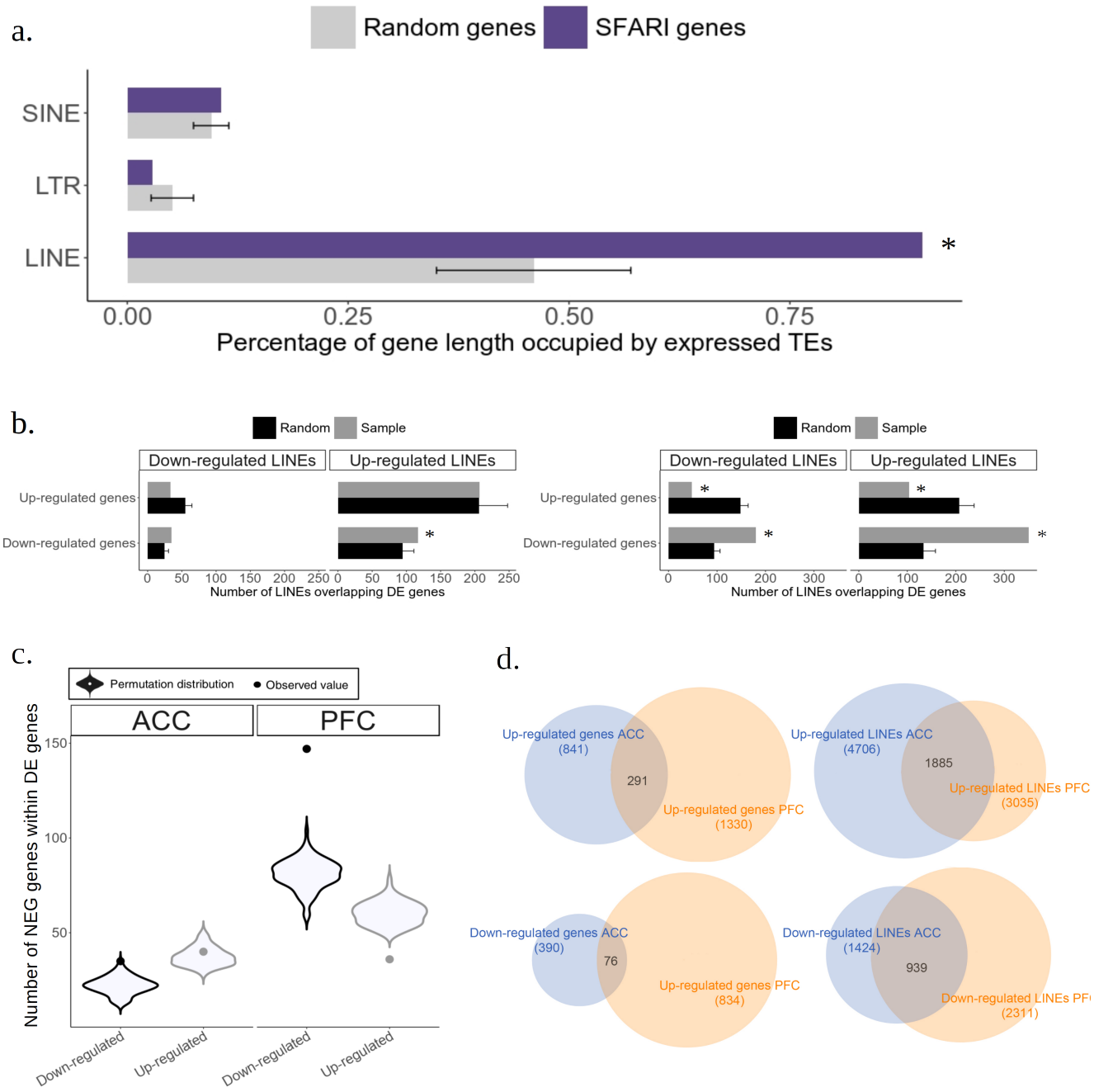

Figure 3.2.5 – a. percentage of the sum of the total length of SFARI genes (score 1, 2, 3 and S) occupied by different classes of transposable elements, this percentage is compared with the average percentage of 100 sets of the same number of the total length of random genes occupied by the same classes of TEs, *Z-score > 3, p-value < 0.01; b. Number of overlaps between different sets of DE LINEs and DE genes, each number of overlaps is compared with the number of overlaps obtained with 100 sets of random expressed genes, *|Z-score| > 3, p-value < 0.01; c. Enrichment for genes whose expression is inversely correlated with respect to the expression of an overlapping LINE element (NEG genes), the dots represent the number of NEG genes in common with each set of DE genes, while the violin represents a random distribution obtained by counting the number of NEG genes occurring within genes present in 100 sets of randomly selected expressed genes of equal number compared to each set of DE genes, *Z-score > 2, p-value < 0.05; d. Venn diagrams displaying the amount of overlap between different sets of DE LINEs and genes in ACC and PFC.

3.3 Functional characterization of DE LINEs loci

ASD_reg samples show a peculiar transcriptional landscape compared to both controls and other ASD cases in ACC and PFC. Interestingly, intronic up-regulated LINEs are enriched within down-regulated genes involved in neuronal functions, while down-regulated LINEs are not. I sought to functionally characterize genomic loci corresponding to up- and down-regulated LINEs in order to possibly trace the observed transcriptional alterations to specific epigenetic modifications.

3.3.1 Down-regulated and up-regulated LINEs overlap different regulatory regions

It is generally possible to imply the regulatory function of a non-coding DNA sequence by assessing which kind of epigenetic marks act upon that specific region in a cell line. I therefore took advantage of already processed chip-seq data from the NIH Roadmap epigenomics mapping consortium. In particular I retrieved chip-seq data for six major histone modification (H3K4me1, H3K4me3, H3K9me3, H3K9ac, H3K27me3 and H3K27ac) from *post-mortem* mid frontal lobe of an healthy individual. Data was retrieved in bed format. I used this data to calculate the enrichment for different histonic marks within DE LINEs. Specifically, I overlapped the genomic coordinates of DE LINEs and histone marks. I then compared, for each histone marks, the number of overlaps with the number of overlaps obtained by 100 random distributions, computed by counting the overlaps of histone marks with sets of non-DE LINEs. Interestingly, up-regulated LINEs are enriched within all histone marks tested: H3K4me1, H3K4me3, H3K9me3, H3K9ac, H3K27me3 and H3K27ac (Z-score > 2, p-value < 0.05) in ACC and all tested marks except from H3K27ac in PFC (Z-score > 2, p-value < 0.05) (figure 3.3.1a). On the other hand down-regulated LINEs are underrepresented within repressive marks: H3K4me1, H3K4me3, H3K9me3 and H3K27me3 (Z-score < 2, p-value < 0.05) and non-significant within activatory marks (H3K27ac and H3K9ac) both in ACC and PFC (figure 3.3.1a). This evidence may suggest that up- and down-regulated LINE elements may mark regulatory regions with different roles in the control of gene expression.

Since specific histone marks are indicative of specific effect on transcriptional regulation, I wanted to elaborate on this result by overlapping DE LINEs with coordinates of enhancers active in the CNS to try to understand whether genomic regions altered in ASD_reg samples may indeed have a functional link to gene expression. I therefore used data from a recent publication[102], where the authors integrated

data from ENCODE, FAMTOM5 and Roadmap and reconstructed the enhancer–target networks in 935 samples of human primary cells, tissues and cell lines.

In particular I retrieved the genomic coordinates of eight brain regions, including angular gyrus, substantia nigra, hippocampus middle, germinal matrix, inferior temporal lobe, anterior caudate, dorso-lateral prefrontal cortex, and cingulate gyrus; and used this information to assess whether DE LINEs were enriched within enhancers which regulate the expression of genes in CNS tissues. Similarly to with previous analyses, the number of overlaps between DE LINEs and enhancers was compared with the number of overlaps resulting from 100 random distributions, computed with the genomic coordinates of sets of non-DE LINEs. Interestingly, for all 8 brain regions, only down-regulated LINEs are enriched within enhancers (figure 3.3.1b). On the other hand, up-regulated LINEs do not show significant enrichment or underrepresentation among enhancers.

Nucleotide sequences embedded within transposable elements are able to provide binding motifs to endogenuos transcription factors, thus taking part in gene expression regulation. Specific transcription factors have different roles in cell development and functions. Enrichment for specific TF binding motifs may therefore functionally characterize a subgroup of genomic segments. I therefore obtained the nucleotide sequences of DE LINE elements and performed TF binding motif enrichment experiments separately on up- and down-regulated LINEs. Enrichments were performed using the AME tool and by using a set of nucloetide sequence relative to random expressed LINE sequences as background. I then compared the top 30 most enriched motifs for both sets of DE LINEs and both tissues analyzed (FDR < 10e-12). Interestingly, both in ACC and PFC, up-regulated and down-regualted LINEs show a substantially different set of enriched TFs, while there is an extensive overlap between the two tissues (figure 3.3.1c). This result further suggests that genomic loci relative to up- and down-regulated LINEs may represent functionally different regulatory elements. The list of the 30 most enriched TF motifs for each set of DE LINEs is reported in table 4.

Interestingly, the top 30 transcription factors most enriched within up-regulated LINE loci include several members of TF families involved in neurodevelopment and immune response. For example binding motifs specific for multiple members of the interferon-regulatory factor (IRF) family (IRF1, IRF2, IRF3, IRF7 and IRF8) are enriched within up-regulated LINEs. These TFs regulate the expression of interferons, important mediators of innate immunity[120]. Alteration in the accessibility of IRF-specific motifs in ASD_reg samples may be in line with the increased innate immunity activation

reported in ASD[88]. Furthermore, TF motifs among the 30 most enriched within up-regulated LINE sequences, include multiple TFs whose activity is crucial for neurodevelopment, such as several members of the FOX family (FOXP1, FOXO4, FOXK1, FOXJ2, FOXJ3), MEF2C and SOX5, whose importance in early brain development has been underlined[121–123]. On the other hand, most of the top 30 most enriched TF motifs within down-regulated LINE sequences are mostly involved in cell cycle and proliferation. However, the most enriched transcription factor binding motif within down-regulated LINEs is specific for the protein FEV. This TF belongs to the ETS transcription factor family. This gene is exclusively expressed in neurons of the central serotonin (5-HT) system, a system implicated in the pathogeny of psychiatric diseases such as depression, anxiety, and eating disorders[124].

Overall the analysis of the functional role of the genomic loci relative to DE LINEs revealed that LINE up- and down- regulation may represent a mark of the alteration of genomic regions with different regulatory purposes whose alteration may be causative for the phenotype.

ACC				PFC			
Down-regulated LINEs		**Up-regulated LINEs**		**Down-regulated LINEs**		**Up-regulated LINEs**	
Transcription factor	**adj_p-value**	**Transcription factor**	**adj_p-value**	**Transcription factor**	**adj_p-value**	**Transcription factor**	**adj_p-value**
FEV	7.4E-39	PRDM6	3.05E-172	NKX25	6.69E-42	PRDM6	5.34E-81
NKX25	1.07E-31	GATA3	9.39E-128	FEV	4.13E-40	SRY	4.7E-70
SMAD3	1.23E-25	SRY	2.55E-125	NF2L1	8.17E-36	GATA3	2.01E-49
ISL1	3.36E-22	IRF1	6.35E-107	SMAD3	1.74E-30	IRF1	2.74E-47
ZBT48	6.17E-22	IRF7	1.27E-94	TGIF1	4.38E-28	SOX5	1.78E-41
THA	5.04E-20	STAT2	3.81E-70	ISL1	1.33E-24	FOXO4	3.97E-37
RORA	1.98E-19	IRF3	1.83E-51	THA	3.09E-23	IRF7	2.55E-35
TBX21	2.01E-19	FOXO4	5.95E-50	IKZF1	2.38E-21	FOXK1	9.38E-26
TGIF1	3.59E-19	FOXK1	2.18E-45	ETS2	7.6E-21	IRF2	1.35E-24
NR1H3	1.83E-18	IRF2	4.37E-44	LYL1	2.2E-20	FOXP1	4.15E-22
E2F3	1.03E-17	SOX5	1.08E-42	NKX21	2.81E-20	FOXJ3	6.4E-22
NF2L1	2.26E-17	FOXP1	3.26E-40	TBX3	6.15E-19	STAT2	6.79E-20
LYL1	1.16E-16	FOXJ3	1.74E-36	E2F3	8.18E-19	NFAC2	1.17E-18
ZN667	1.78E-16	ZFP28	2.4E-32	FOXO1	5.26E-18	NFAC4	3.33E-18
AP2B	1.86E-16	NFAC4	1.93E-31	SUH	7.16E-18	TBP	6.53E-18
RXRB	2.08E-15	NFAC1	1.04E-30	RORA	1.01E-17	PIT1	8.08E-18
ZBT7A	5.83E-15	MEF2B	2.18E-30	NR1H3	2.31E-17	CDX1	2.9E-15
ZN322	1.1E-14	NFAC2	9.31E-29	TBX21	6.04E-17	IRF3	1.64E-13
TAF1	5.14E-14	MEF2A	2.66E-28	PBX1	1.16E-16	MEF2A	3.36E-12
PBX3	5.32E-14	MEF2C	3.7E-24	FOS	1.02E-15	HXA10	4.36E-12
NKX21	5.46E-14	STA5B	5.14E-20	TAF1	2.02E-15	PRRX2	7.95E-12
TBX3	2.52E-13	PRDM1	5.58E-20	ZN322	3.67E-15	HXA13	3.14E-11
IKZF1	2.67E-13	IRF8	2.55E-18	FOSB	2.08E-14	MEF2C	5.9E-11
PBX1	2.78E-13	PIT1	1.74E-16	AP2B	3.47E-14	FOXJ2	3.19E-10
NR4A1	3.32E-13	HXA13	1.43E-15	HIF1A	6.01E-14	ALX1	1.61E-09
SUH	3.74E-13	TBP	2.84E-15	ETV5	1.22E-13	NFAC1	1.94E-09
ZIC1	2.52E-12	STA5A	2.22E-13	MYB	1.37E-13	ZFP28	3.89E-09
MAFB	1.03E-11	CDX1	7.39E-13	ZBT48	2.08E-13	NKX32	5.06E-08
ATF1	2.11E-11	FOXJ2	8.91E-13	ZIC1	2.36E-13	MEF2B	1.57E-07
NKX31	2.35E-11	GATA6	1.3E-12	RXRB	5.98E-13	MEF2D	3.27E-07

Table 4 – Transcription factors whose binding motifs are most enriched within nucleotide sequences relative to up- and down-regulated LINEs in ACC and PFC

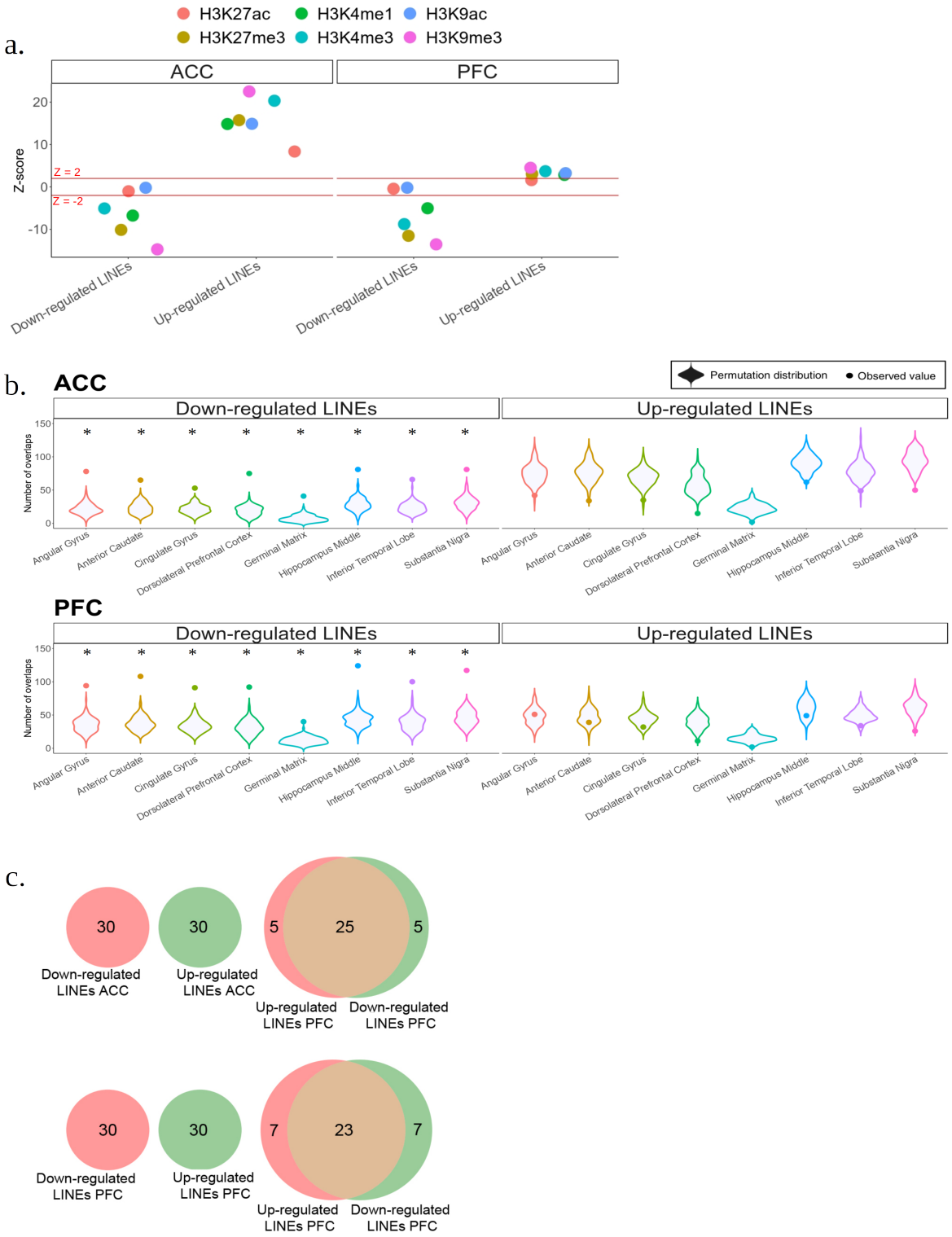

Figure 3.3.1 – a. Z-scores relative to the enrichment of each group of DE LINEs within different types of histone marks for ACC and PFC, enrichments have been performed by comparing the number of DE LINEs overlapping each histone marks (data from Roadmap epigenomic project) with the number of overlaps between the genomic coordinates of 100 random sets of non-DE LINEs and the genomic coordinates of histone marks, significance line: |Z-score| > 2; b. Overlap between the genomic coordinates of DE LINEs and enhancers active in 8 brain regions (Cao et al, 2017), the dots represent the number of enhancers overlapping at least one LINE, while the violin represents the number of overlaps obtained by overlapping 100 sets of random non-DE LINEs with the genomic coordinates of enhancers, *Z-score > 2; c. VENN diagram displaying the number of transcription factors specifically binding motifs relative to nucloetide sequences relative up- and down-regulated LINEs in ACC and PFC, the list of TF is reported in table 4.

3.4 Results reproducibility

All analyses performed suggest the dysregulation of specific regulatory regions, leading to a change in the repertory of LINEs transcription and to the down-regulation of genes involved in synaptic functions in the brain of a subset of ASD patients. However due to the heterogenic nature of ASD, these results need to be reproduced in additional datasets in order to provide a convincing hypothesis.

3.4.1 Reproducibility of results in a dataset of neurons differenciated from patient-derived iPSC

In order to test my hypothesis on a different model, I retrieved whole RNA-seq data from neurons resulting from the differentiation of induced pluripotent stem cells (iPSCs) and neuronal precursor cells (NPCs) obtained from the de-differentiation of fibroblasts extracted from ASD individuals with early brain overgrowth and non-ASD controls with normal brain size[98]. iPSC-derived NPCs from ASD individuals display increased levels of proliferation. This is likely an effect of the dysregulation in the canonical WNT pathway detected in such individuals. Furthermore ASD-derived neurons display reduced synaptogenesis and overall defects in neuronal activity, which may be linked to the gene expression pattern changes observed at the NPC and neuronal stages of fibroblast-derived cells of ASD individuals. In particular, the authors identified 154 genes significantly differentially expressed in ASD compared with controls at the neuronal stage[98]. At the neuronal stage, the up-regulated genes in ASD were enriched for the GO categories related to extracellular matrix, whereas the down-regulated genes were significantly enriched for the GO categories of cilium and axoneme, consistent with the observed synaptic dysregulation. Additionally the authors performed WES on fibroblast obtained from the ASD patients. Then they applied the IlluminaVariantStudio pipeline to call and annotate genomic variants. Subsequently, variants were filtered using custom parameters, resulting in 98 potentially damaging common and rare variants within ASD-related genes.

In order to faithfully reproduce my analyses, I applied my custom (and more stringent) filters to this set of variants. This resulted in a total of 16 putatively deleterious variants within the 8 ASD individuals. The list of variants per individual is reported in table 5. Manual curation of the genes involved in the mutation revealed that, according to my model, only one sample, characterized by a putatively deleterious mutation in the gene KMT2A, belongs to the ASD_reg category; whereas all other ASD samples belong to the ASD_other category.

SFARI gene	Sample	Coordinante	Variant	CADD_phred score	SIFT_pred	Amino Acids	Codons
CSMD1	able	chr8:3351217	A>A/T	25.9	D	L/Q	cTg/cAg
FRMPD4	aero	chrX:12734922	G>G/A	28.2	D	D/N	Gac/Aac
CGNL1	ahoy	chr15:57839612	C>C/T	23.4	D	T/M	aCg/aTg
DDX53	ahoy	chrX:23019571	C>C/A	22.6	D	P/H	cCc/cAc
SCN8A	ahoy	chr12:52164391	C>C/A	24.4	D	T/N	aCc/aAc
MET	apex	chr7:116409748	G>G/A	26.9	D	G/E	gGa/gAa
SERPINE1	apex	chr7:100775203	C>C/T	24.1	D	R/W	Cgg/Tgg
ROBO2	aqua	chr3:77530358	C>C/A	25.7	D	L/M	Ctg/Atg
CTNNB1	arch	chr3:41266229	C>C/T	37	D	Q/*	Caa/Taa
FAT1	arch	chr4:187542431	T>T/C	23.7	D	Y/C	tAt/tGt
KCNJ10	arch	chr1:160011823	G>G/T	25.4	D	A/E	gCg/gAg
NRXN2	arch	chr11:64435099	G>G/A	26.2	D	P/L	cCc/cTc
ROBO1	arch	chr3:78676621	G>G/A	26.2	D	P/L	cCt/cTt
KMT2A	avid	chr11:118370553	C>C/A	28.2	D	S/Y	tCt/tAt
MYT1L	avid	chr2:1921102	C>C/A	32	D	R/I	aGa/aTa
NR3C2	avid	chr4:149073659	C>C/A	28.1	D	S/I	aGc/aTc

Table 5 – putatively deleterious variants associated to each sample according to my custom filters

In order to estimate the amount of differentially expressed genes and transposable element fragments within each sample, I performed gene and TE differential expression comparing two replicates of each of the 6 samples including biological replicates with the controls. Similarly to the DE analyses performed on the samples of the dataset from Velmeshev et al, the differential expression analyses have been performed with DEseq2 upon counting the amount of reads mapping to genes and TEs quantified respectively with HTseq and SQuIRE.

Differential expression analyses resulted in a higher number of DE TEs in the neuron of the ASD_reg sample (avid) compared to all other samples (figure 3.4.1a). Indeed, 2135 transposable elements are differentially expressed in neurons derived from the sample named avid, while most of the other samples are characterized by less than 300 total DE TEs, with the exception of the sample named aqua, which is characterized by 897 total DE TEs (figure 3.4.1b). Differently from analysis performed on *post-mortem* brain tissue samples from Velmeshev et al, I did not observe an enrichment in up-regulated LINE elements. The total number of DE transposable elements for each sample and each class is reported in table 6.

Sample	LINE	SINE	LTR	DNA	Other	Total	Group
ABLE	19	11	16	4	4	54	Down-regulated
AERO	21	24	27	4	4	80	Down-regulated
AHOY	65	77	61	9	16	228	Down-regulated
AQUA	248	179	128	26	10	591	Down-regulated
ARCH	24	17	32	4	8	85	Down-regulated
AVID	458	470	230	77	23	1258	Down-regulated
ABLE	1	6	0	0	0	7	Up-regulated
AERO	1	1	0	0	0	2	Up-regulated
AHOY	24	15	0	0	0	39	Up-regulated
AQUA	128	178	0	0	0	306	Up-regulated
ARCH	4	3	0	0	0	7	Up-regulated
AVID	422	455	0	0	0	877	Up-regulated

Table 6 – number of DE TEs compared to controls for each samples and each class

Gene-related differential expression analyses resulted in a very similar outcome. Indeed, I identified 2386 total DE genes (1128 up-regulated and 1258 down-regulated) in avid, whereas I found 871 total DE genes in aqua and less than 300 DE genes in all other samples (figure 3.4.1c,d). These results suggest that the genome-wide transcriptional alterations characteristic of ASD_reg patients are happening primarily in one sample carrying a potentially deleterious variants within regulatory gene KMT2A.

Interestingly, as in the case of the analyses performed on the dataset from Velmeshev et al, up-regulated LINEs are enriched within down-regulated genes (Z-score > 3, p-value < 0.01) (figure 3.4.1e). Moreover I observed a significant depletion in up-regulated LINEs within up-regulated genes (Z-score < 3, p-value < 0.01) and an enrichment of down-regulated LINEs within up-regulated genes (Z-score > 3, p-value < 0.01) (figure 3.4.1e). These results suggest that the overexpression of specific LINE fragments may impair the expression of overlapping genes. On the other hand, up-regulated and down-regulated LINE elements relative to the sample named AQUA are significantly enriched (Z-score > 3, p-value < 0.01) respectively within up-regulated and down-regulated genes (figure 3.4.1f). This outcome suggests that the cause underling the presence of differentially expressed genes and TEs in the sample named AQUA may be different from that characterizing the sample name AVID.

Recent literature showed how the the level of LINE expression is controlled during cell differentiation and how this regulation may influence gene expression[57]. In particular, LINE expression negatively

influences gene expression. In order to test whether LINE expression would be anti-correlated with gene expression, I exploited the RNA-seq data from three stages of neuronal development (iPSC, NPC and neurons) available in the Marchetto et al.[98] dataset. Namely, I calculated the normalized expression level of genes and LINEs among the three cell stages within all the samples. LINE expression level of all the three experimental groups follows the same trend, with a decrease in overall expression going from iPSC to NPC, and a subsequent increase from NPC to mature neurons (figure 3.4.1g). Interestingly, the ASD_reg sample follows the same trend among the three cell stages, however LINE expression in the ASD_reg sample is significantly higher in iPSC and neurons compared to controls and ASD_other samples (p-value < 0.05) (figure 3.4.1g). A similar but opposite scenario is observed when plotting the normalized gene expression among cell stages. Indeed, the ASD_reg sample shows a significantly lower overall gene expression level compared to controls and ASD_other samples in iPSC and mature neurons (p-value < 0.05) (figure 3.4.1g). Moreover, gene expression is significantly higher in the ASD_reg sample compared to the other experimental groups in NPCs (figure 3.4.1g).
These results suggests that, in ASD_reg samples, LINE expression normally occurring in healthy developing cells, is enhanced in iPSC and neurons. These alterations may drive an aberrant CNS development.

Overall, in line with analysis performed on *post-mortem* brain tissues from ASD patients, the only ASD_reg sample is characterized by a pervasive differential expression of genes and transposable elements compared to controls to a much higher degree with respect to the other samples. Furthermore, I observed a similar interaction between up-regulated LINEs and down-regulated genes. However, the precise asset of DE TEs is different, especially since I do not report an enrichment of LINEs within up-regulated TEs. It is not clear whether these differences may be due to the experimental model or to the poor resolution given by the statistical analysis on the data available. Nevertheless, these results may be comprehensively in line with my model.

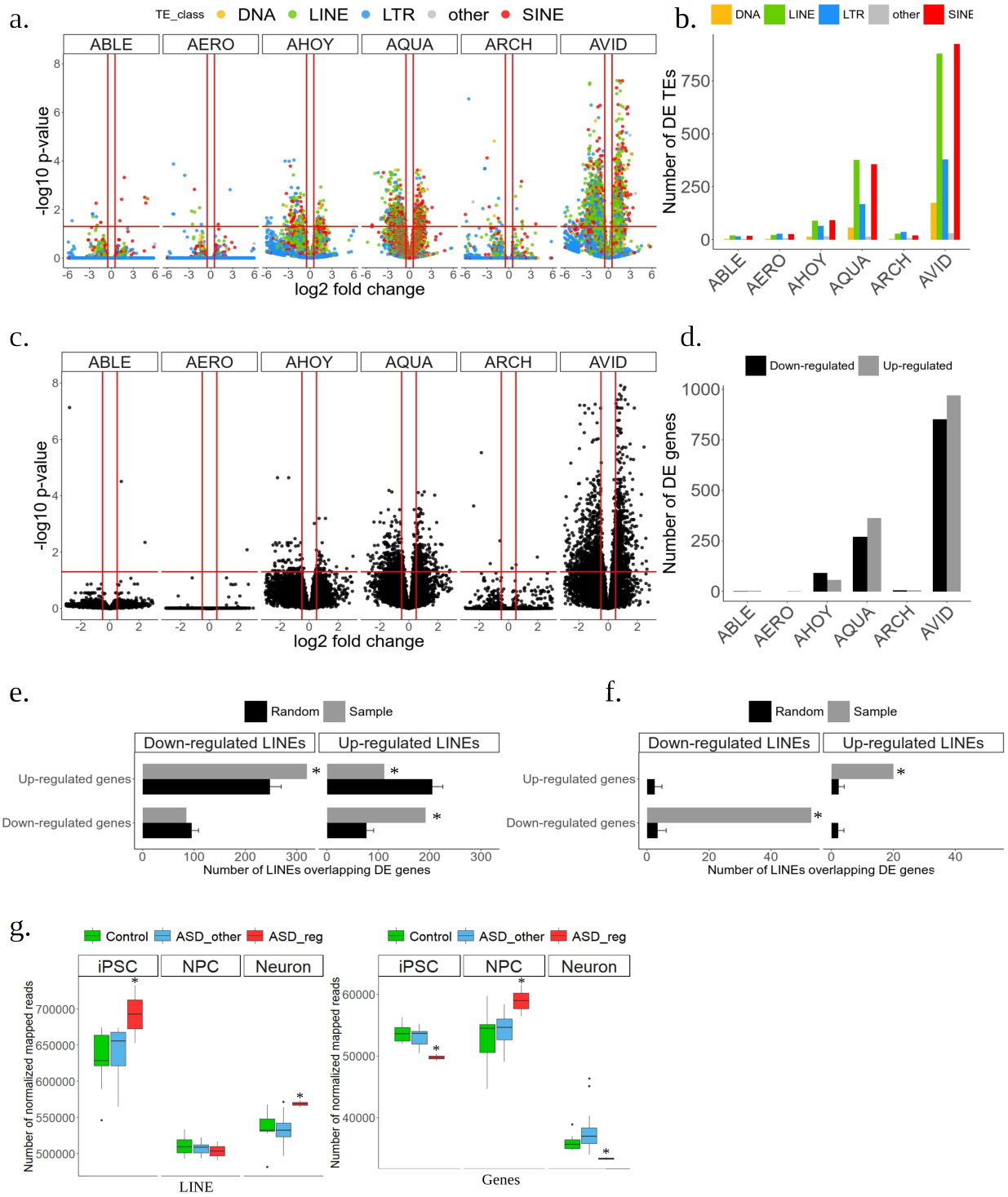

Figure 3.4.1 – a. Volcano plot representing differentially expressed TE fragments quantified with SQuIRE between each sample and controls, each dot represents a TE fragment plotted with respect to the inverse logarithm of its FDR value and the logarithm of its fold change expression with respect to controls, significance line: FDR = 0.05, |fold-change| > 0.5; b. Number of DE TE fragments quantified with SQuIRE in each sample, DE elements are divided into up-regulated (fold-chance > 0) and down-regulated (fold-change < 0); c. Differentially expressed genes quantified with HTseq between each sample and controls, each dot represents a gene plotted with respect to the inverse logarithm of its FDR value and the logarithm of its fold change expression with respect to controls, significance line: FDR = 0.05, |fold-change| > 0.5; d. Number of genes quantified with HTseq in each sample, genes are divided into up-regulated (fold-chance > 0) and down-regulated (fold-change < 0); e. Number of overlaps between different sets of DE LINEs and DE genes relative to the sample named AVID, each number of overlaps is compared with the number of overlaps obtained with 100 sets of random expressed genes, *|Z-score| > 3, p-value < 0.01; f. Number of overlaps between different sets of DE LINEs and DE genes relative to the sample named AQUA, each number of overlaps is compared with the number of overlaps obtained with 100 sets of random expressed genes, *|Z-score| > 3, p-value < 0.01; g. Number of normalized reads mapped onto LINEs and genes among three developmental stages: indiced pluripotent stem cells (iPSC), neuronal precursor cells (NPC) and mature neurons, and among the three experimental groups, *p-value < 0.05.

4. Discussion

The study of thousands of genomic data from ASD patients and controls demonstrated that hundreds of genes are involved in ASD. Interestingly, from a functional perspective, ASD genes may be mainly divided into: a) genes which exert a crucial role in synaptic function and b) genes involved in transcription regulation and/or chromatin remodeling. However, ASD genetic etiology has not been fully explained. Indeed, in most cases it is impossible to trace a definitive link between genetic mutations and the pathogenic phenotype. Several works highlighted a large number of differentially expressed genes commonly dysregulated among different ASD cases, suggesting a convergence in transcriptome dysregulation.

I speculated that the ASD cases which I called ASD_reg, who carry a deleterious mutation within a gene involved in transcription regulation and/or chromatin remodeling (regulatory genes), may be characterized by a different transcriptional landscape compared to ASD cases with causative mutations in genes unrelated to chromatin remodeling.

To my knowledge this kind of sample stratification has never been performed on ASD. Therefore, it needs to be noted that the choices at the basis of the definition of regulatory genes are the result of a manual, and therefore arbitrary, review of genes functions. However, I believe that the mutated genes characteristic of the ASD_reg experimental group, may represent indeed a functionally coherent class of genes. Some of the samples belonging to the ‘ASD_other’ experimental group carry suggestive deleterious variants within SFARI genes responsible for other biological roles, especially synaptic functions. However, I consider appropriate to the scope of my work to put all ASD cases devoid of a deleterious mutation within a regulatory gene in the same experimental group.

As of now, ASD is considered to be a disorder characterized by pervasive transcriptional alterations. Indeed, an extensive literature supports the presence of a great number of DE genes between ASD cases and healthy controls. I detect the presence of genes and transposable elements dysregulation only between ASD_reg and control samples. My stratification process may have therefore granted the possibility to separate a group of ASD cases with no trascriptional differences in comparison to

controls. I believe this may be a novel result of keen importance for the understanding of ASD heterogeneity.

On the other hand, in ASD_reg samples, I detected a pervasive up-regulation of LINE elements especially within introns of genes involved in synaptic functions, which are mostly down-regulated with respect to controls. This scenario would involve the presence of a set of intronic LINE elements particularly susceptible to be retained during transcription within the human genome, and it could explain why a defect in chromatin regulatory mechanisms common to almost any cell type would specifically lead to a neurodevelopmental phenotype. Young and transcriptionaly active LINE elements are enriched within neuron-specific genes[57]. This might be a result of the domestication of LINEs as host regulatory sequences occurred along human evolution. The pervasive presence of intronic LINEs within neuronal genes may be crucial for the tight level of gene expression regulation necessary for the development and function of the CNS, which is undoubtedly higher compared to that needed for any other biological organization. On the other hand, the presence of intronic active LINEs may imply that even a slight drift in the genome regulation program may result in substantially deleterious phenotypes in the CNS. The higher rate of intron retention I detect is also consistent with the increased level of splicing defect in ASD reported in literature[89]. To my knowledge, a causative role of specific intronic LINE elements retention has never been proposed as a causative mechanism for the dysregulation of synaptic genes transcription in ASD.

My results may have a relevant clinical implication concerning the treatment of ASD. In fact, if deleterious mutations within a defined set of genes are indeed at the basis of the molecular etiology of a subset of ASD patients, WES-derived data may be used to specific choices in terms of clinical intervention. As of now, effective medications for the treatment many of ASD core symptoms are lacking. Most of the drug used for ASD derive from knowledge of genes implicated in monogenic disorders associated with altered neurodevelopmental trajectories and autistic symptoms such as fragile X sydrome, Landau–Kleffner syndrome and Rett syndrome[125,126]. As a consequence, medication in patients with ASD has traditionally targeted associated conditions such as distress, aggression, irritability, stereotyped behaviors, anxiety, hyperactivity, and sleep difficulties[125] that occur in the context of ASD, with poor impact for the core symptoms of the condition. These drugs typically target genes associated to synaptic pathways such as dopaminergic and glutamatergic receptors[125]. However

currently, there is no treatment for the primary symptom of ASD: social impairment. Interestingly however, a new research revealed that treatment with a low dose of romidepsin, an anti-cancer drug, restored social deficits in animal models of autism[127]. Romidepsin inhibits the activity of the enzyme histone deacetylase[127], thus allowing genes involved in neuronal signaling and down-regulated in ASD to be expressed normally. To note, there is an extensive overlap in risk genes involved in genome regulation for autism and cancer. The most important implication of this result is that it prompts for the use of existing epigenetic drugs effective in cancer treatment as targeted treatments for specific autism cases. My results may prove their usefulness in stratifying ASD cases in order to pinpoint the ones most suitable for potentially experimental treatments with epigenetic drugs. However, more extensive studies need to be performed on larger cohort of patients in order to better understand the mechanisms underlying the transcriptional alterations observed in a subset of ASD cases.

5. Conclusion

I propose a method which can be used to stratify ASD cases into two main categories, one of which recapitulates the major transcriptional alterations characteristic of ASD, while the other does not present transcriptional differences compared to healthy controls. Despite several limitations, the main results have been reproduced in a different experimental model.

6. Acknowledgments

I am thankful to my supervisor Prof. Remo Sanges for allowing me to work in his laboratory. Without his guidance and suggestions all the work I produced during these years would not have been possible.
I owe much of my improvement as a thinker to Prof. Stefano Gustincich, whose insightful comments and enthusiasm have helped me throughout these years.
A special thanks to Dr. Aldamaria Puliti and her lab, as the expertise you shared has been one of the starting points for the work of this thesis.

Ringraziamenti

Ma tuuu… hai fumat'?
Grazie, perché per affrontare il dottorato ci vogliono persone di alta qualità!

Bomboclat
Grazie, per avere, a mometi, trasformato un gruppo di dottorandi stressati in una spensierata classe di liceo. Aaaaahhhhnnn….

Mi ci sfascio
Grazie, perché spesso è meglio "buttarla in caciara". *Non prendere la vita troppo sul serio, non potrai mai uscirne vivo ...*

Dove sta?
Grazie all'unico membro del Sanges lab abbastanza dignitoso per l'istituzione del matrimonio. Grazie per avermi fatto capire che nel dottorato, come nella vita coniugale, l'essenziale è la pazienza. Non l'amore: la pazienza.

Mr. Worldwide
Grazie per essere stato un esempio di serietà ed impegno. Silenziosamente mi hai mostrato come dovrebbe essere un vero Ph.D. student.

Una vita piena

Grazie, perché il miglior insegnamento che un senior può dare è che si può avere il coraggio di affrontare qualcosa di terrificante e meraviglioso come la paternità durante la follia del dottorato.

Un carattere complicato

Grazie perché non solo io ma tutti avevamo bisogno della tua energia. Spero che tu non smetta di farti valere. Ricordati che tu sarai sempre l'originale, non la pezzotta.

Senpai 先輩

Certe persone ci dovrebbero essere dall'inizio. Grazie per i ragionamenti, i consigli ed i litigi; mi mancheranno tutti. Grazie per l'entusiasmo per il mio progetto che mi hai trasmesso e che mi ha consentito di non mollare nei momenti peggiori. Grazie per aver capito sempre *quello che ti sto dicendo*.

***Signori* si nasce**

Grazie per essere stato il migliore esempio di ricercatore e di persona buona e responsabile. Prima o poi la grande idea arriverà ...

Anche le capre vanno in paradiso

Grazie, perché hai fatto la differenza. La persona più inaspettata ed influente di questo percorso. Sarai sempre una parte della mia musica.

Ed infine grazie a te che hai avuto il coraggio e la pazienza di stare al mio fianco per tutti questi anni. Con te il futuro non mi spaventa.

I nomi non servono, vi siete riconosciuti. Credo sia la prova che ognuno di voi è stato importante a modo suo e che questi anni non sarebbero stati gli stessi con altri compagni di viaggio.

Conosco la metà di voi solo a metà e nutro per meno della metà di voi metà dell'affetto che meritate.

Ulteriori ringraziamenti:

Tutti i frat' ingiustamente carcerati, Savvatore, Giampà, Papa Ratzinger, Silvana pezzotta, Luiggi=o, pat't' o' ricottar', Di Caprio ordinato su Wish, Andrea Alongi, Biogioggero, Uva, i Damiano's samples, la letteratuuuuura, Mascia, Rebecca Nicole Kidman, tutti gli early embrios, il ministro della salute Paolo Vatta, Picci, Geuvadis, i clinicians, i cittadini extracomunitari, i gene panels, i minigenes, 'a mamm' 'e Mauro, Speranza, i neuroni sparati, i motorini rubati, i big data, le special issues, i lab users, la zona Bigiarini, gli schizo e tutto ciò che nel mondo "è BUON'!".

7. Bibliography

1. Schmutz, J. *et al.* Quality assessment of the human genome sequence. 365–368 (2002).
2. Taft, R. J. Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences.
3. Monaghan, F. & Corcos, A. On the origins of the Mendelian laws. 67–69 (1984).
4. Kutschera, U. & Niklas, K. J. The modern theory of biological evolution : an expanded synthesis. 255–276 (2004). doi:10.1007/s00114-004-0515-y
5. Margulies, E. H. *et al.* Identification and Characterization of Multi-Species Conserved Sequences. 2507–2518 (2003). doi:10.1101/gr.1602203.emerged
6. Csete, M. E. & Doyle, J. C. Reverse Engineering of Biological Complexity. **295**, 1664–1669 (2002).
7. Mattick, J. S. Non-coding RNAs : the architects of eukaryotic complexity. **2**, 986–991 (2001).
8. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* 29–35 doi:10.1038/s41576-018-0089-8
9. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat. Publ. Gr.* **17**, 487–500 (2016).
10. Bernstein, B. E. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. 315–326 (2006). doi:10.1016/j.cell.2006.02.041
11. 1000 Genomes Project Consortium, {fname} *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
12. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141 , 456 humans. **581**, (2020).
13. Karczewski, K. J. *et al.* The ExAC browser : displaying reference data information from over 60 000 exomes. **45**, 840–845 (2017).
14. Elements, D. N. A. An integrated encyclopedia of DNA elements in the human genome. (2012). doi:10.1038/nature11247
15. Abugessaisa, I. *et al.* FANTOM5 transcriptome catalog of cellular states based on Semantic MediaWiki. 1–10 (2016). doi:10.1093/database/baw105
16. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium complex. *Nat. Publ. Gr.* **28**, 1045–1048 (2010).

17. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

18. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–11 (2013).

19. Farrall, M. Quantitative genetic variation: a post-modern view. *Hum. Mol. Genet.* **13**, 1R – 7 (2004).

20. Mérot, C., Oomen, R. A., Tigano, A. & Wellenreuther, M. A Roadmap for Understanding the Evolutionary Signi fi cance of Structural Genomic Variation. *Trends Ecol. Evol.* 1–12 (2020). doi:10.1016/j.tree.2020.03.002

21. Ramírez-bello, J. Role of genetic variability in Mendelian and multifactorial diseases. 463–470 (2019). doi:10.24875/GMM.M20000333

22. Huang, Y., Yu, S., Wu, Z. & Tang, B. Genetics of hereditary neurological disorders in children. **3**, 108–119 (2014).

23. Thapar, A., Cooper, M. & Frcpsych, M. R. Neurodevelopmental disorders. *The Lancet Psychiatry* **0366**, 1–8 (2016).

24. Lasalle, J. M. Autism genes keep turning up chromatin. **1**, 1–13 (2014).

25. Vallianatos, C. N., Iwase, S. & Arbor, A. Disrupted intricacy of histone H3K4 methylation in neurodevelopmental disorders. **7**, 503–519 (2016).

26. Corley, M. J., Pang, A. P. S., Lum-jones, A. & Li, D. Epigenetic Delay in the Neurodevelopmental Trajectory of DNA Methylation States in Autism Spectrum Disorders. **10**, 1–14 (2019).

27. Zahir, F. R. & Brown, C. J. Epigenetic Impacts on Neurodevelopment : Pathophysiological Mechanisms and Genetic Modes of Action. **69**, 92–100 (2011).

28. McClintock, B. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci.* **36**, 344–355 (1950).

29. Notwell, J. H., Chung, T., Heavner, W. & Bejerano, G. A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. *Nat Commun.* (2015). doi:10.1038/ncomms7644.A

30. Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* **8**, 272–285 (2007).

31. Elements, R. & Evolution, G. Impact of transposable elements on the evolution of mammalian gene regulation. **352**, 342–352 (2005).

32. Medstrand, P. *et al.* Impact of transposable elements on the evolution of mammalian gene regulation. *Cytogenet. Genome Res.* **110**, 342–352 (2005).

33. Mills, R. E., Bennett, E. A., Iskow, R. C. & Devine, S. E. Which transposable elements are active in the human genome? **23**, (2007).

34. Eric M. Ostertag andHaig H. Kazazian Jr. Biology of mammalian L1 retrotransposons. (2001).

35. Raiz, J. *et al.* The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res.* **40**, 1666–1683 (2012).

36. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* **24**, 363–367 (2000).

37. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. (1989).

38. Ran, C. *et al.* Mobile Interspersed Repeats Are Major Structural Variants in the Human Genome. *Cell* **141**, 1171–1182 (2010).

39. Ostertag, E. M., Goodier, J. L., Zhang, Y. & Kazazian, H. H. Report SVA Elements Are Nonautonomous Retrotransposons that Cause Disease in Humans. *Am. J. Hum. Genet* **73**, 1444–1451 (2003).

40. Ono, M., Kawakami, M. & Takezawa, T. A novel human nonviral retroposon derived from an endogenous retrovirus. *Nucleic Acids Res.* **15**, 8725–8737 (1987).

41. Sundaram, V. & Wysocka, J. Transposable elements as a potent source of diverse cis -regulatory sequences in mammalian genomes. (2020).

42. Trizzino, M., Kapusta, A. & Brown, C. D. Transposable elements generate regulatory novelty in a tissue-specific fashion. *BMC Genomics* **19**, 1–12 (2018).

43. Kapusta, A. *et al.* Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. *PLoS Genet.* **9**, (2013).

44. Manuscript, A. Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenet Genome Res.* **110**, 333–341 (2007).

45. Maka, W. Genomic scrap yard: How genomes utilize all that junk. *Gene* **259**, 61–67 (2000).

46. Conley, A. B., Piriyapongsa, J. & Jordan, I. K. Retroviral promoters in the human genome. *Bioinformatics* **24**, 1563–1567 (2008).

47. Jordan, I. K., Rogozin, I. B., Glazko, G. V & Koonin, E. V. Origin of a substantial fraction of human regulatory sequences from transposable elements. **19**, 68–72 (2003).

48. Trizzino, M., Park, Y., Holsbach-beltrame, M. & Aracena, K. Transposable elements are the primary source of novelty in primate gene regulation. *Renome Res.* (2017).

49. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. **42**, 6–8 (2010).

50. Conley, A. B. & Jordan, I. K. Cell type-specific termination of transcription by transposable element sequences. *Mob. DNA* **3**, 1–13 (2012).

51. Jeyakani, J. & Bourque, G. The Majority of Primate-Specific Regulatory Sequences Are Derived from Transposable Elements. **9**, (2013).

52. Sorek, R., Ast, G. & Graur, D. Alu -Containing Exons are Alternatively Spliced. *Genome Res.* **12**, 1060–1067 (2002).

53. Gal-mark, N., Schwartz, S. & Ast, G. Alternative splicing of Alu exons - Two arms are better than one. *Nucleic Acids Res.* **36**, 2012–2023 (2008).

54. Sundaram, V. *et al.* Widespread contribution of transposable elements to the innovation of gene regulatory networks. 1963–1976 (2014). doi:10.1101/gr.168872.113.

55. Wang, T. *et al.* Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. **104**, 18613–18618 (2007).

56. Lea, A. J. *et al.* Genome-wide quantification of the effects of DNA methylation on human gene regulation. 1–27 (2018).

57. Lu, J. Y. *et al.* Genomic Repeats Categorize Genes with Distinct Functions for Orchestrated Regulation Article Genomic Repeats Categorize Genes with Distinct Functions for Orchestrated Regulation. *CellReports* **30**, 3296-3311.e5 (2020).

58. Kaneda, M., Okano, M., Hata, K. & Sado, T. Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. **429**, 2–5 (2004).

59. Bestor, T. H. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. **431**, 2–5 (2004).

60. Garcia-Perez, J. L., Widmann, T. J. & Adams, I. R. The impact of transposable elements on mammalian development. *Development* **143**, 4101–4114 (2016).

61. Muotri, A. R. *et al.* Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. **435**, 903–910 (2005).

62. Peng, G. E. *et al.* L1 retrotransposition in human neural progenitor cells. **460**, (2009).

63. Baillie, J. K. *et al.* Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**, 534–537 (2011).

64. Sukapan, P., Promnarate, P., Avihingsanon, Y., Mutirangura, A. & Hirankarn, N. Types of DNA methylation status of the interspersed repetitive sequences for LINE-1 , Alu , HERV-E and

HERV-K in the neutrophils from systemic lupus erythematosus patients and healthy controls. *J. Hum. Genet.* **59**, 178–188 (2014).

65. Castelijns, B. *et al.* Hominin-specific regulatory elements selectively emerged in oligodendrocytes and are disrupted in autism patients. *Nat. Commun.* doi:10.1038/s41467-019-14269-w

66. Koberstein, J. N. *et al.* Learning-dependent chromatin remodeling highlights noncoding regulatory regions linked to autism. **6500**, 1–10 (2018).

67. Bogu, G. K., Reverter, F., Marti-renom, M. A. & Michael, P. Atlas of transcriptionally active transposable elements in human adult tissues. (2019).

68. Kitkumthorn, N. & Mutirangura, A. Long interspersed nuclear element-1 hypomethylation in cancer : biology and clinical applications. 315–330 (2011). doi:10.1007/s13148-011-0032-8

69. Saleh, A., Macia, A. & Muotri, A. R. Transposable Elements , Inflammation , and Neurological Disease. **10**, (2019).

70. Bundo, M. *et al.* Report Increased L1 Retrotransposition in the Neuronal Genome in Schizophrenia. *Neuron* 1–8 (2014). doi:10.1016/j.neuron.2013.10.053

71. Li, S. *et al.* Hypomethylation of LINE-1 elements in schizophrenia and bipolar disorder. *J. Psychiatr. Res.* **107**, 68–72 (2018).

72. Amir, R. E. *et al.* Rett syndrome is caused by mutations in X-linked MECP2 , encoding methyl-CpG-binding protein 2. **23**, 185–188 (1999).

73. Jacob-hirsch, J. *et al.* Whole-genome sequencing reveals principles of brain retrotransposition in neurodevelopmental disorders. *Nat. Publ. Gr.* 1–17 (2018). doi:10.1038/cr.2018.8

74. Shpyleva, S. *et al.* Overexpression of LINE-1 Retrotransposons in Autism Brain. (2017). doi:10.1007/s12035-017-0421-x

75. Tangsuwansri, C. *et al.* Investigation of epigenetic regulatory networks associated with autism spectrum disorder (ASD) by integrated global LINE-1 methylation and gene expression profiling analyses. 1–27 (2018).

76. Lai, M.-C., Lombardo, M. V & Baron-Cohen, S. Autism. *Lancet* **383**, 896–910 (2014).

77. Sven Sandin, PhD1; Paul Lichtenstein, PhD2; Ralf Kuja-Halkola, P. et al. The Heritability of Autism Spectrum Disorder Analysis method B. **318**, 1182–1184 (2017).

78. Sanders, S. J. Next-Generation Sequencing in Autism Spectrum Disorder. (2020). doi:10.1101/cshperspect.a026872

79. Williams, T. SFARI Gene : An evolving database for the autism research community Genomics offers new possibilities for global health through international collaboration. (2017). doi:10.1242/dmm.005439

80. Devlin, B. & Scherer, S. W. Genetic architecture in autism spectrum disorder. *Curr. Opin. Genet. Dev.* **22**, 229–237 (2012).

81. Kumar, R. A. & Christian, S. L. Genetics of Autism Spectrum Disorders. (2009).

82. Manuscript, A. & Impact, B. Histone Lysine Methylation Dynamics: Establishment, Regulation, and Biological Impact. **48**, 1–32 (2013).

83. Bayraktar, G. & Kreutz, M. R. Neuronal DNA Methyltransferases : Epigenetic Mediators between Synaptic Activity and Gene Expression ? (2018). doi:10.1177/1073858417707457

84. Liao, J. *et al.* Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. **47**, 469–478 (2015).

85. Cukier, H. N. *et al.* The Expanding Role of MBD Genes in Autism: Identification of a MECP2 Duplication and Novel Alterations in MBD5, MBD6, and SETDB1. **5**, 385–397 (2013).

86. Gauthier, J. *et al.* De Novo Mutations in FOXP1 in Cases with Intellectual Disability , Autism , and Language Impairment. 671–678 (2010). doi:10.1016/j.ajhg.2010.09.017

87. Chloe C.Y. Wong1, Rebecca G. Smith2, Eilis Hannon2, Gokul Ramaswami3, Neelroop N. Parikshak3, Elham Assary4, Claire Troakes1, Jeremie Poschmann5, Leonard C. Schalkwyk6, Wenjie Sun7, Shyam Prabhakar7, Daniel H. Geschwind3, 8, 9, J. M. Genome-wide DNA methylation profiling identifies convergent molecular signatures associated with idiopathic and syndromic autism in post-mortem human brain tissue.

88. Gokoolparsadh, A. *et al.* Searching for convergent pathways in autism spectrum disorders : insights from human brain transcriptome studies. *Cell. Mol. Life Sci.* (2016). doi:10.1007/s00018-016-2304-0

89. Parikshak, N. N. *et al.* Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nat. Publ. Gr.* **540**, 423–427 (2016).

90. Parikshak, N. N. *et al.* Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. *Cell* **155**, 1008–1021 (2013).

91. Gandal, M. J. *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. **362**, 1–32 (2019).

92. Muhle, R. A., Reilly, S. K., Lin, L. & Fertuzinhos, S. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. **155**, 997–1007 (2014).

93. Nardone, S. *et al.* DNA methylation analysis of the autistic brain reveals multiple dysregulated biological pathways. *Transl. Psychiatry* **4**, e433-9 (2014).

94. Hansen, K. D., Briem, E., Kaufmann, W. E. & Feinberg, A. P. Common DNA methylation alterations in multiple brain regions in autism. *Mol. Psychiatry* 1–10 (2013). doi:10.1038/mp.2013.114

95. Alex, A. M., Korammannil, R. & Banerjee, M. Genetic Association of DNMT Variants Can Play a Critical Role in De fi ning the Methylation Patterns in Autism. 901–907 doi:10.1002/iub.2021

96. Taube, J. H. *et al.* The H3K27me3-demethylase KDM6A is suppressed in breast cancer stem-like cells , and enables the resolution of bivalency during the mesenchymal-epithelial transition. **8**, 65548–65565 (2017).

97. Wang, M. *et al.* Increased Neural Progenitor Proliferation in a hiPSC Model of Autism Induces Replication Stress- Increased Neural Progenitor Proliferation in a hiPSC Model of Autism Induces Replication Stress-Associated Genome Instability. *Stem Cell* 1–13 (2020). doi:10.1016/j.stem.2019.12.013

98. Marchetto, M. C. N., Belinson, H., Francisco, S., Freitas, B. & Vadodaria, K. Altered proliferation and networks in neural cells derived from idiopathic autistic individuals. (2016). doi:10.1038/mp.2016.95

99. Ziats, M. N. & Rennert, O. M. Aberrant Expression of Long Noncoding RNAs in Autistic Brain. 589–593 (2013). doi:10.1007/s12031-012-9880-8

100. Al-saad, S., Mukaddes, N. M., Oner, O., Al-saffar, M. & Balkhy, S. Mutations in Human Accelerated Regions (HARs) Disrupt Cognition and Social Behavior. **167**, 341–354 (2017).

101. Velmeshev, D. *et al.* Single-cell genomics identifies cell type – specific molecular changes in autism. **689**, 685–689 (2019).

102. Cao, Q. *et al.* Reconstruction of enhancer – target networks in 935 samples of human primary cells , tissues and cell lines. *Nat. Publ. Gr.* **49**, (2017).

103. Lu, L. & Liu, X. Robust Hi-C Maps of Enhancer-Promoter Interactions Reveal the Function of Non-coding Genome in Neural Development and Diseases ll Resource Robust Hi-C Maps of Enhancer-Promoter Interactions Reveal the Function of Non-coding Genome in Neural Development an. 521–534 (2020). doi:10.1016/j.molcel.2020.06.007

104. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

105. Wang, K., Li, M. & Hakonarson, H. ANNOVAR : functional annotation of genetic variants from high-throughput sequencing data. **38**, 1–7 (2010).

106. Landrum, M. J. *et al.* ClinVar : public archive of relationships among sequence variation and human phenotype. **42**, 980–985 (2014).

107. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. **29**, 15–21 (2013).

108. Harrow, J. *et al.* GENCODE : The reference human genome annotation for The ENCODE Project. 1760–1774 (2012). doi:10.1101/gr.135350.111.

109. Yang, W. R., Ardeljan, D., Pacyna, C. N., Payer, L. M. & Burns, K. H. SQuIRE reveals locus-specific regulation of interspersed repeat expression. **47**, 1–16 (2019).

110. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update , a database of repetitive elements in eukaryotic genomes. *Mob. DNA* 4–9 (2015). doi:10.1186/s13100-015-0041-9

111. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).

112. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. 1–21 (2014). doi:10.1186/s13059-014-0550-8

113. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

114. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).

115. Bailey, T. L. *et al.* MEME S UITE : tools for motif discovery and searching. **37**, 202–208 (2009).

116. Middleton, R. *et al.* IRFinder : assessing the impact of intron retention on mammalian gene expression. 1–11 (2017). doi:10.1186/s13059-017-1184-4

117. He, J. *et al.* Transposable elements are regulated by context- speci fi c patterns of chromatin marks in mouse embryonic stem cells. (2019).

118. Huda, A., Mariño-ramírez, L. & Jordan, I. K. Epigenetic histone modifications of human transposable elements : genome defense versus exaptation. 1–12 (2010).

119. Bantysh, O. B. & Buzdin, A. A. Novel family of human transposable elements formed due to fusion of the first exon of gene MAST2 with retrotransposon SVA. *Biochem.* **74**, 1393–1399 (2009).

120. Manuscript, A. The IRF family, revisited. **89**, 744–753 (2007).

121. Bacon, C. *et al.* Brain-speci fi c Foxp1 deletion impairs neuronal development and causes autistic-like behaviour. 632–639 (2015). doi:10.1038/mp.2014.116

122. Li, H. *et al.* Transcription factor MEF2C influences neural stem / progenitor cell differentiation and maturation in vivo. 1–6 (2008).

123. Martinez-morales, P. L., Quiroga, A. C., Barbas, J. A. & Morales, A. V. SOX5 controls cell cycle progression in neural progenitors by interfering with the WNT–b-catenin pathway. *EMBO Rep.* **11**, 466–472 (2010).

124. Krueger, K. C. & Deneris, E. S. Serotonergic Transcription of Human FEV Reveals Direct GATA Factor Interactions and Fate of Pet-1- Deficient Serotonin Neuron Precursors. **28**, 12748–12758 (2008).

125. Ghosh, A., Michalon, A., Lindemann, L., Fontoura, P. & Santarelli, L. Drug discovery for autism spectrum disorder : challenges and opportunities. doi:10.1038/nrd4102

126. Press, D. What ' s in the pipeline ? Drugs in development for autism spectrum disorder. 371–381 (2014).

127. Qin, L. *et al.* Social deficits in Shank3-deficient mouse models of autism are rescued by histone deacetylase (HDAC) inhibition. *Nat. Neurosci.* **21**, 564–575 (2018).