PhD degree in Systems Medicine

(curriculum in Computational Biology)

European School of Molecular Medicine (SEMM),

University of Milan and University of Naples “Federico II”

Settore disciplinare: MED/04

Multi-omic deconvolution of the regulatory networks underlying neurodevelopmental and autism spectrum disorders: a multidimensional analysis for a new disease modelling paradigm

Alessandro Vitriolo

IEO, Milan

Matricola n. R11136

Supervisor: Prof. Giuseppe Testa

IEO, Milan

Added Supervisor: Pierre-Luc Germain PhD

IEO, Milan

Anno accademico 2017-2018

Table of Contents

List of Abbreviations

7DupASD; 7q11.23 micro-duplication syndrome
ADHD; Attention-Deficit/Hyperactivity Disorder
ADNP-ASD; Helsmoortel van der AA syndrome

ASD; Autism Spectrum Disorder

AtWBS; Atypical Williams Beuren Syndrome patients with partial 7q11.23 deletion

aWBS; ASD diagnosed Williams Beuren Syndrome patients

BP; Biological Process

ChIP-seq; Chromatin ImmunoPrecipitation followed by Sequencing

CNS; Central Nervous System

CNV; copy number variation

CTL; control

DEA; Differential Expression Analysis

DEG; Differentially Expressed Gene

FC; fold-change

GaDeVS; Gabriel De Vries Syndrome

GO; Gene Ontology

hESC; human Embryonic Stem Cell

HVDAS; Helsmoortel van der AA syndrome

ID; Intellectual Disability

iPSC; induced pluripotent stem cell

KS; Kabuki Syndrome

MSC; mesenchymal stem cell

NCSC; neural crest stem cell

NDD; NeuroDevelopmental Disorders

PCA; Principal Component Analysis

PCN; pluripotency core network

RNA-seq; RNA Sequencing

SFARI; Simons Foundation Autism Research Initiative

TAD; Topologically associating domains

VZ; Ventricular Zone

WBSCR: Williams Beuren Syndrome Chromosomic Region

WBS; Williams Beuren Syndrome

WS; Weaver Syndrome

ABSTRACT

Recent literature has highlighted that mutations causing neurodevelopmental syndromes are particularly enriched in genes related to chromatin regulation and synaptic functions. While the latter could be easily predicted, the former shed seeds to the flourishing of epigenomic studies focused on this type of disorders. Intriguingly, most of such disorders couple different shades of intellectual disabilities with peculiar cranio-facial features and systemic defects which are shared, opposite or unique across them. I have set up a dynamic framework of analysis, that encompasses the comparison of multiple disorders, cell types and cultures, to highlight cell-type specific and developmentally-relevant paths of transcriptional deregulation. Building on my lab's expertise to harness potency and stability of induced pluripotent stem cells (iPSCs), I identified two main axes of development through which we could characterize on the one hand cerebral cortex related dysregulations and on the other hand cranio-facial features associated traits, peripheral nervous system- and cardiovascular system-related dysregulations. The former is based on the production of adult glutamatergic cortical neurons through ectopic expression of NGN2 in iPSCs and, in parallel, through production of brain organoids: 3D cultures obtained by neuronal differentiation and patterning via sequential exposure to small molecules. The latter is based on differentiation of iPSCs to neural crest stem cells (NCSCs) and mesenchymal stem cells (MSCs).

I collected and standardised transcriptomic data coming from controls- and patient-derived iPSCs accounting for six disorders, NCSCs for five disorders and MSCs for two disorders; NGN2 neurons for two disorders and brain organoid for one disorder. During my research I helped define new standards for RNA-seq experiments tailored for differential-expression analysis and developed or implemented tools to make cross-disorder and cross-tissue comparisons in a connectable way. This work let me identify regulatory circuitries shared by all disorders or by subgroups characterized by shared phenotypes; symmetric deregulations in disorders caused by mutation of opposite histone modifiers; unexpected subgroups that will require further investigation. For most disorders, my work confirms previously published evidence that dysregulations identified at the pluripotent stage can be inherited and amplified in disease-relevant tissues in a tissue-specific fashion. Thus, I was capable of identifying disease-specific dysregulations at the pluripotent stage and in

disease-relevant tissues; I drew conclusions on iPSCs cross-disorder transcriptional dysregulations through the definition of transcriptional modules; I implemented an analytical framework to boost the ability of identifying the effect of knocking down a certain gene on transcriptional and epigenetic landscapes; I identified sets of genes whose deregulation at the pluripotent stage reverberates and amplifies along development, funnelling and filtering several analyses to converge on a small set of actionable targets; I identified a small set of potential direct targets of PRC2 complex involved in brain development and on the onset of Weaver Syndrome; I identified BAZ1B-specific transcriptional dysregulations in NCSCs that confirm its importance for migration and craniofacial morphogenesis but more in general for chromatin remodelling and human evolution; I helped in the molecular characterization of YY1 mutations, which led to the identification of Gabriele-de Vries Syndrome; I contributed to the molecular characterization of Kabuki Syndrome in neural crest and adult cortical neurons.

INTRODUCTION

Chromatin modelling and gene transcription regulation: a brief overview

It is nowadays common knowledge that genetic information is propagated from mother to daughter cells and from parents to offspring via DNA. Epigenetics is that field of biology that studies how some traits can instead be acquired and inherited without changing DNA sequence. The term itself comes from the greek *epi*, which means “above”, and *genetikos*, which is the adjective of genesis/origin and obviously stands for “genetic”. The first use of this word can be ascribed to Conrad Waddington, in 1942, when he used it to define “the causal interactions between genes and their products which bring the phenotype into being” (Figure 1A). Since then a variety of meanings have been associated to the word epigenetics, such as:
a) the set of phenomena responsible for inheritance of a phenotype across generations - of cells or individuals - that does not require changes in DNA sequence (Boniolo and Testa, 2012; Meloni and Testa, 2014);
b) those mechanisms by which a phenotype is stably maintained across the lifespan of an organism, including non-dividing cells (Beck et al., 2010);
c) that set of processes of gene regulation that requires chromatin modifications (“On the use of the word Epigenetics” Ptashne 2007).

Chromatin is the eukaryotic heterogeneous combination of DNA and proteins, built up as a set of nucleosomes, which is responsible for genetic and epigenetic inheritance. Nucleosomes are made by DNA wrapped with histones octamers, each

carrying ~147bp, and canonically intended to be made by pairs of histone 3 (H3), histone 4 (H4), histone 2A (H2A), histone 2B (H2B) (Luger et al., 1997). Humans bear multiple isoforms of genes coding for these proteins, distributed in multiple copies along the genomes, constituting clusters that are expressed differently depending on cell type and context, with H2A being the richest in terms of isoforms, by counting 19 forms coded by 26 genes. H2B is instead present in 19 variants coded by 23 genes; H3 variants are 6, coded by 18 genes; intriguingly H4 is coded by 14 genes all made of the same variant and equally expressed ubiquitously (Kamakaka and Biggins, 2005; Khare et al., 2012; Talbert and Henikoff, 2010). Among histone proteins we find also H1, which is not included in nucleosomes and it is defined as the “linker” histone, for its role in binding and thus stabilising the two tails of DNA wrapped around nucleosomes, further favouring higher orders of chromatin organization. Indeed, nucleosomes pack into chromatin fibers that dynamically condense to constitute chromosome territories, also binding the nuclear lamina, which can be open or closed to respectively favour or repress transcription. These processes require chromatin remodelling, which is not only accomplished by alternation of histones variants, depending on the cellular stage, type or condition, but also via ATP-consuming remodellers (Goodwin and Picketts, 2018) action and through histones post-translational modification (HPTMs). These modifications can thus be written, read and eventually erased to remodel chromatin (Arrowsmith et al., 2012); Figure 1B). Key developmental processes including cell specification are obtained through histone regulation (Lawrence et al., 2016). Histones modifications (elsewhere “marks”) include primarily acetylation, methylation, phosphorylation and ubiquitination and mostly attain the n-terminal tails of each protein protruding from the nucleosome core (Figure 1C). Gene expression regulation is thus tightly modulated by modifying chromatin structure and accessibility or by recruitment of

transcription factors and histone modifiers at specific genomic locations, including promoters and enhancers. Enhancers are short sequences of DNA, between 10 and 10K bp, typically located on the same chromosome of their target genes (*cis*-regulatory regions) and they can be found at any distance from the relative gene promoter (from very close proximity up to megabases; Long et al., 2016). In fact, enhancers largely outnumber genes – they are estimated to be hundreds of thousands - and the activation of specific ones is correlated with cell-type specific transcriptional schemes. Their position in the genome can be recognized by tracking H3-Lys4 mono-methylation (H3K4me1) and their activity is defined by further histone mark deposition. H3-Lys27 and H3-Lys122 acetylation (H3K27ac and H3K122ac respectively; Tropberger et al., 2013) are the most common HPTMs associated with enhancer/transcription activation, while H3-Lys27 tri-methylation is associated with transcription suppression both at enhancers and promoters. In general, enhancers are defined by identifying genomic regions enriched for H3K4me1 and depleted in H3-Lys4 tri-methylation (H3K4me3), which instead is a recognized mark of active promoters (Rada-Iglesias, 2018). It is to be noted that, up to the moment of this thesis writing, the dilemma on the correlation or causation between H3K4me1 deposition and enhancer recognition by the cell machinery has not been solved yet. In fact, H3K4me1 can be considered sufficient to identify enhancers but it has been proven insufficient for enhancer activation. Moreover, expression of genes can escape the catalytic inhibition of KMT2C and KMT2D which are the main H3K4 mono-methyltransferases (Dorighi et al., 2017; Rickels et al., 2017). H3K4me1 presence is instead acknowledged both at poised and active enhancers (Rada-Iglesias et al., 2011). They work by constituting loops made by physical proximity with their target genes promoters, and these loop structures can be inherited at cell division or upon differentiation, to ensure that correct coupling is

maintained in the right cell-type (Ghavi-Helm et al., 2014; Jost et al., 2014). These structures are kept together by proteins such as cohesin, those of the mediator complex, YY1 and by accessory chromatin remodellers and histone modifiers (Hu and Tee, 2017; Weintraub et al., 2017a). Only in recent years we have come to the awareness that chromosome structure is crucial for gene expression regulation. Topologically associating domains (TADs) are the building blocks of genome three-dimensional organization that lead to the formation of chromosome territories: regions of the nucleus where subsets of chromatin are active and closely regulated (Figure 1D,(Matharu and Ahituv, 2015). Among chromatin remodellers CTCF and YY1 have been identified to regulate crucial stages of TADs formation. The former is the main insulator, which binds specific genomic sites to define TADs boundaries. The latter works inside TADs to define which enhancer is bound to which promoters (Figure 1E). Poised enhancers are regions marked both with H3K4me1 and with H3K27me3, in order not to be active but immediately actionable by sequential demethylation and acetylation upon specific stimuli. Active enhancers are commonly defined as marked both by H3K4me1 and H3K27ac. Moreover, chromatin looping appears to be stabilized by active transcription (Chen et al., 2018) and enhancers are themselves transcribed, up to the point that some studies have entertained the idea that genomic locations can possess shades of promoterness and enhancerness (Tippens et al., 2018). Finally, long stretches of H3K27ac have gathered the attention on peculiar region of the genome defined as “super-enhancers”, for their prevailing H3K27ac enrichment and for their sheer size when compared to majority of enhancers (Pott and Lieb, 2015; Weintraub et al., 2017a).

Other marks associated with repression of transcription and epigenetic silencing are H4K20 mono- and di-methylation (H3420me1, H4K20me2), H3K9 di- and tri-methylation (H3K9me2, H3K9me3). These histone marks have been stably

identified to be enriched in heterochromatin and, in general, with chromatin regions associated to the nuclear lamina (Kouzarides, 2007; van Steensel and Belmont, 2017). Finally, another histone mark worth mentioning is H3K36 methylation (mediated by SETD2), which is generally associated with active transcription, mostly towards 3' of coding regions, and to the recruitment of human mismatch repair machinery (Li et al., 2013; Schmidt and Jackson, 2013). Methyltransferases and histone binding proteins involved in H3K36 methylation (H3K36me2 and H3K36me3) and binding of so methylated regions have initially been proven capable of antagonizing transcription silencing by Polycomb complexes. Only recently H3K36me regions have been more precisely defined as a hallmark of splicing sites and spurious transcription start sites (TSS), on which they act negatively, by inhibiting their engagement by RNA Pol II complexes to fine tune and safeguard transcription (Huang and Zhu, 2018).

Besides cell-type specific, and enhancer mediated, transcription regulation can also be the result of DNA methylation or of the acquisition of specific polymorphisms. In fact, among DNA covalent modifications, one of the most studied and predominant in the human genome is 5-methyl-cytosine (5mC), which happens mostly at CpG sites and is capable of negatively influencing gene expression. Cytosine residues are methylated on the pyrimidine ring by DNMT3A and DNMT3B and methylation is maintained by DNMT1. At promoters 5mC results in downregulation of expression, which is diluted passively through cell division. It can only be removed, after sequential oxidations, by the base excision repair machinery (BER)(Wallace, 2014). Indeed, spontaneous deamination of 5mC is a source of mutations since it can convert the C nucleotide into thymine resulting in a T:G mismatch that may further result in an incorrect substitution on the other strand, to form a T:A pair.

Other genetic elements capable of modifying expression, without changing gene sequence, are expression quantitative trait loci (eQTLs). These genomic objects, like enhancers, can be found at variable distances from the associated gene and are generally distinguished in local (*cis*) and distant (*trans*) eQTLs. Like enhancers they are also generally recognized to work in a tissue-dependent fashion. More specifically, they are single-nucleotide variations often located inside enhancers and promoters, suggesting a positional identification of the interaction between genes and their regulatory regions. Variations in these locations would modulate or hamper the interaction and hence genes transcriptional read-out (Croteau-Chonka et al., 2015; Imprialou et al., 2017; Lee, 2018; Võsa et al., 2018).

Furthermore, an important role on the axis between transcription and translation is played by nonsense mediated decay (NMD): the process by which eukaryotic cells degrade transcripts with premature termination codons (PTCs), which can be introduced by transcription errors and mutations. Transcripts targeting to the NMD pathway occurs through identification of upstream open reading frames (ORFs) or PTCs caused by alternative splicing (Kurosaki and Maquat, 2016). Among other consequences, pathological genetic mutations can exert their effect through induction of NMD (Lloyd, 2018).

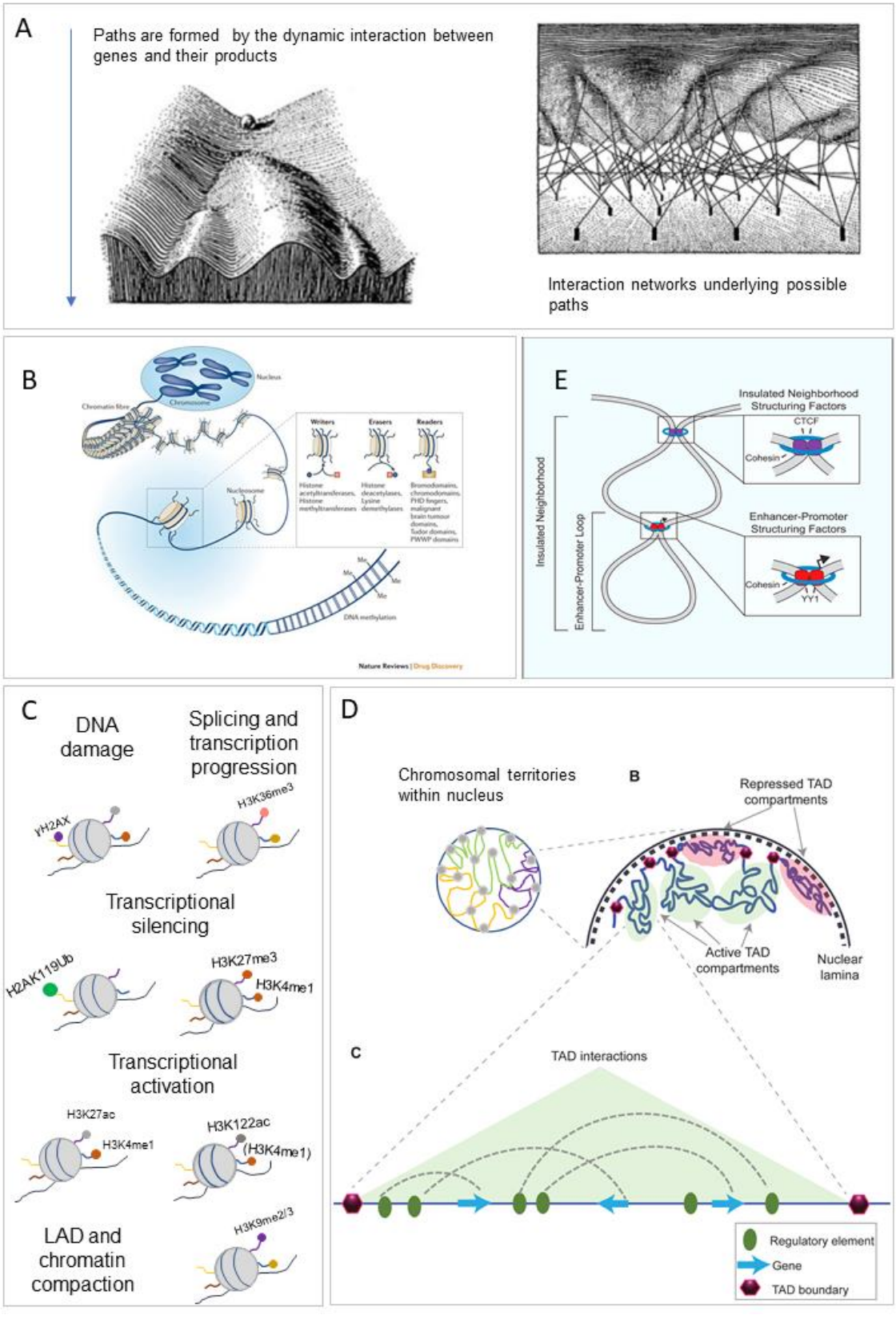


Figure 1 . A) Waddington epigenomic landscape; adapted from the original 1942; It depicts a surface with slopes - comparable to a potential energy surface – under which connections determine interactions between possible path, shaping the final outcome; B) Hierarchy of DNA, chromatin and chromosomes coupled with schematics of nucleosomes with histone tails modifications, readers and modifiers; adapted from Arrowsmith et al. 2012; C) Examples of chromatin modifications and their outcome; D) Hierarchical depiction of chromatin loops forming TADs and chromosomal territories; E) Graphical abstract from Weintraub et al. 2017 representing the alternative roles of CTCF and YY1 in chromatin looping for isolation and enhancer formation.

Human development: from the zygote to neural crests and neurons

Human development starts with fusion of maternal and paternal gametes to form the zygote, the primal totipotent cell, and proceed through asynchronous sequential mitotic divisions. Cells arising from this first step of differentiation are called blastomeres. They further divide for four days (E4) to form the Morula: a 16/32-cell embryo constituted by the specification of these cells into two topological entities: the embryoblast and the trophoblast. At day 5 there is the formation of a blastocyst through a cavitation process, during which the trophoblast secretes liquids inside the embryo forming a cavity called blastocoel. At this point trophoblast cells are finally structured in a single-cell thin layer, clearly separated from the inner cell mass (ICM) cells, which are pluripotent and, after implantation, are destined to give rise to the fetus. Indeed, trophoblast cells will give rise to the chorion and the embryonic portion of placenta, while ICM cells will develop into the fetus, yolk sac, allantois and amnion.

This first commitment step is regulated by the mutually exclusive expression of NANOG and POU5F1 (OCT4) in the ICM, and CDX2 in the trophectoderm (Ralston and Rossant, 2005). After implantation, at about E9 there is the second commitment event during which the ICM will divide into epiblast and hypoblast. The former cells are those keeping pluripotency features and they give rise to all the three germ layers (endoderm, mesoderm, ectoderm) and the amniotic ectoderm: the latter will become the yolk sac. During this commitment the hypoblast express GATA6, which is capable of repressing the self-sustaining pluripotency core network (PCN) composed by NANOG, POU5F1, and SOX2 (Boyer et al., 2005) which instead will

be expressed in the epiblast (Li and Belmonte, 2017). Between E13 and E20 the blastula is reorganized into a three layers structure, formed by cell migration across the primitive streak and towards the rostral part of the embryo, in a process called gastrulation.

Neural Crest Stem Cells

At the end of the third gestational week (3 GW) the embryo completes gastrulation and starts developing the central nervous system (CNS). Cells from the dorsal mesoderm send molecular cues to the dorsal ectoderm, which responds by elongating its cells into columnar ones thus forming the neural plate, which has stopped expressing NANOG and POU5F1 but still expresses SOX2 and PAX6 (Osumi et al., 2008). This process is fundamental to begin neuronal differentiation and essential for maintenance of neuronal progenitors (Zhang and Cui, 2014). The regions flanking the neural plate are called neural ridges and at E21 they start to form, fold and fuse to form the neural tube, in a process called neurulation. The neural tube closes, starting from the middle region of the embryo and continues towards both the rostral and caudal regions. Thus, it is located above the notochord and below the epidermal ectoderm, with which it exchanges morphogenetic signals. This is the embryo component that gives rise to neural crest stem cells (NCSC), which immediately start to migrate ventrally in multiple regions of the embryo depending on their rostral-caudal fate. NCSCs are at the origin of many tissues and organs (Figure 2), and this property has recently gained them the epithet of “fourth embryonic layer” (Dupin et al., 2018). Among the fates of NCSCs there is the peripheral nervous system; the enteric nervous system; cranial nerves; melanocytes; the anterior facial cartilages and bones (from cranial NCSCs); the adrenal gland; the muscle-connective tissue wall of the large arteries, the septum

separating pulmonary circulation from the aorta (from cardiac NCSCs). Furthermore, they participate to the development of teeth, thymus, parathyroid, and thyroid glands (Vega-Lopez et al., 2018).

Disorders and clinical conditions leading to clear neural crest-dependent phenotypes are defined neurocristopathies. Given the paramount importance of the NC in the anatomical foundation of the human body, it is not surprising to find recurrent clinical manifestations such as craniofacial abnormalities, neurological, cardiac, and immunological disfunctions in genetic disorders.

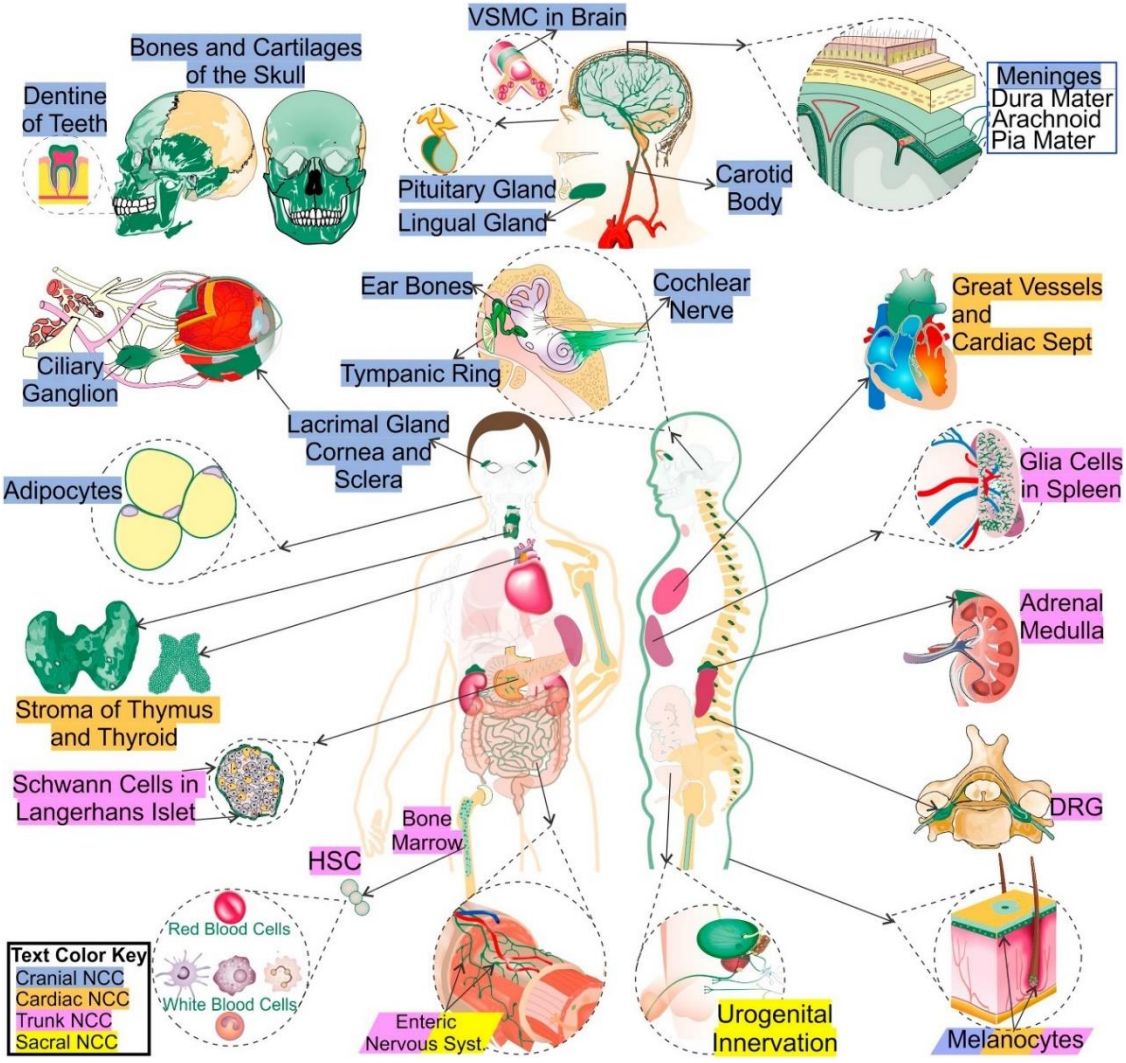


Figure 2. Summary rapresentation of neural crest multipotency, taken from Vega-Lopez et al., 2018

Central nervous system development

After completion of the neural tube, neuronal progenitors gather into a single layer forming a structure with a hollow core. In this setting, rostral progenitors are those that will give raise to the brain, while neuronal progenitors in the caudal neural tube will give rise to the hindbrain and the spinal cord. This cavity will become the brain ventricular system and thus it is called ventricular zone (VZ). Before neural tube closure, the anterior end starts to develop the three brain vesicles: prosencephalon, mesencephalon, and rhombencephalon, which will respectively become the forebrain, the midbrain, and the hindbrain (Figure 3). At this stage PAX6 is expressed only in forebrain, hindbrain, and spinal cord (Osumi et al., 2008), while OTX2, another important transcription factor already expressed in the morula, is active only in forebrain and midbrain (Beby and Lamonerie, 2013).

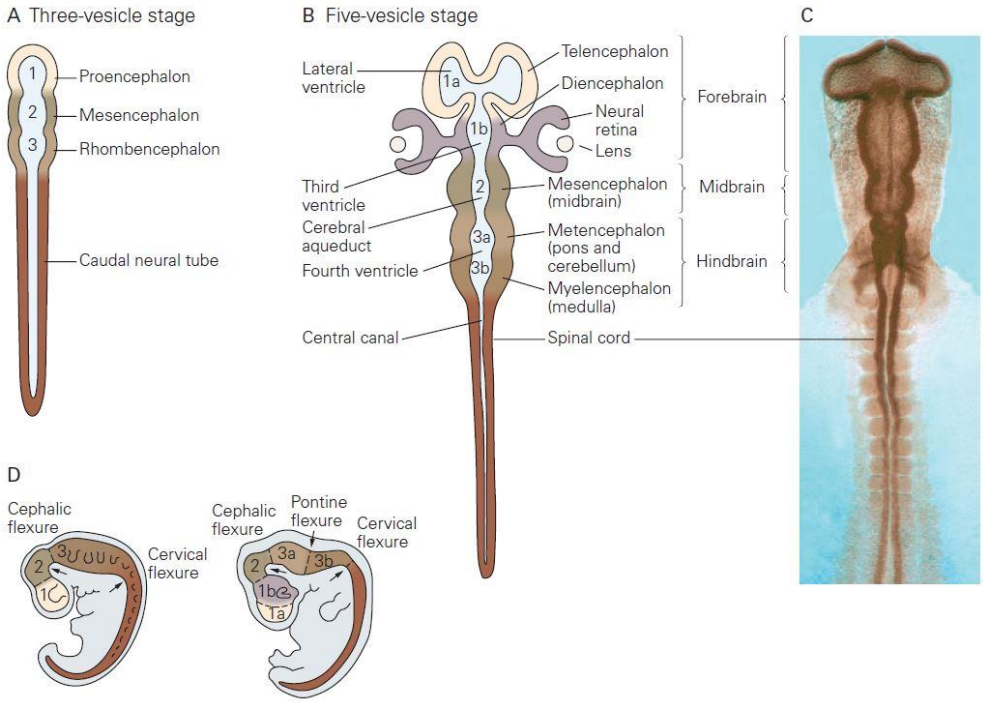


Figure 3. Stages of central nervous system development from the neural tube. Adapted from Kandel – Principles of Neural Science 5th edition.

Brain morphology development happens mostly in the fetal period (i.e. between GW9 and the end of gestation) during which gyri and sulci are formed. During this

period neural progenitors expand clonally to then differentiate into neurons. Most neurons in the adult brain are formed within mid-gestation (spanning between GW18 and GW24). Neural progenitors divide symmetrically starting from the end of gastrulation (~GW3) up to GW6. Afterwards, a switch to asymmetrical division takes place, thanks to which each mother cell gives rise to a neuronal progenitor and a post-mitotic neuron.

Neocortex development

The neocortex is that part of the CNS generally considered responsible for behaviour and cognition. It is composed of six stratified layers of organized neurons, generated by proliferative waves of migrating neurons in contact to the VZ. The neocortex is indeed one of the most complex areas of human brain and it is remarkably different from other species, including primates (Molnár and Pollen, 2014). This structure seems to have appeared in reptiles during the Carboniferous Period and it has increased in size, connectivity and complexity during evolution. Anatomically and functionally the human neocortex is probably the part of the brain that has evolved the most and distinguished itself, even from the most studied mammalian model (i.e. *Mus musculus)*. These differences are evident, especially in the prefrontal cortex structure, which constitutes almost one third of the neocortex and it is responsible for complex cognitive tasks (Carlén, 2017). Neurons of the neocortex are generated in sequential migratory waves during which they migrate from the proliferative zone in the VZ towards the pial surface. In this process, the first migratory neurons reach the pial surface and generate the preplate (PP). The second migratory wave allows the division of PP into the subplate (SP) and the marginal zone (MZ), which includes Cajal-Retzius cells (CR). CRs cells produce Reelin, which controls the positioning of migrating neurons by stopping their

migration. Consequently, every following neuronal migratory wave will get through the most upper layer taking its place in the developing neocortex. Thus, apart from the very first migratory wave, the neocortex develops an inside-out structure in which the oldest neurons are the closest to the VZ. Radial glial guides, initially thought to support neuronal migration from the VZ to the pial surface, thanks to their long processes, extended from the VZ to the pial surface (Rakic, 1972), derive from asymmetric neuronal progenitors, which possess the ability of replicating both int the VZ and in the sub-ventricular zone (SVZ) (Noctor et al., 2001, 2004).

To completely form, the neocortex requires a class of interneurons which integrate this tissue after migration from a proliferative region of the lateral and medial ganglionic eminences, situated between the VZ and the striatum (Stiles and Jernigan, 2010). Inside the neocortex, GABAergic neurons emerge from the ventral telencephalic region, migrating from medial and lateral ganglionic eminences (MGE and LGE, Figure 4), also thanks to Sonic Hedgehog (Shh) transcriptional inhibition (Wilson and Rubenstein, 2000). Glutamatergic cortical neurons differentiation is instead favoured by activation of Ngn1 and Ngn2 transcription factors, which repress the GABAergic phenotype (Schuurmans et al., 2004).

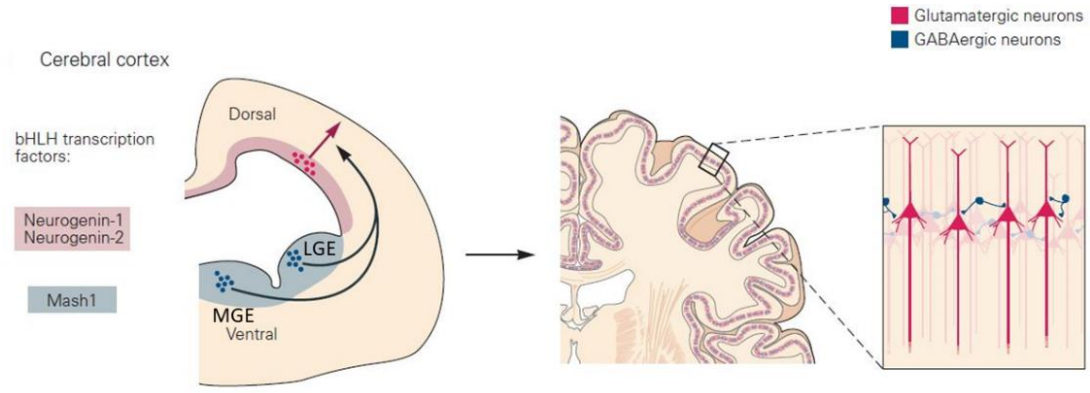


Figure 4 Transcription factors involved in the generation of glutamatergic or GABAergic interneurons. MGE: medial ganglionic eminence; LGE: lateral ganglionic eminence. Adapted from Kandel – Principles of Neural Science 5th edition.

Neurodevelopmental disorders as an outcome of genetic mutation: chromatinopathies and neurocristopathies under study

Neurodevelopmental disorders (NDDs) are a vast class of early onset neurological diseases that encompass intellectual disabilities (ID), autism spectrum disorders (ASD), attention-deficit/hyperactivity disorder (ADHD), schizophrenia, bipolar disorder, learning disabilities, and major depressive disorder. NDDs are caused by both environmental- and genetic factors. It is difficult to precisely estimate the incidence and prevalence of NDDs, especially for those resulting in ID, since it is mainly identified through IQ tests, which are heterogenous and often not accurate (Casaletto and Heaton, 2017; Ropers, 2010). It has been estimated that *de novo* mutations arising in coding sequences explain 42% of severe NDDs cases, with an average prevalence of 1 in 295 birth, of which 59% operate by loss of function and 41% by altered function (Deciphering Developmental Disorders, 2017). Concerning *de novo* single nucleotide variants (SNVs) in non-coding elements such as fetal brain-active elements and highly conserved elements it has been estimated that, genome-wide, 1–3% of patients without a diagnostic coding variant carry a pathogenic *de novo* mutation in active regulatory elements of fetal brain, but only 0.15% of all possible mutations within highly conserved fetal brain-active elements are effectively causing neurodevelopmental disorders with a dominant mechanism (Short et al., 2018). All these observations highlight the importance of regulatory elements in the pathogenesis of these diseases. Moreover, taking into account the list of 910 genes responsible for ASDs (January 2018), provided by the Simons

Foundation Autism Research Initiative (SFARI)[1], “chromatin binding proteins” emerge as the fourth gene ontology (GO) enriched category, preceded only by three terms (in the molecular functions domain) clearly pointing to neurobiology (Figure 5, Gabriele et al., 2018). At the current state of knowledge, it is known that germline mutations in genes coding for chromatin remodellers, histone post-translational modifiers, and DNA methylases causes NDDs (Figure 6).

[1] https://www.sfari.org/resource/sfari-gene/

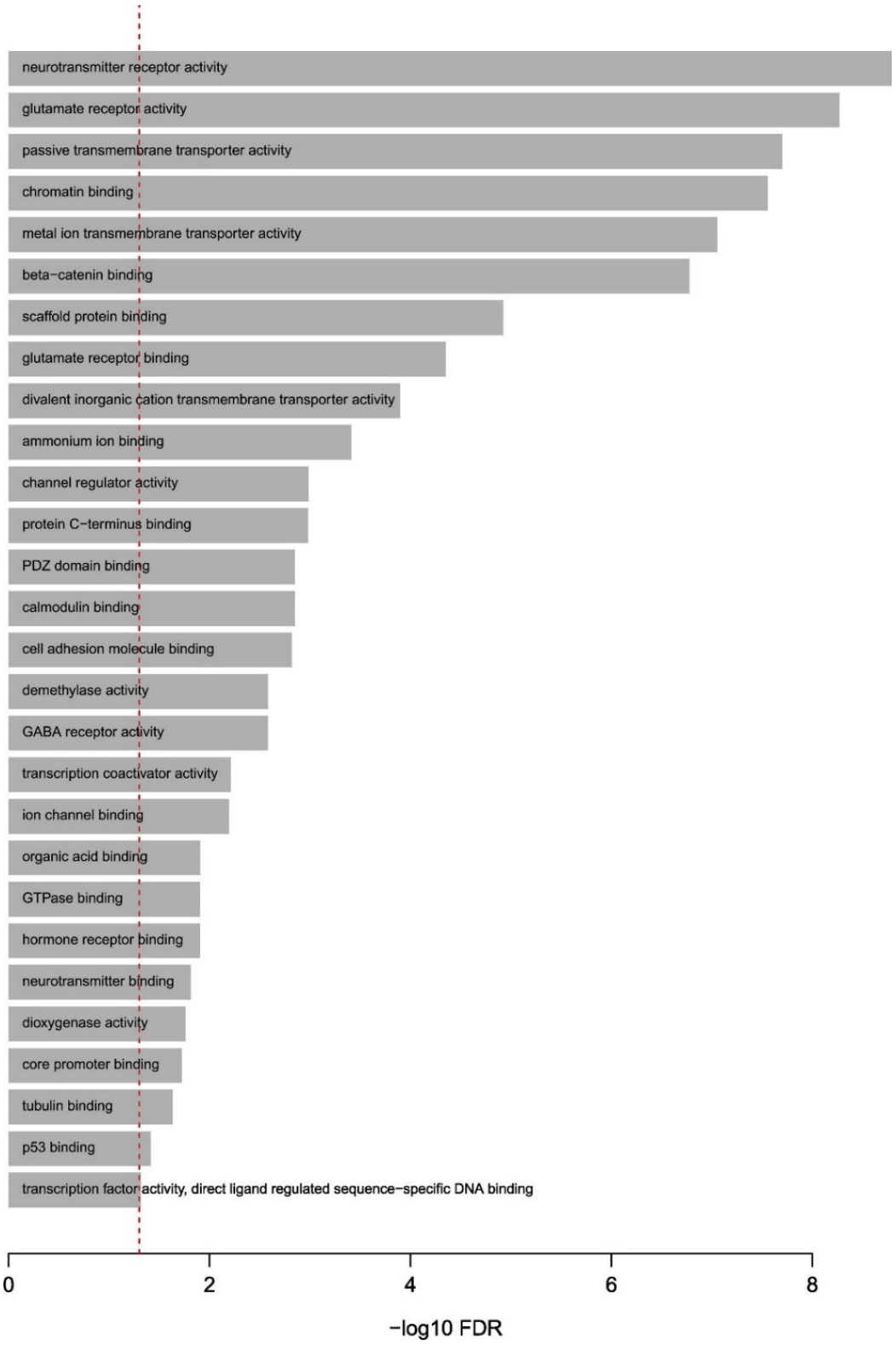


Figure 5. Gene Ontology analysis of SFARI genes responsible for ASDs, taken from Gabriele et al., 2018

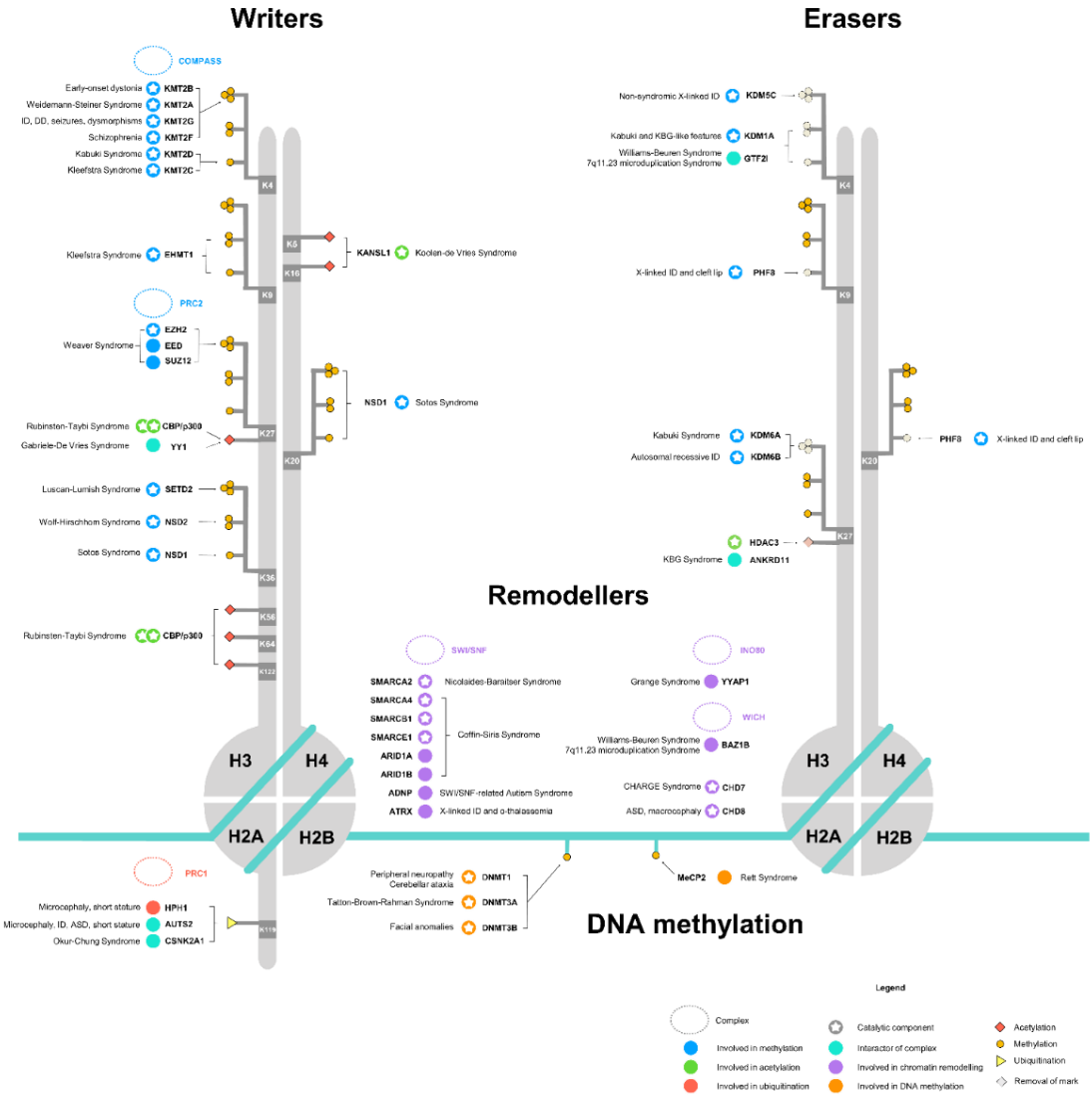


Figure 6. Representation of several genes coding for protein involved in chromatin remodelling, DNA methylation, and histone post-translation modifications and relative syndromes caused by germline mutations. Taken from Gabriele et al., 2018

There are several regions of the brain expressing these genes but apparently the highest expression patterns are located in the brain cortex, in particular in the Brodmann area 9 (located in the prefrontal cortex), which is known to be involved in high cognitive functions and frequently disrupted in several NDDs. Their temporal expression pattern clearly underscores their importance for neurodevelopment, since most of these genes show higher expression during the gestational period and mostly during the fetal period, from GW9 to GW24 (Figure 7; Gabriele et al., 2018). Indeed, these are thought to be the most susceptible developmental stages for neocortex development (Stiles and Jernigan, 2010).

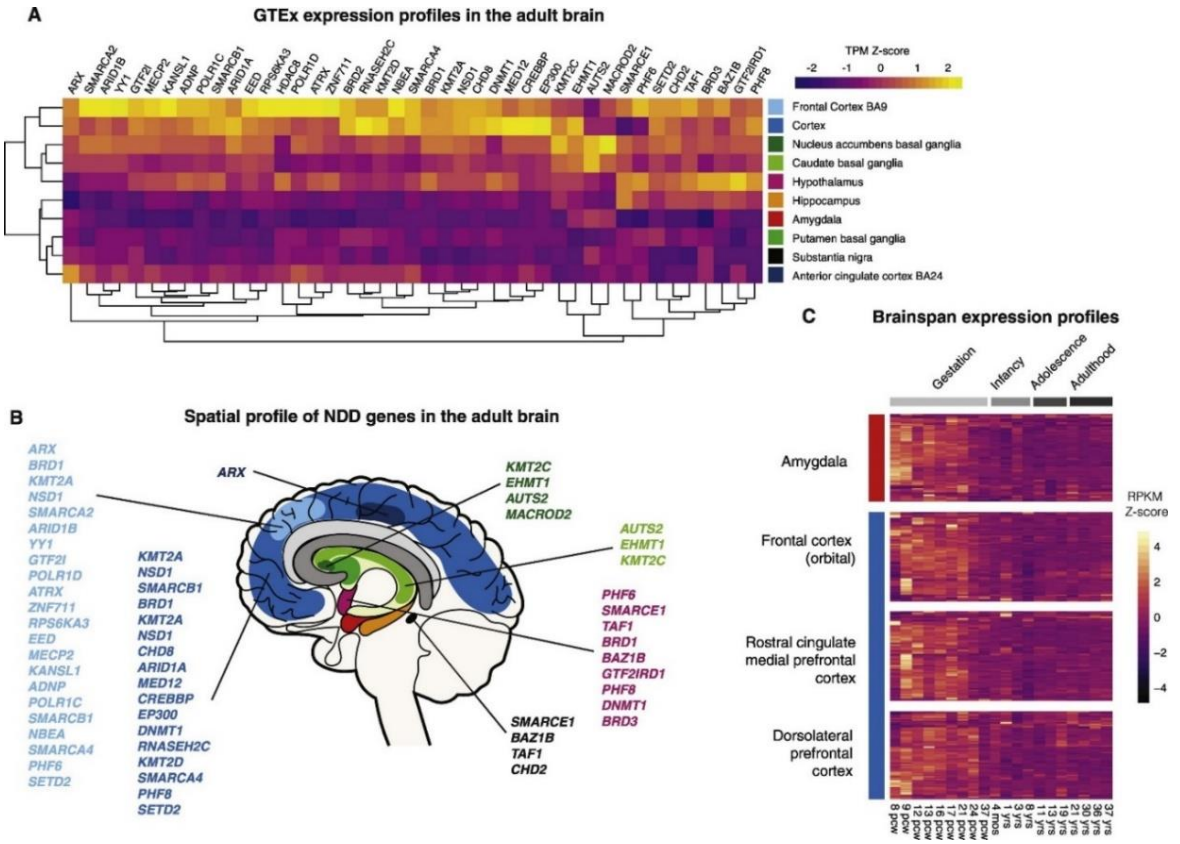


Figure 7. A) Spatial expression pattern in GTEx brain tissues; B) Alternative representation of spatial expression pattern; C) Temporal expression pattern according to BrainSpan atlas. Adapted from Gabriele et al., 2018

Neurodevelopmental disorders caused by mutations or copy-number variation in chromatin modulator genes show both phenotypic convergences and divergences.

A significant proportion of the human population is affected by rare diseases (globalgenes.org). Indeed, neurodevelopmental disorders encompass a diversified set of human health burdens that affect the lives of many patients and their families, in many instances for an extensive amount of time. To reduce such a burden stands as a crucial goal for biomedicine, which can only be achieved by understanding causes and finding cures, or at least treatments capable of reducing it. Moreover, while being generally unique in their genetic origin and makeup, these disorders show crucial phenotypic and molecular commonalities, a remarkable feature that warrants the effort of studying them as a whole, in a dynamic framework (Neurodevelopmental Disorders Across The Lifespan, K. Farran and A.K. Smith). In fact, several shared, opposite or unique phenotypes and endophenotypes, ranging from cranio-facial features to cardiovascular abnormalities, have been identified across neurodevelopmental disorders, including some in the autism spectrum. Moreover, as stated in the previous chapter, a large number of them is caused by mutations in epigenetic modulators and transcription factors(Gabriele et al., 2018). Mutations in genes that regulate transcription, both affecting specific gene regulatory networks (GRNs) and in a broader setting (such as the bulk deposition of a histone mark) are likely to have an immediate transcriptional impact, which reverberates through effects that can dramatically impinge on proper cell fate acquisition.

The majority of epigenetic modulators causing neurodevelopmental disorders is highly expressed during a “window of vulnerability”, during the early phases of

neuronal development. Their expression in the fully developed central nervous systems mostly favours the cortex, which contains neuronal circuits deputed to higher functions such as language processing, visuo-spatial construction and social cognition (Gabriele et al., 2018). Until now, the effort of a sizable part of the scientific community has been devoted to elucidating the regulation of neuronal development through systematic profiling of different stages and areas of the central nervous system. Nevertheless, the presence of recurrent anomalies in many other districts and organs that also overlap among different disorders, as mentioned above, allows to hypothesize that such windows exist also in the development of those organs.

The advent of somatic cell reprogramming (Takahashi, 2014; Takahashi et al., 2007) allows to ask fundamental questions regarding shared and specific vulnerabilities during development (Acab and Muotri, 2015; Flaherty and Brennand, 2017; Ilieva et al., 2018; Quadrato et al., 2016). More specifically, we harness the power of patient-derived induced pluripotent stem cells (iPSCs) and their differentiated derivatives. By analyzing in a concerted way, thanks to *in vitro* differentiation, a diverse set of mutations which cause developmental disorders with shared clinical manifestations, we aim at identifying critical shared and exclusive nodes of transcriptional (and later on functional) dysregulation caused by mutations in epigenetic modulators.

Among neurodevelopmental disorders caused by mutations of copynumber variations affecting chromatin regulators we selected some for their partially overlapping clinical manifestations.

Namely, *ADNP* is mutated in Helsmoortel van der AA syndrome (ADNP-ASD or HVDAS), with a 0.17% estimated prevalence of pathogenic variants in individuals diagnosed with autism spectrum disorders (ASD), thus defining it as one of the most commonly mutated single genes causing autism (Helsmoortel et al., 2014). Patients with mutations in this gene also share facial characteristic with high hairline,

prominent forehead, eversion or notch of the eyelid, broad nasal bridge and thin upper lip. *BAZ1B* and *GTF2I* are both located in the Williams Beuren Syndrome Chromosome Region (WBSCR), spanning around 27 genes on chromosome 7. Copy number variations of this region, which is flanked by highly repeated sequences, lead, during chromosome crossing-over at meiosis, to partial or complete hemizygous deletion or duplication of the WBSCR, causing respectively Williams Beuren Syndrome (WBS) (Barak and Feng, 2016; Bellugi et al., 2000; Meyer-Lindenberg et al., 2004) or 7q11.23 micro-duplication disorder (7DupASD) (Sanders et al., 2011; Van der Aa et al., 2009). WBS features broad forehead, bitemporal narrowing, wide mouth, full lips, small jaw, and prominent earlobes. Usually WBS affected individuals also show small spaced teeth and Iris stellate, which is a very common characteristic. Differently from WBS, 7DupASD can be subtler in terms of cranio-facial phenotype and it is claimed to be underdiagnosed (Bellugi et al., 2000; Meyer-Lindenberg et al., 2004; Osborne and Mervis, 2007). Nevertheless, when spotted, 7DupASD patients present facial features that are usually opposite to those of WBS patients. Intellectual abilities are only partially impaired in both disorders in a symmetric way, with WBS individuals showing hypersociability, good grammar skills compared to individuals with comparable overall intellectual disability, and 7DupASD showing social withdrawal and severe language impairment. *EED*, *EZH2* and *SUZ12* mutations have been found causative of Weaver Syndrome (WS) while *KMT2D* and *KDM6A* mutations cause Kabuki Syndrome (KS). Noteworthy, these two syndromes are respectively caused by mutations in genes either implicated with transcription silencing (Polycomb repressive complex 2) or enhancement (Compass Complex) and show several opposite traits, such as tall versus short stature and peculiar cranio-facial features (Gibson et al., 2012; Imagawa et al., 2018; Kuniba et al., 2009; Tatton-Brown et al.,

2017; Van Laarhoven et al., 2015; Weaver et al., 1974). WS features macrocephaly, tall stature and micrognathia. Intellectual disability is generally mild while Its facial features can include a broad forehead, hypertelorism, large low-set ears and dimpled chin. KS features long and narrow lid fissures and the lateral third of the lower lids are often everted. Eyebrows are highly-arched and broad with some sparsity especially in the lateral portion and eyelashes are thick. Ptosis and strabismus are both quite recurrent. Cleft lip and palate are seen in about a third of patients and the palate is highly arched in most of them. Like in WBS teeth are usually small and widely spaced but they can also be malformed. Finally, YY1 is mutated in the recently characterized Gabriele-De Vries Syndrome (GADEVS) (Gabriele et al., 2017), which features low-set ears and general fullness of the periocular area. Like in WBS and KS patients lid fissures slant downward and strabismus is often present.

Table 1.Neurodevelopmental disorders show shared and unique clinical features

Clinical features	Williams Beuren Syndrome	7q11.23 µDup	Atypical WBS	Kabuki Syndrome	Weaver Syndrome	YY1 ID	ADNP ASD	Summary
Intellectual disability	+	+	**Moderate**	+	+	+	+	*******
Visuospatial impairment	+	+	+	+	-	+	+	******
ASD features	**rare**	+	-	**some**	-	-	+	****
Hypersociability	+	-	+	-	-	-	-	**
Craniofacial dysmorphism	+	+	**Mild**	+	+	+	+	******
Cardiovascular defects	+	+	+	+	+	-	+	******

Helsmoortel van der AA syndrome: ADNP function and mutations

The *ADNP* gene is located on chromosome 20 and it includes five exons, of which only the last three are coding. Most of the reported pathogenic mutations are nonsense or frameshifts mapping on the last exon, which codes for 9 zinc fingers, a nuclear localization signal (NLS), a homeobox domain, and an HP1 binding motif (PxVxLx) (Vandeweyer et al., 2014). Nevertheless, these mutations do not trigger nonsense mediated decay (NMD)(Helsmoortel et al., 2014; Van Dijck et al., 2018). ADNP has been shown to have an anti-apoptotic role during neurodevelopment and a neuroprotective function that was associated to its NAP motif, which is also located in the fifth exon (Pinhasov et al., 2003). Homozygous inactivation of *Adnp* in mouse causes embryonic lethality due to defective neural tube closure at day E8.5/9.5(Mandel et al., 2007), while *Adnp*+/- heterozygous mice display tauopathy, neuronal cell death, abnormalities both in social behaviors and in cognitive functioning (Pinhasov et al., 2003; Vulih-Shultzman et al., 2007). ADNP was initially found to interact, through its carboxy terminal domain (CTD), with the catalytic domains of SMARCA4 and SMARCA2, two components of the BAF (BRG1- or HBRM-associated factors) chromatin remodelling complex (Mandel and Gozes, 2007). More recently, ADNP was found to interact with chromatin remodeler CHD4 through its N-terminal domain (NTD) and with HP1γ/β with its CTD, thereby establishing H3K9me3-independent local heterochromatic domains (Ostapcuk et al., 2018). Notably, mutations in its interactors SMARCA4, SMARCA2 and CHD4 are all established causes of other neurodevelopmental disorders (NDD)(Gabriele et al., 2018; Sifrim et al., 2016; Weiss et al., 2016). Moreover, ADNP knock-out mouse embryonic stem cells (ESC) fail in early neural induction, suggesting that ADNP regulates cell fate decisions. Finally, a recombinant ADNP protein carrying a pathogenic mutation that eliminates its CTD loses the capability of binding HP1γ but retains DNA

and CHD4 binding (Ostapcuk et al., 2018). These results place ADNP at the apex of chromatin regulation and become extremely relevant for its pathogenic role. Our ADNP patient cohort includes several individuals with common mutations: a selection is reported in Table 2 and Figure 8A. These mutations were chosen in light of their distribution over the whole gene and protein structure with the expectation of being representative of the entire cohort of patients.

Table 2. Cohort of ADNP-ASD individuals, with genetic mutations and their protein coding outcome.

Patient	**cDNA mutation**	**Protein mutation**
AD1	c.1222_1223delAA	p.Lys408Valfs*31
AD2	c.2496_2499delTAAA	p.Asp832Lysfs*80
AD3	c.1211C>A	p.Ser404*
AD5	c.2491_2494delTTAA	p.Lys831Ilefs*81
AD6	c.539_54	p.Val180Glyfs*17
AD7	c.2130insC	p.Ser711Lysfs*24
AD9		p.Asn832Lysfs*81

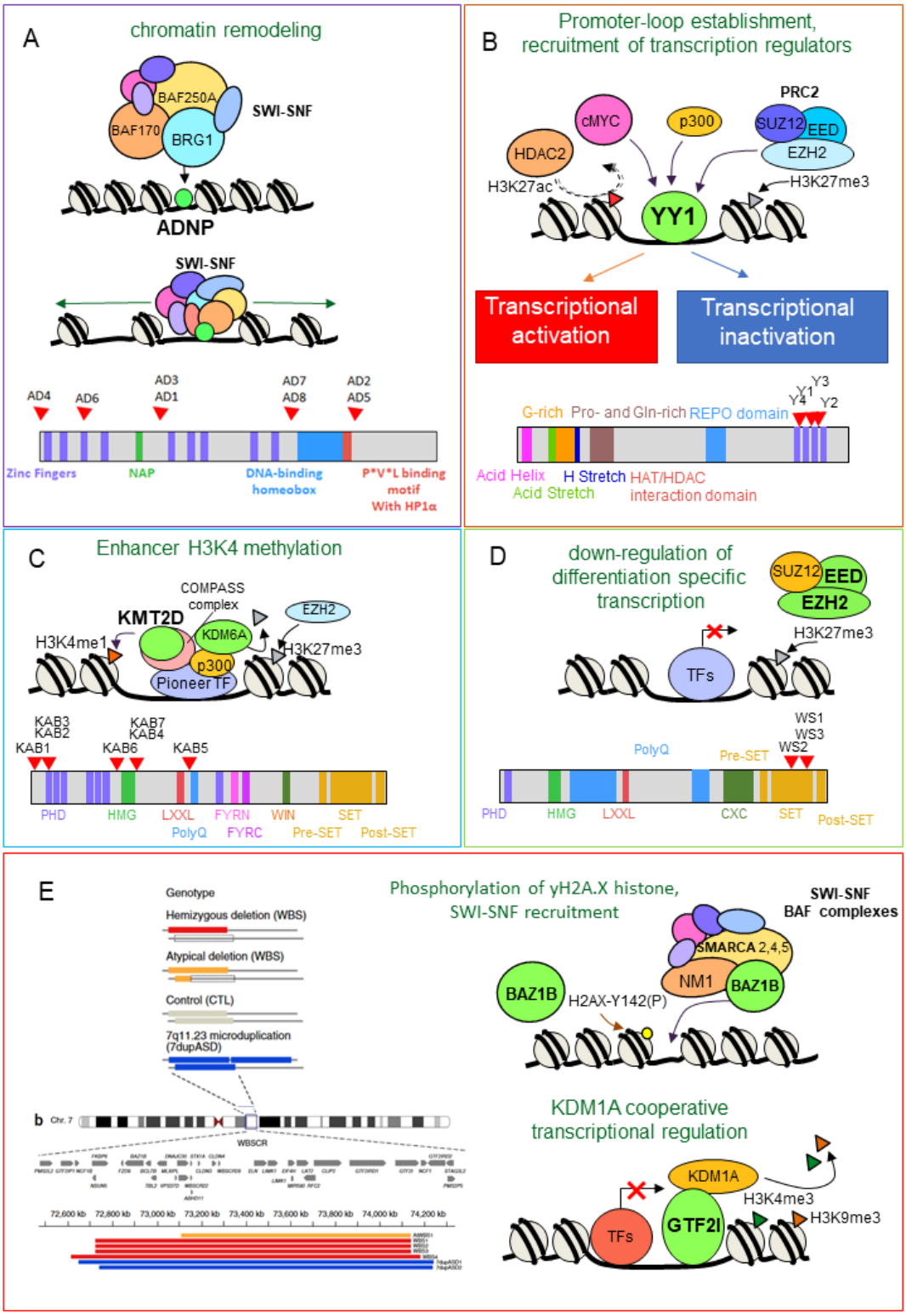


Figure 8 Mapping mutations on the schematics of each gene structure and function. A-D) the depiction of genes involved in ADNP-ASD, GaDeVS, KS and WS with indication of their selected mutations locations. E) depiction of the size of deletion and duplication

Gabriele de Vries Syndrome: YY1 function and mutations

The *YY1* gene is located on chromosome 14 and it codes for an ubiquitous zinc finger protein of 414 amino acids (Figure 8B) (Shi et al., 1991, Patten et al., 2018). Its N-terminal region is responsible for RNA-binding, and enables binding both enhancer RNAs (eRNA) of active enhancers and DNA of promoters. For this reason, it has been proposed that RNA contributes to stabilize YY1 occupancy in regulatory elements (Sigova et al., 2015). Elsewhere it has been shown that also the Zinc Finger domain is capable of binding RNA, even if with low specificity (Wai et al., 2016). The *YY1* homologous in *D. melanogaster* is *Pleiohomeotic (PHO)*, which codes for a transcription factor capable to mediate the recruitment of Polycomb Group (PcG) proteins to DNA Polycomb Responsive Elements (PRE) (Brown et al., 1998).

YY1 was initially found to function both as activator and repressor and was thus named after *Yin-Yang*, the Lao Tsu concept of equilibrium (Shi et al., 1991). Given its evolutive connection with HPO YY1 has initially been studied as a PRC2 interactor and was thought to mediate its repressive function. However, most literature shows YY1 as an activator, especially in nervous system development (He and Casaccia-Bonnefil, 2008). Indeed, genome-wide analysis conducted on mouse ESCs suggested this gene to have a major involvement in activation of highly transcribed genes, and a negative role in nuclear and nucleolar small non-coding RNAs biogenesis (Vella et al., 2012). Moreover, in the same study, no protein of the Polycomb-group was found to immunoprecipitate with YY1 (Vella et al., 2012). Nowadays, the functional role of YY1 seems to have been clarified and it can be compatible both with findings reporting its activation or repression functions. Indeed, YY1 mediates the formation of the structural loop between enhancer and promoters(Weintraub et al., 2017). Experiments showing YY1 as a repressor can be

coupled with the recent description of poised enhancer (Calo and Wysocka, 2013). For example, in myoblasts, together with PRC2, YY1 occupies the MyoD promoter, which is then expressed in myotubes. YY1 removal, even if it is associated with H3K27me3 decrease, did not trigger gene activation (Caretti et al., 2004). Indeed, it is not YY1 function *per se* to activate or repress a gene, but the interactors associated with a chromatin loop. The main interactors of YY1 are p300/CBP (Lee et al., 1995), the INO80 chromatin remodelling complex, as well as the two RNA helicases Ddx5 and Ddx3x (Cai et al., 2007; Vella et al., 2012; Wu et al., 2007). A recent study, in which eQTL were analyzed in 25 tissues, showed YY1 to have predominantly an activating role (Reshef et al., 2018). Furthermore, YY1 has been shown to mediate X-chromosome inactivation by binding *Xist* RNA (Jeon and Lee, 2011) and promoting its expression (Makhlouf et al., 2014). While YY1 homozygous deletion results in embryonic lethality, mouse models have shown a pleiotropy of YY1 haploinsufficiency outcomes. This degree of imbalance leads to serious growth retardation, proliferative and neurological defects such as exencephaly, pseudoventricles, and asymmetry of the developing brain, even though at incomplete penetrance (Donohoe et al., 1999). Other lines of research have underscored the direct relevance of YY1 for neuronal development (He and Casaccia-Bonnefil, 2008). Indeed, YY1 activity is necessary for global nerves myelination of oligodendrocytes (He et al., 2007) and in Schwann cells it mediates the neuregulin-dependent program of myelination genes expression (He et al., 2010). Finally, *YY1* knockdown specifically impairs enhancer-promoter interactions of neuronal progenitor cells regulators (Beagan et al., 2017). YY1 mutations causing GADEVS can be missense and mostly cluster on the zinc fingers domain, and far from its N-terminal region which, as depicted by EXAC database, is instead enriched for non-pathological missense mutations suggesting a robustness of the N-terminal

domain and a higher susceptibility of the C-terminal one (Figure 9, Gabriele et al. 2018).

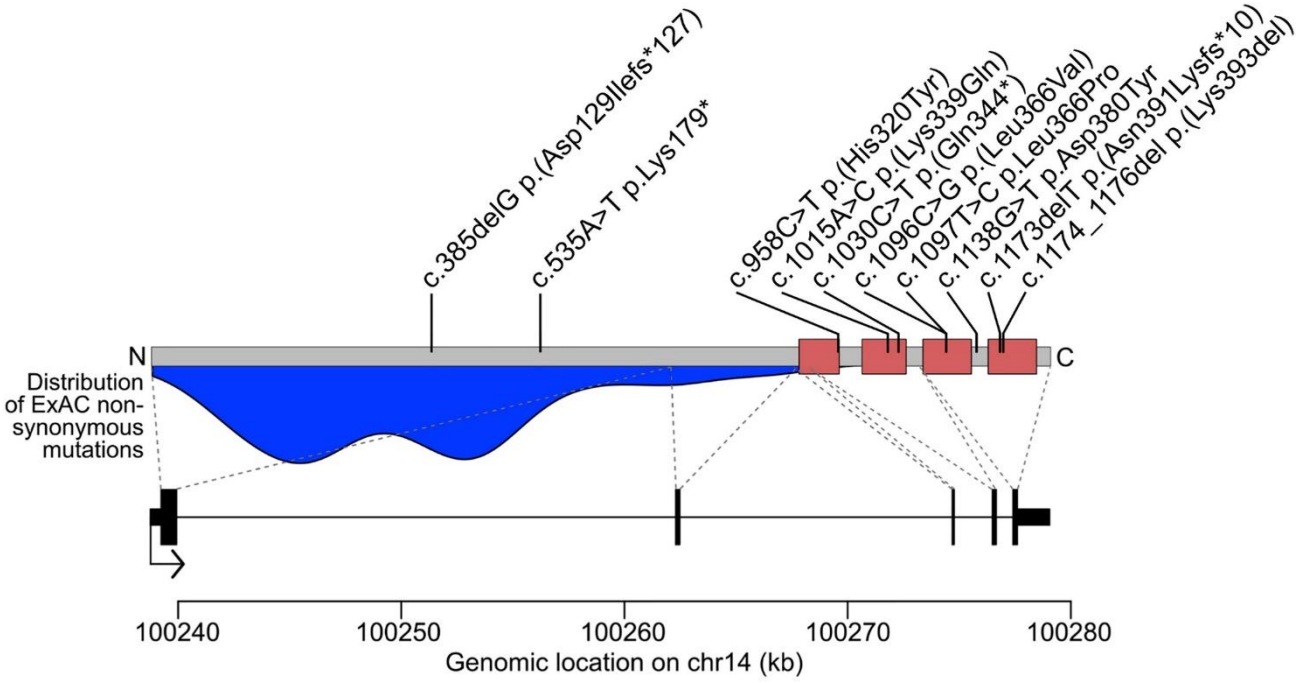


Figure 9. Mapping of YY1 mutations considered in this work, adapted from Gabriele et al. 2018

Kabuki Syndrome: KMT2D and KDM6A functions and mutations

The *KMT2D* and *KDM6A* (Figure 8C) code for the catalytic subunits of Compass Complex involved in K4me1 deposition and K27me3 removal to favour enhancer activation and target genes transcription. *KMTD2* is located on chromosome 12 and codes for the H3K4 methyltransferase KMT2D, a 5537 aminoacids protein containing 2 plant homeotic domain (PHDs) clusters, each made of three PHDs, at the N-terminal domain, (Ruthenburg et al., 2007). At the C-terminal domain (CTD) the protein contains a catalytically active SET domain, another PHD and two FY-rich domains (depending on their primary sequence: FYRN and FYRC). Its structure includes nine nuclear receptor interacting motifs (LXXLLs) and a high mobility group (HMG-I) (Froimchuk et al., 2017). *KDM6A* (also known as UTX) is located on the X chromosome; its expression escapes X inactivation (Greenfield et al., 1998). The 1401 aminoacidics protein has three tetratricopeptide (TPR) repeats at the N-terminal, which are known to mediate protein-protein interactions (Smith et al., 1995), and a treble-clef zinc finger at the C-terminal, which may be involved both in DNA binding and in protein-protein interactions (Grishin 2001, Ginalski et al., 2004). KDM6A is one of two demethylases (KDM6B[2] being the other one) capable of counteracting the function of PRC2, by catalysing H3K27me2 and H3K27me3 demethylation through the JmjC-domain (Agger et al., 2007; Lan et al., 2007). Intriguingly, the Y chromosome hosts a KDM6A homologue named UTY, which is known to be catalytically inactive (Hong et al., 2007). KS samples considered in this thesis are exclusively mutated in KMT2D (Table 3).

[2] Also known as JMJD3

Table 3. Kabuki Syndrome individuals and their mutations. Along mutations it is also reported for which samples we have half-matched controls

individual	Half-matching controls	Mutations in KMT2D
KS1	CTL5	p.C197AfsX11
KS2	CTL12	p.Y223X
KS3	CTL11	p.235PfsX26
KS4	CTL6	p.R263X
KS5	CTL7	p.R3321X
KS6	-	p.K1885QfsX18
KS7	-	p.R2635X

Weaver Syndrome: EZH2 and EED functions and mutations

EZH2 is the main catalytic subunit of the polycomb repressive complex 2 (PRC2) and it is responsible for post-translational di- and tri-methylation of lysine 27 of histone H3 (H3K27me2/me3), a key mark associated to transcriptional repression. Together with SUZ12, EED is one of the three indispensable subunits of PRC2, with the role of enhancing PRC2 processivity. Indeed, EED allosterically favours EZH2 methyltransferase activity by binding specifically H3 tails already carrying trimethylated Lys27 residues (Margueron et al., 2009). Trimethylation of regulatory regions to repress gene expression is well conserved across species and it was well-characterized as responsible for the fine tuning of time-dependent repression of transcription during differentiation (Margueron and Reinberg, 2011; Sparmann and van Lohuizen, 2006). In pluripotency, EZH2 contributes to define a chromatin configuration known as bivalency, in which H3K27me3 and H3K4me3, coexist to maintain the expression of genes at a minimum level: poised for either final activation or repression (Mas et al., 2018; Minoux et al., 2017). Modulation of the expression of PRC2 components is pivotal for the de-repression of cell-fate factors maintained silenced during the early stages of development(Boyer et al., 2006; Lee et al., 2006).

As discussed in the previous chapters, cortical neurogenesis occurs during development in a strictly defined spatiotemporal fashion, so that the smallest perturbation of this timing may turn out to have severe consequences in the final organization of the cortex. Cortical development is characterised by a highly cell-autonomous nature, posing the perfect basis to study the regulation mediated by PRC2 in this context (Testa, 2011). Indeed, upon knock out of EZH2, it is possible to observe an imbalance between self-renewal and differentiation, which favours

accelerated differentiation, reflecting the chronological order of events while showing alteration both of timing and of cellular types proportions (Pereira et al., 2010). Moreover, in mature hippocampal neurons, regulation of the expression of post-synaptic density proteins highly depends on PRC2 direct targeting (Henriquez et al., 2013). More recently, it was shown that knock-down (KD) of EZH2 during neurogenesis causes alterations in the migration of the maturating neurons due to a de-repression of Reelin (Zhao et al., 2015). In adult mouse brains instead, the expression of EZH2 was only observed in neurogenic zone such the adult subventricular zone. Conditional deletion of EZH2 appears to impair the neurogenic capacity of the astroglia, within this structure, to start neurogenesis. Finally, KD of EZH2 in neural tube from chicken embryo led to defects in the apical-basal polarity of neuroblasts with alterations in the neural tube organization and to premature differentiation (Akizu et al., 2016). In this context, p21WAF1/CIP1, a regulator of the cell cycle and of the Rho family members, was identified as a direct target of EZH2 partially explaining the observed phenotype (Akizu et al., 2016). Notably, in the context of the Weaver Project, Sebastiano Trattaro managed to introduce in the genetic background of a control sample (CTL2) the homozygous deletion of EZH2 by means of CRISPR/CAS9 (Cong et al., 2013; Deltcheva et al., 2011; Jinek et al., 2012; Ratan et al., 2018). The obtained line has been included in the Weaver samples cohort (WS5) and its phenotype is currently under investigation. Table 4 includes EZH2 and EED patients considered in this thesis together with their mutation and whether we obtained a half-matched sample from parents.

Table 4. Weaver Syndrome lines with indicated whether any of their sample was coupled with a half-matched control. Mutation at the protein aminoacidic level is indicated on the third column

patient	half-matched control or iso-genic line	Protein; mutation
WS1		EZH2; p.Arg684Cys
WS2		EZH2; p.Val626Met
WS3		EZH2; p.Ala738Thr
WS4	CTL9 (half-matched)	EED; p.His258Tyr
WS5	CTL2 (isogenic)	EZH2 -/- by CRISPR/CAS9
WS6		EZH2; p.Arg684Cys

Williams Beuren Syndromes and 7q11.23 microduplication syndrome: the role of GTF2I and BAZ1B

The General Transcription Factor II-I (GTF2I) takes its name from the helix-loop-helix “I-repeats” domains which characterise the TF-II-I family. At 7q11.23 we can find two paralogs of this gene: GTF2IRD1 and GTF2IRD2. GTF2I is a basal transcription factor, able to bind Initiator elements at core promoter to start transcription (Roy, 2001). It exerts its function also by integrating extracellular signals that trigger its phosphorylation, which induces its nuclear translocation. In the nucleus phosphorylated GTF2I can bind E-boxes and both upregulate or downregulate specific genes expression (Hakre et al., 2006). More recently it has been proposed to directly regulate a relevant proportion of genes differentially expressed in WBS and 7DupASD iPSCs, through cooperation with KDM1A (H3K4me3 and H3K9me3 demethylase) (Adamo et al., 2015).

Three out of five coding isoforms of *GTF2I* (α, β, γ, δ, ε) are expressed in the brain (α, γ, ε), with γ being the most expressed. Links between this gene and human intelligence and behaviour have been provided by several scientific works. First, atypical deletions of the WBSCR appear to cause a very rare WBS form that has no impact on IQ (Ferrero et al., 2010), while several different deletions always encompassing GTF2I neurocognitive defects (Antonell et al., 2010; Dai et al., 2009; Edelmann et al., 2007; Morris et al., 2003). Moreover, two single nucleotide polymorphisms (SNPs) found on *GTF2I* have been associated to ASD within WBS and 7DupASD (Malenfant et al., 2012). One of these SNPs has been further associated with social anxiety, reduced social communication and amygdala activation by avverse stimuli in the healthy population (Crespi and Hurd, 2014; Swartz et al., 2017). Hypersocial behaviour has been observed in mice bearing a

deletion of *Gtf2i* (Martin et al., 2018), and various previous works have associated heterozygous and homozygous KO - but also duplication - of this gene with craniofacial, neurological and behavioural traits reminiscent of either of the two WBSCR CNV dependent syndromes (Mervis et al., 2012; Osborne, 2010). More recently *GTF2I* epigenetic silencing, among few other WBSCR genes, has been associated to variation in social behaviour in wolves (vonHoldt et al., 2018).

BAZ1B is a chromatin remodeller which has been identified as responsible of regulating several cellular processes: from transcription to DNA repair and replication. It has been shown to be important for Vitamin D metabolism (Lundqvist et al., 2013) and to favour epithelial to mesenchymal transition (EMT) (Meng et al., 2016). It is able to phosphorylate histone H2A on Lys142, to signal DNA damage and to recruit Topoisomerase I at replication forks (Aydin et al., 2014a; Barnett and Krebs, 2011; Ribeyre et al., 2016; Yang et al., 2015). It regulates chromatin remodelling by interacting with ISWI and SWI/SNF complex (Aydin et al., 2014b; Goodwin and Picketts, 2018).

Critically, BAZ1B loss has been associated to heterochromatin instability coupled with severe changes in gene expression (Culver-Cochran and Chadwick, 2013). BAZ1B was also shown to be specifically expressed and to play a crucial role in neural crest maintenance and migration in Xenopus Laevis (Barnett et al., 2012). Indeed, its KO confers the acquisition of craniofacial defects in mice (Ashe et al., 2008). Intriguingly, WBS patients bearing a partial deletion of the region that spares few genes, including *BAZ1B*, display milder craniofacial dysmorphisms (Ferrero et al., 2010), further pointing to its involvement in this specific phenotype. *BAZ1B* KO has also been shown to cause heart defects in mice (Kitagawa et al., 2011). Notably, it regulates reward-related behaviours in response both to positive and negative

emotional stimuli in nucleus accumbens (Sun et al., 2016). Our cohorts of patients include four WBS, three 7DupASD, one atypical WBS (AtWBS1) sparing few WBSCR genes (NSUN5, TRIM50, FABP6, FZD9, BCL7B, NCF1, GTF2IRD2), and three WBS diagnosed with ASD (aWBS) Table 5.

Table 5 Williams Beuren and 7q11.23 microduplication syndromes patients. Availability of a half-matched control and ASD condition in single individuals is indicated on the second and third column respectively

patient	Half-matched control	Genotype; condition
WBS1	CTL1	7q11.23 deletion; WBS
WBS2		7q11.23 deletion; WBS
WBS3		7q11.23 deletion; WBS
WBS4		7q11.23 deletion; WBS
WBS5		7q11.23 deletion; WBS with ASD
WBS6		7q11.23 deletion; WBS with ASD
WBS7		7q11.23 deletion; WBS with ASD
AtWBS1		7q11.23 atypical deletion; mild cranio-facial and moderate intellectual delay features
7DupASD1		7q11.23 duplication; ASD
7DupASD2		7q11.23 duplication; ASD
7DupASD3	CTL4	7q11.23 duplication; ASD
7DupASD4		7q11.23 duplication; ASD

Embracing complexity: the power of integrating data and domains to compare neurodevelopmental disorders with overlapping phenotypes

Many models have been described by psychiatry and neuropsychology on the way humans gain their unique cognitive functions while developing their adult brain structure (Burgess and Stuss, 2017; Fletcher and Grigorenko, 2017). Our brain is peculiar in many ways: it has a long post-natal developmental phase; it reaches high levels of modularity and specializations along development; inside it, different cortical regions collaborate to support “higher” cognitive functions (Dekker and Karmiloff-Smith, 2011; Karmiloff-Smith, 2006, 2007; Sakurai and Gamo, 2018). Neurobiologists know and study how our brain development is triggered and guided by internal and external cues generally coming from individual genetics and parental epigenetic heredity together with environmental interactions. On the other side, the most supported theories about psychophysical development - up to few decades ago - were based on the “skill learning” and “maturational” paradigms (Johnson, 2011).

The former, in a nut shell, describes human cognitive development as a process in which individuals develop certain skills mainly by facing, recognizing and training them, and becoming more proficient as they train each skill. The latter, to remain in an equally short and simplified scheme, supports the claim that each cognitive function appears, and brain regions are developed, in a blueprinted process, with slightly different pace in each individual but still, in everybody in the same order. These theories attain to neurodevelopment and have not interacted much with biology up to recent times - also because of the historical moment in which they were developed - while adult psychiatry has been the major field of study to correlate

clinical and molecular with psychological knowledge. Thus, in a way, all our notions have been historically biased, at least for the fact that adults often lose specific domains of cognition/regions of the brain, as outcome of injuries or neurodegenerative conditions (Karmiloff-Smith, 2007; Thomas and Karmiloff-Smith, 2002). In the effort of putting together the skill learning and the maturational viewpoints, in a more biologically friendly (if not grounded) theory, Annette Karmiloff Smith and her coworkers built on the Interactive Specialization (IS) framework (Johnson, 2011) a new integrative paradigm (Dekker and Karmiloff-Smith, 2011; Karmiloff-Smith, 2012). IS is a neuroconstructivist framework that overcomes the two theories by reaching a coproductive intermediate capable of taking into account genetic and molecular knowledge:

-modularity of domains: there is a blueprinted modularity in brain development, but its molecular basis keeps the plasticity to adapt to situations in which something is missing or abnormally present (for instance, upon mutations and gene dosage imbalances);

- progressive disruption model: in the case of neurodevelopmental disorders, this theory takes into account the possibility of a progressive disruption in which certain regulatory circuitries, at the cellular and tissue level, may be more or less important in certain developmental stages, and that impairments may progressively sum up, along the course from embryonic to postnatal stages, to the moment of their blooming/diagnosis;

- general dysfunction impacting primarily a certain domain: it is still possible that a specific mutation or reactions to environmental cues, being they internal or external (i.e. inter cellular/tissue interaction vis a vis bio/chem/physical insults from out of the cell/body), might affect a specific domain;

-indirect dysfunctions: always in the case of neurodevelopmental disorders, some dysfunctions may arise indirectly, for instance: children might need diverse amounts of time to learn to speak, or be variably impaired in learning it, because of problems in earing or understanding other people (in various manners);

The new proposed paradigm builds on the necessity of integrating data and skills from different domains, from genetics to imaging to psychology with the clear aim of connecting the dots of cognition and brain development, with a peculiar focus on language and visuo-spatial recognition (D'Souza and Karmiloff-Smith, 2017). Their approach reveals the limits of studying the brain through focusing on adult fully developed specimens and models, and shifting scientific attention both towards the importance of initial stages of physical and mental formation and to recent evidences of a constant developmental nature of the brain. Thus, on the one hand it points out the importance of considering the difference between late life brain disorders (Karmiloff-Smith, 2009; Karmiloff-Smith et al., 2012) but it also primes neuroscientists to recast commonly defined neurodevelopmental disorders (Karmiloff-Smith, 2010), such as Alzheimer and Parkinson diseases (Wiseman et al., 2015), into late onset neurodevelopmental ones (Edgin et al., 2015).

Characterization of Human transcriptome and epigenome via Next Generation Sequencing.

Next-generation sequencing (NGS) technologies have been developed and implemented in several fields of biology in the last ten years. Along this stretch of time great improvements have been achieved in terms of speed, read length, and throughput, but also in terms of costs. These results have fostered the development

of new NGS applications in basic science as well as in translational research areas (Dijk et al., 2014; Yohe and Thyagarajan, 2017).

Transcriptomic characterization by RNA sequencing

Among NGS technologies, qualitative and quantitative measurements of gene expression can be achieved by mean of RNA sequencing (RNA-seq). From a technical point of view it is worth mentioning that the material being sequenced is actually cDNA, obtained through retrotranscription and amplification of the original RNA material (Berge et al., 2018). Thus, RNA-seq helps take a snapshot of the set of RNA molecules present in a biological system, the transcriptome is their subset, which can be seen as the primary cellular proxy to actualize cell-type specific and developmental genetic programs.

The vast majority of data analysed in this thesis has been produced by Illumina sequencing. In this case libraries of cDNA are loaded onto a flow cell where the sample sequences bind to short oligonucleotides complementary to the adapter sequence. Bridge amplification creates dense “clonal clusters” of each cDNA loaded (Bentley et al., 2008). sequencing by synthesis determines the sequence of each cluster (Ju et al., 2006): single stranded templates are then read while the complementary strand is generated. At each step a single fluorescently labeled deoxynucleoside triphosphate (dNTP) is added. The so formed label has the function of terminating and preventing further dNTP incorporation. At this point the fluorescent label is imaged, and immediately after it is enzymatically cleaved, so that the next dNTP can bind to the elongated chain. Individual base composition is thus inferred directly from the intensity of the fluorescent signal. Moreover, sequencing of cDNA libraries can be Single- or paired-end. In the former case, only one end of cDNA inserts is sequenced, as opposed to the latter, which yields two reads in

opposite orientation, posing a great advantage for future reconstruction of the original fragments (Berge et al., 2018). RNA-seq enables the comparison of gene expression between genotypes, conditions (e.g. stimulation or knock-down), different tissues or cell types, genotypes, time points, et cetera. Such comparisons are undertaken with the intention of identifying sets of so-called differentially expressed genes (DEGs), which are supposed to be specifically affected in one or more of the considered states. The characterization of DEGs can be thus pursued to understand the nature and magnitude of the observed deregulation, and differences across states. The outcome of these analyses is the identification of affected pathways and the reconstruction of gene-regulatory networks underlying the transcriptional phenotype and, by inference, of the condition under study.

Coupling NGS with Chromatin Immunoprecipitation

Chromatin Immunoprecipitation (IP) followed by high-throughput sequencing (ChIP-seq) is used to characterize cellular states by enriching and amplifying fragments coming in this case from chromatin extracts. The objective of ChIP-seq is typically to identify HPTMs, transcription factors DNA binding sites, and generally any ChiP-able functional genomic element (Germain et al., 2014). The first step of this procedure is chromatin crosslinking, fractionation, IP of the target mark/protein with an antibody, library preparation (single- or paired-end) as described for RNA-seq. In most cases conventional ChIP-seq can be done on nuclear extracts or on entire cells with similar results; usually it requires 100 thousand to millions of cells; each IP requires a separate experiment. Sensitivity and specificity of distinct antibodies may require different numbers of cells and amount of starting material, and eventually nuclear extraction. Thus, to yield quantitative results it needs standardization and titrations to produce a robust experiment that requires the characterisation of several conditions and observables (i.e. histone marks etc) (Ryan and Bernstein, 2012) . To identify enriched regions for a certain observable their signal needs to be compared to a null reference. This is usually called “input”: a portion of each sample of chromatin that has not undergone IP with any antibody. Another strategy, mostly useful to study histone marks genomic distribution, is to ChIP the total H3 of each sample. Several other strategies have been recently implemented to have quantitative references by incorporating in each chromatin sample small amounts of exogenous DNA or synthetic spike-in controls, to help the comparison across different marks and cell-types (Bonhoure et al., 2014; Grzybowski et al., 2015; Orlando et al., 2014). Moreover, classic formaldehyde crosslinking introduces strong false positive signals, so many laboratories have

developed native ChIP-seq methods that are able to capture antibody/target interactions without the need to fix nuclei or cells (e.g. Brind'Amour et al., 2015). Coverage Depth is another parameter that must be seriously taken into account when defining an experimental setup which aims at characterising many marks and chromatin binding events at the same time. Large studies and communal efforts have been published with the aim of identifying tissue and development-specific epigenomic makeups, and most of them show ranges of coverage from 10M to 40M reads, and read length ranging from 15 to 100 bp (www.encodeproject.org, www.roadmapepigenomics.org,Hansen et al., 2015; Jung et al., 2014; Nakato and Shirahige, 2017). Nevertheless, recent research has provided evidence that common ChIP-seq setups, much more in terms of coverage than read-length, are underestimating the depth required for a good representation of certain marks, in particular, which have broad and wide distribution along the genome, such as H3K27me3 (Carelli et al., 2017) . Following recent literature, the vast majority of H3K27me3 ChIPseq are not able to recapitulate more than 40% of the global H3K27me3 distribution in human samples by not reaching ChIPseq saturation(Jung et al., 2014). Finally, experiments based on small amounts of cells (from 100 to 100 thousand cells), for instance organoids or developing brain structures, show a detrimental strong decrease in resolution in classic ChiPseq experiments (van Galen et al., 2016).

For these reasons our lab has implemented MINT-ChIP and CUT&RUN protocols (van Galen et al., 2016; Skene and Henikoff, 2017). The former is based on the use of a double barcoding - one sample- and one mark-specific - and its major advantage is to let the user do many ChIP at the same time, on the very same chromatin sample, with several antibodies. It is based on the use of total H3 as reference signal, and it has been published providing evidence of it being able to

recapitulate, with high fidelity, marks distributions in samples as small as 500 cells (van Galen et al., 2016). CTU&RUN is a native method based on the use of an MNase tagged with Protein A. Nuclei are permeabilized and treated with the modified Protein A and an antibody of choice to ChIP a specific target. Protein A binds the heavy chain of the antibody and poses MNase close to the ChIP target. Only small fragments of chromatin enriched for the target are released from the cell. Libraries produced from these fragments have been proven capable of recapitulating correctly marks distribution at coverages as low as 5M reads(Skene and Henikoff, 2017).

Disease modelling with induced pluripotent stem cell and their differentiated derivatives: an experimental design with two main axes.

Studying neurodevelopmental disorders (NDDs) at the molecular level poses challenges and limitations due to ethical and technical questions that have only partly being answered in recent times. I have very briefly discussed previously how and why the human brain and mind are complex to build and understand. The molecular basis of brain development, in terms of regionalization and acquisition of cognitive functions, such as those of learning and memory, still need to be clearly understood (Adolphs, 2015; Day and Sweatt, 2011; Kaang and Kim, 2017; Stiles and Jernigan, 2010). Here I present the basis and the logic of disease modelling on which this thesis has been designed. We have resorted to human models to precisely avoid the limits of animal ones and specifically to identify gene-regulatory networks that are unique to our species. Clearly, direct access to patients' or disease specific brain tissues cannot be achieved: first without a clear consent from patients' and their families and second without facing strong limitations on the actual acquisition of cells and portions of the brain without harming individuals. In fact, direct access to human brains and their use as specimens is almost exclusively possible post-mortem or along brain surgeries motivated by cancer or traumatic events. This fact poses limit to the study of fine dynamic processes such as cellular, epigenetic and transcriptional ones. Moreover, on the need of human models, the NDDs we have considered, given their genetic basis, are likely not only to be included in enhanceropathies (Rickels and Shilatifard, 2018; Smith and Shilatifard, 2014) but also, given their secondary traits, - such as craniofacial features, peripheral nervous system impairment and cardiovascular defects - in

neurocristopathies. Indeed, as described earlier, neural crest is an important domain of the human body that is affected in these disorders, and notably, it does not survive in the adult individual. In facts, neural crest stem cells (NCSCs) migrate in the very early stages of development to populate and constitute several districts of the body after several specific differentiation stages (Bhatt et al., 2013; Mishina and Snider, 2014; Santagati and Rijli, 2003; Spokony et al., 2002).

Developmental and stem cell biology have made crucial steps in recent years to favour the use of human embryonic stem cells (hESCs) to differentiate into those relevant cell types that cannot be accessed in vivo (Craft and Johnson, 2017). hESCs-based experiments have been extremely useful but, if one considers the historic timeframe and the scope of this thesis, they have clear limits in terms of availability and ability to model human genetic heterogeneity and disease specific phenotypes. In terms of reproducibility and use in a translational environment, hESCs are not suitable to predict drug responses: their genetic homogeneity limits their ability to sample human variation. They cannot be easily transplanted to patients, because of their inherent limited immunohistocompatibility. Finally, and specifically in the context of this thesis, these cells cannot have any genotype-phenotype association (Colman, 2008). For all these reasons adult somatic cell reprogramming can represent a solution to gather patients' specific induced pluripotent stem cells (iPSCs), which are capable to regenerate all three embryonic layers (Nakagawa et al., 2008; Okita et al., 2007; Takahashi and Yamanaka, 2006). Somatic cell reprogramming has initially been based on the ectopic overexpression of the so called "Yamanaka factors": Oct4, Sox2, Klf4, and c-Myc. The original protocol was based on the use of integrative vectors and c-Myc overexpression, which was shown to confer oncogenic potential (Okita et al., 2007). The two main solutions to this limitation were i) the use of non-integrative methods, such as Sendai

self-replicating RNA virus (Fusaki et al., 2009) and the backbone of Venezuelan Equine Encephalitis (VEE) (Yoshioka et al., 2013), ii), with the removal of c-Myc and the introduction of GLIS1 (Maekawa and Yamanaka, 2011; Nakagawa et al., 2008). Before and along the proceedings of this thesis, my hosting lab has implemented few reprogramming methods, depending on cogent reasons, such as availability, resistance to specific reprogramming methods by refractory lines, and following need for inter- and intra-project consistency. All reprogramming methods entail the use of non-integrative self-replicating RNA viruses and the exact method is reported in Table 6, together with individuals' genotype. Reprogramming Method 1, as referred to in Table 6, is still based on the use of the original Yamanaka Factors, thus including cMyc, the second and third methods use the same 4 factors (with Glis1 in place of cMyc) but they differ by transfection methods. Indeed, reprogramming method 2 is based on the use VEE virus backbone, while method 3 is based on the leverage of Sendai Virus (Zhang, 2013).

Table 6. iPSCs cohort studied in this thesis, with anonym indication of Genotype, Sample name, ASD diagnosis, eventual half-matched control sample, reprogramming method, and color code used in following graphs (see RColorBrewer)

Genotype	Sample name	ASD	Parental sample	Reprogramming method	graphs color
WS	WS4	NO	CTL9	2	chartreuse1
WS	WS4	NO	CTL9	2	chartreuse1
AD	ADA3	YES		2	darkorchid
AD	ADA3	YES		2	darkorchid
AD	ADA3	YES		2	darkorchid
AD	ADA32	YES		2	darkorchid
WS	WS2	NO		2	chartreuse1
DEL	WBS6	YES		2	red
DEL	WBS6	YES		2	red
DEL	WBS7	YES		2	red
DEL	WBS7	YES		2	red
AD	ADA29	YES		2	darkorchid
DEL	WBS8	YES		2	red
AD	ADA6	YES		2	darkorchid
AD	ADA6	YES		2	darkorchid
WS	WS3	NO		2	chartreuse1
WS	WS3	NO		2	chartreuse1
DUP	7DupASD4	YES		2	blue
DEL	WBS1	NO	CTL1	1	red
DUP	7DupASD2	YES		1	blue
CTL	CTL2	NO		1	black
CTL	CTL1	NO		1	black
CTL	CTL1	NO		1	black
CTL	CTL1	NO		1	black
CTL	CTL2	NO		1	black
CTL	CTL3	NO		1	black
AD	ADA1	YES		2	darkorchid
CTL	CTL5	NO		2	black
CTL	CTL5	NO		2	black
CTL	CTL5	NO		2	black
KS	KS1	NO	CTL5	2	deepskyblue
CTL	CTL6	NO		2	black
CTL	CTL6	NO		2	black
CTL	CTL6	NO		2	black
CTL	CTL6	NO		2	black

KS	KS2	NO	CTL6	2	deepskyblue
KS	KS2	NO	CTL6	2	deepskyblue
KS	KS2	NO	CTL6	2	deepskyblue
CTL	CTL7	NO		2	black
CTL	CTL7	NO		2	black
KS	KS3	NO	CTL7	2	deepskyblue
KS	KS3	NO	CTL7	2	deepskyblue
KS	KS3	NO	CTL7	2	deepskyblue
AD	ADA2	YES		2	darkorchid
AD	ADA2	YES		2	darkorchid
AD	ADA28	YES		2	darkorchid
WS	WS1	NO		2	chartreuse1
WS	WS1	NO		2	chartreuse1
WS	WS1	NO		2	chartreuse1
WS	WS2	NO		2	chartreuse1
WS	WS2	NO		2	chartreuse1
CTL	CTL4	NO		2	black
CTL	CTL4	NO		2	black
DUP	7DupASD3	YES	CTL4	2	blue
DUP	7DupASD3	YES	CTL4	2	blue
DUP	7DupASD2	YES		1	blue
DEL	WBS5	NO		1	red
DEL	WBS4	NO		1	red
DUP	7DupASD1	YES		1	blue
DUP	7DupASD1	YES		1	blue
DEL	AtWBS1	NO		1	red
DEL	WBS1	NO	CTL1	1	red
DUP	7DupASD2	YES		1	blue
DEL	WBS1	NO	CTL1	1	red
CTL	CTL2	NO		1	black
DEL	AtWBS1	NO		1	red
DUP	7DupASD1	YES		1	blue
DUP	7DupASD1	YES		1	blue
DUP	7DupASD2	YES		1	blue
DUP	7DupASD2	YES		1	blue
DEL	WBS3	NO		1	red
DEL	WBS3	NO		1	red
DEL	WBS4	NO		1	red
DEL	WBS4	NO		1	red
DEL	AtWBS1	NO		1	red
DEL	AtWBS1	NO		1	red

CTL	CTL8	NO		2	black
CTL	CTL9	NO		2	black
WS	WS5	NO	CTL2	2	chartreuse1
WS	WS2	NO		2	chartreuse1
YY1	YY1.1	NO		3	coral
YY1	YY1.2	NO		3	coral
CTL	CTL10	NO	CTL10	3	black

Having my hosting lab the tools and expertise to harness potency and stability of iPSCs, I identified two main axes of development through which it would be possible to characterize on the one hand, cerebral cortex related dysregulations and on the other hand, neural-crest derived deregulations (hinting at cranio-facial features associated traits, but also peripheral nervous system- and cardiovascular system-related dysregulations). The former is based on the production of adult glutamatergic cortical neurons through ectopic expression of NGN2 in iPSCs (Zhang et al., 2013) and, in parallel, through production of brain organoids (Paşca et al., 2015). The latter was obtained through differentiation of iPSCs to neural crest stem cells (NCSCs) and then to mesenchymal stem cells (MSCs) with established protocols (Menendez et al., 2013).

The cerebral cortex axis has the advantage of having rapid and stable production of a homogeneous culture of adult cortical neurons, with respect to the established but older and more heterogenous dual SMAD inhibition protocol (Chambers et al., 2009; Muratore et al., 2014). Thus, I was able to identify transcriptional deregulations associated to mature and functional neurons. On the same axis, starting from IPSCs the lab has obtained three-dimensional cultures of human cortical organoids by implementing and refining one of our collaborators protocol (Pasca et al. 2015), which is able to recapitulate early and mid-gestational stages of human cortical development both in terms of transcription and in terms of cellular composition in the first 50 days of culture (Figure 10A and Figure 10B respectively). Indeed, our organoids are 3D cultures obtained from iPSCs by neuronal differentiation and patterning via sequential exposure to small molecules (Figure 10C). Two days of

culture are required for iPSCs to form embryoid bodies (EB). In this phase cells are plated as single cells in E8 medium with ROCK inhibitor. After 48 hours EB are formed and transferred to ultralow attachment dishes. The first differentiation cue is transmitted under a dual-SMAD inhibition, mediated by TGFβ inhibitor and dorsomorphin, which takes 5 days. On the sixth day floating spheroids are moved to be cultured with EGF and FGF. A first adaptation of the original Pasca protocol is applied from day 12 of culture, when organoids are placed on a shaker to favour growth and oxygen intake. The third phase of neural induction starts after day 25, in which organoids media changes to of BDNF and NT-3 for around ~20 days. Organoids are maintained on EGF and FGF from then on.

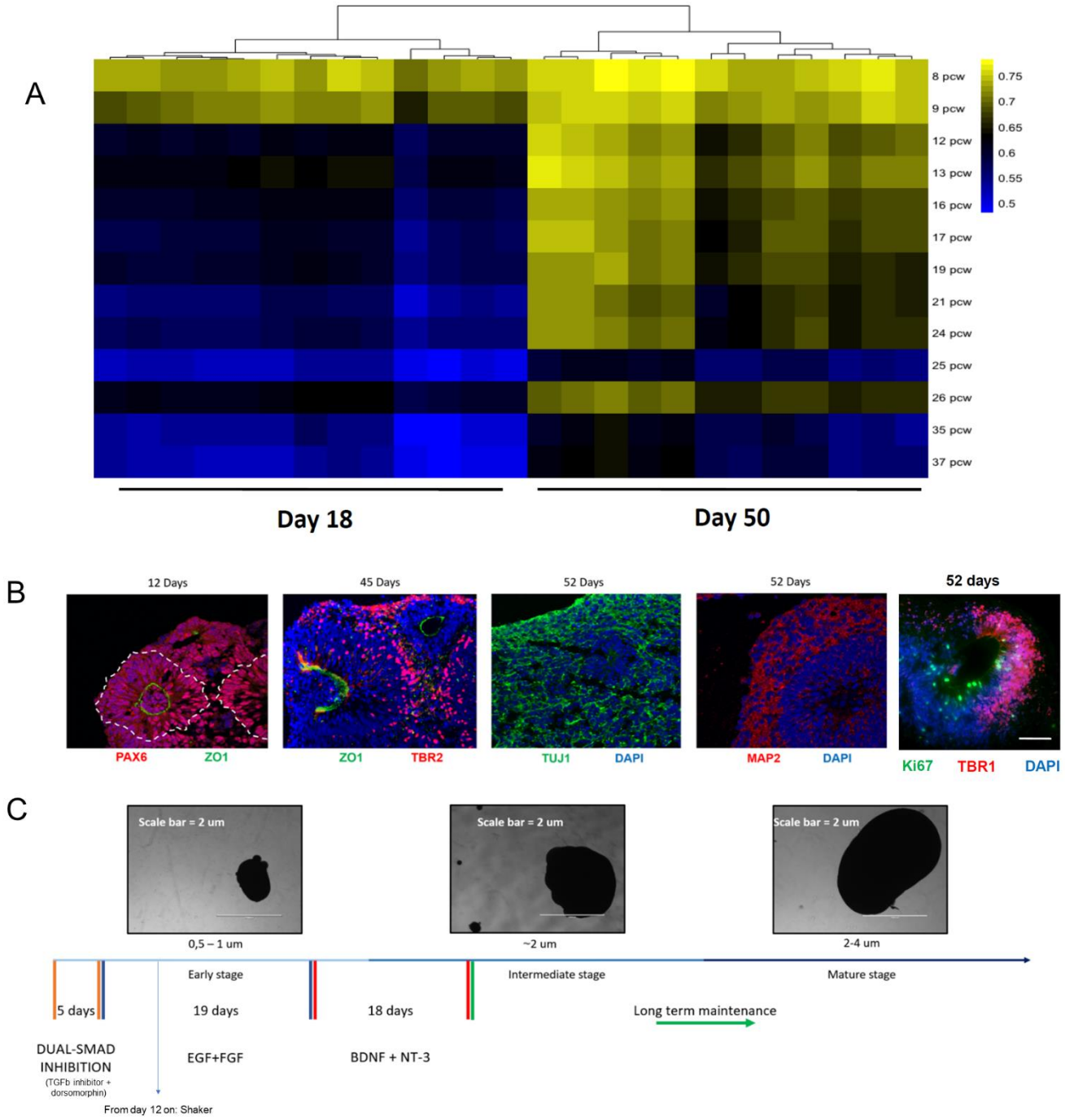


Figure 10. Brain Organoids derivation protocol. A) Correlation of gene-expression levels between organoids and Brainspan tissues; B) cellular composition of brain organoids at 18 and 50 days of cell culture. C) Differentiation and maintenance protocol of human cortical brain organoids up to day 200.

Moreover, I analysed organoids at different stages of culture: between day 18 and day 25, around day 50, and around day 100. This gave us unprecedented resolution to identify modules of genes changing expression along the first stages of

development and disease-relevant ones, both along development and in a stage-specific fashion.

iPSCs lines have been obtained from controls and patients of all six disorders, NCSCs have been obtained for five disorders and MSCs have already been obtained for two disorders. Glutamatergic adult neurons have been obtained for four disorders and organoids, at different stages of differentiation, have been obtained for three disorders.

Previous literature has shown that in WBS and 7DupASD a good proportion of transcriptional dysregulation identified at the pluripotent stage is inherited and amplified in disease-relevant tissues in a tissue-specific fashion (Adamo et al., 2015). Thus, on an expanded cohort of individuals and disorders, I identified disease-specific dysregulations at the pluripotent stage and verified to what extent those are kept and amplified in disease-relevant tissues (Figure 11A). Moreover, upon identifying GRNs dysregulated at the pluripotent stage, either disease-specific or shared by groups of disorders (Figure 11B), I characterised their expression in derived lineages (Figure 11C).

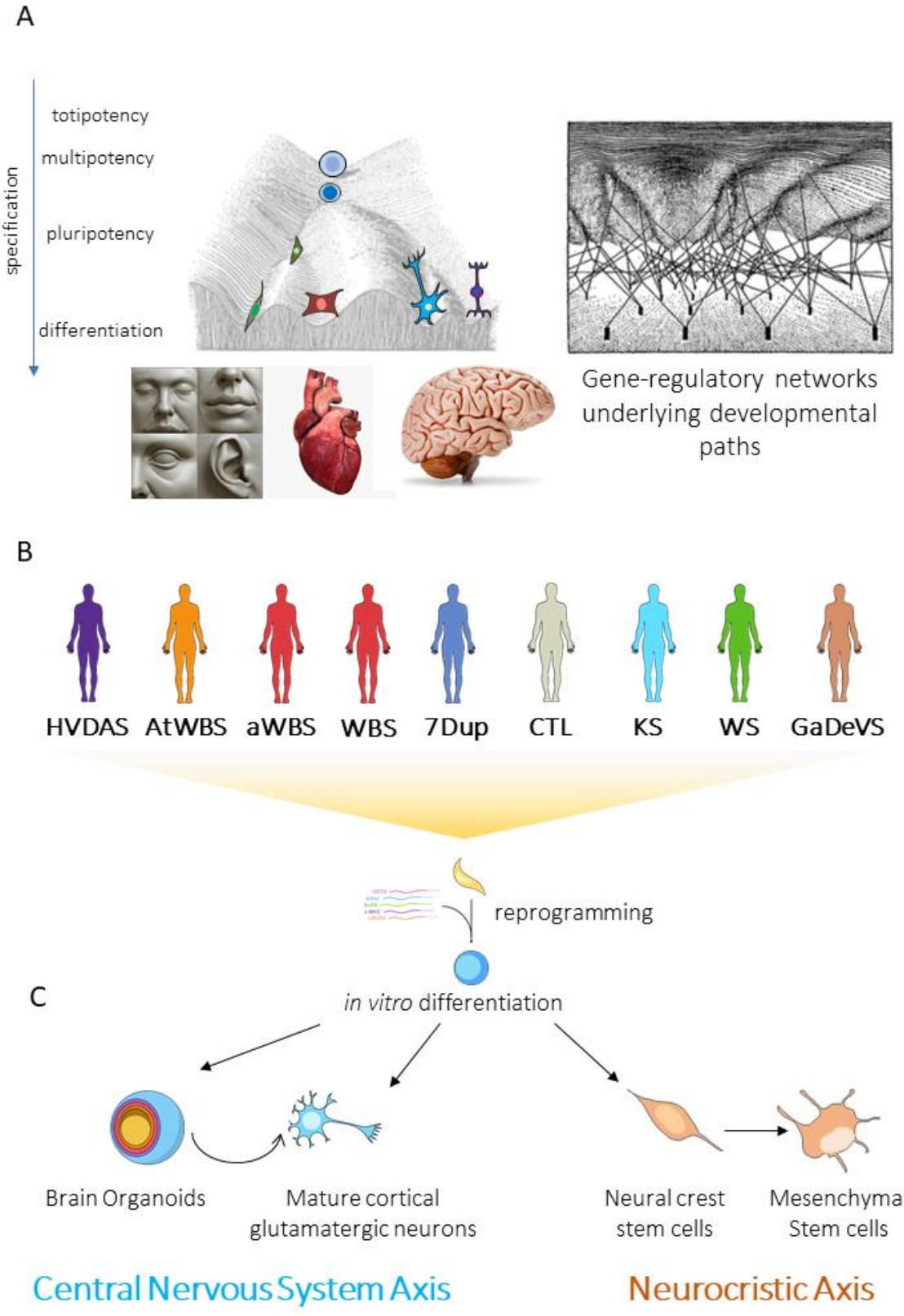


Figure 11. Experimental model: A) iPSCs pluripotent state is the most potent we can achieve in vitro and it shows funding deregulations amplified in differentiated lineages; B) set of disorders among which we want to identify unique and common (div- or conv-ergent) deregulations; C) set of lineages obtained from patient-derived iPSCs forming the two deregulated axis of development affected by NDDs.

AIM 1. To identify gene-regulatory networks underlying developmental paths

The main aim of this thesis is to identify collections of genes responsible for the regulation of the developmental paths under investigation. Taking advantage of several patient-derived cell lines produced in the lab and characterized via NGS I aim at identifying transcriptional modules affected across disorders and differentiation trajectories. In this context target GRNs are those set of genes proposed either as regulators of a differentiation process or as specific of a certain cell type. To identify such GRNs I will start by analysing iPSC data, to identify genes differentially expressed in genotype- or ASD- specific fashion and genes dysregulated across disorders. Then, while analysing data from the other cell types I will verify our previously published observation that iPSCs show dysregulation further retained and amplified in a tissue-specific manner (Adamo et al., 2015).

Following our experimental model, I expect NCSCs to be predictive of mesenchymal deregulation, and to identify clusters of WBS- or 7DupASD-specific dysregulated genes (DEGs). Given the developmental and physiological proximity of neural crest and neural precursor, I expect to find mild commonalities among deregulations in Brain Organoids and NCSCs. Moreover, our organoids, at day 18 and at day 50, respectively show levels of gene expression comparable to 9 and 24-26 pcw embryonic brains as depicted on Brainspan[3] (Hawrylycz et al., 2012; Oldre et al., 2014) (Figure 10A). Considering the comparison between organoids and NGN2 neurons data, the fact that I will be dealing with bulk RNA-seq data in all conditions

[3] © 2016 Allen Institute for Cell Science. Project Overview. Available from: alleninstitute.org/what-we-do/cell-science/our-research/project-overview/

(Table 7), and that organoids are expected to include heterogenous populations of cells - following a slower and more physiological process of differentiation - the overlap between DEGs found in Brain Organoids, across stages, and DEGs in NGN2 neurons could be limited. Nevertheless, I expect NGN2 neurons to be largely represented in later stages of Brain Organoids (d100 and later).

Table 7. List of RNA-seq datasets constituting our cohort

	RNA-seq datasets	
Tissue type	genotypes	conditions/treatments
fibroblasts	CTL, KS, WS	
iPSCs	CTL, AtWBS, aWBS, WBS, 7Dup, KS, WS, YY1, ADNP	KDM1A INH/KD, GTF2I KD, BAZ1B KD
NCSCs	CTL, AtWBS, WBS, 7Dup, KS, WS, ADNP	BAZ1B KD
MSCs	CTL, AtWBS, WBS,7Dup	
NGN2	CTL, WBS,7Dup, KS, WS	
Organoids	CTL, WS, WBS, 7Dup	d18, d25, d50, d100

AIM 2. Definition of a new paradigm: querying convergences and divergences, from the genetic to the molecular to phenotypes.

One major conceptual novelty included in this thesis stems from the original A. Karmiloff-Smith's idea of comparing, on the one hand, overlapping neurodevelopmental disorders phenotypes and, on the other hand, considering all those patients as a single group (in contrast with "normal" individuals). Upon discussion with Pierre-Luc Germain, Michele Gabriele and the lab's head Giuseppe Testa, I decided to move further from Karmiloff-Smith's work and to take in to account disorders with opposite traits, defined at the genetic or at the phenotypical

level (WBS vs 7Dup-ASD, ASD vs non-ASD, KS vs WS), and to ask how they differ at the epigenomic and transcriptional level. I thus present, on each cell-type, how patients iPSCs, and derived tissues, differ in a genotype-specific way or with respect to ASD eventual diagnosis. Afterward I focus on tissue-specfiic and developmental-wise deregulation that appear to be either shared by multiple disorders or opposite as revealed by their genotype, phenotype or cognitive traits. Indeed, I have defined two main subgrouping strategies: 1) disorder specific: considering each of them independent and 2) ASD/nonASD, based on our knowledge on each patient (Table 6).

AIM 3. Defining layers of epigenetic and transcriptional deregulation across disorders

To address questions on the developmental-wise deregulations affecting each or specific groups of disorders, I combine data from iPSCs, NCSCs and MSCs to identify genes that better recapitulate transcriptional changes associated with the Neurocristic axis (neural crest derived); I perform the same type of analyses on the combination of iPSCs and organoids data at different stages; I try to predict and make testable hypothesis on how the developmental-wise deregulation affect a specific district/cell type of the human body involved with NDDs and I try to detect gene-regulatory network involved in interesting pathways or subset of the characterised transcriptomes. To do so I will show how I realized (or intend to) strategies to leverage few interesting public databases such as OMIM, HPO, the HipSci, 4D Genome, ENCODE, Roadmap Epigenomics, BrainSpan, Allen Brain Atlas, Cortecon, High-Resolution Epigenomic Atlas of Human Embryonic Craniofacial Development, and GTEx. These tools and methods will be made

publicly available to be potentially applied on similar or simpler experimental models, and I intend to develop a web interface specifically designed to query our results and produce new ones.

Aims achievement schematics

Here I report a brief schematic of the logics behind this thesis and the main processes through which its aims have been achieved (Figure 12). After assembling RNA-seq or ChIP-seq data and upon quality control, I focused on three main lines of research: 1) detection and characterization of pluripotent state (iPSCs) dysregulations;

2) detection of Neurocristic Axis specfic dysregulation and reconstruction of the GRNs regulating the developmental process leading from iPSCs to MSCs, going through NCSCs;

3) detection of Central Nervous System Axis specific dysregulation, and identification of putative direct targets of KMT2D and EZH2, by taking advantage of WS and KS lines.

iPSCs have been analysed with several approaches and genesets have been characterised using common and novel methods (described later in the text). The Neurocristic Axis has been analysed to identify GRNs regulating it. Moreover, *BAZ1B* KD lines, in combination with ChIP-seq data of the protein and of HPTMs, have been used to reconstruct BAZ1B dependent GRNs in NCSCs, and BAZ1B regulatory function in the context of the Neurocristic Axis. Central Nervous System Axis data has been used to identify i) early brain development transcriptional modules (building on Brain Organoids data) and EZH2 targets in that context, ii) adult cortical neurons specific transcriptional dysregulations in WS and KS lines,

and, taking advantage of HPTMs ChIP-seq and external ChIP-seq data, KMT2D targets in adult cortical neurons.

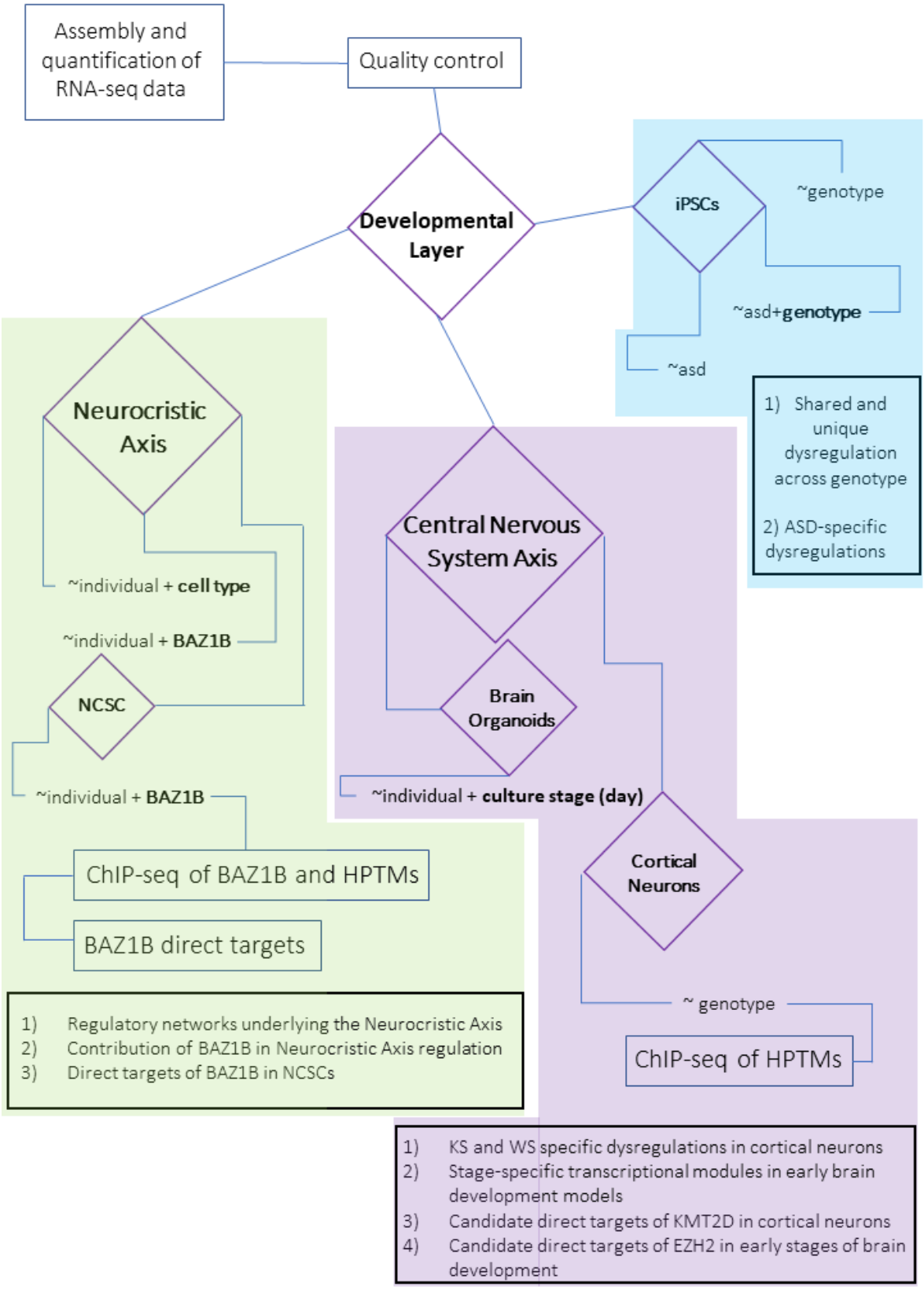


Figure 12. General Scheme of the main aims of this thesis. After assembling and quality check, three groups of analyses have been conducted: iPSCs specific (blue), Neurocristic Axis specific (green), and Central Nervous Axis specific (violet). The ~ symbol anticipate the experimental design of each differential expression analysis performed, with the independent variable in bold when combined with dependent ones.

METHODS

Development of pipelines for RNAseq-based differential expression analysis

RNA-seq gives the advantage of potentially reconstruct the entire transcriptome of a certain sample. Different experimental models can be realised depending on research interests, samples availability and experiments feasibility. In order to identify disease-specific gene expression levels, and underlying gene-regulatory networks, I needed to set up *ad hoc* differential-expression analyses, starting from those already developed in my hosting lab.

Previously published work showed the capacity to identify, in iPSCs, gene expression imbalances across genetic conditions that were further specifically amplified in differentiated disease relevant tissues (Adamo et al., 2015). Moreover, Nanostring data (Germain et al., 2016) for 150 genes - differentially expressed and house-keeping - derived from the same iPSCs lines were available. This suggested the possibility to test different platforms of quantification and differential expression analysis on the ability to reconstruct well established, *bona fide* verified, transcriptomic deregulations *vis à vis* those already implemented. Notably, RNAseq libraries from these lines were generated including also ERCC ExFold spike-ins.

Development of strategies to test differential-expression analysis pipelines required review of available methods and pipelines. The recent introduction of alignment-free tools to quantify gene expression starting from RNA-seq data(Bray et al., 2016; Patro et al., 2017), and the contemporary absence of gold standards and benchmarks on these methods, favoured the idea of realising a new, leveraging the

advantages in hour hands. Moreover, we realised a web platform to test, on the same RNA-seq/Nanostring datasets, evolving and newly implemented pipelines (Germain et al., 2016). It is necessary to mention that the work presenting this effort includes a complete reanalysis of the SeQC dataset and the generation and leverage of simulated transcriptomes, which amplified the power and scope of the benchmark, on which I did not take any direct part.

In this context I manage to help verify that for the purpose of identifying relative quantification across samples, an exact or extremely accurate quantification is not required. Following the same aim as above, count-based methods are superior at the gene level and, in this context, elaborated algorithms (Trapnell et al., 2013) do not outperform more basic tools (Liao et al., 2014) (Germain et al., 2016). Moreover, pseudo-alignment algorithms are equal to alignment-based ones and provide the same performances in an extremely smaller amount of time (minutes vs tens of hours).

Quantification and differential expression methods of choice

Depending on the biological question or conditions to test, a different model matrix must be provided. Model matrices are schematics of the experiment that usually define independent variables that are supposed to drive the variation across groups of samples. They can include dependent variables capable of explaining part of data variation, against which the statistical test must be corrected.

Following on the results of our benchmarking, as a lab, we identified Salmon as the elected read count alignment-free quantifier, using the quasi-mapping algorithm (Patro et al., 2017). During the time of this thesis production, since before its publication several subversions were released, we gradually used latest versions of

the software. I finally re-quantified both at transcript- and at gene- level all available datasets with Salmon v0.91, to reduce potential biases derived from bugs affecting each previous software versions and to optimise inter-analyses comparability. To quantify human-derived data I resorted to NCBI GRCh38 Homo sapiens genome assembly. To perform differential expression analysis and TMM normalization of read-counts I selected the R package edgeR. All statistical analyses presented on the following sections has been performed via R 3.3.3.

Principal component analysis and heatmaps of gene expression

All along this thesis Principal component analyses and heatmaps of gene expression levels will be extensively produced. Unless specified in individual sections, all PCA will be performed on log-normalized read-counts. Normalization, in particular, will be performed with edgeR TMM function on RNA-seq data, while ChIP-seq data will be normalized on library size. Gene expression heatmaps will be presented in terms of z-scores, measured gene-wise on log-normalized read-counts, to identify in each row (gene), the range of expression of each gene across samples.

ChIP-seq alignment and quantification

ChIP-seq data has been aligned with Bowtie 1 (Langmead et al., 2009), using “-v 2” and “-m 1” parameters. The former permits end-to-end alignment of all reads with not more than two mismatches. The latter suppress all alignments for reads more than one is possible. Raw data has been aligned onto the same NCBI hg38 genome used for RNA-seq data. Upon alignment I proceeded with Peak Calling with MACS

v2.1. In the case of peak calling for histone marks I resorted to “—broad” and “–broad-cutoff 0.1” options and for transcription factor binding sites I resorted to default narrow peak calling, with “-q 0.1”, unless specified along the text. The main output of Peak Calling is a bed format files, which has many variants but strictly keep the peak positional information at its core. The first three lines of bed files are “chromosome” were the peak is found, “start” and “end” indicating the two genomic positions at which the signal is confidentially recognised. Genomic regions are generally used to annotate features onto specific genomic locations and they can be used to do a first qualitative analysis of ChIP-seq data. For instance, one can ask whether a certain region of the genome is characterised by the presence of a certain mark only in one of two samples, groups or conditions. Intersection and manipulation of bed files have been conducted with bedtools v2.23 (http://bedtools.readthedocs.org, Quinlan, 2014; Quinlan and Hall, 2010). Once a set of reference regions is defined, I proceed with quantification of raw read counts per region. Historically this has been pursued in the lab using “bedtool coverage”. Latest changes in the software development but most importantly the possibility of conducting coverage calculations in a multi-threaded fashion convinced me to resort to DeepTools 3.0.2 (Ramírez et al., 2016). This software permits coverage estimation but also several alternative normalisation methods and further manipulation/visualisation of relevant data features. Specific features will be mentioned and more precisely referred to along the result session.

K-means clustering

K-means clustering is an unsupervised machine learning algorithm aimed at identifying groups (clusters) of observations into an a priori given number of clusters (k). This number is usually given by visual inspection of data distribution (observations), for instance a PCA. To avoid biases in the identification of the correct number of clusters I proceeded with the so-called “elbow method”. By iterating k-means clustering with increasing k, it is possible to use metrics such as betweenness (between-cluster sum of squares) and “withinness” (within-cluster sum of squares) to measure the ratio between the two, with respect to the total sum of squares in the clusters. This is done to estimate the error made by the clustering and to minimise the error, considering that the sum of squared errors equals 0 only when the number of clusters is equal to the number of observations.

Cortecon Database

Cortecon has been a wide project to assess temporal transcriptional expression during *in vitro* cortical development (van de Leemput et al., 2014). After adpting existing dual-SMAD inhibition protocols, the authors selectively obtained prefrontal neurons from hESCs. After measuring and assessing expression levels along such developmental model, they produced comprehensive data, and further developed a easily queryable database, realised as a publicly available online resource (http://cortecon.neuralsci.org/), including long non coding RNAs and alternatively spliced transcripts species. With such tool they were able to identify genes which expression change along corticogenesis and enriched for neurological and

psychiatric disorders, including autism. Our lab has implemented such tools and retrieved Cortecon data to primarily characterize transcriptomes and genesets.

Allen Brain Atlas

Allen institute is a relatively young non-profit, independent, research organisation funded by Paul and Jody Allen, focused on the study of human brain to support collaborative and multidisciplinary medical, and neuroscientific, research. Among its main subjects are syndromes such as Parkinson and Alzheimer but also Autism Spectrum Disorders. Since its foundation the main goal has been the achievement of a comprehensive atlas of the human brain, spanning from molecular to physiological characterisation, including transcriptional and anatomical data. Initial efforts have been spent to produce the still highly relevant and useful atlas of the Mouse Brain. Data of both Mouse and Human Brain Atlas are today publicly accessible (https://portal.brain-map.org/) and it includes: Gene expression data - from RNA-seq and exon microarray- for tissues and structures also characterised via *in situ* hybridization (ISH), histology, MRI and anatomic data. Microarray data is available for ~300 laser microdissection derived structures (obtained along midgestational development, i.e. 15-21 pcw). Nowadays Brainpan (Atlas of the developing human brain, http://www.brainspan.org/) contains RNA-seq derived transcriptomes data ranging from 8 pcw to 40 years of life. It has thus become a reference to study neurodevelopment and neurodevelopmental disorders also in my hosting lab.

GTEx Project

Genotype-Tissue Expression (GTEx) Project has been designed and developed at Broad Institute, to study genetic variants and their effect on human health. GTEx database includes genome wide association studies (GWAS) and RNA-seq data, obtained from different tissues (also available in a parallel tissue bank) of human individuals, with the primary aim of identifying eQTLs. Nowadays, GTEx data is available through its own portal (https://gtexportal.org/home/). It is worth noticing that also RNA-SeQC has been developed in the preliminary stages of GTEx initiative (DeLuca et al., 2012).

Prior Lab Knowledge: software produced in the lab and applied in this thesis while not being yet published

Basic enrichments and visualization tools

During my PhD I have been supported by few experienced bioinformaticians. Among them my co-supervisor Pierre-Luc Germain and Giuseppe D'Agostino developed independently - and often taking into account my opinions - toolkits, useful wrappers and relatively small R scripts, which have become R packages commonly used in the lab and strongly adapted to this thesis writing. When visualising and depicting results fostered or supported by their work, I will respectively refer to them as PLG and GD. In particular, recursively shown GO enrichments have been carried out using one PLG's wrapper for topGO (Alexa A, Rahnenfuhrer J (2018). topGO: Enrichment Analysis for Gene Ontology. R package version 2.34.0). Master regulator analysis (elsewhere "TF enrichment"), when not explicitly indicated, has been conducted with another PLG R function that leverages the human TFBS tools database (Tan and Lenhard, 2016) to identify, via hypergeometric test, which known targets of transcription factors are enriched in a given geneset, considering a certain universe, which must be given as input.

Brainmaps (GD): leveraging GTEx to visualise and characterise genesets on the basis of their expression across brain regions

This tool requires “grImport2”, “viridis”, ”pheatmap” and “grid” libraries. I obtained these libraries at their latest versions on Bioconductor with R 3.4. First publication of its results are present in (Gabriele et al., 2018b). It takes as input a single gene name and it plots TPM expression levels on a map of the brain (e.g. Figure 13). My contribution was a mere *beta-testing*.

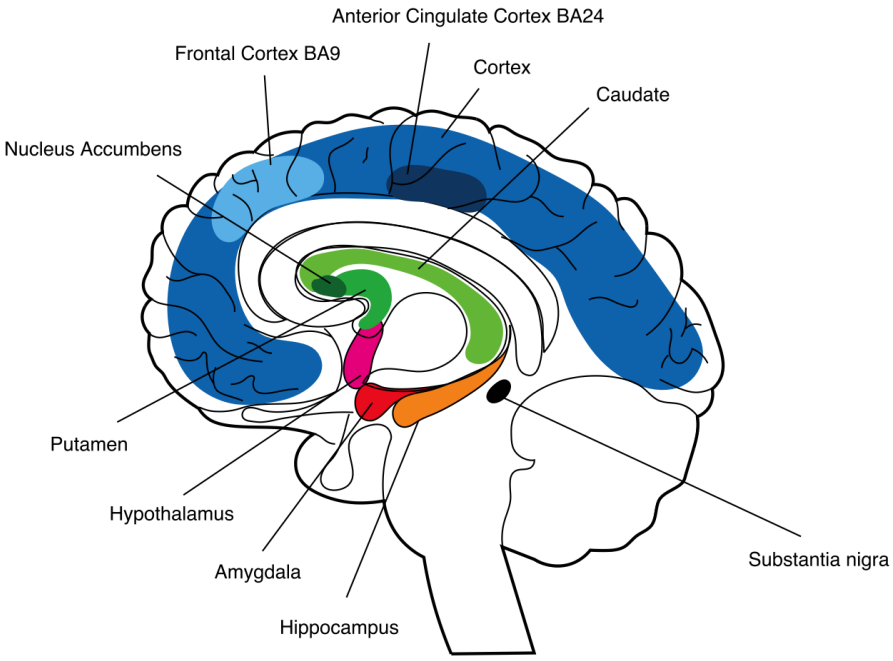


Figure 13. Coloured Brainmaps regions from GTEx. Designed by GD.

Volcano plots (GD+AV)

Volcano plots are commonly used to summarize differential expression analyses by plotting together fold-changes and significance of tested genes. This tool requires “viridis”, “edgR” and “plotrix”, all available in Bioconductor and working with R 3.4. Differential expression analysis conducted with DeSeq2 or edgR output a practical table with genes as rows name and several columns containing the information produced. The software takes as input the DEA table, and thresholds for log2(FC) and FDR values (which are fetched from the same table). Viridis is used for color palettes. Plotrix is used to define a “top signature” by identifying genes with high FC and low FDR, in not crowded areas, that will have their labels shown on the volcano plot. Original input as designed by GD was a DESeq2 dataframe which, with respect to edgeR, mostly changes in terms of column names and ordination. My contribution was a mere adaptation of the code to edgR.

RESULTS

Assembly of iPSCs RNA-seq cohort

For our iPSCs cohort I assembled RNA-seq data from 39 individuals with at least two iPSCs clones, including 10 controls, 8 WBS (out of which 3 diagnosed with ASD and 1 carrying an atypical deletion), 7 ADNP, 5WS, 4 7DupASD, and 3 KS. Only 2 YY1 samples are included in this work. Moreover, in the specific case of YY1 samples, we resorted to single iPSCs clones after the outcome of the above mentioned work by PLG (Germain et al., 2017).

Our cohort of iPSCs has been profiled with two different RNA-seq library preparation methods (RiboZero and PolyA[4]), which have been partially published on GEO (GSE63055; Adamo et al., 2015). The RiboZero (also “Ribo0” later in the text) dataset contains 108 libraries including short-hairpins for GTF2I and BAZ1B and lines interfered with inhibitors of relevant partners of the gene under study, such as KDM1A. In this cohort we observe a strong presence of rRNAs, of which RNA45S is the most abundant. Of this cohort we removed 4 samples which were showing less than 5 million reads after removing rRNAs (1 ADA6, 1 ADA1 and 1 ADA2 clones, see Table 6). Yet, globally rRNAs account for a negligible number of reads in most samples, with 4.09% read-counts at gene-level quantification as a median and an average level of 13.11%, with 88 samples having less than 25% of reads eroded by rRNAs presence (Figure 14A) and rare situations in which the amount of

[4] The former is based on rRNA depletion while the latter is based on enrichment of poly-adenylated transcripts (Zhao et al., 2014)

rRNAs reaches 82.87% (Figure 14B). Most samples show a coverage higher than 10M reads, excluding rRNAs (Figure 14C).

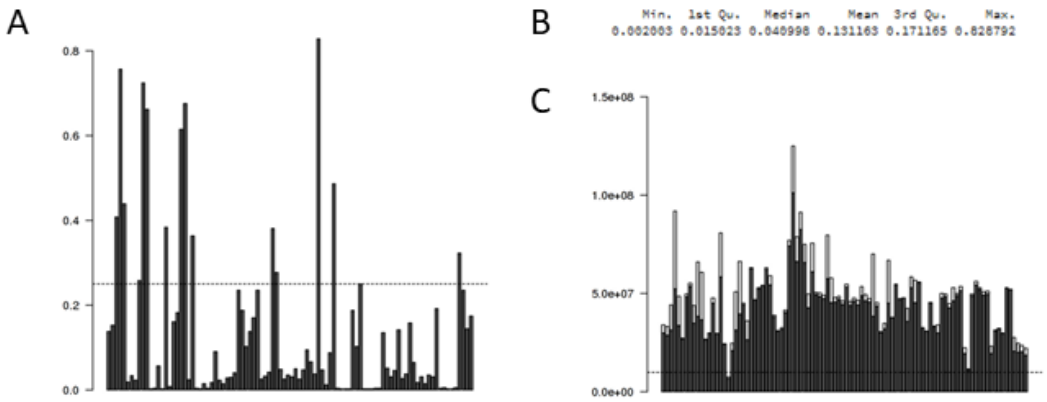


Figure 14. Preliminary analysis of iPSCs RNA-seq data and rRNAs presence in Ribo0 data. A) Most samples show less than 25% of reads to be removed due to rRNAs contamination; B) Summary statistics of percentages of RNA45S reads over total reads count per sample; C) Most samples show a coverage higher than 10M reads after filtering rRNAs;

Out of previously published 48 iPSCs lines, 21 were polyA. Principal component analysis of log-TMM normalized-counts show a net separation on the first principal component (which accounts for 21% of the global variation) by library type (Figure 15A). Given the plan disproportion in number of polyA and Ribo0 libraries I decided to conduct the following analyses only on Ribo0 data, accounting for a wealth of 131 lines.

To conduct a multifactorial analysis on iPSCs across disorders, using control lines as a base line of expression I decided to remove all RNA-seq libraries where samples were interfered with short-hairpin or treated with inhibitor molecules (i.e. 40 lines). Moreover, a deeper analysis of this data composition highlighted the presence of a control sample (CTL1) for which 3 different clones had been sequenced 10 times in total (Figure 15B). To avoid over-representation of one sample and a bias towards CTL1 in terms of controls representation I decided to

aggregate all sequencing runs in a clone-specific fashion. Both before and after aggregating reads from those 10 sequencing runs into 3 clone-specific ones, correlation levels are extremely high (Figure 15C).

As I mentioned in the introduction, iPSCs lines considered in this work have been reprogrammed with 3 different methods, posing a potential technical bias. Principal component analysis shows a general overlap between the two latest reprogramming methods (i.e. 2 and 3), and an acceptable global distribution (Figure 15D). Still, besides showing a general clustering on the first two principal components (PCs), the two most abundant reprogramming methods (in terms of sample size) appear to show a general overlapping distribution (Figure 15D). Noteworthy, several genotypes are well represented in the two largest groups (Figure 15E) and RPKM normalization shows an even higher overlap among the three methods (Figure 15F).

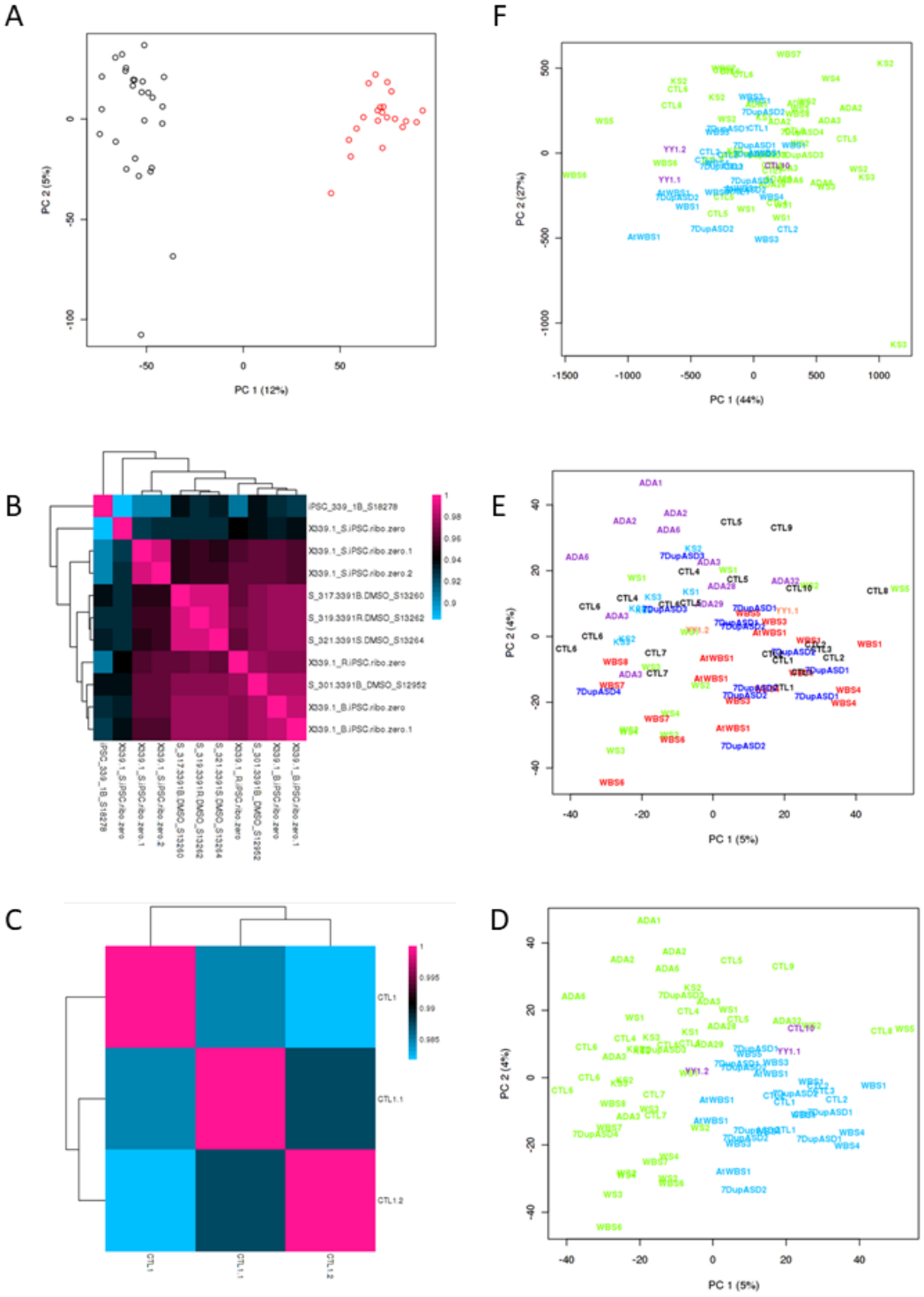


Figure 15. A) Principal component analysis on the 48 samples data published in 2015 shows a sizeable technical bias introduced by library types (Ribo0 in black and polyA in red); B) Correlation heatmap of CTL1 RNA-seq libraries (log-TMM read-counts); C) Correlation heatmap of aggregated CTL1 RNA-seq libraries; D) Principal Component Analysis of iPSCs lines show a separation by reprogramming method (Method1 in blue, Method 2 in green, Method 3 in purple); E) Principal Component Analysis of iPSCs lines show a decoupling of "genotype" and "reprogramming method" factors; F) Principal Component Analysis of gene expression (RPKM) show global accordance across line reprogrammed with different methods (same colours as in E).

Transcriptional characterization at the pluripotent stage: shared and unique deregulations across disorders and ASD-specific signatures

iPSCs specific deregulation

In order to identify genes differentially expressed across disorders I set up a multifactorial model matrix taking into account two factors: genotype and ASD diagnosis. In order to verify whether part of transcriptional deregulations happening at the pluripotent stage was connected to ASD I envisioned four different types of analysis: testing against a single factor (i.e. ~Genotype, ~ASD) or for both (i.e. ~ASD+Genotype), using one or the other as independent variable. Moreover, while testing for genotype, I also considered doing a linear regression either with all genotype coefficients at once or using one at a time (dropping the other coefficients/hypotheses, to define disease-specific deregulations). For the sake of simplicity, I show here four types of analysis: the first has been carried out testing for all genotype coefficients as the independent factor, correcting for ASD ("genotype wise"; ~ASD+Genotype); the second has been carried out testing for Genotype as in the first case but without correcting for ASD (~Genotype); the third has been carried out testing for ASD without any blocking/correction; the fourth is actually a set of differential expression analyses (DEAs) testing each time against a specific genotype ("genotype specific", e.g. ~WBS).

Genotype wise differential expression in iPSCs (without correcting for ASD)

After aggregating different CTL1 sequencing runs into 3 clone specific ones, I initially conducted a differential expression analysis on the 91 selected lines, accounting for 39 individuals and 7 genotypes, considering as one aWBS, AtWBS and WBS (~Genotype). I identified 151 genes differentially expressed in at least 4 disorders with FDR 0.05 (Figure 16A), 475 are shared by at least 3 disorders. These two types of selection are done without specifying any group of four disorders. Thus, subsets could be shared by different groups of disorders. GO enrichments of these genes point to few categories, which account for: cellular response to zinc ion, and its regulatory processes, regionalization, multi-organism reproductive process, bone mineralization and biomineral tissue development, osteoblast differentiation and sensory perception (Figure 16B). Most genes with FDR < 0.05 show a very small fold-change (Figure 16C). Thus, I selected, as bona fide differentially expressed across disorders, those with an average FC of at least 1.5 (Figure 16D). These 372 genes show an interesting enrichment for forebrain neuron differentiation (p-adj ~$2.5e^{-04}$), which is actually attributed to only 8 genes. Nevertheless, 6 of these genes are downregulated in most disorders and 2 upregulated, showing two cluster, with one including mostly CTLs samples (Figure 16E). Finally, few WBSCR genes are differentially expressed in disorders which are not caused by CNV of the 7q11.23 region (Figure 16F). Among them, MLXIPL is a transcription factor involved in metabolism, involved in Fatty Liver Disease; ELN has been associated with late arterial morphogenesis and to WBS heart defects; CLDN3 and CLDN4 are integral membrane proteins involved in cell junctions dynamics and blood-brain barrier transmigration by immune cells; STX1A is involved in osteogenesis and cystic

fibrosis but it has also been implicated in Asperger Syndrome (Durdiaková et al., 2014).

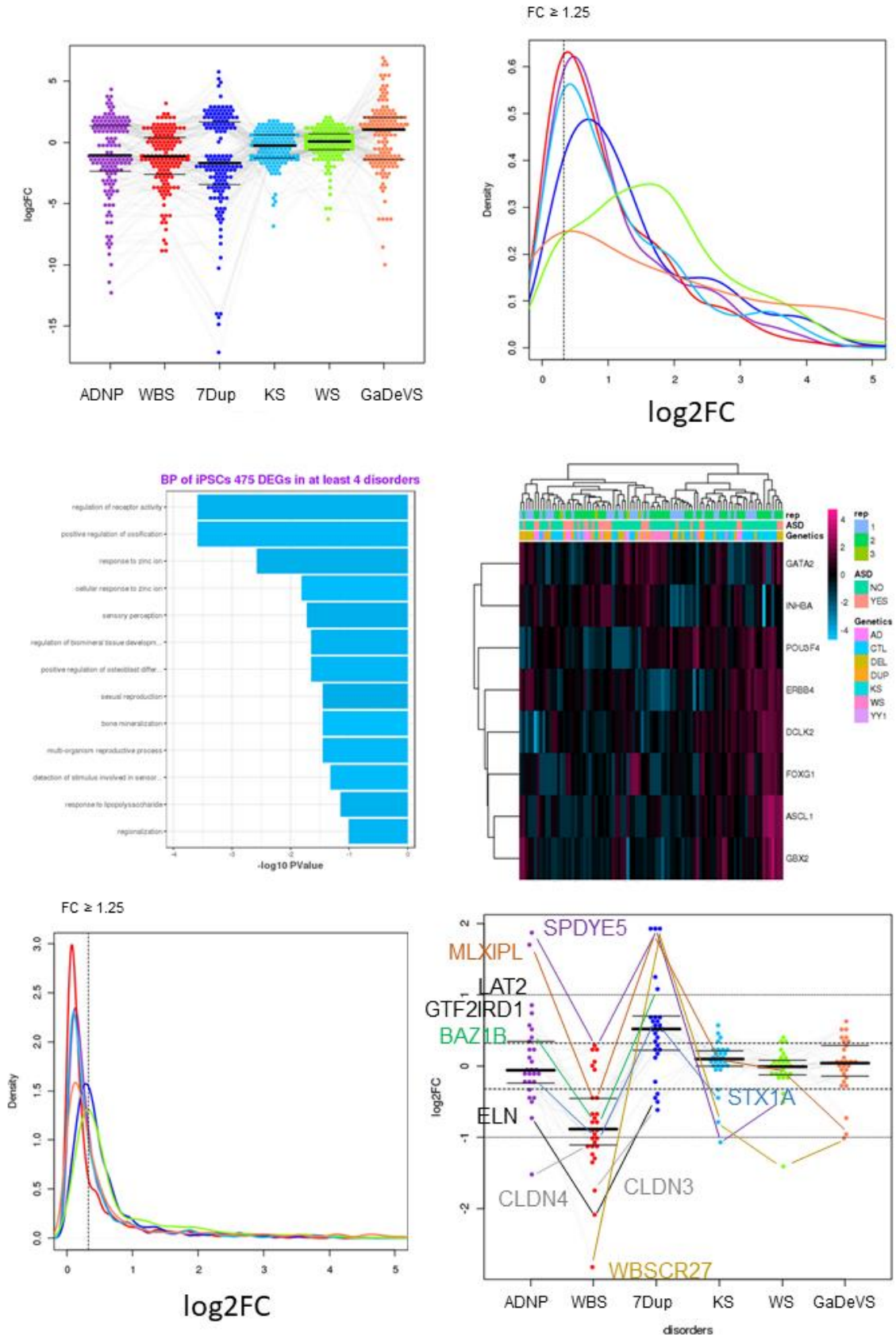


Figure 16. A) 151 genes differentially expressed in at least 4 disorders (Overlay of Boxplot and Beeswarm of logFC; disorder-specific fold-changes are respectively reported for ADNP-ASD, WBS, 7DupASD,KS,WS,GaDeVS); B) Barplot of GO enrichment for genes in section A; C) kernel density distribution of log2FC of genes with FDR < 0.05; D) kernel density distribution of log2FC of DEGs across disorders with FDR < 0.05 and meanFC > 1.5; E) heatmap of DEGs enriched for forebrain neuron differentiation; F) WBSCR genes differentially expressed across disorders

Genotype wise differential expression in iPSCs (correcting for ASD)

This analysis has been conducted following the same approach as the previous one, only changing the model matrix (~ASD+Genotype) and considering Genotype as the independent variable while dropping ASD. Remarkably, looking at the logFC distribution of genes passing FDR < 0.05 we observe that the three most affected iPSCs lines are those of GaDeVS, 7DupASD and ADNP-ASD (Figure 17A). In this case significantly enriched GO categories - “biological process”- account for anterior/posterior pattern specification, (embryonic) skeletal development, regionalisation (including genes for “neuron fate specification”), limb development, myoblast proliferation and many others (Figure 17B). The same list of DEGs show significant enrichments (p-adj < 0.001) among “molecular function” GO categories such as “RNA polymerase II transcription factor activity”, “double-stranded DNA binding”, “core promoter proximal region DNA binding” and similar. Genes enriching these categories were coarsely the same (Figure 17C) and in Cortecon database they account for few clusters, being important for pluripotency state, Neuronal differentiation, Cortical Specification, Deep Layers or Upper Layers. Notably, among them PAX7, RFX4, ZIC1, NR2F1 and NR2F2, ARID3C and LHX2 appear to be crucial along all layers and phases of cortical development. Genes enriched for skeletal development appear to be expressed also in the brain (Figure 17D). In this case I found 3 clusters that are expressed respectively i) in Upper Layers; ii) Neural differentiation and Cortical Specification; and iii) in pluripotency. Genes enriched for anterior/posterior specification (Figure 17E) do not cluster clearly, are mostly expressed in pluripotency and never expressed in Upper Layers. Intriguingly, four differentially expressed genes determine an enrichment for myoblast proliferation (Figure 17B, F). Among these genes is PAX7 which is included both in Muscle Cell

Differentiation and Neural Crest Differentiation categories and it is highly expressed in the frontal cortex (Figure 17G). ANKRD2 is associated with Dilated Cardiomyopathy and muscle stress response (in mice) (Belgrano et al., 2011; Nagueh et al., 2004), but it appears also to be highly expressed in the frontal cortex (Figure 16G). KCNA5 is a Potassium channel, highly expressed in Nucleus Accumbens and Hypothalamus (Figure 17G,H), and it has been associated with familial atrial fibrillation (Christophersen et al., 2013). IGF1 is an Insuline Like Growth Factor, generally important for growth and development, associated with Pituitary Gland Disease. Finally, PITX2 is a homeodomain protein, that acts as a transcription regulator associated to Axenfeld-Rieger Syndrome, which causes congenital malformations, and abnormal corona morphology (Seifi and Walter, 2018). All these genes, while being commonly associated to, and enriching for, myoblast differentiation and proliferation appear to be highly expressed in several brain regions. Moreover, as observed in the Cortecon database, while ANKRD2 appears to be important for pluripotent stages of brain development, and PAX7 generally present during Cortical Specification and Neuronal differentiation, IGF1, PITX2 and KCNA5, appear to be upregulated in latest stages of brain development (Figure 17I).

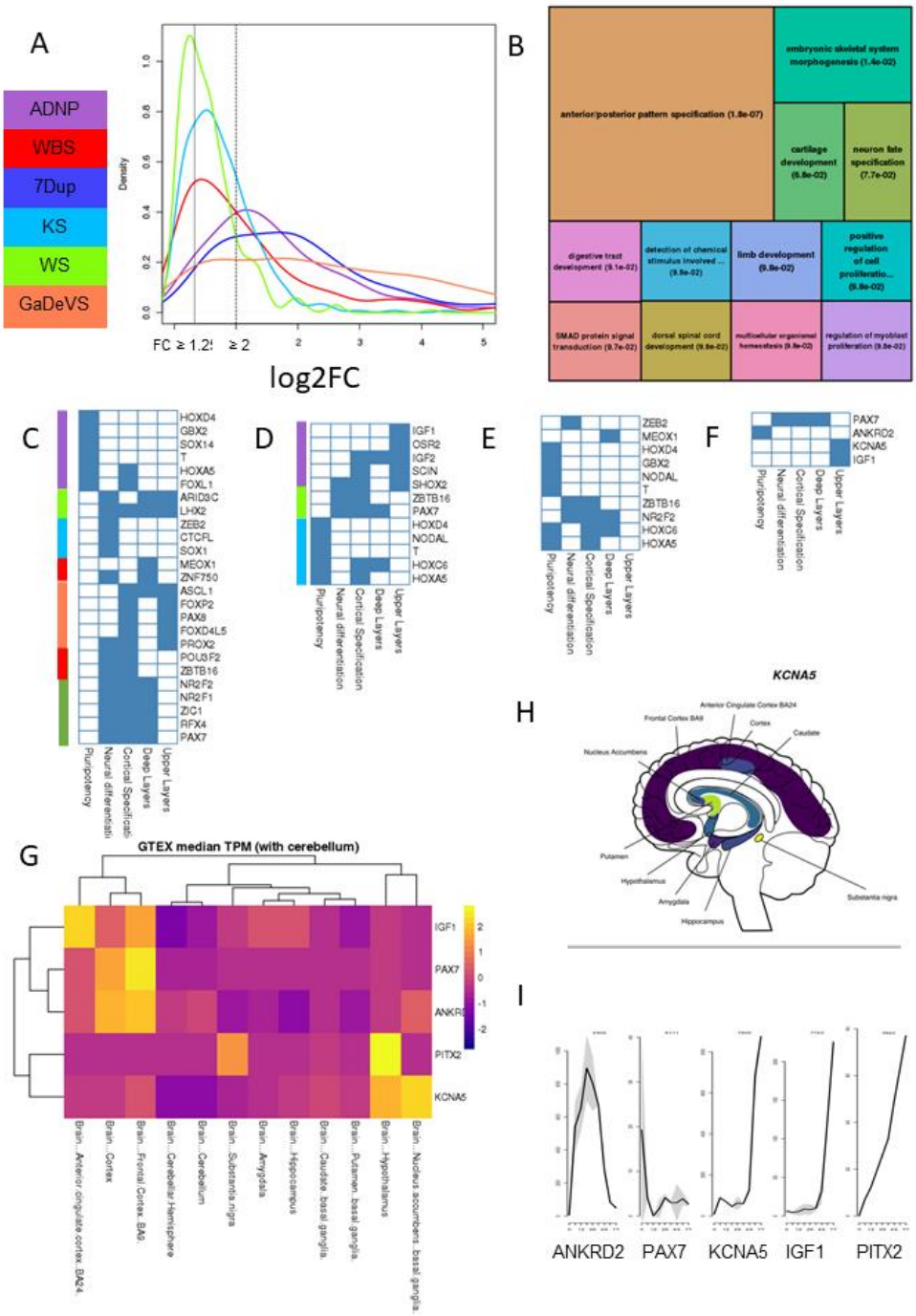


Figure 17. Genes differentially expressed across genotypes, correcting for ASD: A) show higher fold-changes in 7DupASD, ADNP-ASD and GaDeVS; B) are enriched for disease-relevant categories. C) differentially expressed transcription regulation genes appear to be expressed in several moment of brain development; D) Genes enriched for skeletal development are selectively expressed either at pluripotency or in differentiation and specification or in brain upper layers; E) genes important for anterior/posterior specification and initial embryo development are mostly expressed in pluripotency and never expressed in upper layers (as expected); F) Myoblast differentiation genes are actually expressed in late phases of brain development and are G) highly expressed in several districts of the brain. H) KCNA5 is highly expressed in hypothalamus and Nucleus Accumbens; I) several genes associated with myoblasts proliferation appear to be upregulated in final stages of brain development.

ASD specific transcriptional deregulation in iPSCs

The question of whether a subset of genes is differentially expressed in ASD patients iPSCs versus non-ASD patient iPSCs can be posed to the same dataset in few ways. Here I show the outcome of three different analyses.

The first analysis has been conducted by correcting for genotype (~Genotype+ASD) including controls in our 91 clones-wise dataset.

Only 174 genes were differentially with FDR < 0.25 and FC higher than 1.5 Figure 18A. Using these thresholds these genes show a (very mild) GO enrichment. GO categories enriched downstream of this analysis refer to tissue and cellular muscle structures and function (Figure 18A,B) generated by 16 genes, mostly upregulated in ASD patients iPSCs and all showing FC > 2. Four of these genes appear to be effectively expressed in the brain (Figure 18C). Among them DCLK2 is a doublecortin CAM kinase, important for calcium regulation. Its family of protein has been linked with lissencephaly and includes proteins expressed in the hippocampus. DCLK2, querying GTEx data, appears to be highly expressed in all cortex and, from our analysis, downregulated in ASD iPSCs (Figure 18B, C). DNM3 is a GTP-binding protein, highly expressed in frontal cortex (Figure 18C), associated to Optic Atrophy and behavioural/neurological phenotypes (MGI, http://www.informatics.jax.org, October 2018). MYL9 is a component of myosin light chain, up to the moment of this thesis writing it has never been associated to any disorder, and it appears to be highly expressed in substantia nigra (Figure 18C). ANXA1 appear to be expressed mostly in the hippocampus (Figure 18C) and the only association found was with Brain edema in MalaCards (Rappaport et al., 2013).

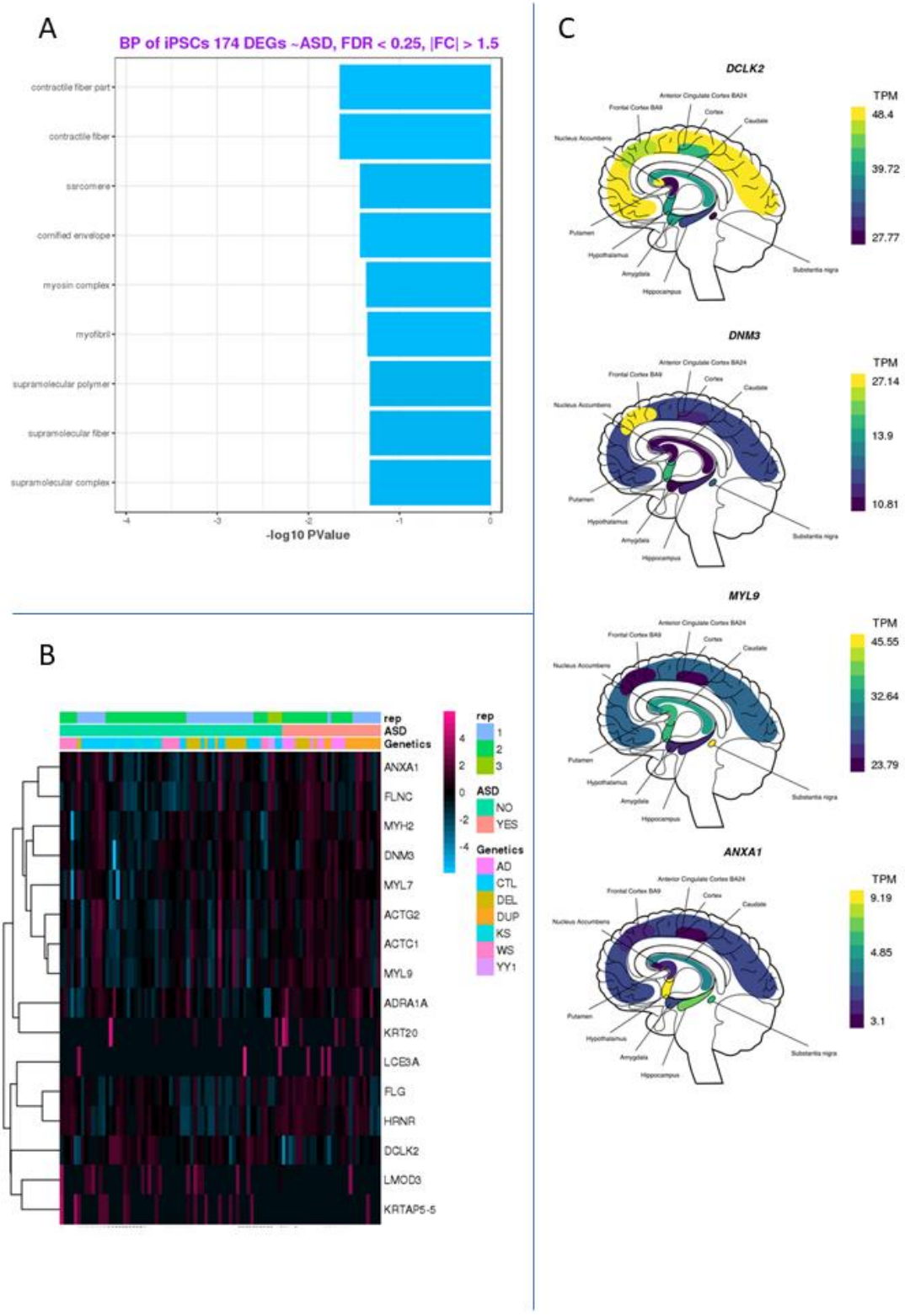


Figure 18. ASD related DEGs across genotype A) GO enrichments B) heatmap of genes enriched in GO categories described in A) (z-scores of log-TMM read-counts); C) Brain maps of 4 out of 16 genes enriching for muscle-related GO categories effectively expressed in brain (GTEx)

DEGs identified in this first analysis show high intra-genotype variability also in control samples (Figure 18B) and, apparently, marginally relevant results: a strict observer could deem them as potentially false-positives.

As observed in recent literature (Germain et al., 2017), to i) reduce the imputation of spurious differences across condition (ASD) or genotype introduced by the variable number of iPSC clones per individual and ii) establish a comparison in which the number of ASD- and nonASD- samples is more equal, I resorted to aggregation of individuals data (summing reads of clones from the same individual). The second and third analysis has thus been conducted on the aggregated data. In the second analysis control lines are included and I introduce a correction by genotype (~genotype+ASD), in the third one I excluded both control samples and blocking (~ASD).

When including controls and blocking for genotype I expect to to highlight differences between ASD and non-ASD samples partially pruned of genotype-wise differences. What I observed, plotting the heatmap of z-score (measured on the log-normalised read-counts) of genes differentially expressed respecting the ~genotype+ASD model, is a dominance of aWBS transcriptional signature (Figure 19A). This is due to aWBS and WBS constituting the only genotype group including both ASD and non-ASD samples. Indeed, almost all 248 genes passing FDR < 0.05 and FC > 1.5 are either down-regulated in aWBS and up- in the other disorders (with internal differences) or up-regulated in aWBS and mildly up- in most WS samples and in 7DupASD4 (Figure 19A). These genes are enriched in GO biological process categories such as “chromatin silencing”, “nucleosome assembly” and other connected with transcription regulation (Figure 19B). Among them, to name a few, I found several histone proteins, *ATN1, EP300, ARID1A, FOXO3, CHD6, CHD7,*

NCOR1 and few zinc finger proteins. *ATN1* is reported in OMIM to cause dentatorubral-pallidoluysian atrophy (DRPLA), which is a syndrome characterised by epilepsy, dementia, ataxia and choreoathetosis, with onset around twenty years of age, and death in the forties.

EP300 is the main H3K27 acetylase, responsible of enhancer activation and included in SFARI genes, since it causes Rubinstein-Taybi Syndrome 2. ARID1A is another SFARI gene (causing Coffin-Siris Syndrome), which is a chromatin remodeller, included in the SWI/SNF family, having both ATPase and helicase function. FOXO3 is a forkhead box transcription factor which regulates apoptosis. CHD6 is also a chromatin remodeller with helicase activity further presenting two chromodomain, which is associated in MGI with behaviour/neurological phenotypes, since it has been associated to motor coordination (impaired rotorod performances) in mouse models (Lathrop et al., 2010). CHD7 is part of the same family of CHD7 and its mutations have been associated with CHARGE syndrome. NCOR1 is a nuclear receptor that inhibit transcription by favouring chromatin condensation and has been implicated in the regulation of cortical progenitor self-renewal in mouse embryos (Jepsen et al., 2007). Plotting fold-changes of these DEGs as measured across genotypes in our previous genotype-wise analysis, reveal that most of them are not simply deregulated in a genotype-specific fashion, but they mostly impact 7DupASD and ADNP-ASD samples (Figure 19C). aWBS specific fold-changes are masked in this comparison because they were included in WBS samples. Moreover, among differentially expressed genes associated with ASD I found some being up- or down-regulated with very similar FC in four disorders (WBS, 7DupASD, ADNP-ASD and GaDeVS) (Figure 19C). *ANKRD1, NAP1L1, NKAPL, PEG3, DDX43* are all involved in DNA-binding, chromatin remodelling and transcriptional regulation. *NPPB* and *IGFBP5* are involved in growth and signaling; *DPP6* is a serine protease

which binds voltage-gated potassium and is associated in MalaCards to mental retardation and microcephaly.

Intriguingly, among genes up-regulated both in 7DupASD and ADNP-ASD are four SFARI genes, namely *ARID1A, CHD7, EP300* and *KMT2A*: all these genes are down-regulated in aWBS. Finally, *CHD7* is also down-regulated in WS and up-regulated in GaDeVS, which also show up-regulation of *KMT2A*.

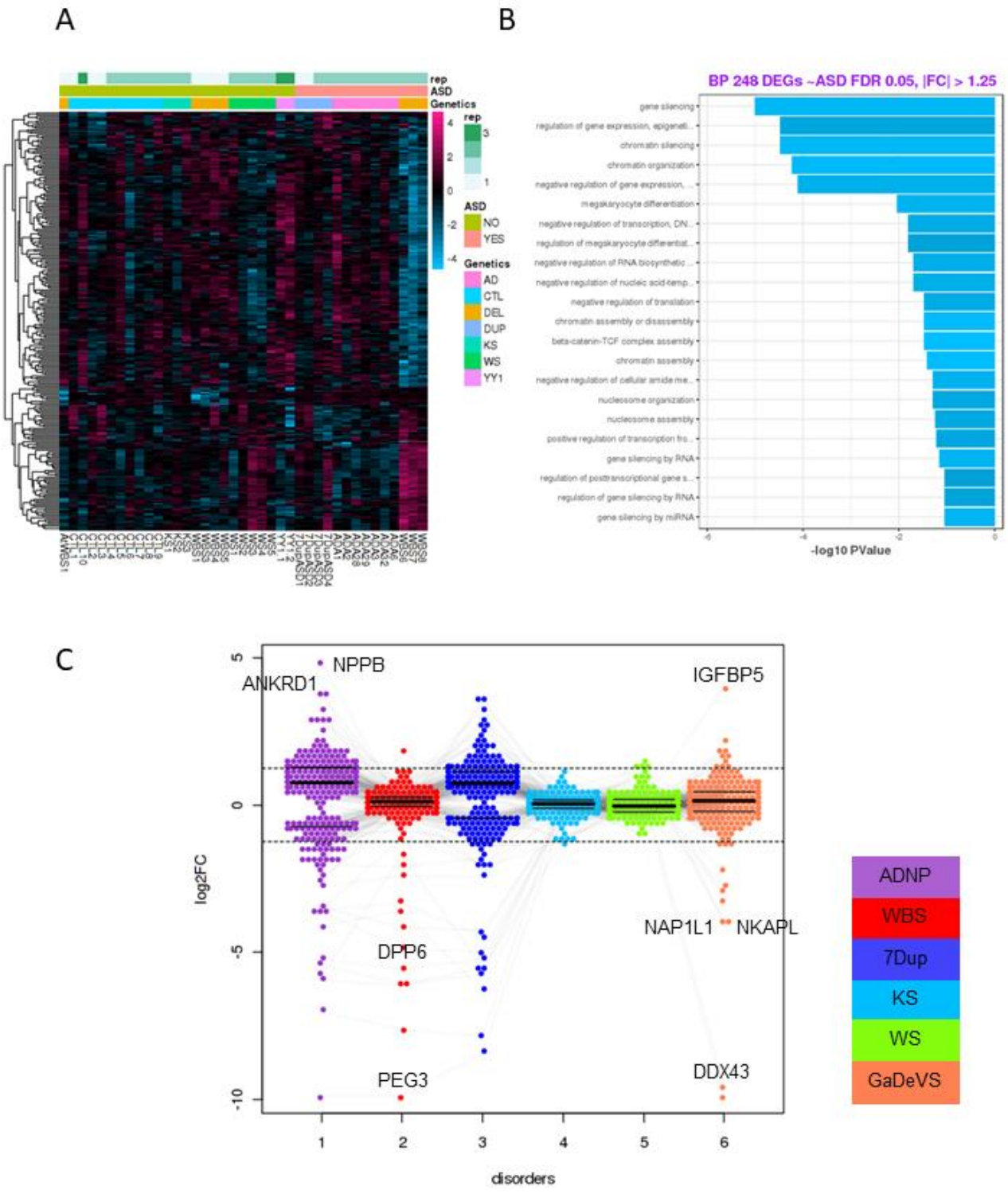


Figure 19. Blocking for genotype and testing for ASD in iPSCs are found A) a large set of genes inversely expressed between aWBS and 7DupASD/ADNP-ASD (z-score log-TMM read-counts); B) GO enrichments referring to transcriptional/chromatin silencing and regulation; C) Several genes with common FC across disorders, excluding mainly KS and WS, reported over a barplot of FC ~Genotype. Given the type of analysis, in this plot aWBS and WBS are averaged.

In my last attempt to identify genes specifically differentially expressed only in ASD-patient-derived iPSCs I resorted to one of the ideas that conceptually funded this thesis. Control samples represent a minoritarian reference subset in a cohort of conditions showing shades of common phenotypic traits originated from plainly distinct genetic origins. What I achieved, by removing control samples, is to remove an outlier and focusing on the only real difference among these disorders: some acquired ASD traits and some did not. Moreover, I excluded from the model matrix the peculiar WBS genotype feature of combining in one group both ASD and non-ASD samples. Doing so a dysregulation that is common among all ASD individuals (Figure 20A) was revealed. This analysis identified 17 genes passing FDR 0.05 with FC > 1.5. Using shallower thresholds, as done in the previous analyses, I could observe 49 genes with FDR < 0.1 and FC > 1.25 (Figure 20A). Overlaying this list of genes on Cortecon data, to identify their spatio-temporal expression in the brain along development, I observed that 35 are expressed in the brain (Figure 20B), 11 expressed in pluripotency, 15 in Neuronal Differentiation and 17 in Cortical Specification. 17 of these genes are expressed in Deep Layers and 15 in Upper Layers. Crossing genes differentially expressed with p-value < 0.05 and the SFARI database I found (in the ASD condition): AUTS2 up-regulated (FDR 0.16), KDM6A down-regulated (FDR 0.06) and SOX11 up-regulated (FDR 0.28) (Figure 20C). Using a common p-value threshold of 0.005 I could identify 254 genes. Among these genes we can find a GO enrichment for “regulation of neurogenesis” (FDR < 0.1), including 27 genes (Figure 20D). Finally, among genes with FDR < 0.25 and FC > 1.5 I found many solute carrier genes (Figure 20E).

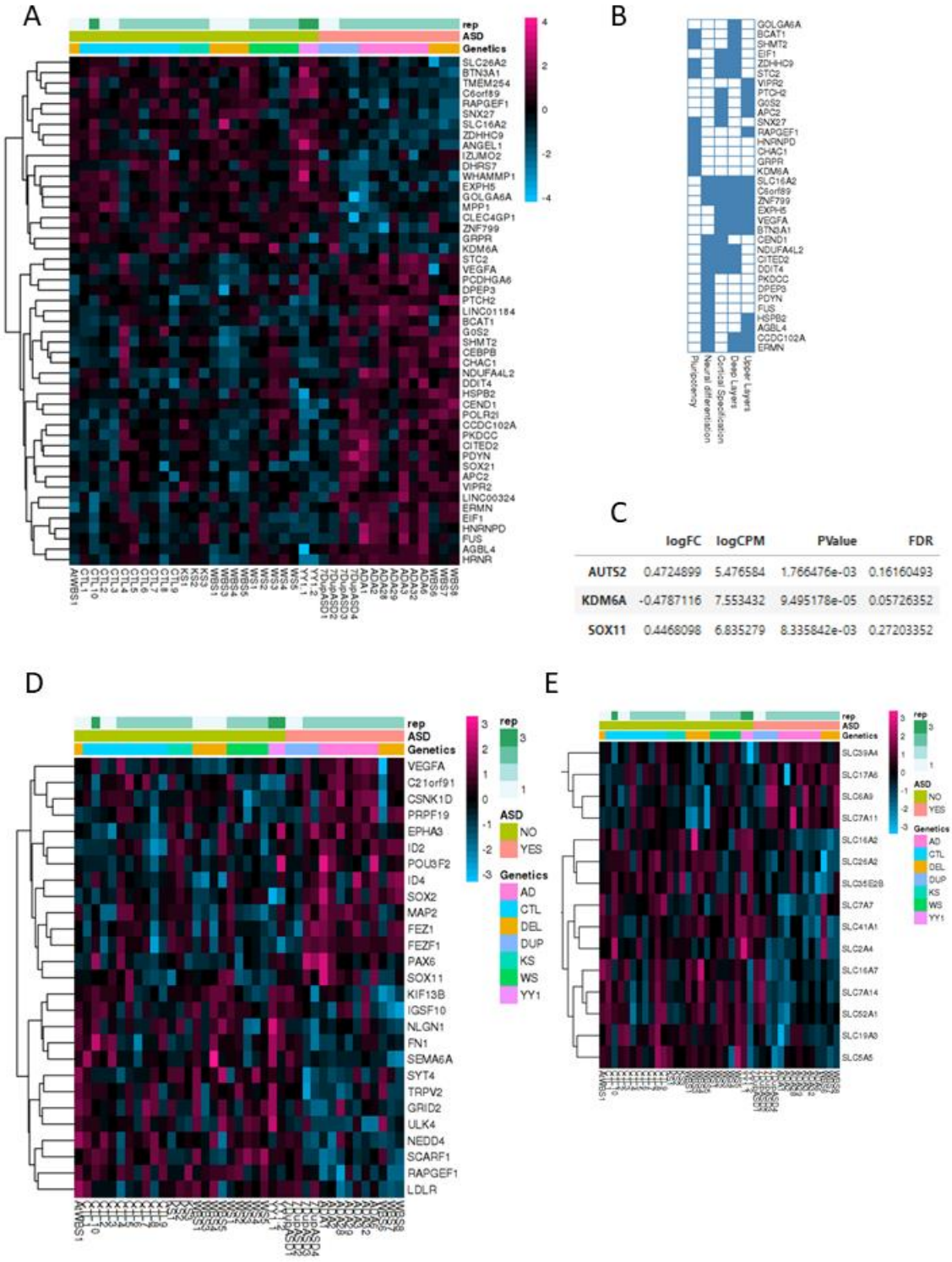


Figure 20. Selective dysregulation across ASD iPSCs. A) Heatmap of genes differentially expressed with FDR < 0.1 and FC > 1.25; B) same genes in A) overlaid on Cortecon data to identify their spatio-temporal expression in the brain along development. C) SFARI genes differentially expressed in ASD lines; D) Heatmap of genes (p < 0.01) significantly enriched for "regulation of neurogenesis"; E) Heatmap of SLC genes differentially expressed in ASD lines. In all heatmaps unit measure is z-score of (log-TMM read-counts)

Identification of modules of differentially expressed genes across disorders in iPSCs.

All analysis -testing for genotype- conducted until here serve the scope of identifying genes differentially expressed across disorders, increasing statistical power by testing several coefficients at once. This approach has served for the identification of genes differentially expressed in a similar manner in groups of disorders. The lists produced up to this phase were then further mined to identify shared modules of deregulation. Among them I concentrated my work on the recognition of genes i) differentially expressed in similar ways among subgroups of disorders or ii) keeping a certain ratio between disorders. Clustering of log2-fold-change values of expression can be set up with this scope.

In order to briefly verify the robustness of analyses conducted with or without aggregating clones coming from the same individual I observed the number of genes identified as DEGs in our ~genotype analyses in both situations. The vast majority of genes passing FDR threshold 0.05 and 0.005 in the analysis with 91 samples are included in both analyses, with the one with aggregated clones showing four times more genes, in terms of sheer number of significant DEGs (FC > 1.5 and same FDR thresholds). Using k-means clustering (as briefly described in the methods) I identified two k values (4 and 10) clustering coherently ~4 thousands differentially expressed genes.

Genes differentially expressed in iPSCs as divided in four subsets by k-means clustering

Using k=4 I identified one cluster of 30 genes (“Kcluster1of4”), including *NKAPL, DPP6, NKAPL, PEG3, DDX43* and few zinc fingers identified as up- or down-regulated with very similar fold-change in WBS, 7DupASD, ADNP-ASD and GaDeVS. This analysis corroborates the observation made previously and gives a first group of genes potentially working in complex (Figure 21).

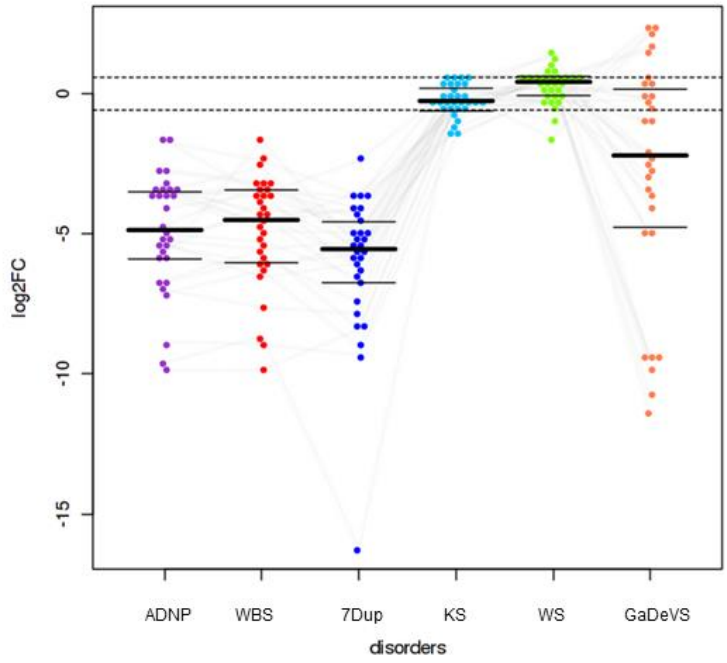


Figure 21. K-means Cluster 1 of 4. 30 Genes out of 4K have a specific set of fold-changes across 4 disorders.

Cluster two of four (“Kcluster2of4””) depict a vast set of genes (~2 thousand), up-regulated in ADNP-ASD, 7DupASD and GaDeVS (Figure 22A). These genes are enriched for GO biological process such as neuron and neuron projection development, limb and sensory organ morphogenesis, ear development (FDR < 0.005) (Figure 22B). Looking closely at the expression levels of these genes, I found

a strong difference between aWBS and WBS samples (Figure 22C). Z-score measured on log-normalized read-counts show i) WBS and aWBS to have a small set of genes down-regulated in (highlighted in Figure 22C with a white square), and to be much more similar to ADNP-ASD, 7DupASD and GaDeVS.

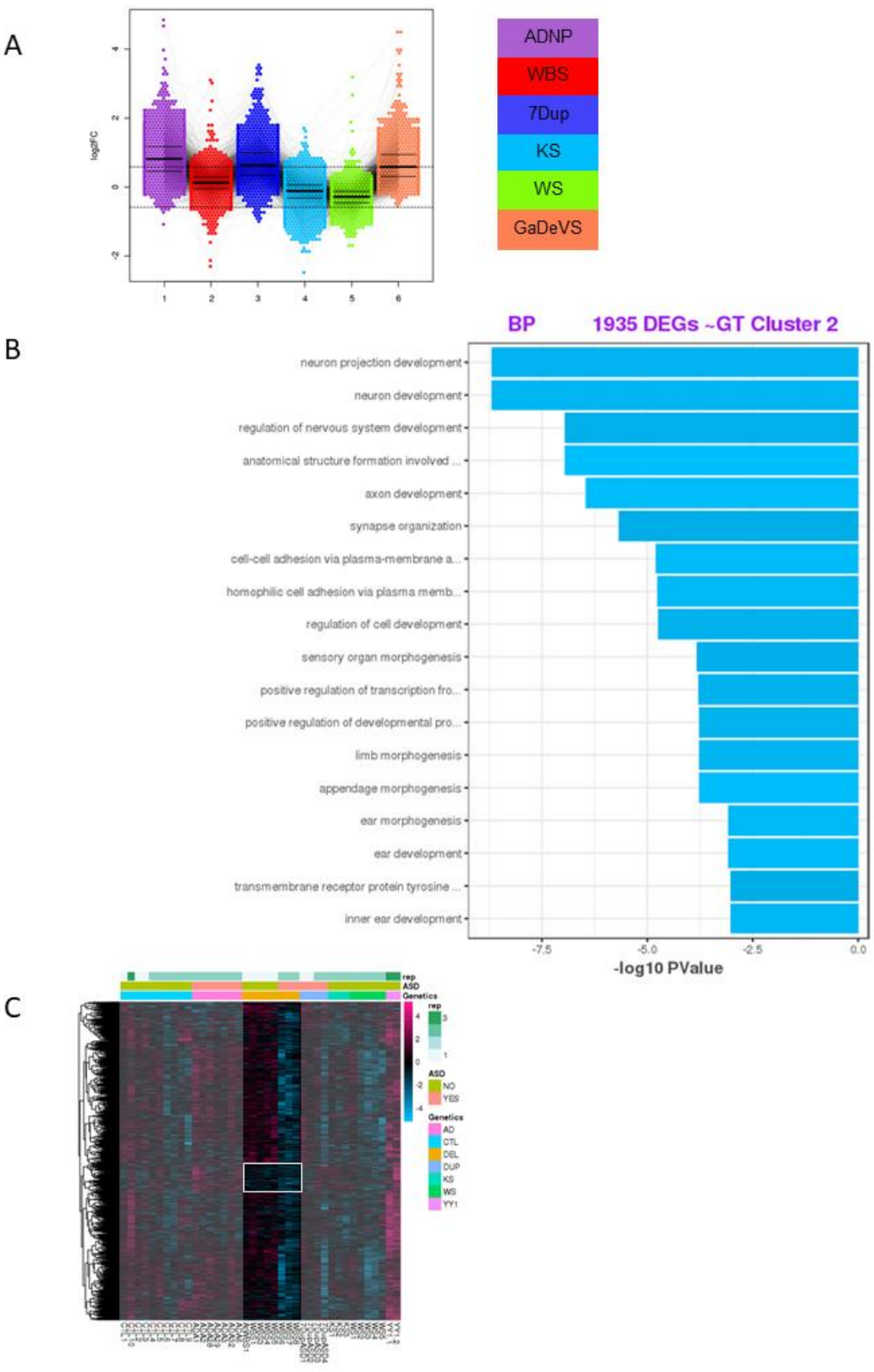


Figure 22. K-means Cluster 2 of 4 of genes differentially expressed ~Genotype. A) beeboplot of logFC in the different disorders measured in iPSCs; B) GO enrichments of Cluster 2 genes; C) The vast majority of Cluster 2 genes are opposedly expressed between aWBS and WBS.

The third cluster (“Kcluster3of4”) obtained with this analysis show 724 genes up-regulated in ADNP-ASD,7DupASD, KS and down-regulated in WBS, WS and GaDeVS (Figure 23A). These genes are enriched for several GO biological process (BP) categories (Figure 23B). Among them I found “positive regulation of transport”, migration genes enriched for “regulation of locomotion”, and many other enriching liver- and immune-system associated categories, which will require further investigation. Among molecular functions (MF) GO categories I found only “mRNA binding” (FDR 0.04), which enriching genes are reported in figure (Figure 23C).

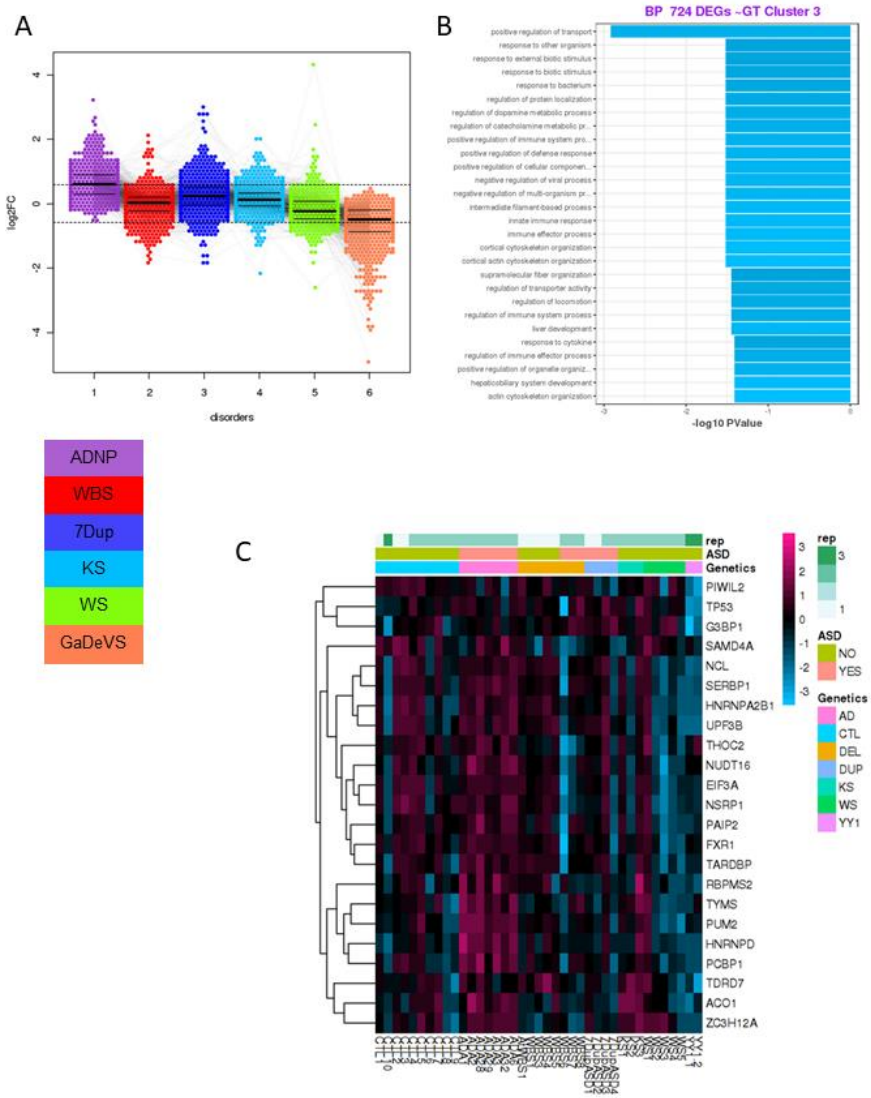


Figure 23. Characterization of cluster 3 of 4; A) boxplot over beeswarm of logFC across different disorders; B) GO enrichments in Biological Process categories; C) z-score Heatmap of log-norm-read counts for genes involved in mRNA binding included in cluster 3.

Also the fourth cluster identified (“Kcluster4of4”) accounts for most DEGs (2079). These genes are generally down-regulated in ADNP-ASD, WBS, 7DupASD, and mostly in GaDeVS(A). Half of them are up-regulated in KS (A). Intriguingly, among these genes are 5 SFARI genes and 9 WBSCR genes. The former group is up-regulated in aWBS, 7DupASD, and WS (B). The latter is down-regulated both in aWBS and WBS, up-regulated in 7DupASD, and up-regulated in most WS, KS and GaDeVS samples (C). Finally, among the whole cluster several GO categories involved in translation are enriched, with “translation” category accounting for 90 genes and FDR < 0.001.

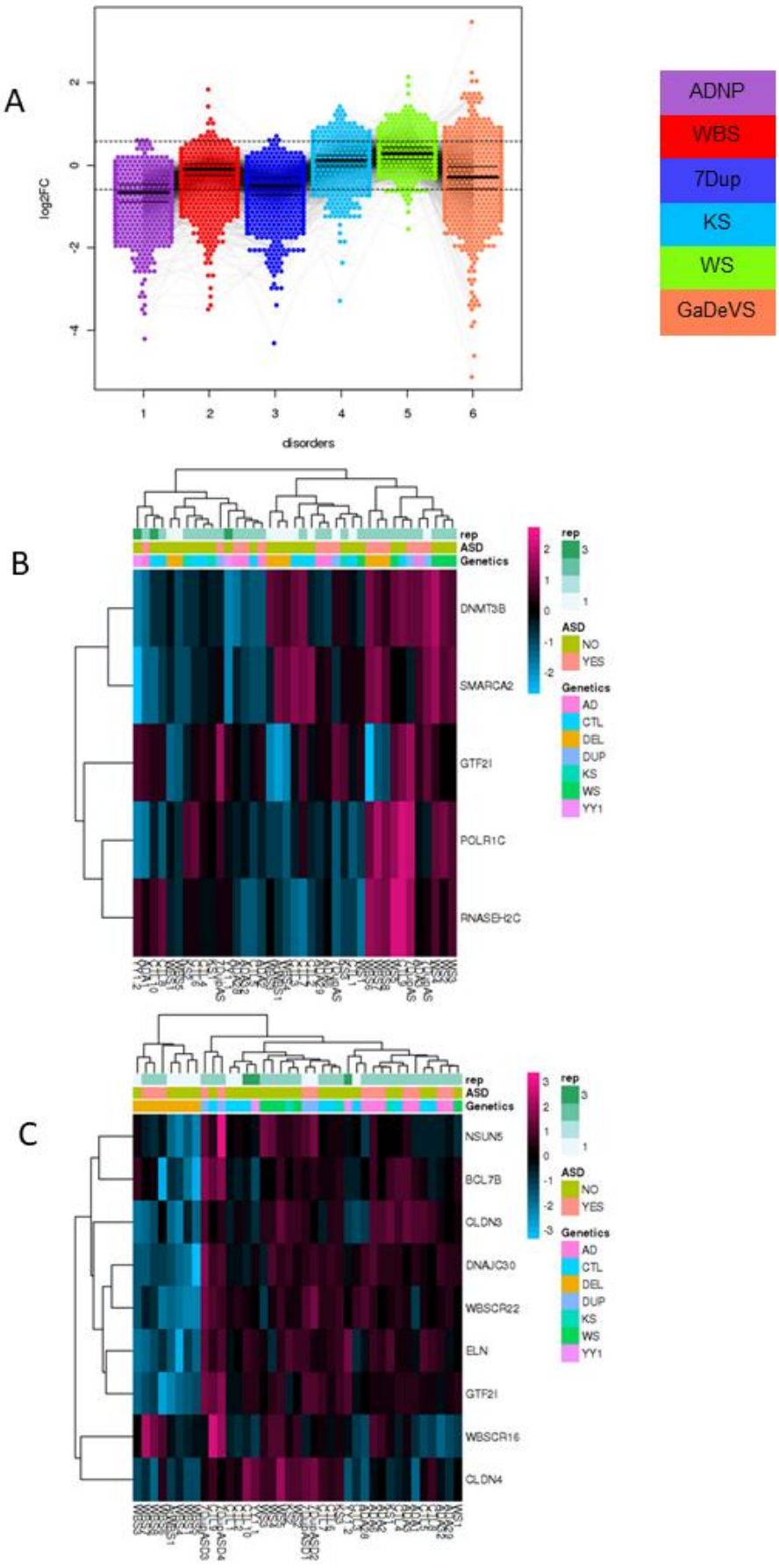


Figure 24. Cluster 4 includes A) genes mostly down-regulated in ADNP-ASD, WBS, 7DupASD and GaDeVS; B) SFARI genes and C) WBSCR genes

Genes differentially expressed in iPSCs as divided in ten subsets by k-means clustering

When the same genes described in the previous chapter are divided into 10 clusters, I found one (Kcluster1of10) of 929 genes entirely included in Kcluster4of4 described in the previous analysis (Figure 24). The same fold-change ratio across genotypes is observed (Figure 24A,Figure 25A)- Moreover, previously observed GO enrichments are confirmed (Figure 25B). Among this shorter list of genes other enrichments become significant, such as HPO and MGI (obtained with Enrichr) (Figure 25C). 24 of these genes are, in fact, enriched for “sever global developmental delay” and “short nose” in HPO, “decreased cranium height” and “abnormal fontanel formation” in MGI.

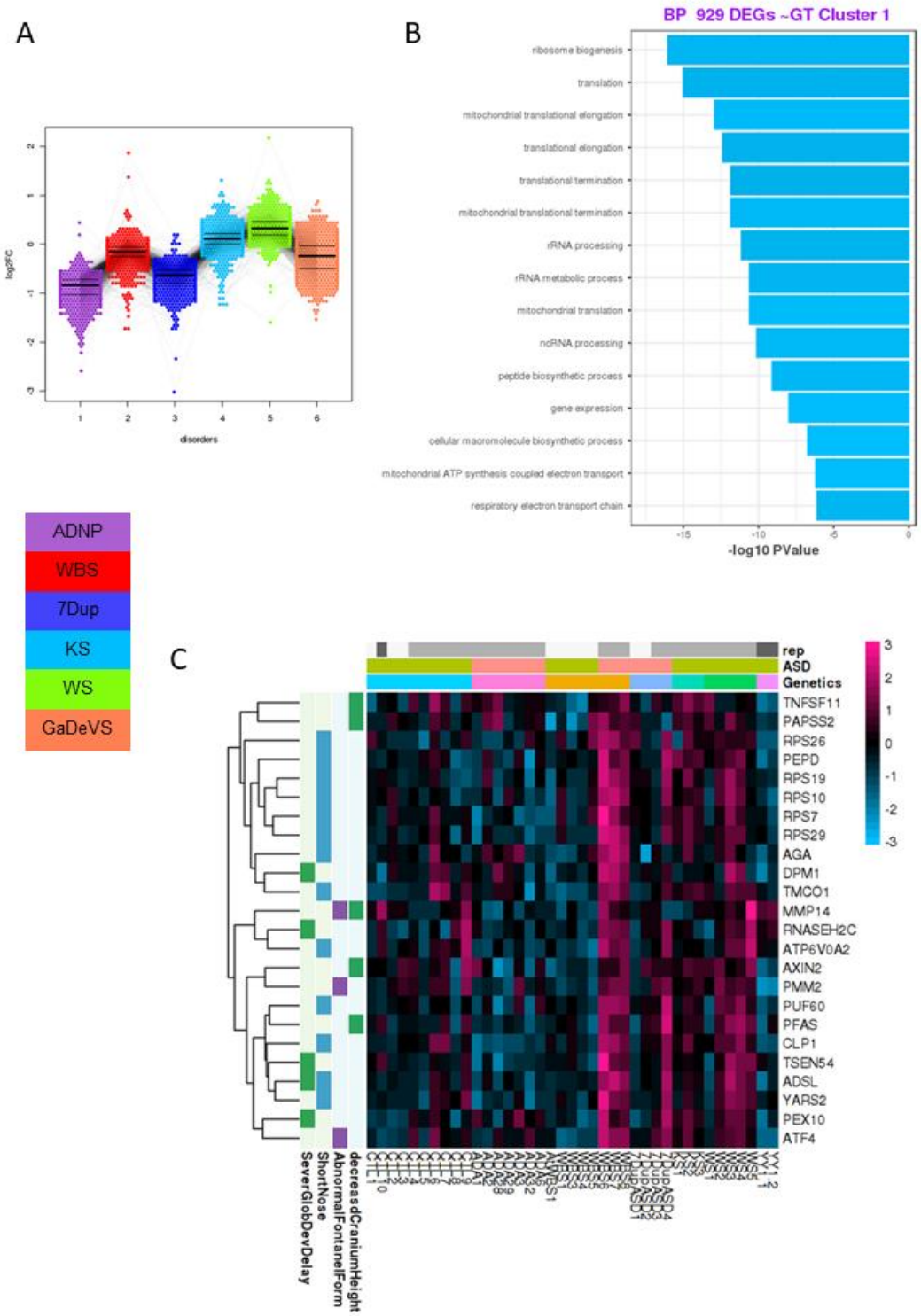


Figure 25. Characterization of Kcluster1of10. A) Being included in Kcluster4of4, it shoes the same fold-change ratio across disorders; B) GO enrichment already observed in Kcluster4of4 are confirmed and even more significant. C) Heatmap of genes enriched for relevant HPO and MGI categories (z-score of log-norm-read counts).

Following the established nomenclature for k-means clusters described until here, KCluster5of10 and KCluster7of10 show the same fold-change ratio of KCluster1of10 (Figure 25), accounting respectively for 99 and 896 genes. Notably, the range of fold-change is wider for KCluster5of10 (Figure 26). In fact, besides showing similar fold-change trends across disorders, k-means clustering has helped further dissection of a larger gene list into functionally interesting groups. GO enrichments of KCluster5of10 refer to “chromatin silencing”, “gene silencing” and its regulation, but also “regulation of ossification” (Figure 27,Figure 28), while GO enrichments of KCluster7of10 account for “Wnt signalling pathway” and related categories (Figure 27,Figure 28). Finally, KCluster8of10 includes the same genes of KCluster1of4, corroborating its uniqueness, probably given by sheer extreme absolute fold-changes. At the same time, the ratio between mean-fold-change across disorder is very similar also to KCluster4of4.

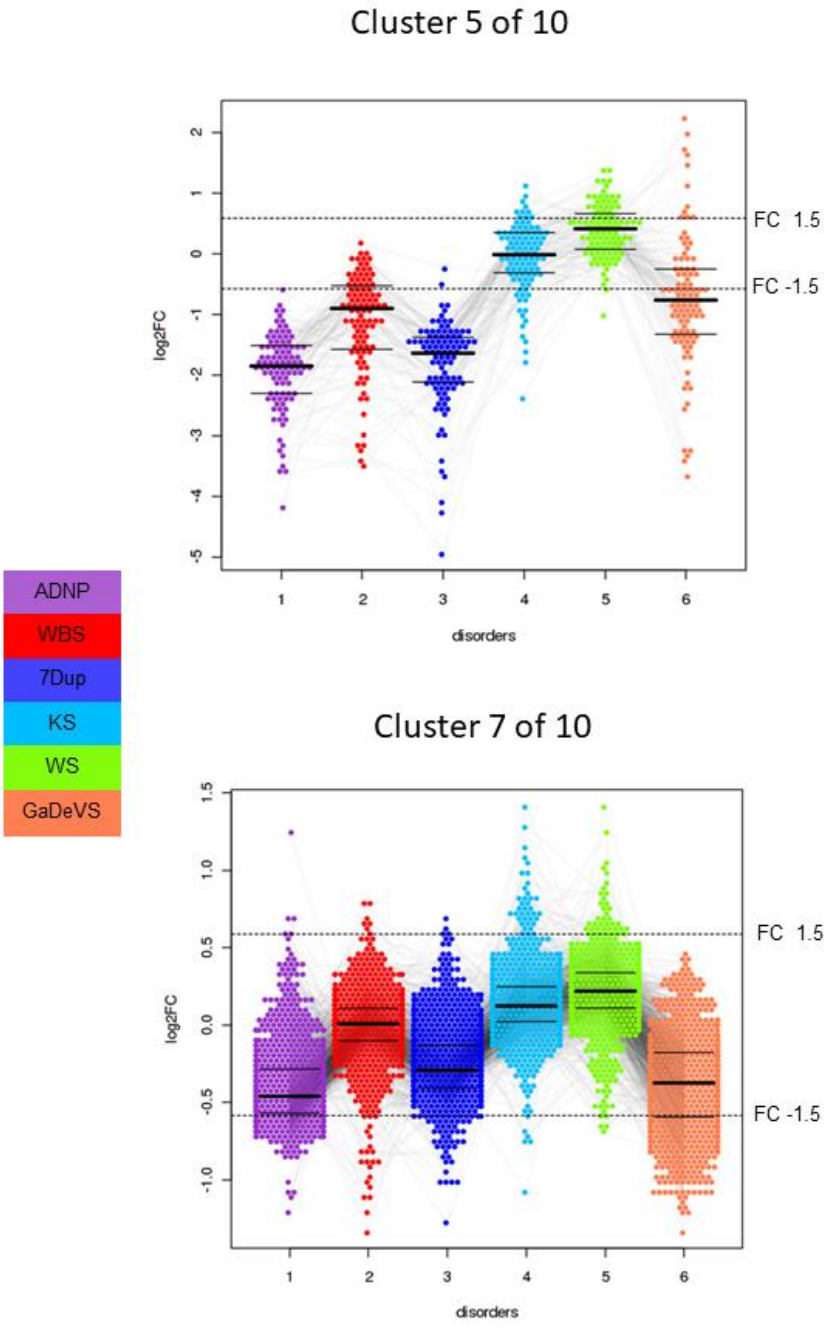


Figure 26. Kcluster5of10 and KCluster7of10 show similar fold-change ratio across disorders like Kcluster1of10 and Kcluster4of4

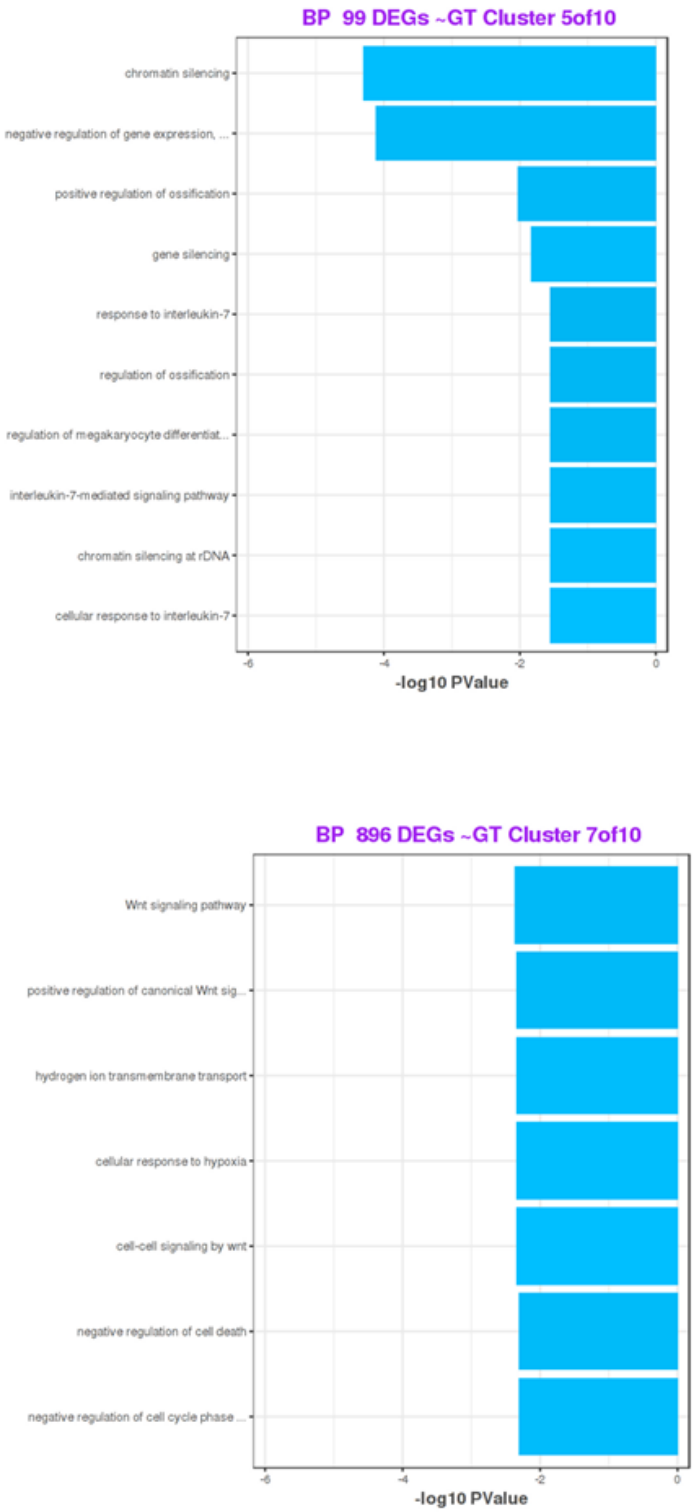


Figure 27. GO enrichments distinguish different biological functions associated with Cluster 5 and 7

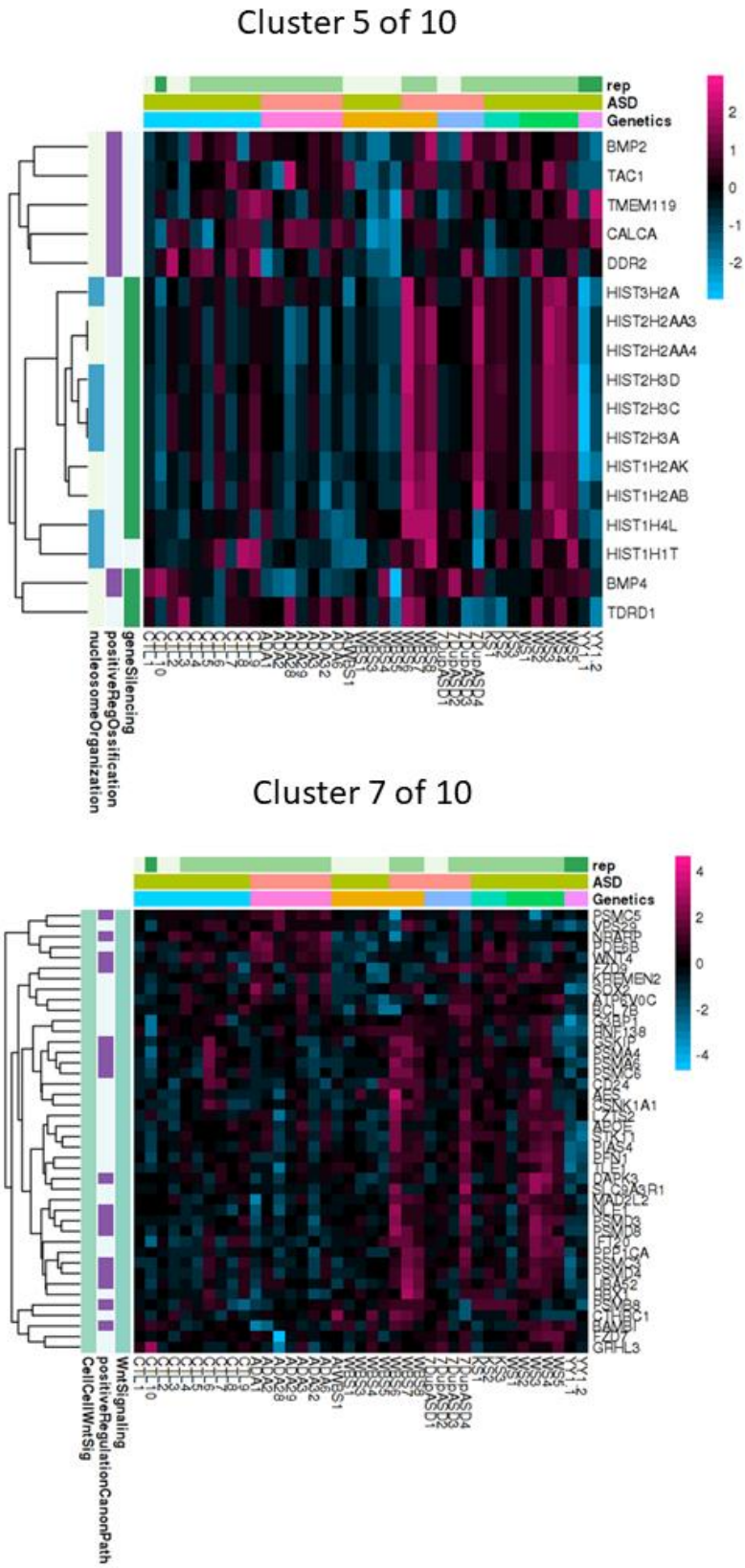


Figure 28. Heatmaps of genes enriched in relevant GO categories (z-score of log-normalized read-counts).

The observation of relative fold-change distributions in genes included in KCluster2of10 but also in the third and ninth of the same analysis, confirms that, by increasing the number of clusters fed to k-means algorithm, we are effectively able to dissect genes lists generally following common trends across disorders while being distinct in FC ranges. Indeed these 3 clusters share same FC across disorders as observed in KCluster2of4 while enriching separately same GO categories. This observation is coupled with the finding that each cluster shows unique GO enrichments that were not exposed by KCluster2of4. Moreover, this advanced clustering strategy show an alternative separation of crucial genes (such as SFARI and WBSCR ones), which were previously separated and now gathered in higher number inside smaller clusters. In this case, 18 SFARI and 4 WBSCR genes are enriched in KCluster3of10. All these genes are upregulated in ADNP-ASD, and 7DupASD. Only *BAZ1B, GTF2IRD1, LAT2, LIMK1* are differentially expressed in WBS (down-regulated). *KMT2D* and *ZNF711* appear downregulated and *ARID1B* up-regulated in KS (all with FC > 1.3 and FDR < 0.05). *ARID1B, ADNP, PHF6, GTF2IRD1, KMT2D, TAF1, RPS6KA3, ATRX, CHD7* and *NBEA* are all down-regulated in WS (all with FC > 1.5 and FDR < 0.05). No WBSRC genes are deregulated in WS. *NBEA, CHD7, RPS6KA3, MECP2, KMT2D, ZNF711, ADNP,* and *HIP1* are all up-regulated in GaDeVS (with FC > 1.5 and FDR < 0.05).

KCluster9of10, on the other hand, include *ZBTB20, CHD2, KMT2A, MBD5, KMT2C, ARID1A, MLXIPL*, and *SPDYE5*, all strongly up-regulated in 7DupASD and ADNP (FC > 2.5). *MDB5, KMT2C* and *ARID1A* appear slightly up-regulated also in WBS (FC > 1.25) while MLXIPL is down-regulated as expected (FC < -1.5). Only *KMT2A* and *SPDYE5* are down-regulated in KS and no other of these genes are affected. In WS *ZBTB20, KMT2C, SPYDE5* and *CHD2* are down-regulated. GaDeVS iPSCs show up-regulation of *ZBTB20, CHD2, KMT2A, KMT2C,* and *MBD5*.

KCluster4of10 has the same trend of KCluster3of4 and 424 of its 460 genes are indeed included in KCluster3of4. Among the 36 spared genes I could find genes (with FC ~3) associated with chromatin binding and modification such as *KLF6, SETD7, BAZ2A, KDM4B, SHANK3,* and *PCGF5*, but also ion channel such as *KCTD12* and *SLC5A12*, and other genes associated with intellectual disability such as *AFF1*, and *ATXN2*. Among these genes I could also find *WWTR1*, which is involved in Hippo signalling and favours PAX8 dependent gene activation.

KCluster6of10 and KCluster10of10 are unique to this analysis. In facts, they are mostly enriched for genes differentially expressed only in GaDeVS. Indeed, the former (A) includes genes upregulated in GaDeVS and the latter down-regulated ones (B). Out of 424 genes in KCluster6of10, 289 have a fold-change higher than 1.5 (FDR < 0.05). Apparently, these genes are not enriched for GO categories, but I could identify an enrichment in master regulators (with a transcription factor enricher made by PL, see Methods). Most genes (> 150) are enriched for TAF1, RBBP5, TBP and POL2, (with FDR < 0.005). Among these genes, 22 have a fold-change higher than 3, including: USP6 (involved in Ubiquitin dependent proteolysis), PAX8, AMER2 (involved in the regulation of neuroectodermal patterning via Wnt Signaling pathway), LRRN3, OTOF (indicated by MalaCards to be involved in Deafness), MAPK10, LIX1, which has been proposed to have an important role for motor neuron maintenance (Fyfe et al., 2006; Moeller et al., 2002), HMX2, SYNDIG1L (down-regulated in animal models of Huntington disease de Chaldée et al., 2006) and HTR2C (associated with Anxiety in MalaCards). Remaining genes were all pseudogenes and long-noncoding RNAs which I did not discuss, since our RNA-seq libraries were not designed to quantitatively recognize and measure small transcripts. KCluster10of10 does not show any GO enrichments. TF enrichments

show a prevalence of *EZH2*, *RAD21* and *CTCF* targets in this list, enriched respectively by 29, 140 and 144 genes, always with an enrichment FDR < 0.005. Comparing these lists with OMIM, and querying for “intellectual disability” causing genes I identified *PIGO*, *TBC* and *PIGL* among up-regulated genes and *TAF13*, *SMARCA2*, *EEF1A2* and *ADAT3* among down-regulated ones.

Genotype specific differential expression in iPSCs

All analyses conducted until now have tried to capture genotype specific or ASD related transcriptional deregulations considering each genotype/condition as independent from the others. In this last analysis on iPSCs only data I tested each genotype against all the others, to identify bona fide unique sets of dysregulations. Indeed, 7DupASD, GaDeVS, KS and WS showed barely detectable differential expression with, respectively, 8, 4, 0 and 1 genes passing FDR < 0.05 and FC > 1.25. Out of these genes, in the original multifactorial analysis, using the same thresholds, edgeR identified 7 in 7DupASD, 2 in GaDeVS and in 1 KS. Now, using a non-corrected p-value threshold of 0.005, edgeR detects 103, 59, 80 and 179 genes differentially expressed in a genotype-specific manner. Of these genes, 66, 31, 66 and 39 were found with FDR < 0.05 in the multifactorial analysis. Considering that the same dataset was used, and the list of expressed genes has been used as a reference background (universe), all these overlaps are statistically significant (p-adj < 0.0001) and show a substantial inclusion of rare genotype-specific dysregulations in the multifactorial analysis. The reasons and the nature of this overlap will be faced in the discussion section of this thesis. Conversely, ADNP-ASD show 439 genes differentially expressed, out of which 373 were identified in the

previous analysis (p-adj < 0.0001). WBS lines show 147 DEGs, of which 97 were already found previously (p-adj < 0.0001).

ADNP-ASD specific deregulation shows significant enrichments FDR < 0.1) for “central nervous system development”, “cell division” and related categories, “artery-” and “aorta development”, ”mitochondrial translation”, ”regulation of catabolic process”, “regulation of nuclear cell cycle DNA repair”. CNS development genes are mostly up-regulated (B). Among them I found four enriching for artery/aorta development*: LRP1, LRP2, SUFU, NF1* and *BMP4* (the only down-regulated one). *SUFU* is a known negative regulator of hedgehog signalling, actively expressed in patterning and development, required for maintenance of neuronal progenitor identity and corticogenesis (Yabut et al., 2015), of which variants have been recently proven to cause Joubert Syndrome, which is characterised by hypertelorism, cranio-facial and skeletal defects (De Mori et al., 2017). *NF1* is mutated in Neurofibromatosis type 1. This disorder affect skin and skeletal development but is often associated with mild cranio-facial features (Visnapuu et al., 2018). BMP4, in mice, has been shown to be necessary for skull and cerebral vein formation, acting downstream of TWIST1, which is known to cause craniosynostosis in infants (Tischfield et al., 2017).

Among genes enriched for DNA repair at cell cycle and CNS development I found *ATRX*, which is mutated both in non-syndromic and syndromic intellectual disability and which defects in expression have been associated with memory deficits in mice (Tamming et al., 2017).

WBS specific transcriptional regulation include, together with most WBSCR genes, *BAZ1B, GTF2I* and *GTF2IRD1*, which are also SFARI genes. *GTF2I* and *GTF2IRD1*, together with *ACTN3*, *IL15, INPP5F, PPP3CA* and *SMAD3*, constitute a small but significant enrichment (FDR < 0.1) for “striated-, and skeletal-, muscle

adaptation” and “regulation of skeletal muscle adaptation”. *GTF2IRD1*, *INPP5F* and *ACTN3* are all represented in Cortecon. The first is associated with Neural differentiation, Deep- and Upper-Layer. The second with Cortical Specification and Deep-Layer. The third appear only in Neural differentiation. When confronted with GTEx data, all seven genes named in this paragraph appear indeed highly expressed in Cerebellum and Frontal Cortex.

7DupASD specific deregulation accounts for as few as 11 genes, 5 of which are in *WBSCR: EIF4H, LIMK1, BAZ1B, GTF2I* and *TBL2*. GO enrichments are absent even when considering the 103 genes with $p < 0.005$. No GO enrichment was also found for KS and WS specific genes.

Characterisation of Neurocristic Axis across disorders.

As anticipated in the introduction and aims chapters, previously published data of the lab demonstrated the ability of iPSCs to recapitulate disease-relevant transcriptional deregulations further amplified in a tissue-specific fashion (Adamo et al., 2015). Building on data published with that work I devised a simplified model of neural crest derived tissues development. In this system, iPSCs represent pluripotent lineages, at the apex of development, which could be impacted by genetic lesions, and develop epigenetic scars inherited in their many derivatives. NCSCs represent the fourth germ layer, upstream of endoderm and mesoderm formation, that is lost during development and which gives rise to tissues impaired by the considered NDDs, crucially including the peripheral nervous system. *In vitro* derived MSCs represent a mesoderm derivative of NCSCs that is still multipotent and potentially able to differentiate into bone, cartilages, muscles, marrow and various connective tissues, effectively impaired in the considered disorders. I selected RNA-seq data coming from iPSCs, NCSCs, MSCs control samples included in our cohort, and filtered out genes that were not expressed in all but one sample with minimum read-count per gene = 20. To clarify, this filter is applied on a gene-wise fashions, thus each gene kept in the analysis could be expressed in different subset of samples but always in at least 21 (out of the 22) samples.

A principal component analysis of control samples from our iPSCs, NCSCs and MSCs cohort shows a clear separation of the three lineages (Figure 29A). Adding fibroblasts coming from the same cohort of individuals shows the same separation, while highlighting a global transcriptional proximity between MSCs and fibroblasts with respect to NCSCs and fibroblasts themselves, suggesting a global developmental-wise shift (Figure 29B).

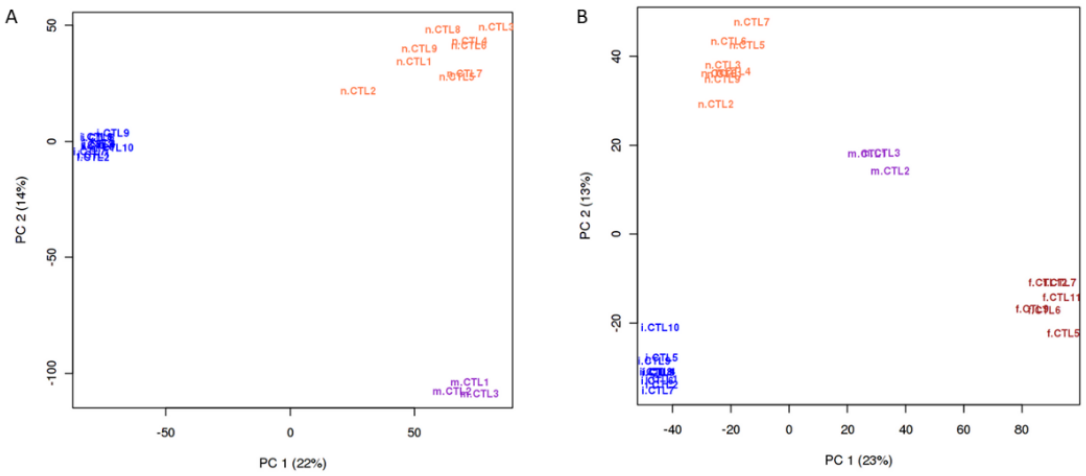


Figure 29. Principal component analysis of iPSCs, NCSCs, MSCs, and fibroblast log-normalized-read counts, filtered on genes expressed in iPSCs, NCSCs and MSCs. A) excluding fibroblasts B) including fibroblasts; iPSCs are labelled with "i." before sample name, in blue; NCSCs are labelled with an ".n" before sample name, in orange; MSCs are labelled with "m." before sample name, in purple; fibroblasts sample names include an "f.", in brown.

Here I present a differential expression performed on partially published data (Adamo et al., 2015, CTLs 1-3), in which I take into account individual genetic background and tissue identity. Indeed, I used iPSCs lines from all 10 CTLs, NCSCs lines from CTLs 1-9, and MSCs from CTLs 1-3, using as model ~Individual+CellType, testing for CellType and dropping Individual. The number of genes showing a significant change in expression were 5282 (FC > 2.5, FDR < 0.001). These genes were divided in 12 clusters using the same K-means protocol described previously (Figure 30).

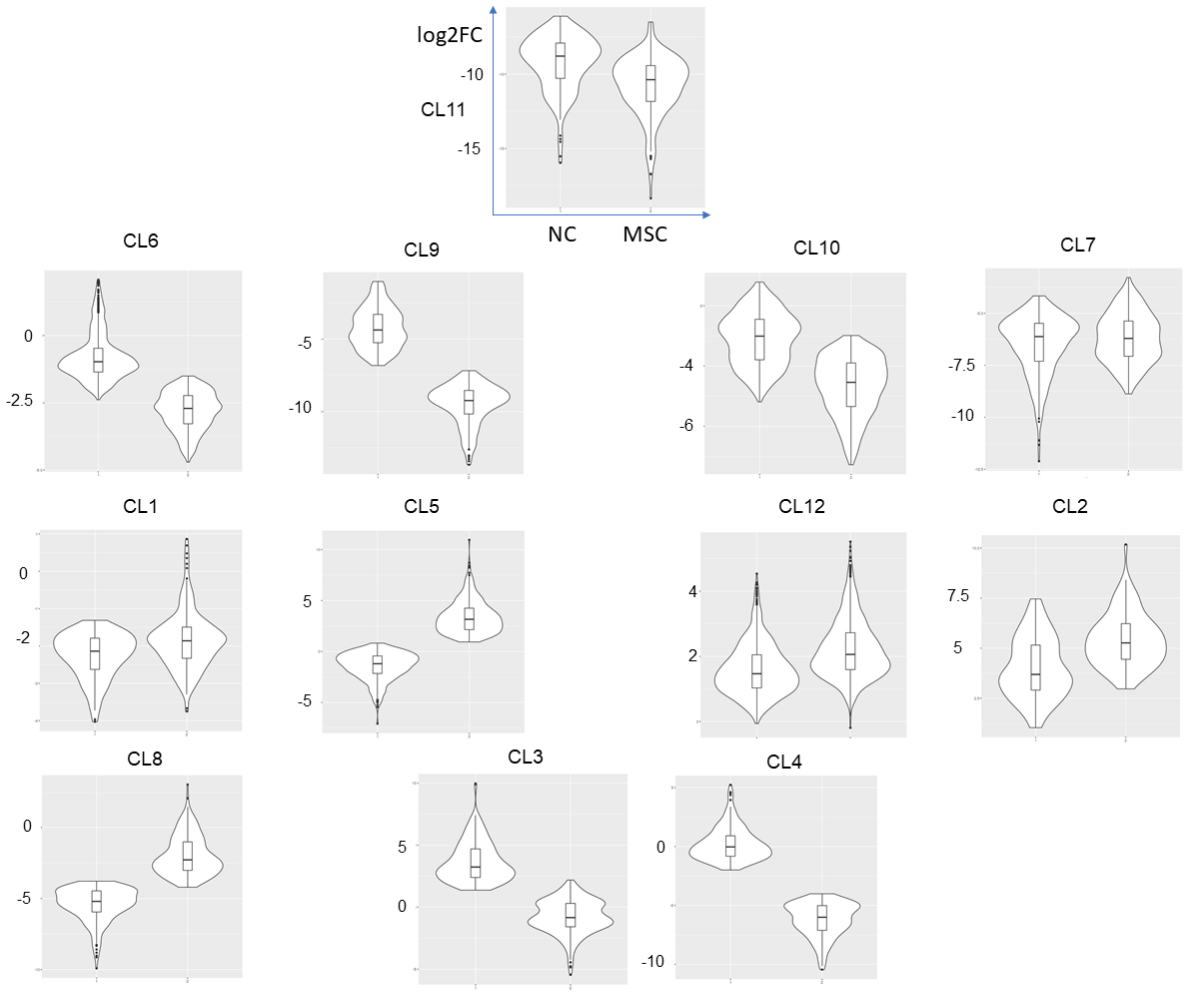


Figure 30. Cluster-specific violin plot of genes differentially expressed in NCSCs and MSCs; Axes are the same shown for CL11

GO enrichment of Biological Process for the first cluster (CL1, 661 genes) list “regulation of organ growth”, “renal system process”, “regulation of cardiac muscle tissue growth”, “muscle cell proliferation” and few others (Figure 31A). The second cluster (CL2, 284 genes), has many more GO Biological Process enrichments, including “cartilage development”, “mesenchyme development”, “positive regulation of locomotion” and “regulation of cell migration”, “ossification”, “embryo development”, “formation of primary germ layer”, “circulatory system development”, “gastrulation”, “fat cell differentiation”, “central nervous system development” and many other with FDR < 0.05. In figure (Figure 31B) I report a treemap, excluding categories with enriched children for visualization reasons.

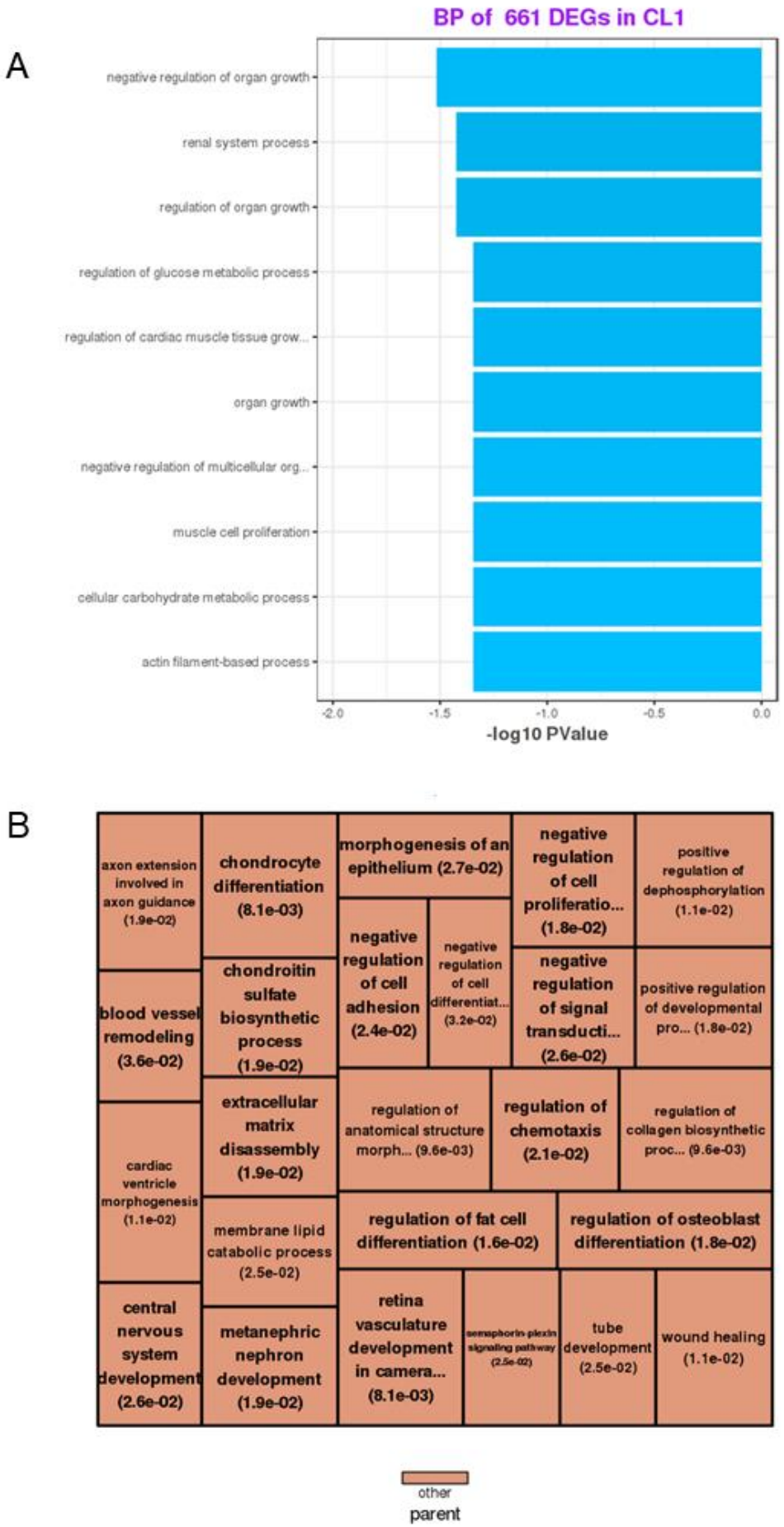
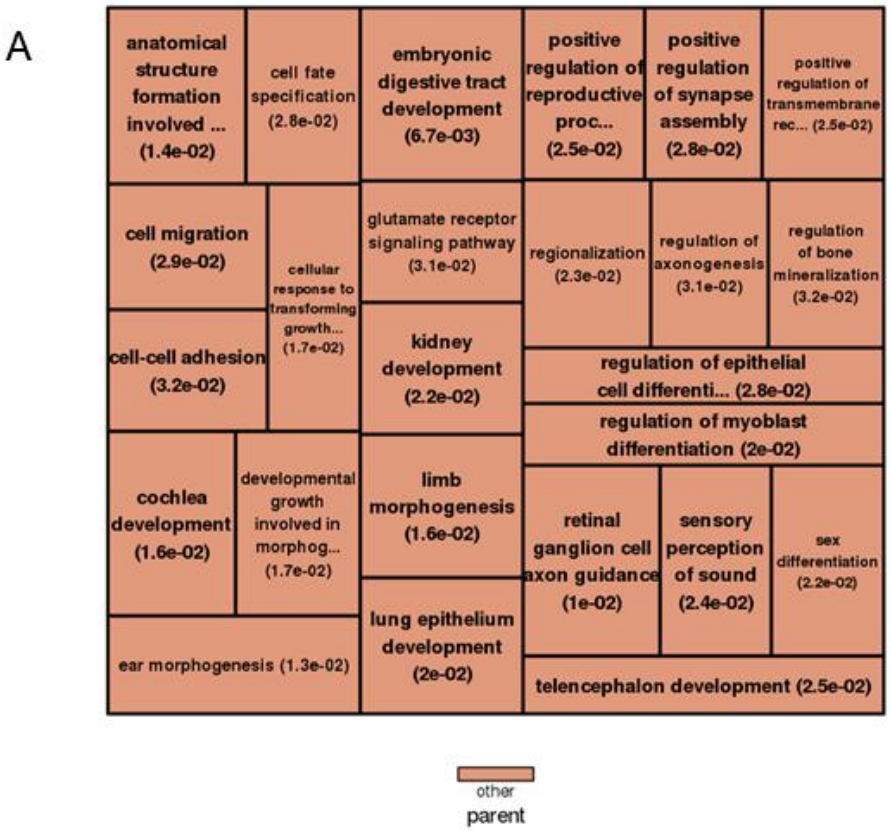


Figure 31. GO enrichments for Biological Process. A) Barplot of significant categories enriched in CL1; B) Treemap of enriched categories in CL2 devoid of children

Cluster 3 accounts for 223 genes, and is enriched for Biological Process GO categories such as “embryonic digestive tract development”, “cell motility”, “retinal ganglion cell axon guidance”, “urogenital system development”, “connective tissue development”, “ear-“ “development” and “morphogenesis”, “neuron projection development”, “regulation of myoblast differentiation”, “behaviour”, “telencephalon development” and several others (Figure 32A). Cluster 4 (177 genes) shows enrichments for Cellular Component GO terms such as “integral component of plasma membrane” and “intrinsic component of plasma membrane” (FDR < 0.05). Despite the small number of enriched categories, among genes included in both of them, I could find *SHANK2, KCNK5, KCNK10* and *ERBB3*. *SHANK2* is part of Shank family, that play a role in excitatory synapses and has been linked to autism and studied in mice models with contrasting results, but still pointing to autism spectrum behaviours (Schmeisser et al., 2012; Won et al., 2012). *KCNK5* and *KCNK10* are potassium channels which may play a role in signal transduction. Indeed, *KCNK5* knock-out in mice causes deafness and defines it as required for hearing maintenance (Cazals et al., 2015). *ERBB3* is an EGF receptor, with a neuregulin binding domain, which has been reported mutated in one patient with lethal congenital contractual syndrome type 2, and has been historically associated to neural crest migration and in particular to sympathetic nervous system formation in mice embryos (Britsch et al., 1998). Cluster 5 (CL5, 401 genes) shows enrichments for “artery development”, “regulation of cardiac muscle contraction”, “epithelial cell differentiation”, “platelet- aggregation” and “degranulation” and many others (Figure 32B).



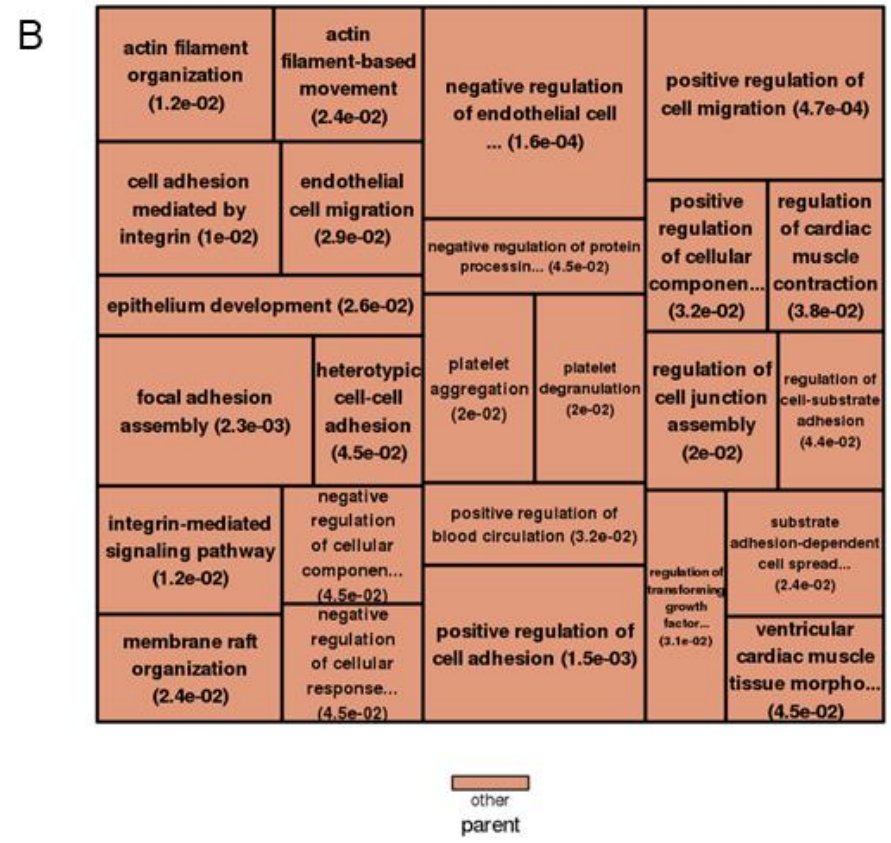


Figure 32. GO enrichments for Biological Process. A) Treemap of enriched categories in CL3 devoid of children; B) Treemap of enriched categories in CL5 devoid of children

Cluster 6 shows only two GO enrichments, for “lipoprotein particle binding” and “protein-lipid complex binding” (FDR 0.059 in both cases). These categories are enriched by *APOE, HSPD1, LRP8, PCSK9. APOE* is an extensively studied gene, which take part in cholesterol transport and has been associated with Alzheimer and late-onset dementia(Riedel et al., 2016). Intriguingly, *LRP8* (also known as ApoER2) is a lipoprotein binding receptor highly expressed in the brain which also has been reported to have, like *APOE*, missense mutations associated with Alzheimer(Ma et al., 2002). PCSK9 (also known as NARC1), is an enzyme, expressed in cerebellum during mice brain post-natal development, capable of inducing LRP8 degradation, as described in a paper studying its role in cholesterol and lipid homeostasis, which suggests PCSK9 importance for brain development hypercholesterolemia and cardiovascular diseases (Poirier et al., 2008). In a previous study the same gene was suggested to play a role in kidney mesenchymal cells but it was also attributed neurogenic functions, after its ectopic expression in E13.5 mouse embryos induced an increase in post-mitotic neurons number (Seidah et al., 2003). Cluster 7 showed no GO enrichment as well as Cluster 8, 9 and 11. Cluster 10 shows significant enrichments (FDR < 0.05) for “appendage- ” and “limb development” (enriched by the same genes), and for “digestive tract development”) but also for “midbrain development” and “regulation of ossification” (FDR < 0.1). Cluster 12 show enrichments in GO Biological Process such as “regulation of cell differentiation”, “regulation of vasculature development”, “skeletal system morphogenesis” and others reported in Figure 33.

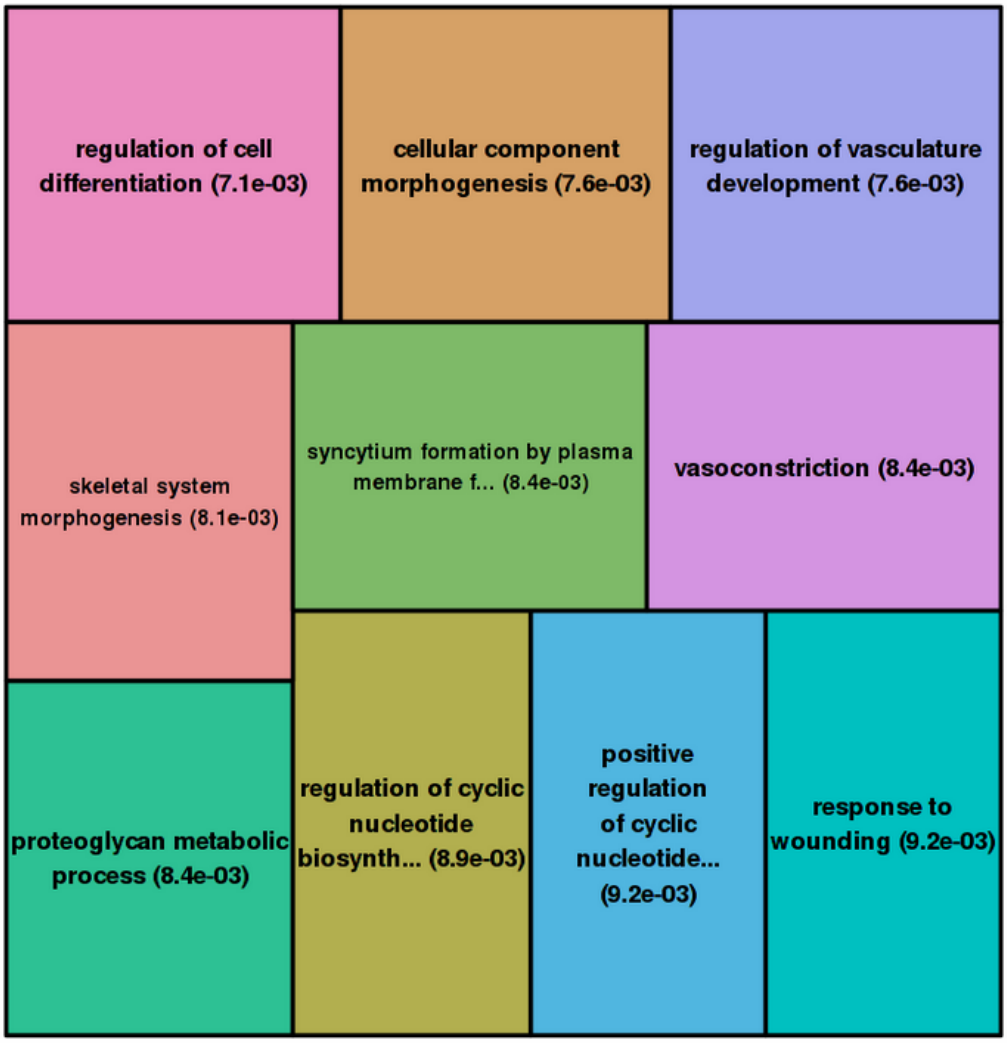


Figure 33. Treemap of enriched categories in CL12 devoid of children

I then proceeded with transcription factor enrichments for all twelve clusters identified, with the known caveat that our TFBS database is based on hESC data. TF enriched in the twelve lists are reported in the following table

Table 8. TF enriched in the 12 cluster identified upon differential expression analysis along the neurocristic axis

Cluster number	TF enriched	FDR (<)
1	BCL11A, NANOG, MAFK, KDM1A, HDAC2,	$9.8e^{-08}$

	GTF2I, TEAD4 and POU5F1	
2	EZH2, CTBP2, TEAD4, MAFK, TCF12 and BCL11A	$8.7e^{-08}$
3	EZH2, MAFK, SUZ12, CTCF, NANOG and RAD21	$< 8.1e^{-05}$
4	GTF2I and TCF12	0.01
5	TEAD4, CTBP2, CJUN, TCF12 and RAD21	$3.8e^{-06}$
6	SP1, KDM1A, POU5F1, USF1, BCL11A and TAF7	0.005
7	N.A	N.A
8	NANOG, MAFK, BCL11A, CJUN	0.005
9	POU5F1, CHD1 and NANOG	0.001
10	NANOG, HDAC2, TEAD4, BCL11A, MAX and EP300	0.001
11	N.A.	N.A
12	EZH2, CTBP2, SUZ12, CTCF, RAD21, TEAD4, MAFK, GTF2I and TCF12	0.0001

Combining information from GO and TF enrichments I concluded that CL8,9 and 11 will require further characterization that could not find place here, both experimentally and in silico. Nevertheless, considering fold-change trends and GO enrichments I grouped clusters into four main ones (Figure 34). CL5 includes MAFK and FOSL1 (also known as AP1) which are predicted as master regulator of several other clusters (Table 9). BACH1 is included in CL12 which is predicted to regulate several others. NANOG and POU5F1 are included in CL11 and are predicted regulators of genes in almost all clusters but CL4, CL8, and CL9 (Table 9). The first group is defined “Pluripotency genes”, including CL6, CL7, CL9, CL10 and CL11, because of their gradual down-regulation along differentiation towards MSCs. Nevertheless, CL10 shows GO enrichments similar to CL3, so I expect it to include genes that could be assigned to the other cluster. Instead, I suppose CL11, which includes NANOG and POU5F1 to play an essential role at the pluripotent stage. I defined the second group as “Neurocristic axis activation”, because of its increasing fold-change along differentiation towards MSCs, and it includes CL2 and CL12. Notably, CL12 includes BACH1, which is a TF enriched in many of the other clusters and especially CL1, CL5, CL6. The third group is called “Neural Crest Specific, because it is made of genes being up-regulated in NCSCs and down-regulated both in iPSCs and MSCs: in includes CL3 and CL4. The fourth group is “Mesenchymal Stem Cell Specific”, because it includes genes up-regulated only in MSCs, and it is made of CL1, CL5, and CL8. Notably, CL5 includes MAFK and FOSL1, which are TF enriched in CL1 and slightly in CL8, but also in CL3, CL6, CL9, CL10 and CL12. Moreover, TFs enriched for CL5 (mesenchymal specific main cluster regulator in light of these results) include CJUN, which plays together with FOSL1 an important role in regulating osteoblast differentiation through Wnt non canonical pathways, further corroborating this gene network deconvolution (Eferl et al., 2004; Krum et

al., 2010). In one work, studying the role of Wnt16 in osteoporosis, upon depletion of *Wnt16* in mouse models, a significant enrichment for FOSL1 binding sites among promoters of deregulated genes is observed (Sebastian et al., 2018). On the other hand, MAFK is still part of base-leucine zipper AP-1 superfamily, and it forms heterodimers with FOS (main FOSL1 interactor). Old studies on mouse models connect *MAFK* to both neural- and mesoderm-lineage commitment, show its expression in heart, osteoblasts and skeletal muscles. Intriguingly, the same author providing evidence of *MAFK* determining neural- or mesoderm- differentiation depending on which of two promoters is activated, proved also *BACH1* to be activated, together with *MAFK*, by TGF-β, to contrast Nrf2 activity (Motohashi et al., 1996; Okita et al., 2013).

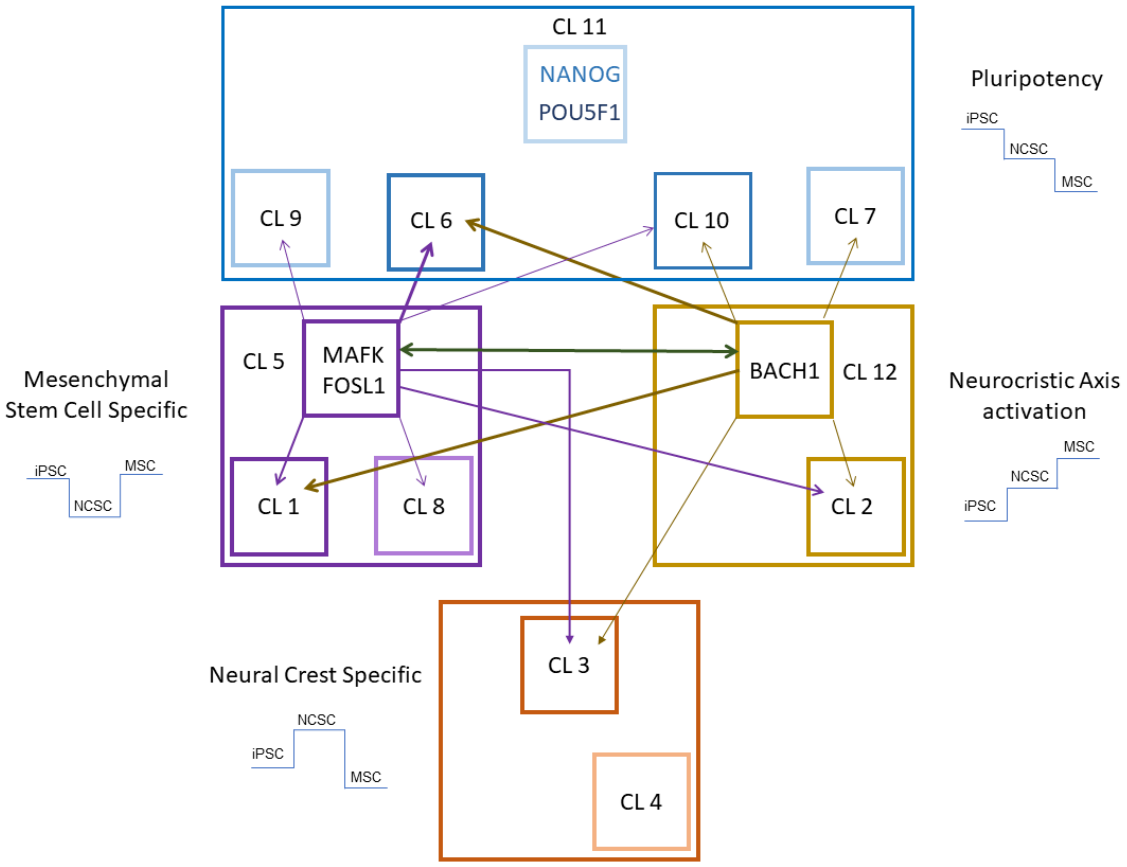


Figure 34. Modules of expression across the Neurocristic Axis. Proposed Model of regulation; Arrows indicate whether genes included in the starting cluster are predicted regulators of receiveing cluster; TF included in the clsuter are indicated;

Apparently, there is a cross-talk between CL5 and CL12, reciprocally composed of genes regulated by those included in the other cluster. Moreover, each cluster includes many targets of the same TF they contain (Table 9). I thus devised a model in which:

- a set of genes, important for pluripotency, are gradually turned off along development (CL11) and some follow the same pattern;
- a group of genes that is important all along Neurocristic Axis development, and could be largely regulated by BACH1;
- a set of genes mostly constituting MSCs transcriptional makeup are regulated by MAFK and FOSL1;
- a set of genes which characterise Neural Crest transcriptional landscape in our cohort and experimental design, which requires further investigation to define a main transcriptional conductor.

A Table with details on the number of genes regulated by transcription factors described in Figure 34 is provided below (Table 9)

Table 9. Number of genes in each cluster predicted to be regulated by the identified master regulator in Figure 30; last row indicates the number of total genes in each cluster; When a single target is included its name is reported

Clusters/TFs	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CL9	CL10	CL11	CL12
NANOG	70	33	26	3	35	69	DNMT3B	6	3	27		197
POU5F1	50	28	15	ZIC3	25	59	0	4	3	17		145
MAFK	93	53	41	6	52	96	0	7	3	26		353
BACH1	101	37	36	7	58	148	DNMT3B	5	2	37		449
FOSL1	9	8	3	SBK1	3	20	0	PAWR	0	2		69
total genes	661	284	223	177	401	725	279	206	222	569	169	1366

Characterisation of neural crest stem cells deregulation across disorders

Neural crest stem cells RNA-seq data includes 9 control, 5 ADNP-ASD, 3 WBS, 1 AtWBS, 3 7DupASD, 5 KS and 4 WS samples. WBS samples in this cohort are all non-ASD related. Experiments to differentiate, culture, extract RNA and produce RNA-seq libraries have been carried out by three different subgroups of the lab, in three different temporal moments. In order to assess the effect of such batches I filtered in two different ways gene-wise read-counts and performed principal component analyses on both datasets after log-normalization (TMM). The first filter excluded genes with less than 20 read-counts in at least 3 samples (Figure 36A). The second excluded genes not expressed in “all samples -1” gene-wise (i.e. each kept gene is expressed in at least “all samples but one”) (Figure 36B). Titrations of these “number of samples” filters show small shades of differences included in the two extreme ones reported in Figure 36. Using the former less stringent filter, CTL3 and AtWBS1 could be considered outliers, even if PC1 and PC2 capture small global variations (8% and 7% respectively). In this context, moderate batch effects between the “orange” team (KS, ADNP-ASD) and the other two can also be observed. Using the latter filter differences between one (“orange”) and the other two (“black” and “purple”) appear exacerbated (Figure 36B). Indeed, this filter, reveals a reduction in AtWBS1 and CTL3 distance from their batch of samples and an increased distance between “orange” and the others (12% on PC1). These last two observations lead me to the conclusion that a too strict filter would have reduced the impact of individuals genetic background, tuned the transcriptomic landscape to genes effectively detected in all samples but probably subject of batch effects (e.g. library preparation), thus excluding individuals’ transcriptional make-up and

indirectly forcing the analysis onto genes which quantification could be simply more able to expose technical batches.

Considering the relevance of neural-crest in the neurocristic axis as devised in this thesis, I reviewed the available literature to identify a signature that would represent genes expressed or relevant for Cranial Neural Crest fate. In particular: NGFR has been used to validate differentiation at FACS, and together with *SOX9, SOX10, SNAI1, SNAI2, TWIST1, TWIST2* and *B3GAT1* is one of the most adopted markers of NCSCs (Spokony et al., 2002); *LIN28A* is a marker of pluripotency expressed in NCSCs; *SMAD4* has been deemed essential for early stage of mouth and face development (including teeth, which often show defects in our NDDs (Ko et al., 2007)); *CUL3, TCOF1* and *NOLC1* have been included precisely in the stage of fate decision between the CNS precursor and neural crest commitment (Werner et al., 2015); I further selected few HOX genes, *MSX2*, *FOXD3*, and *DLX* depending on their expression and localization during embryonic development (Bhatt et al., 2013; Mishina and Snider, 2014; Santagati and Rijli, 2003). I provide an heatmap in which selected genes are reported in terms of z-scores measured gene-wise on log-normalized read counts. In this picture, samples clustering by Euclidean distance reveal no major difference across NCSCs lines depending on batch effects or genotype (Figure 35).

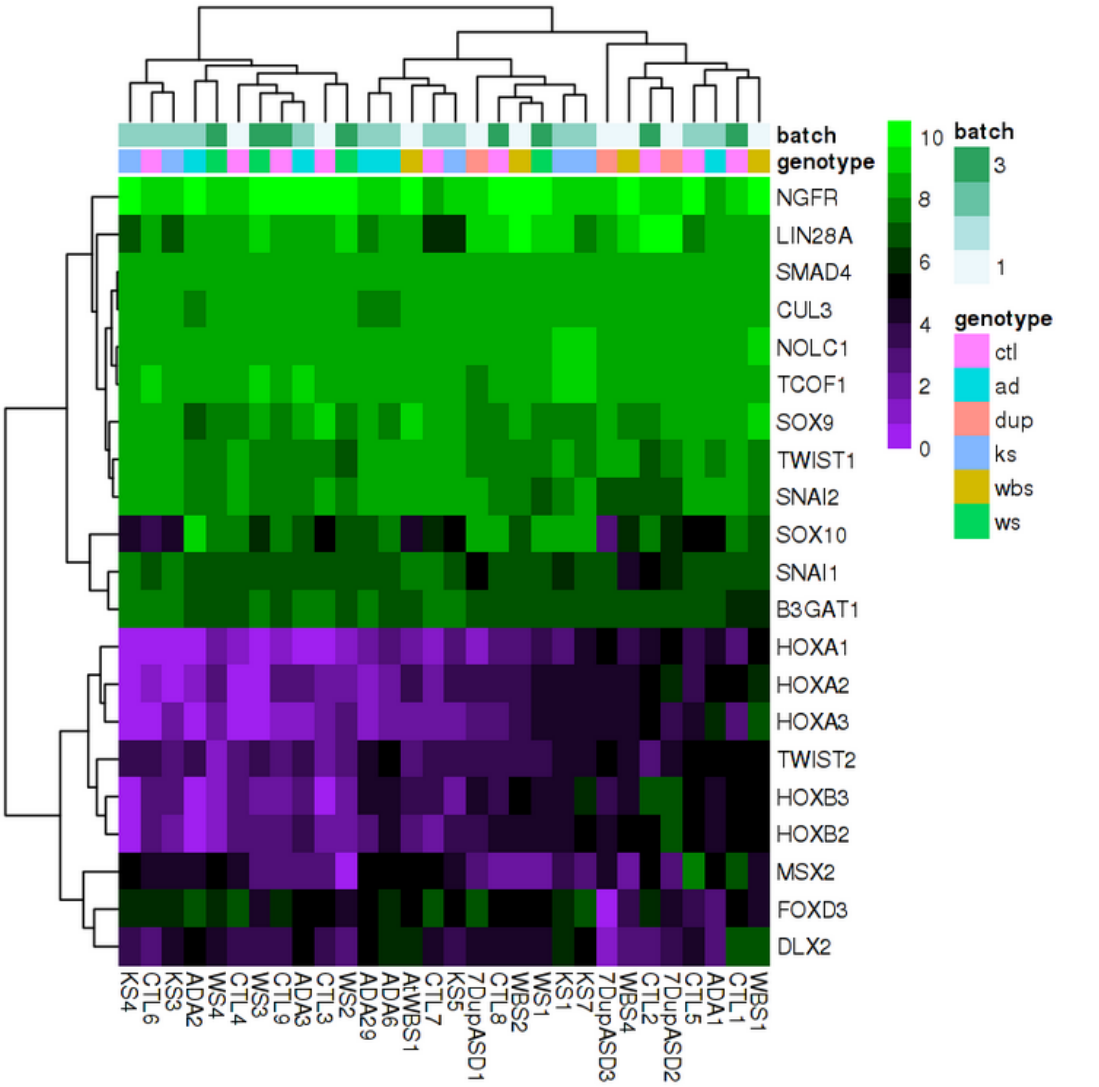


Figure 35. Heatmap of Cranial Neural Crest fate signature genes (z-score of log-normalized (TMM) read-counts)

Considering the presence of at least 3 genotypes per batch, and the presence of at least 3 controls per batch (Figure 36C) I decided i) to use a more conservative threshold, keeping genes expressed in at least 3 samples: the minimum number of samples in each genotype (i.e. 7DupASD), and ii) to treat the technical batch as a

coefficient in my differential expression model matrix (~batch+Genotype), testing for Genotype

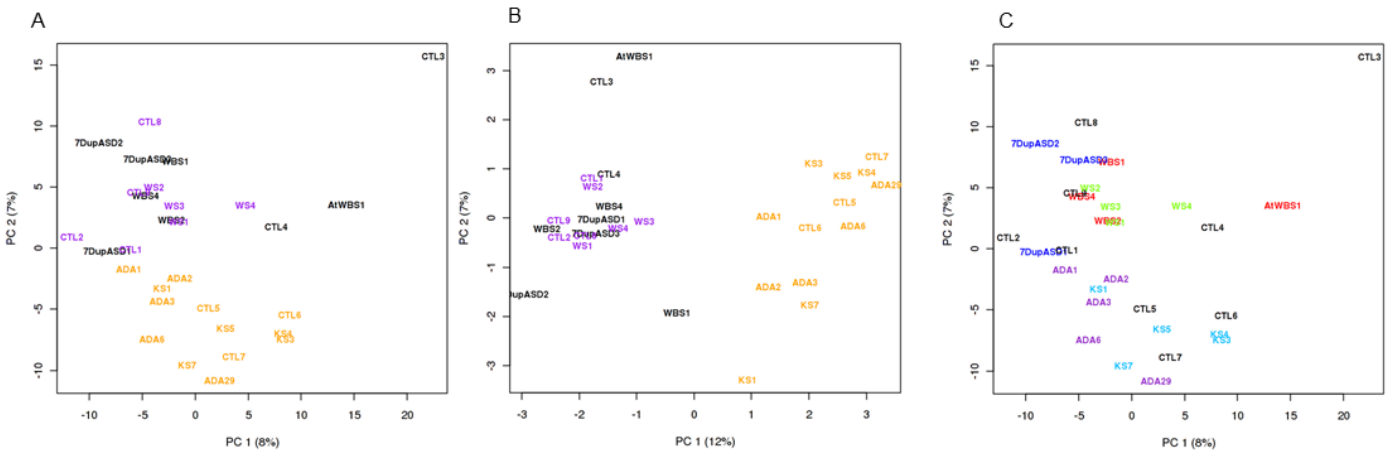


Figure 36. Principal Component Analysis of Neural Crest Stem Cells RNA-seq data. A) PCA of log-normalized read-counts after filtering genes expressed in at least 3 samples (colours represent technical batches as per described in main text); B) PCA of log-normalized read-counts after filtering genes expressed in at least 29 samples (same colouring of A); C) PCA of log-normalized read-counts after filtering genes expressed in at least 3 samples. WBS samples are reported in red, 7DupASD in blue, KS in cyan, WS in green, ADNP-ASD in purple, CTLs in black

Downstream of such an analysis I obtained 457 genes with mean-fold-change > 1.5 and FDR < 0.05. These genes show a clear clustering per genotype (Figure 37), with one cluster including most controls, from all batches, with 3 out of 5 KS, 1 WS and the atypical WBS samples (which spares few genes of the WBSCR and show mild cranio-facial features, which are expected to be NC-derived traits). A second cluster includes 3 controls, 2 WS (out of 3) and all 3 canonical WBS. A third cluster includes 2 out of 5 KS and 4 out of 5 ADNP-ASD. The fourth cluster includes all 3 7DupASD, 1 ADNP-ASD and 1 control samples. Notably, CTL2, which is found within the fourth cluster, appears both in Figure 29A and Figure 29B as the closer to iPSCs, eventually pointing, for instance, to a slightly incomplete differentiation or, less likely, to a unique epigenetic signature that could be individual- or reprogramming- derived.

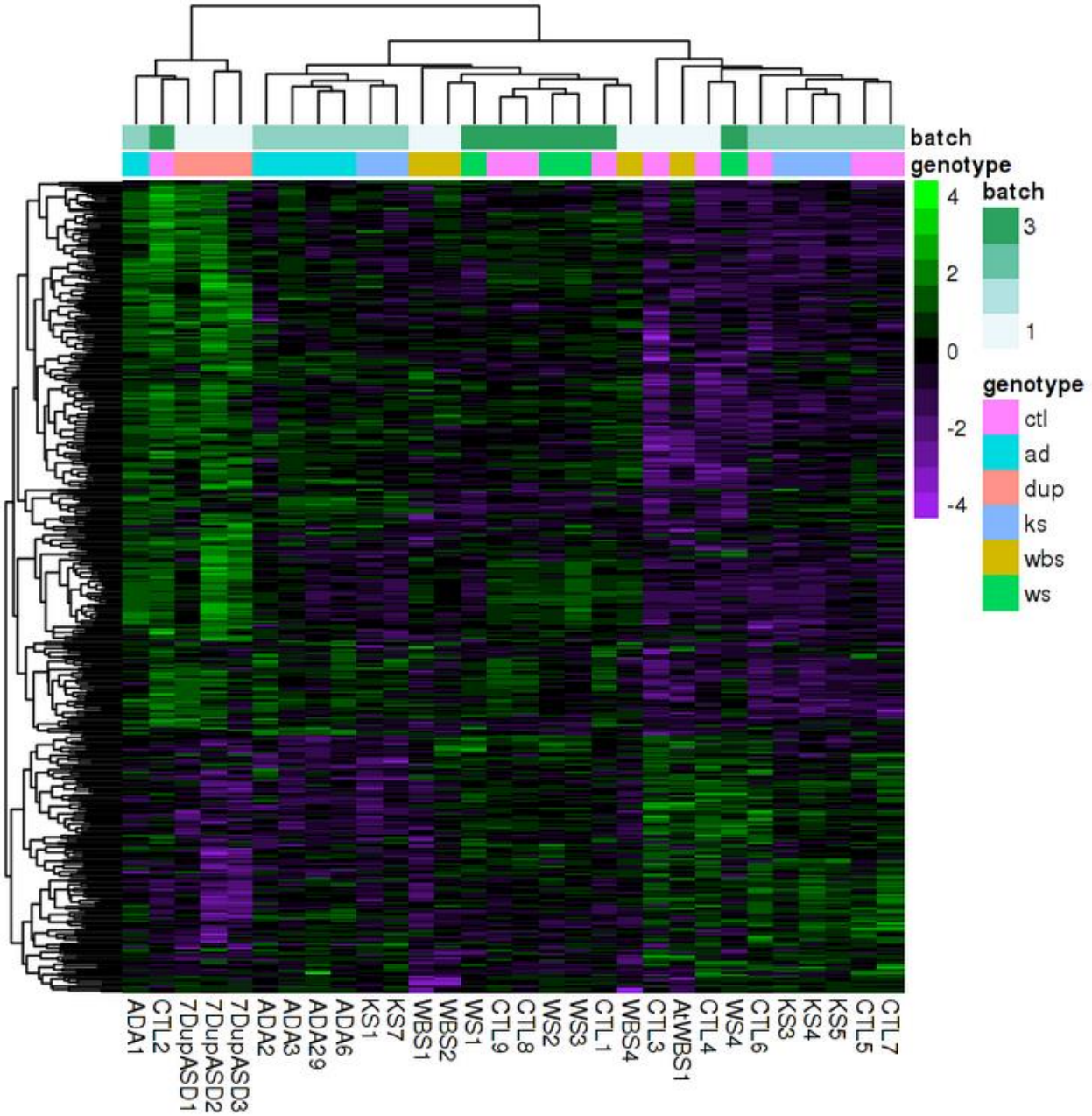


Figure 37. Heatmap of log-normalized read-counts (z-scores) of 457 genes differentially expressed across genetic conditions

Globally, observing the heatmap of all 457 genes, most of these genes appear gradually up-regulated going from WBS and WS to ADNP-ASD and 7Dup-ASD, while a second smaller set of genes appears to go in the opposite direction. Out of 457 genes, 152 were differentially expressed with FDR < 0.005 and mean-FC > 2, and 219 were differentially expressed in all disorders (Figure 41).

These 457 genes are divided, in the heatmap in Figure 37, by row (gene z-scores) into 3 large clusters: one down- and two up-regulated in 7DupASD and ADNP-ASD). I conducted K-means clustering on these genes using 3 as number of centroids to identify and separate the three lists in an accurate way. The first cluster is made of 154 genes (NC.CL1), the second is made of 246 genes (NC.CL2), the third accounts for 57 genes (NC.CL3).

These three clusters show distinct GO enrichments. The first includes genes down-regulated in 7DupASD, WBS and ADNP-ASD and seems more connected with development, enriching terms such as “anatomical structure formation involved in morphogenesis”, “smooth muscle tissue development”, “blood vessel morphogenesis”, “gastrulation”, “negative regulation of developmental process” (Figure 38).

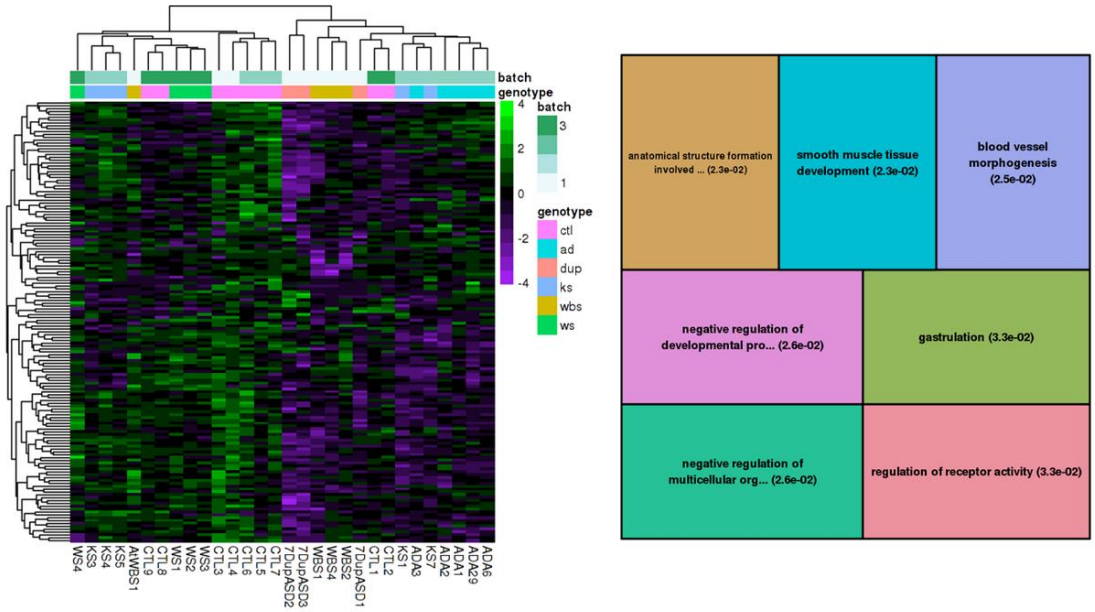


Figure 38. Heatmap and GO enrichments of BP categories for NC.CL1 cluster of genes differentially expressed in NCSCs across disorders

The second cluster includes genes up-regulated in 7DupASD and ADNP-ASD, and less strongly up- in WBS. NC.CL2 appear to be enriched for signal transduction, accounting for BP processes such as “action potential”, “cation transport”,

“regulation of postsynaptic membrane potential”, “chemical synaptic transmission”, but it includes also genes enriched for “cardiac muscle tissue development”, “diencephalon development”, and many others, including “behaviour” Figure 39.

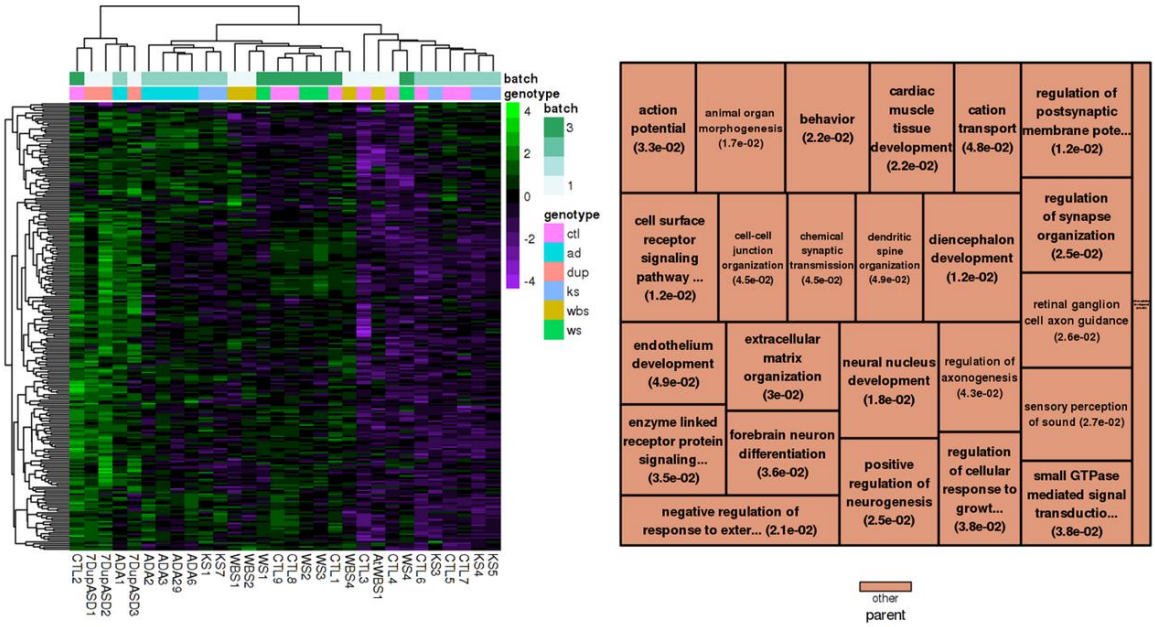


Figure 39. Heatmap and GO enrichments of BP categories for NC.CL2 cluster of genes differentially expressed in NCSCs across disorders

The third cluster, which groups controls, the AtWBS, all KS, and half of WS individuals, show BP categories strongly related to brain development, with “neuron migration”, “ganglion development”, “cerebellum development”, “neuron fate specification” and “autonomic nervous system development” but also “musculoskeletal movement”, ”neuromuscular junction development”, “regulation of skeletal muscle tissue development” (Figure 40), eventually revealing targets of GRNs shared by neural crest lineages with cranial nerves, motor neurons and sensory ganglia.

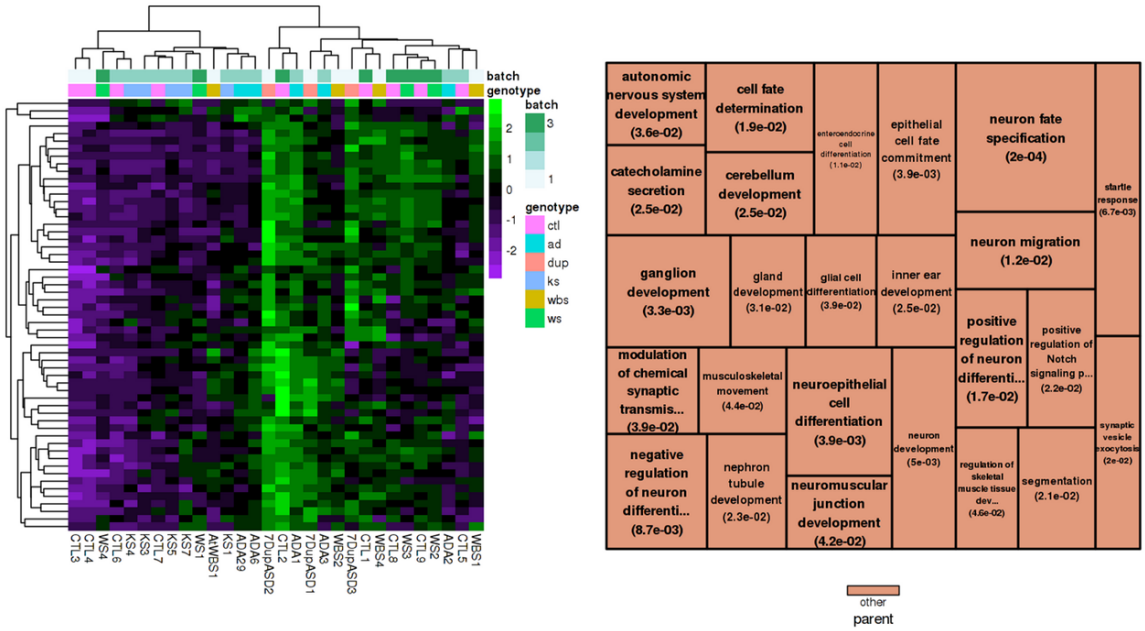


Figure 40. Heatmap and GO enrichments of BP categories for NC.CL3 cluster of genes differentially expressed in NCSCs across disorders

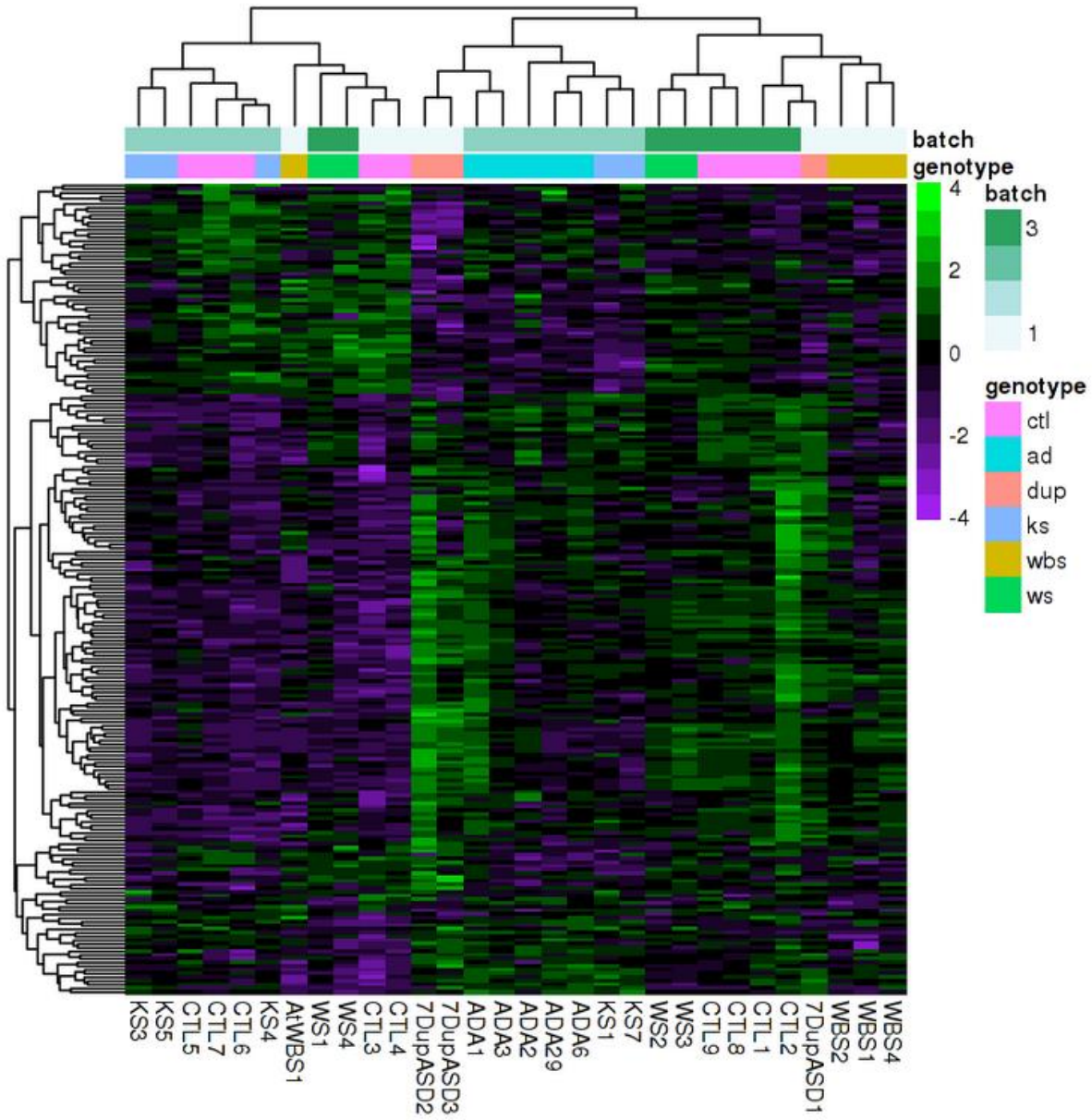


Figure 41. Heatmap of 219 genes differentially expressed in all disorders with FC > 1.5 and FDR < 0.05

Genes differentially expressed in all disorders show neural-crest related disease-relevant GO categories enrichments. Among them, I highlight 52 that enrich for “neuroepithelial cell differentiation”, “neural precursor differentiation”, “forebrain neuron differentiation”, “glutamatergic synaptic transmission”, “behaviour”, “chemotaxis”, “skeletal development”, “skeletal muscle organ development”, “segment specification”. Given their partial overlap in terms of genes enriching each category, they are all reported in an annotated heatmap (Figure 42).

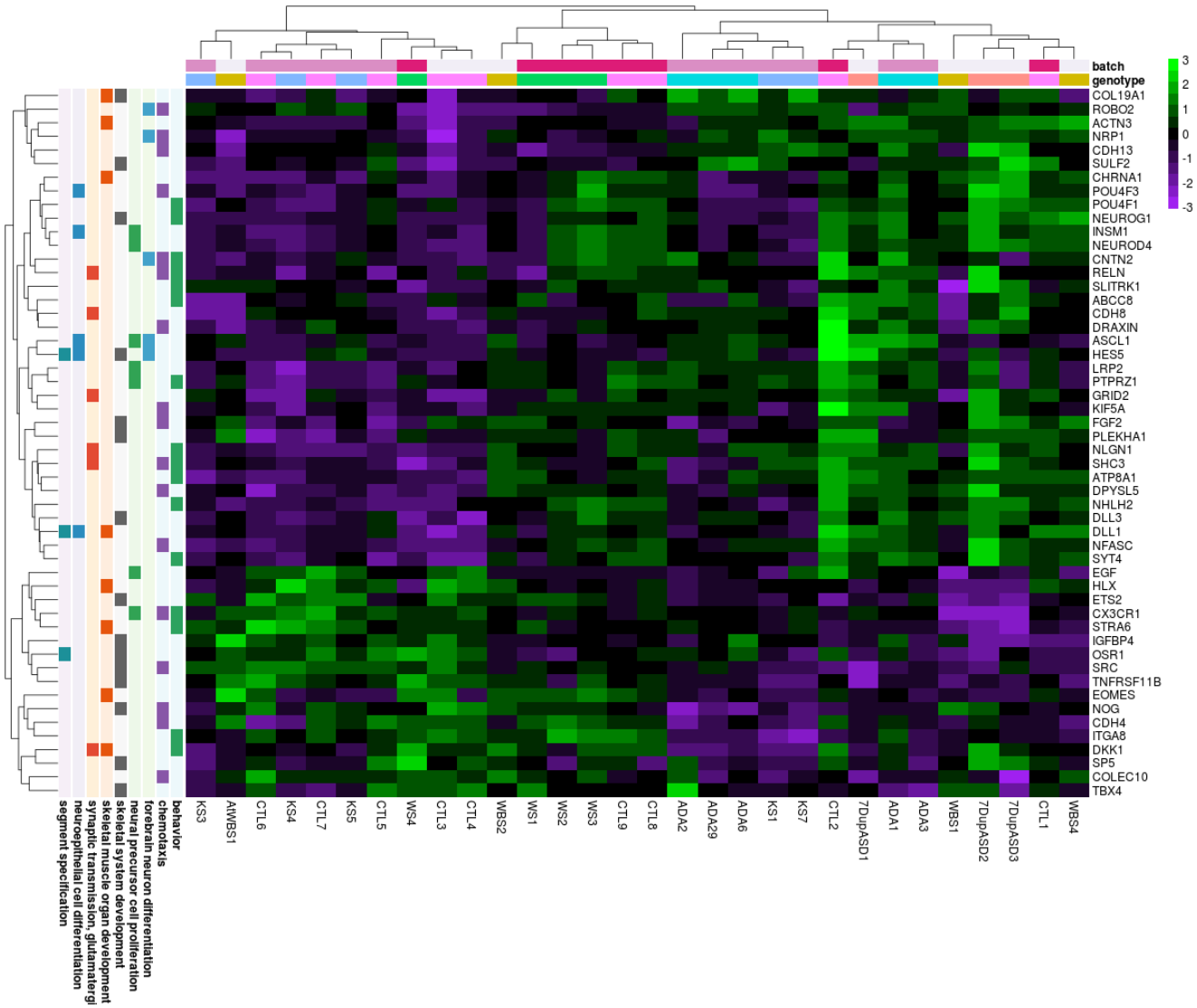


Figure 42. GO annotated heatmap of 52 neural-crest related genes differentially expressed in NCSCs in 5 disorders.

Notably, these 52 genes show a very similar clustering to the one of all 457 genes identified across disorders (Figure 37). Moreover, most genes enriched in "behaviour" and "chemotaxis" and all genes enriched for "forebrain neuron differentiation" categories are down-regulated in 7DupASD, ADNP-ASD, WBS and slightly down-regulated in 3 out of 4 WS individuals.

The same transcription factors are enriched both in the 52 and in the 457 genes lists: EZH2, REST, CTBP2, MAFK, SUZ12, RAD21, GTF2I, TEAD4, CTCF, NANOG, CEBPB and ZNF143 (with FDR < 0.005). EZH2, SUZ12, GTF2I, MAFK, CTCF,

NANOG relevance has already been discussed extensively. REST requires a special mention for its well-known role in mouse adult neural stem cells (NSCs) maintenance, regulated via recruitment of CoREST and mSin3A to chromatin of proneural genes (Gao et al., 2011).

Convergences and divergences among disorders with opposite genetic and phenotypic features in the neural crest.

In the previous section I described results of a multifactorial differential expression analysis including samples from ADNP-ASD, WBS, 7DupASD, KS and WS. Here I focus on four of those disorders, WBS and 7DupASD, KS and WS. As anticipated, one important theme of this thesis is the comparison of genotypically different, and phenotypically opposite disorders and conditions. Indeed, the first two disorders are caused by symmetrical copy number variations of the 7q11.23 region while the second two are caused respectively by mutations in subunits of KMT2D/Compass complex and PRC2.

Among genes differentially expressed across disorders, in WBS and 7DudASD I could identify 261 genes up-regulated in both disorders and 127 down-regulated in both disorders. Genes up-regulated in WBS and down-regulated in 7DupASD were 26, unexpectedly including one WBSCR gene (TRIM74), while 46 were up- in 7DupASD and down- in WBS (including 7 WBSCR genes[5]). In the context of WBS/7DupASD comparison, several clinical and cranio-facial similarities are shared between the two syndromes. Thus, finding a large concordance in terms of differential expression was expected, unless one hypothesised that expression levels of genes in WBSCR were directly correlated with most of the identified transcriptional deregulation. Interestingly, 25 out of 46 genes down- in WBS and up-regulated in 7DupASD enriched the list of REST targets (FDR < 0.0025), which corroborates its putative role as master regulator of neural-crest deregulation across

[5] GTF2I, EIF4H, LIMK1, TBL2, WBSCR22, RFC2, BCL7B

disorders. At the opposite spectrum of transcriptional dysregulation 16 out of 26 genes up-regulated in WBS, and down-regulated in 7DupASD, include *EOMES* and *TIMELESS* and are enriched for SP1. *EOMES* (more frequently called TBR2) encodes for a transcription factor, which plays a crucial role both in mesoderm- and in central nervous system development. *TIMELESS* is a gene included in GO category “DNA binding”(no reference associated), involved in circadian rhythm and in cell survival after DNA damage (Ünsal-Kaçmaz et al., 2005). SP1 is a zinc finger transcription factor that preferentially binds promoters at GC-rich genomic locations which, to my knowledge and literature understanding, has never been implicated in neural crest regulation. The two lists of genes going in opposite direction in WBS and 7DupASD neural crest stem cells will be used along the rest of the thesis and eventually referred to as “WBSCR correlated genes in neural crest”.

All four possible patterns of differential expression were observed also comparing KS with WS. In fact, I found 10 DEGs up- in both disorders, 24 down- in both disorders, 68 up- in WS and 122 up-regulated in KS. These genes will be eventually referred to in the next chapters as “KS/WS axis genes”. For what concerns WS and KS, starting from the different genetic alterations, with coarsely opposite enzymatic functions, and going to the physical contrasts between the two disorders (macro- vs microcephaly, overgrowth vs dwarfism, etc) identifying a major opposition in terms of gene expression level was more expected. Nevertheless, being PRC2 a major regulator of development, via direct down-regulation of its targets, and being KMT2D/KDM6A complex involved in specific activation of gene expression via enhancer modification, a less heterogenous pattern could have been expected. Among these genes, those down-regulated in KS are indeed enriched targets of EZH2/SUZ12 (FDR < $3.5e^{-04}$), and the vast majority (59) were enriched targets of RAD21 (FDR < $5.3e^{-04}$), which is an important player of cohesion complex,

responsible for Cornelia De Lange Syndrome (a syndrome characterised also by cleft palate and microcephaly). 30 genes were found enriched for BACH1 targets, including *CDKN1A, EOMES* and *TNFRSF11B*. Notably, CDKN1A (more probably known as p21) is a cyclin dependent kinase, which has been recently associated with EZH2. Crucially, CDKN1A, which is up-regulated in WS patients - who are supposed to bear a haploinsufficiency of EZH2 - has been proposed as target of EZH2 specific down-regulation (Akizu et al., 2016). In this work *EZH2* depletion, by knock down in mouse embryos, leads to formation of smaller and less organised neural tubes. Notably, in the same work *EZH2* reduction leads to a balanced number of genes up- and down-regulated (similarly to what I observed here). *CDKN1A* promoter is enriched both in H3K27me3 and in EZH2 binding in mouse embryos neural tube and *CDKN1A* depletion lead to neural tube defects equivalent to those observed upon *EZH2* knock-down (KD). *TNFRSF11B*, is implicated in bone homeostasis and osteoblast maturation, where its expression appears to be repressed by *RUNX2* activity (Komori, 2018). Finally, also genes up-regulated in KS (and down- in WS) were mostly enriched for REST (82, FDR < $6e^{-15}$) and EZH2 (32 genes, FDR < $7e^{-11}$), but also for CTBP2 (52 genes, FDR < $4e^{-07}$) and MAFK (60 genes, FDR < $2e^{-06}$). Moreover, these genes showed interesting GO enrichments including “cerebellum development”, “chemical synaptic transmission”, “forebrain

development", "neuron migration", "dendritic spine morphogenesis" and many others Figure 43.

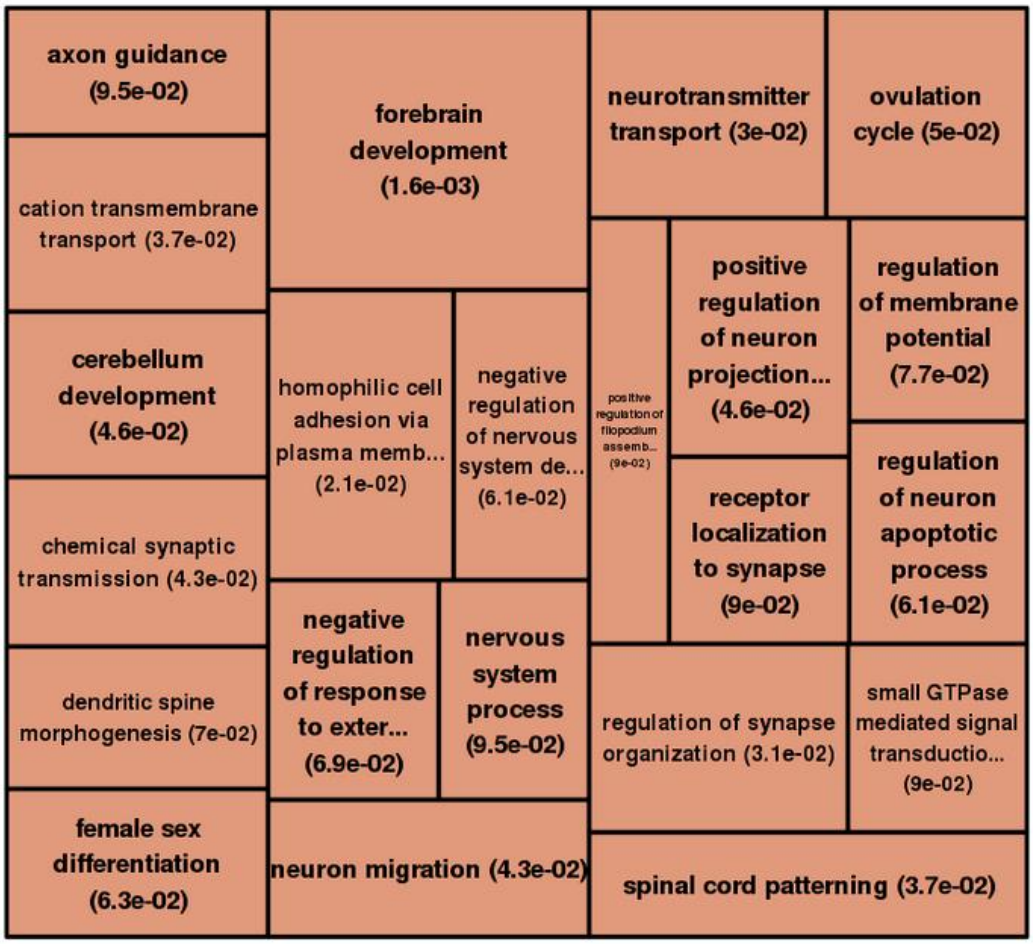


Figure 43. Biological Process GO enrichments of genes up-regulated in KS and down-regulated in WS, in neural crest stem cells

Crucially, among genes enriching these categories I found *RELN* (Reelin), which has been discussed in the introduction for its crucial role in regulating neuron migration and whose expression is altered upon EZH2 KD during neurogenesis (in mice), leading to defects in neurons migration. Finally, among genes up-regulated in KS and down- in WS I could find one SFARI genes (SOX3), which is associated with cell-fate determination, embryonic development and mental retardation.

Given the TF enrichment for BACH1 targets in genes down- in KS and up-regulated in WS, I compared them with cluster 12 of the Neurocristic Axis and found 9 genes in common: *CASP7, TNFRSF10D, MDFI, CALHM2, CDKN1A, RPS6KA4, KCNQ5*,

CLEC11A, PRKCDBP. Comparison of genes up- in KS and down- in WS with cluster 5 of Neurocristic Axis DEGs, I found other 9 genes*: CDH13, AXL, COL5A3, SLC8A1, ARHGAP18, YPEL2, PDLIM3, LINC01137, PLD1*. Moreover, I found 47 genes that were differentially expressed in KS both at the iPSCs stage and in NCSCs, out of which 15 were down- and 16 were up-regulated in both tissues. Unfortunately, these genes have no GO enrichment, but most of them significantly enrich (FDR < 0.005) the list of REST (28 genes), EP300 (26 genes), MAFK (23 genes), and CEBPB (27 genes) targets.

Convergences and divergences among disorders with opposite genetic and phenotypic features in mesenchymal stem cells.

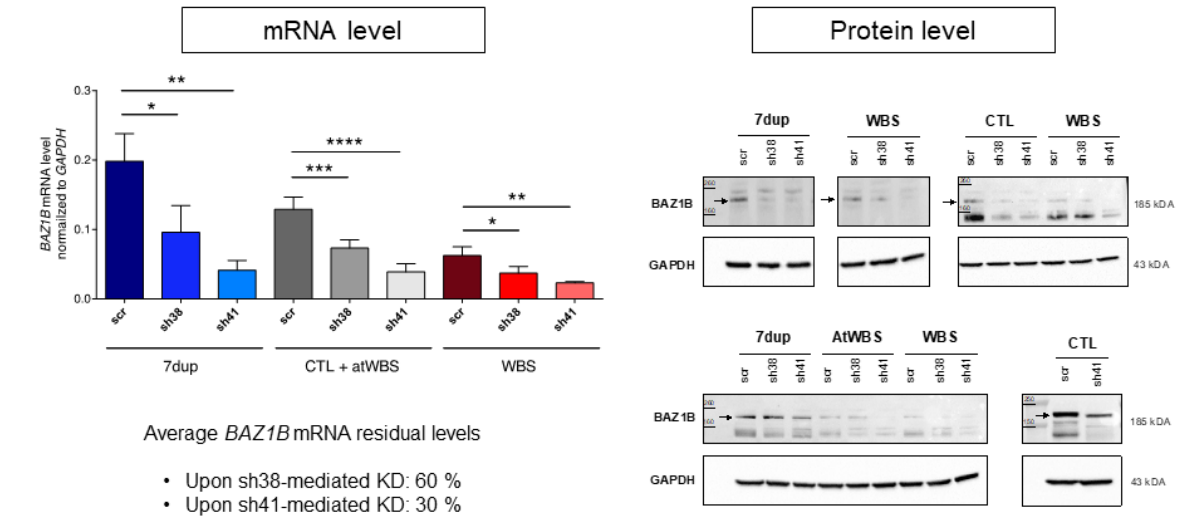
Our cohort of mesenchymal stem cells includes 3 controls, 3 WBS, 1 AtWBS and 2 7DUP-ASD samples. Having these data already been introduced in the previous chapter I briefly present a Principal Component Analysis of MSCs transcriptional data, which presents a peculiar separation of samples. PCA has been conducted on log-normalized read-counts, devoid of genes expressed in less than 2 samples with less than 20 read-counts. Controls are displaced on a diagonal separating WBS from 7DupASD, with 7DupASD1 and AtWBS1 in near proximity of CTL1 and CTL3. PCA MSCs.

In this context, I performed a differential expression analysis using only one factor (~Genotype) to compare WBS and 7DupASD as independent groups with respect to controls. Among ~12300 expressed genes, 134 and 129 were found differentially expressed in WBS and 7DupASD, respectively. Most of them (95) were DEGs in both conditions. The vast majority (50) were down- in WBS and up-regulated in 7DupASD, while only 6 were up- in WBS and down-regulated in 7DupASD. 15

genes were up-regulated and 24 down-regulated in both conditions. Thus, while on NCSCs I observed a general concordance in differential expression direction, in a more differentiated lineage, a transcriptional discordance between the two disorders was found. Moreover, among genes up- in WBS and 7DupASD 26 were also in cluster 1 of the neurocristic axis genes and 17 were up-regulated already at iPSCs. Among genes down- in both WBS and 7DupASD 18 were already down- at iPSCs stage and 18 were present in cluster 2 of Neurocristic Axis.

Deconvolution of BAZ1B dependent regulatory networks in the neural crest

Among WBSCR genes, BAZ1B is by far the most characterised -in animal models- for playing an important role in neural crest. It has already been associated to neural crest cells migration regulation and cranial morphogenesis in X. lævis and M. musculus, respectively. To reveal whether such roles are maintained in humans and to dissect the GNRs underlying such cellular and phenotypical traits we devised a strategy based on BAZ1B RNA-silencing in neural crest stem cells, followed by transcriptional and epigenomic characterisation. The cohort designed for this experiment includes 3WBS, 1 AtWBS (having a wild-type copy number of BAZ1B), 3 7DupASD, and 4 control lines. 10 out of 11 samples have separately been transfected with a scramble short hairpin (“shSCR” or “scr” along text and figure legends), a short-hairpin that has shown mild BAZ1B down-regulation (sh38), and a short-hairpin inducing strong BAZ1B knock-down (sh41) (Figure 44). One control (CTL3) has been interfered only with scramble and sh41.



Student's *t*-test (ns: not significant; *: p-value < 0.05; **: p-value < 0.01; ***:p-value < 0.001 and ****:p-value < 0.0001)

Figure 44. BAZ1B RNA- and Protein levels measured experimentally by means of qPCR and WesternBlot. Experiments performed by Matteo Zanella.

RNA-seq was performed on 32 lines:

- 11 individuals: 3 controls, 1 AtWBS, 3 WBS and 3 7DupASD;
- One scramble and two knock-down short hairpin per 10 samples
- One control was interfered only with scramble and sh41 (CTL3)

ChIP-seq has been performed on 27 lines (all but CTL3), against BAZ1B, H3K4me3, H3K4me1, H3K27ac and H3K27me3. It is worth mentioning that for ChIP-seq experiments, instead of cell lines transfected with shSCR, chromatin has been taken by non-interfered lines.

Transcriptional analysis of BAZ1B knock-down across four genotypes.

According to Principal Component Analysis (PCA), the effect of *BAZ1B* KD did not induce a major alteration in NCSCs transcriptomes, since several KD lines tend to cluster with their respective shSCR, thus supporting a larger effect due to genotype and individual genetic backgrounds. Indeed, looking at shSCR lines (Figure 45), controls show higher variability than WBS and 7Dup, which instead tend to cluster at very narrow PC1 coordinates. If CTL2, which has already been deemed as a potential outlier in previous chapters, is excluded, I could support the claim that controls cluster towards the left side of the space defined by PC1 and PC2, together with the AtWBS, while 7DupASD and WBS lines cluster towards the right side (Figure 45).

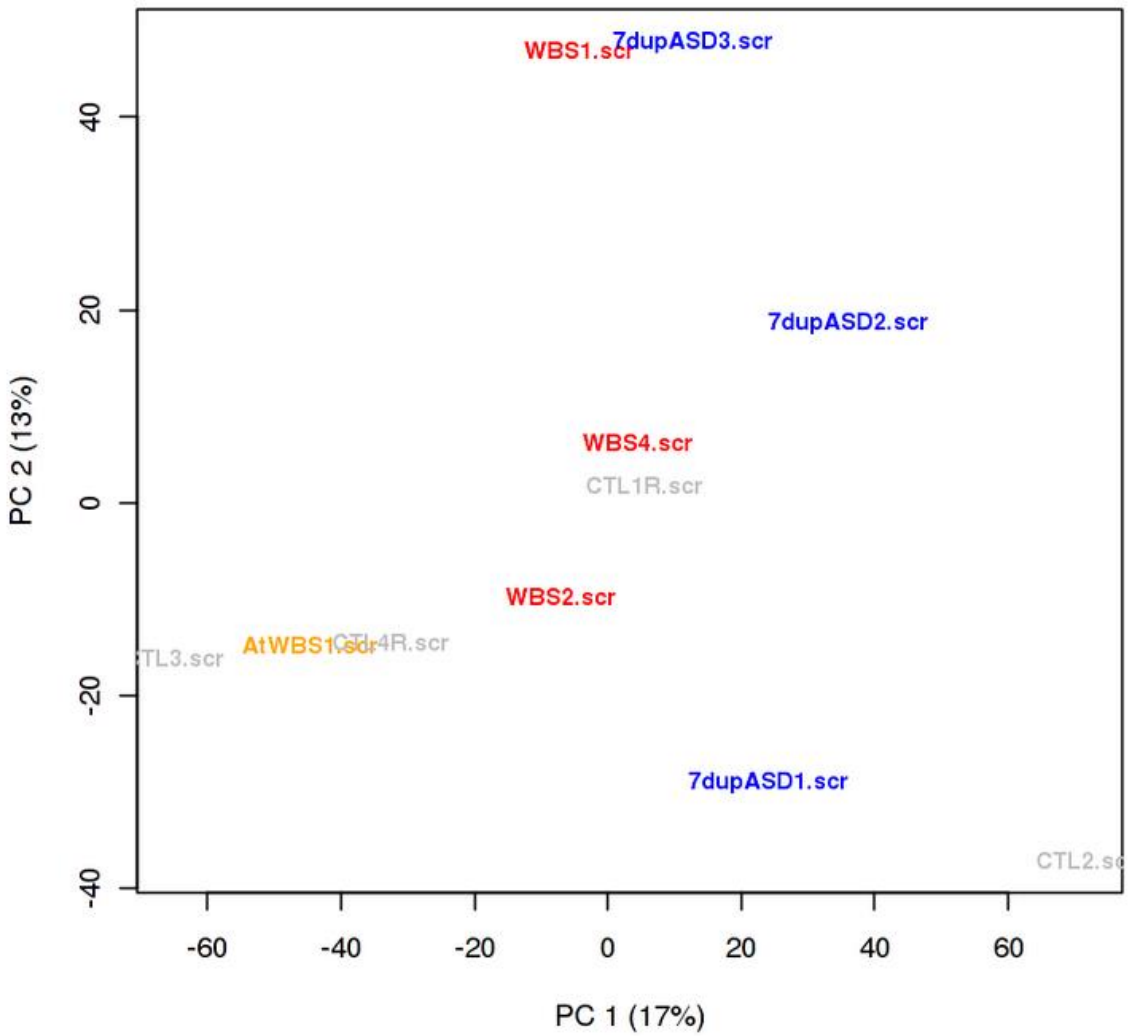


Figure 45. Principal Component Analysis performed on scramble lines for all 11 samples. (calculated on log-normalised read counts)

When I performed pair-wise analyses to identify differences between couples of genotypes (WBS vs CTL, WBS vs DUP, CTL vs DUP), excluding the atypical sample (AtWBS1) and the outlier “CTL2”, I could observe substantial differential expression across genotypes. Moreover, including AtWSB1 transcriptional levels only in the moment of plotting heatmaps of DEGs found with this strategy, I could observe a substantial similarity between AtWBS1 and CTLs, corroborating the idea that genes responsible for neural-crest wise differences between WBS and CTLs are caused by genes spared in the atypical sample (Figure 46).

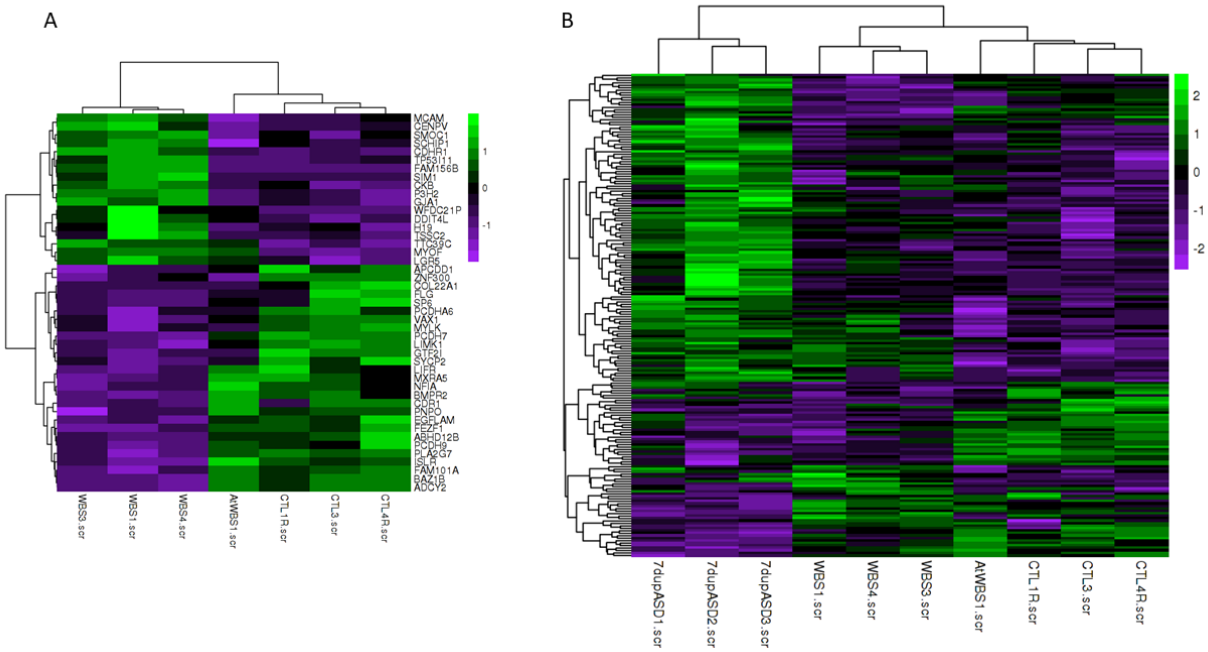


Figure 46. Heatmap of Z-scores measured on log-normalized read counts; A) Genes differentially expressed between WBS and CTLs; B) Union of DEGs found in all 3 paired-genotype comparisons

Doing a principal component analysis including knock-down lines reveals i) a larger occupation of the space produced by plotting the first two principal components ii) KDs clustering close to their shSCR reference and iii) a larger overlap across genotypes (Figure 47).

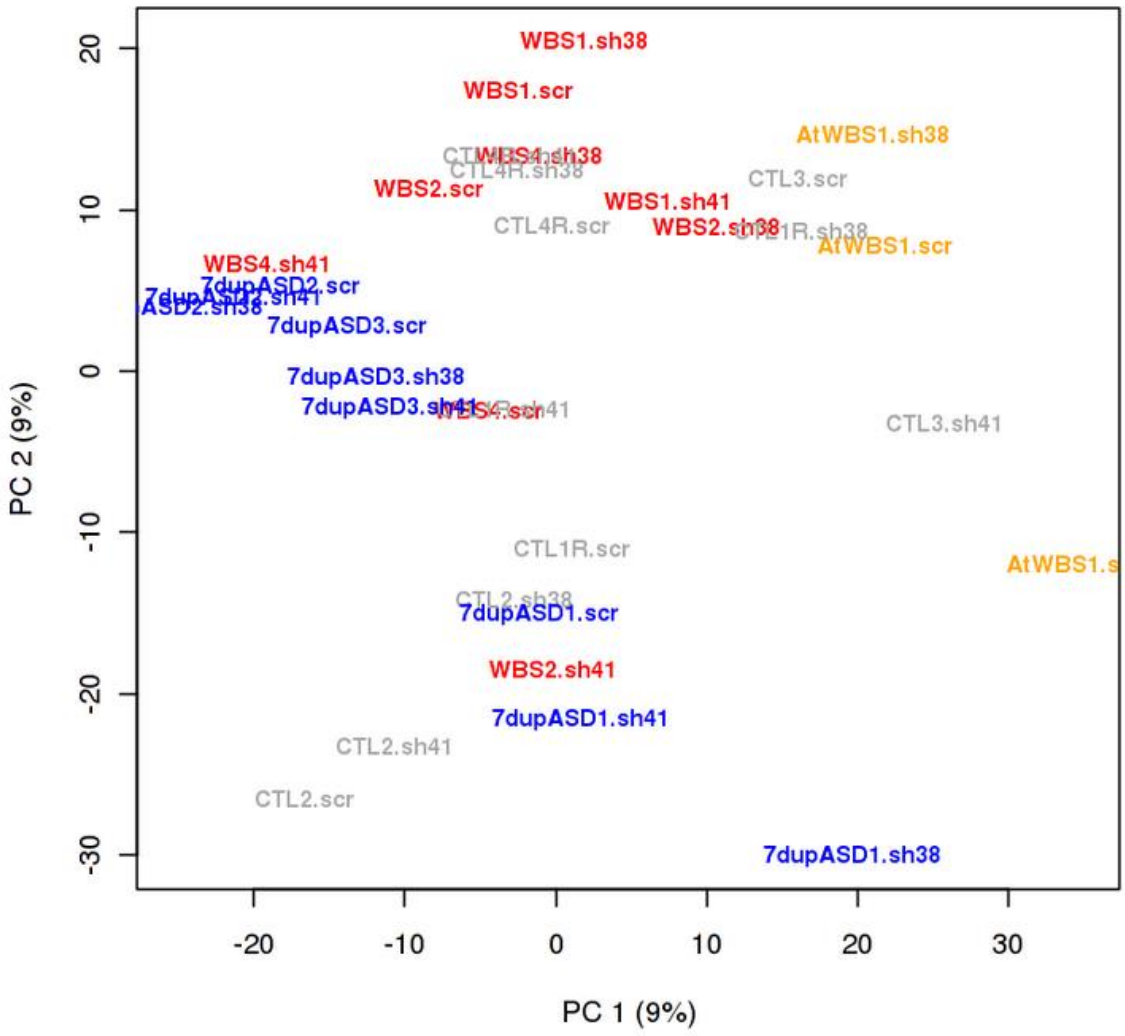


Figure 47. Principal Component Analysis of log-normalised read counts calculated on 32 lines derived from 11 invidiuals (1 scr and 2 KD).

Performing a comparative analysis between KD lines (including in a single group both short-hairpins) and their respective controls (scr) (~individual+knock-down), leads to the identification of 29 DEGs with FDR < 0.1. Notably, in this short list, down-regulated genes include: *MKRN1, RCOR1, LSM6, ENG*, and *HMGB1*. MKRN1 is a peculiar zinc finger with E3 ubiquitin ligase function, so it is both able to bind nucleic acids and to send target proteins towards degradation by proteasome. Indeed, it has been proven to regulate both p53 and p21 by targeting them to the proteasome (Lee et al., 2009) and to have both transactivation and transrepression activity by working in concert with RNA polymerase II (Omwancha

et al., 2006). RCOR1 is the gene coding for CoRest which is recruited by REST to chromatin to regulate neurogenesis genes and determine NSCs fate. LSM6 is involved in splicing regulation. ENG is a gene that gives rise to neurological disorders when mutated: Hereditary Hemorragic Telengiectasia (MalaCards ID: HRD008). HMGB1 is an important chromatin remodeller deleted in 13wq12.3 microdeletion syndrome (which can lead to microcephaly).

Given the paucity of genes identified with a loose significance threshold, I consider the possibility that this analysis design could be subject to the known differences in strength of the two short-hairpins, or that some batch effect could impinge on the ability of differential expression algorithm. By considering library preparation and RNA-extraction history I identified a batch effect (Figure 48), which distinguish as both “operator” and “year” in which experiments had been conducted. Including “batch” to differential-expression model matrix (~batch+individual+knock-down) did not change the results with respect to the previous differential expression analysis.

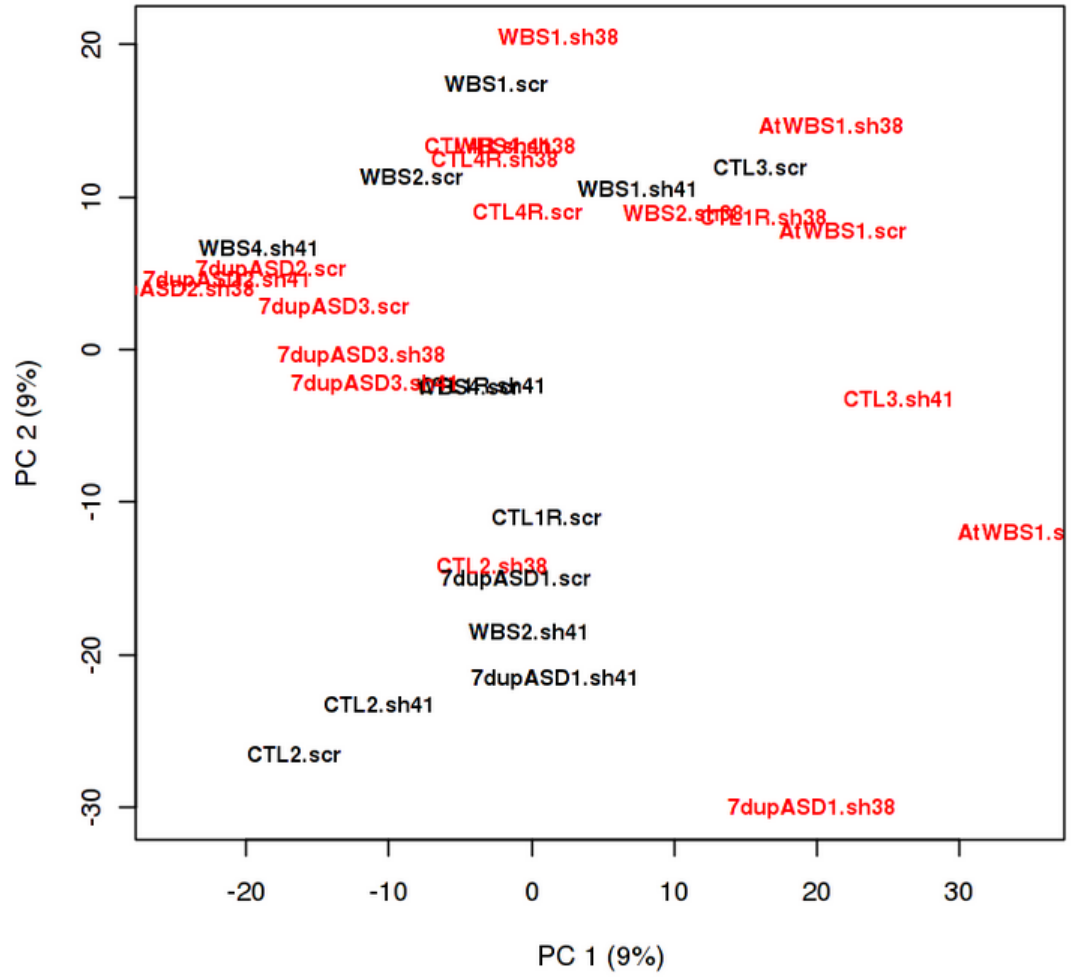


Figure 48. Principal component Analysis of NCSCs RNAseq including WBS, CTLs and 7DupASD interfered with "scr" "sh38" and sh41"; sample coloured by batch

Thus, to overcome potential limits of this canonical analysis I decided to take advantage of the consistent different efficacy of short-hairpins in reducing levels of BAZ1B in all samples, and performed a regression analysis on BAZ1B levels, considering dropping individuals and batch and testing for BAZ1B levels (~batch+individual+BAZ1B). The large number of individuals and the availability of 3 different conditions per individual boosted the power of the analysis (as shown at page 180) and secured the analysis reliability from using a numerical number as independent variable.

This analysis identified a total number of 448 genes with FDR < 0.1 (1192 with FDR < 0.25) whose levels followed *BAZ1B* levels, either in a direct (202; 539 with FDR < 0.25) or inverse (246; 653 with FDR < 0.25) fashion. In addition, the genes identified in the regression analysis alone included most differentially expressed genes (DEGs) found in the previous comparative analysis (Figure 49).

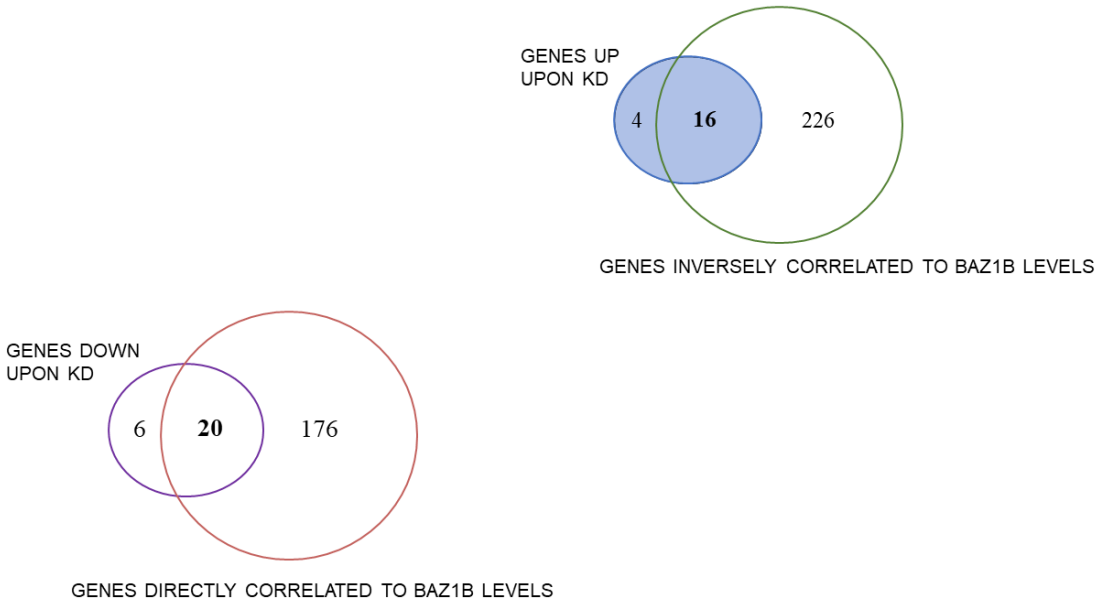


Figure 49. Comparison between paired analysis (~batch+KD) and linear regression on BAZ1B levels (~batch+individual+BAZ1B)

Notably, given the type of analysis, fold-changes refer to the ratio between the considered gene level change and the change observed in BAZ1B (i.e. genes having the same fold-change of BAZ1B across samples will have the same fold-change measured for BAZ1B). Thus, I chose a small fold change threshold and kept the same FDR previously adopted (FC > 1.25, FDR < 0.1) to consider genes in terms of correlation with BAZ1B levels across samples. Among genes inversely correlated with *BAZ1B* levels (i.e. up-regulated upon BAZ1B knock-down) I identified: i) *POSTN* which plays a crucial role in cranial NC-mediated soft palate development, and is responsible for heart valve defects and periodontal disease-like phenotypes in KO mice; ii) *ERBB4* which is important for skeletal muscle

development and NC migration, and causes heart defects and aberrant cranial NC migration in deficient mice.

Genes directly correlated to BAZ1B levels include: i) *OLFM1* whose overexpression promotes an increased and continued production of neural crest cells; ii) *TNFRSF11B* involved in bone resorption and osteoclast activation (target of BACH1, down-regulated in KS and up- in WS, as described earlier in the text) and iii) *CUL3*, which is an essential regulator of early fate determination of cranial neural crest, branching from the CNS developing lineages (Werner et al., 2015).

GO enrichment analysis conducted on directly correlated genes shows specific enrichments for RNA processing and splicing (which includes LSM6), while inversely correlated genes show enrichments for biological processes particularly relevant for NC and NC-derivative function, such as cell migration and cardiovascular and skeletal development (Figure 50).

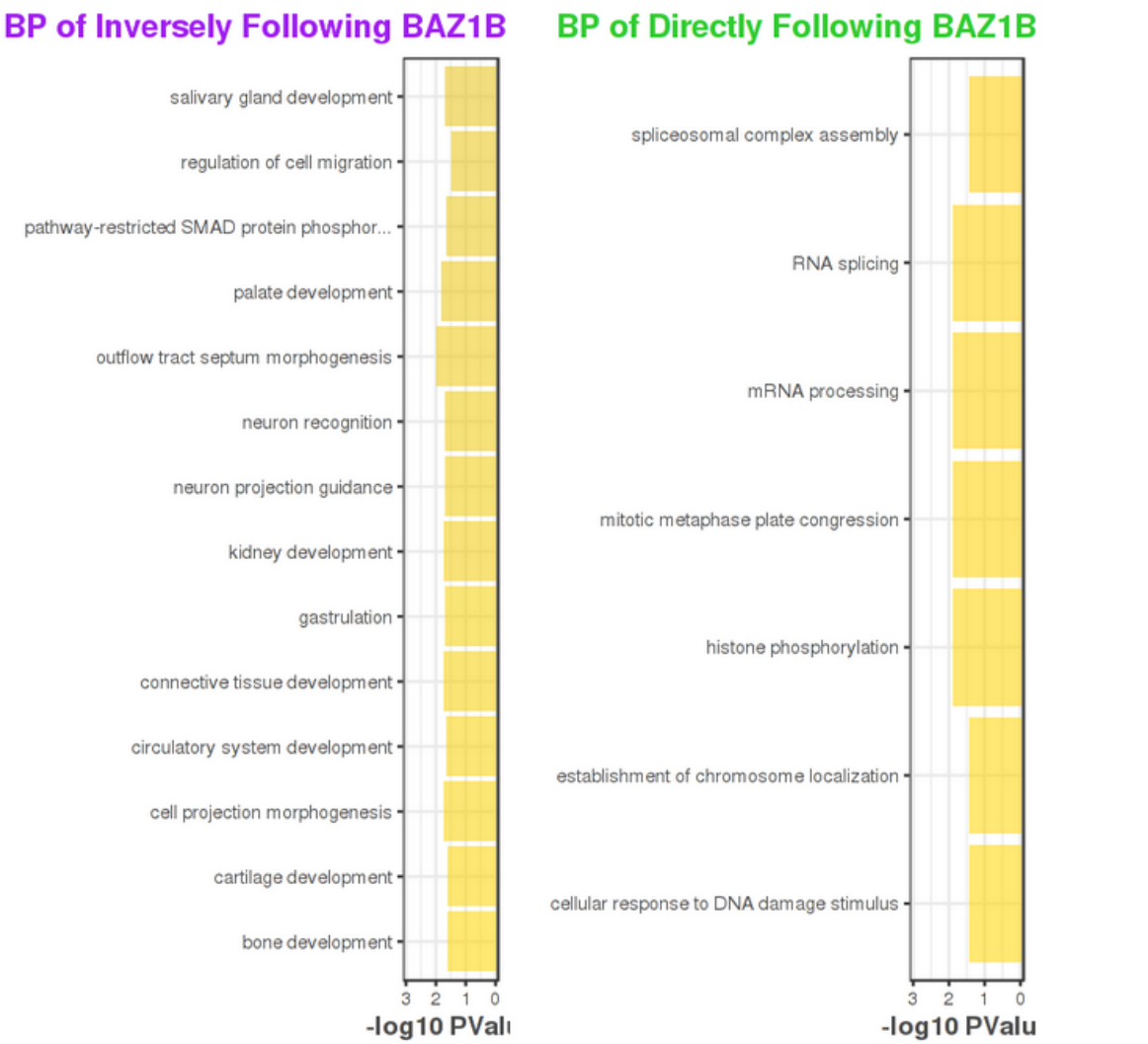


Figure 50. GO biological processes enriched by genes following BAZ1B levels.

Interestingly, intersecting DEGs with OMIM with omimCrawler (see page 184) I found several genes responsible for genetic disorders (Figure 51), which phenotypes were including keywords such as “mental retardation”, “intellectual disability”(Figure 52) and/or “facial dysmorphisms” (Figure 53).

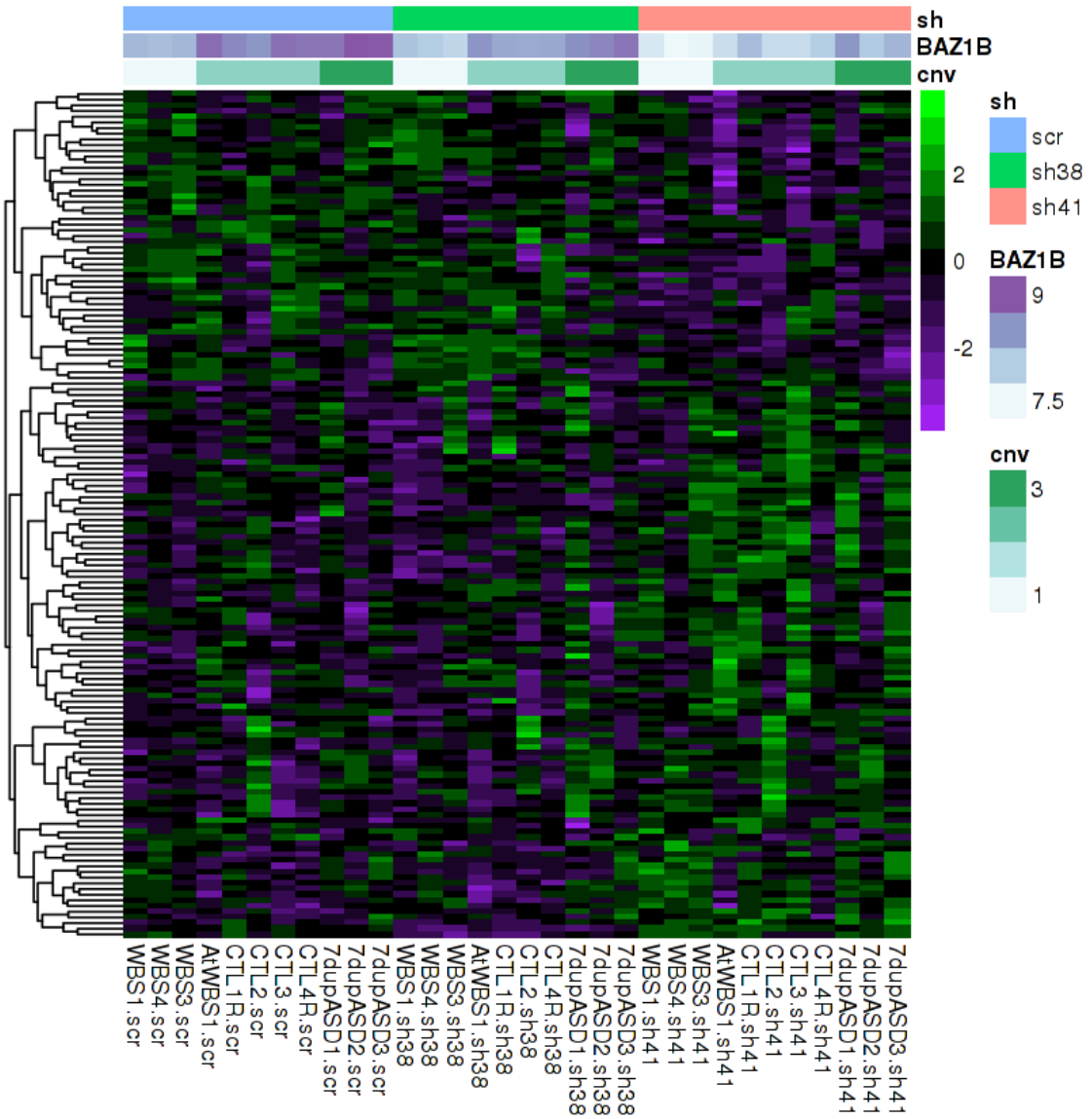


Figure 51. OMIM disorders genes following BAZ1B levels across NCSCs interfered with sh38 and sh41 (BAZ1B KD).

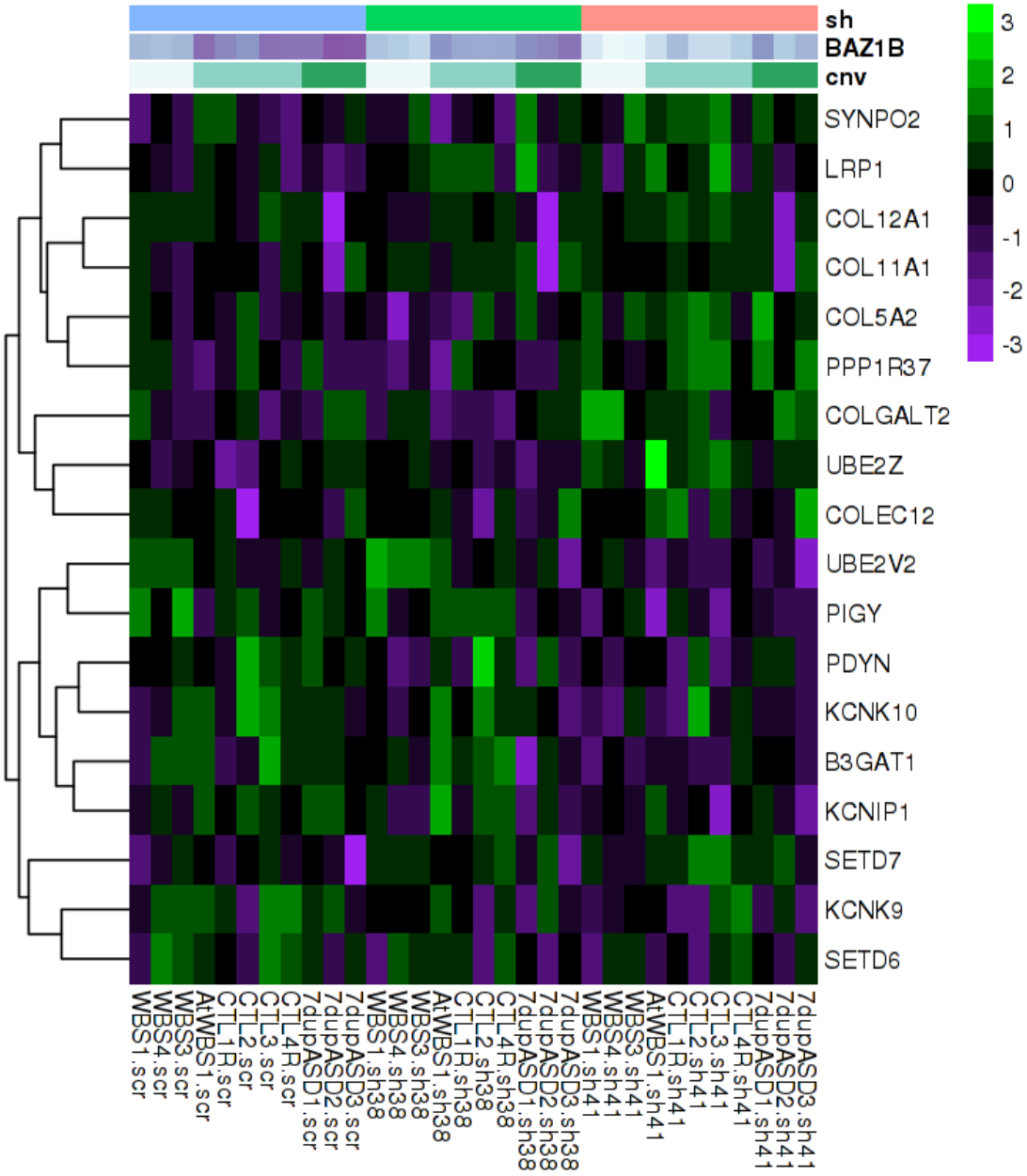


Figure 52. Heatmap of genes associated with "mental retardation" or "intellectual disability" in OMIM. (z-scores of log-normalized read counts) (in the sh legend blue refers to “scr”, green sh refers to “sh38” and pink refers to “sh41”)

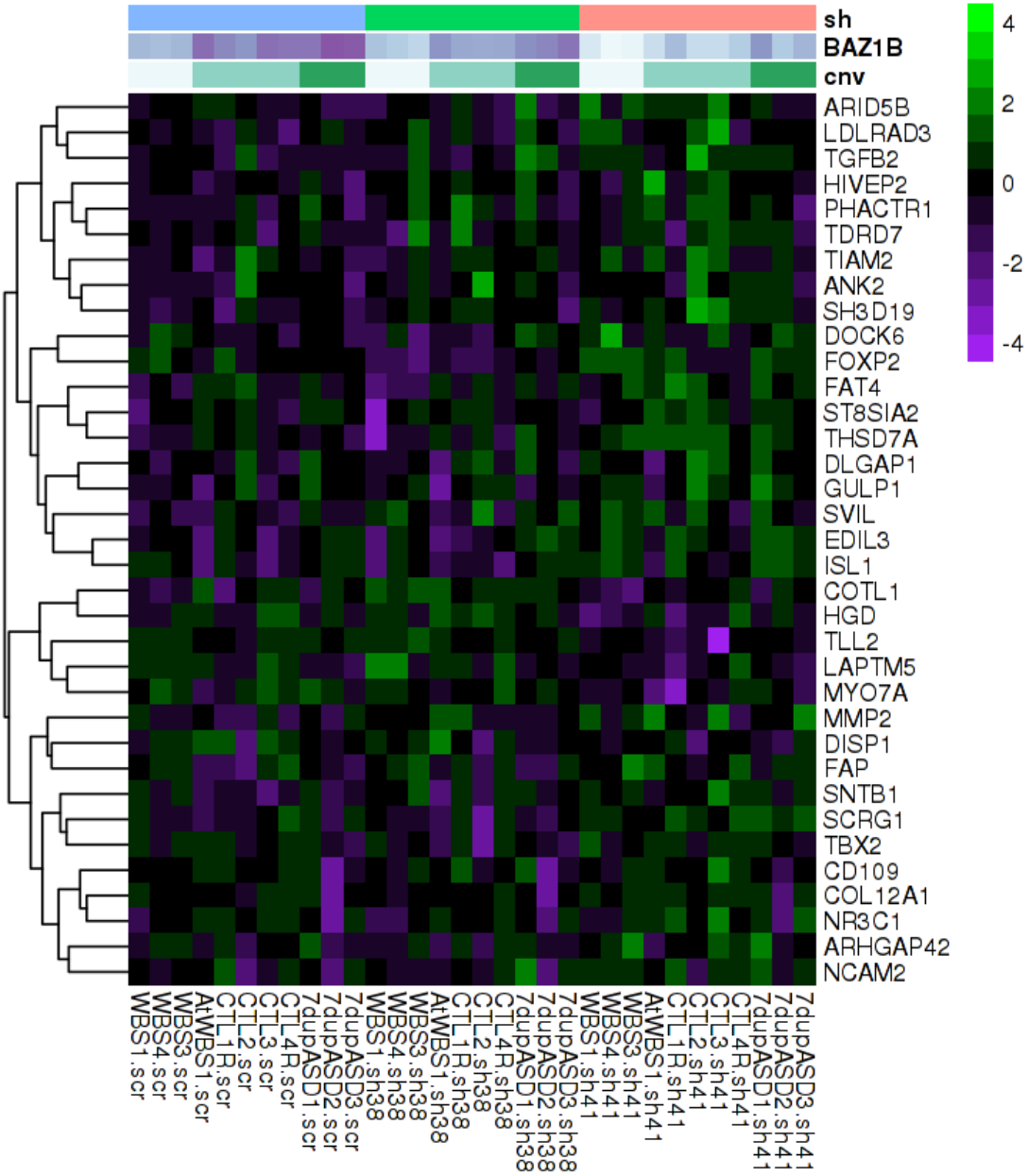


Figure 53. Genes following BAZ1B levels in NCSCs, associated with Facial morphogenesis and dysmorphisms in OMIM.

Moreover, a Master Regulator Analysis identified several putative regulators of genes following BAZ1B levels (elsewhere also “BAZ1B DEGs”), including factors involved in enhancer mark modification (SEBP2, p300, RBBP5, HDAC2, KDM1A and TCF12), chromatin remodelling (CTCF, YY1 and RAD21) and promoter activators (TBP, TAF1 and POL2). Among such defined domains of transcription regulations, chromatin remodelling largely appears as the most enriched.

Several molecular functions have been assigned to BAZ1B, as described in the introduction, ranging from chromatin marking of DNA damage sites, to transcription regulation, to chromatin remodelling. The balance between numbers of genes up- or down-regulated upon BAZ1B knock down, together with the greater overlap - and sheer size and significance - of enrichments in chromatin remodelling with respect to the other identified domains of transcription regulation, corroborates the inclusion of BAZ1B in the domain of factors acting upstream enhancer and promoter modulation, but also confirms a modularity in its action (Figure 54).

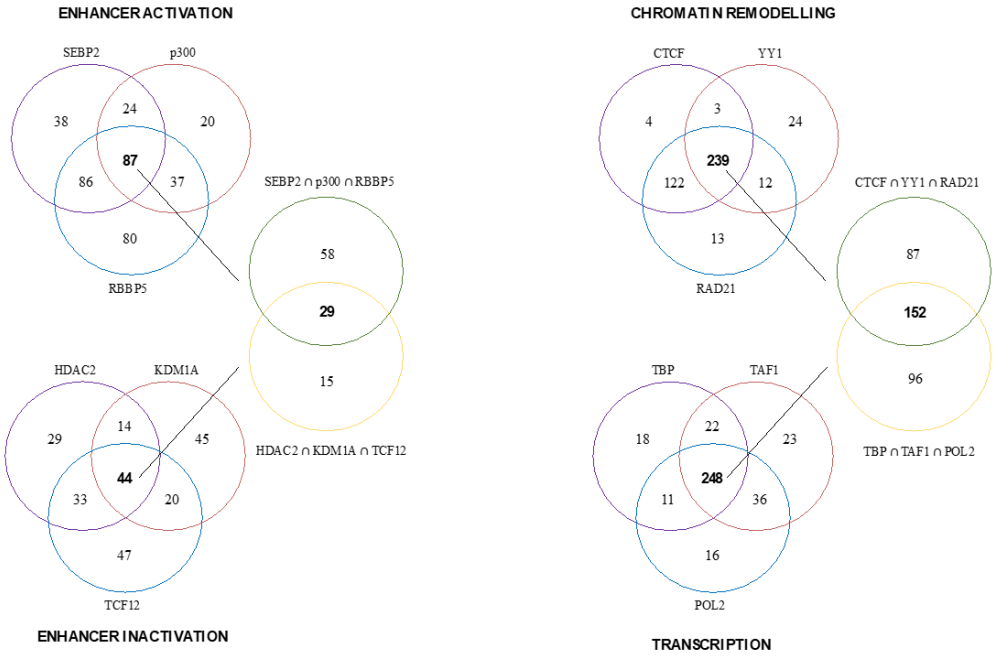


Figure 54. Venn diagrams of TF enrichments of genes following BAZ1B levels

Moreover, among transcription factors enriched for genes following BAZ1B levels I found also two DEGs: *EGR1* and *MXI1* (both inversely correlated with BAZ1B levels and significantly enriched as master regulator of directly correlated BAZ1B DEGs). Notably, among MXI1 targets I found *TGFB2* and *NFIB*, which have already been associated to intellectual disability and craniofacial defects. Moreover, among genes following BAZ1B levels I identified *TBX15*, whose mutation has also been recently associated with cranio-facial features (Claes et al., 2018), *AMMECR1*, which is

associated with midface hypoplasia and hearing impairment, *ZMYND11*, which is an H3K36me3 reader that, when mutated, gives rise to facial features and mild intellectual disability. Other genes associated with craniofacial features in OMIM and included in BAZ1B DEGs are: *FOXP2, TARDBP, MFN2, SH3TC2, FBLN5, KIF1A, HECW2, FAT4* and *DST*.

Characterisation of BAZ1B dependent chromatin landscape in neural crest

In order to define BAZ1B direct targets and to investigate, at the same time, changes in their chromatin states at both promoter and enhancer regions upon KD, we performed ChIP-seq of BAZ1B and histone marks. Given the absence of ChIP-seq grade antibody, we resorted to a targeted genome editing strategy: by taking advantage of the CRISPR/Cas9 system we added a 3xFLAG tag to endogenous BAZ1B gene sequence in iPSCs (Fig. 3A). We obtained one homozygously tagged iPSC line for three samples (1 WBS, 1AtWBS and 1 control) and a heterozygously tagged 7DupASD. Such lines were then differentiated to NCSCs and ChIP-seq was performed against the tag (Figure 55).

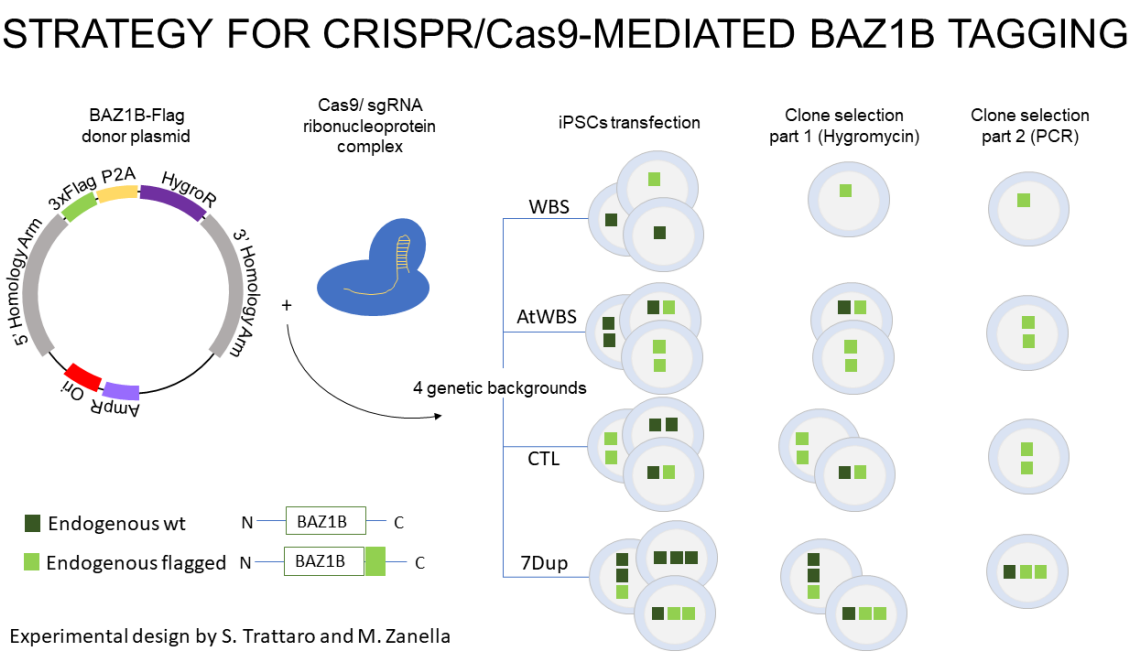


Figure 55. Strategy for CRISPR/Cas9 mediated BAZ1B triple tagging

ChIP-seq libraries have been produced using our sequencing facility robot. Quality controls, provided from the sequencing facility, highlighted scarce enrichments and low coverage of DNA libraries produced by our lab (resulting in ~20M single-end reads aligned out of 35M per each sequencing and 20% PCR duplicates). Thus, to proceed with peak calling using MACS2, I resorted to aggregation of reads from all sequencing experiments. After aggregation I obtained an estimated coverage of 150M reads (two sequencing replicates per genotype). Peak Calling on such regions was performed imposing a specific fragment size that was measured experimentally before library preparation and resulted in the identification of ~23 thousand regions bound by BA1B with FDR < 0.25. On such a set of chromosomal locations, I performed a quantitative analysis with DeepTools, to count reads per region in each experiment (in a sample-wise fashion). PCA analysis showed a clear separation of samples based on BAZ1B copy-number (independently from the number of flagged copies) with CTL and AtWBS samples clustering together and WBS and 7Dup

samples occupying opposed position with respect to the control. Here I report both PCAs performed on the separate sequencing replicates and on the genotype-wise aggregated one. The former shows a smaller variability on PC2 which is also correlated with copy-number variation.

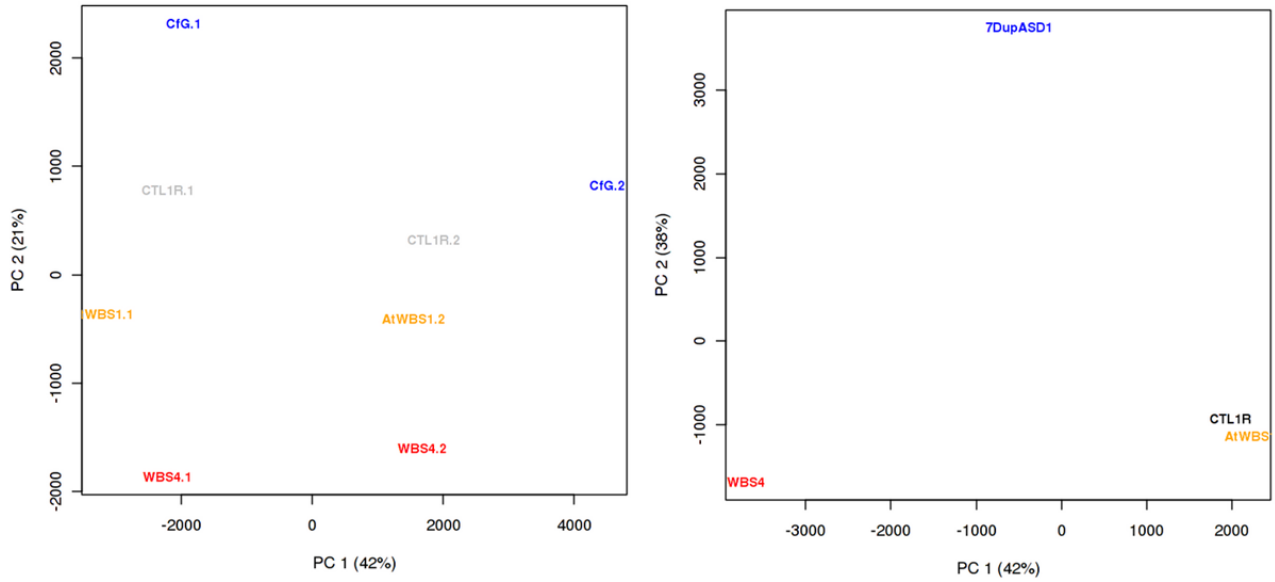


Figure 56. PCA of BAZ1B ChIPseq quantitative analysis on 23 thousand regions bound by the protein. On the left panel: all sequencing runs separated (WBS in red, AtWBS in orange, 7DupASD in blue); On the right panel: PCA of genotype-aggregated samples

Genes bound by BAZ1B, and expressed in our cohort of neural crest stem cells, are enriched for several BP terms such as “axon guidance”, “tube development”, “dendrite development”, “outflow tract morphogenesis”, “odontogenesis”, “wound healing”, “endochondral bone morphogenesis” (Figure 57)

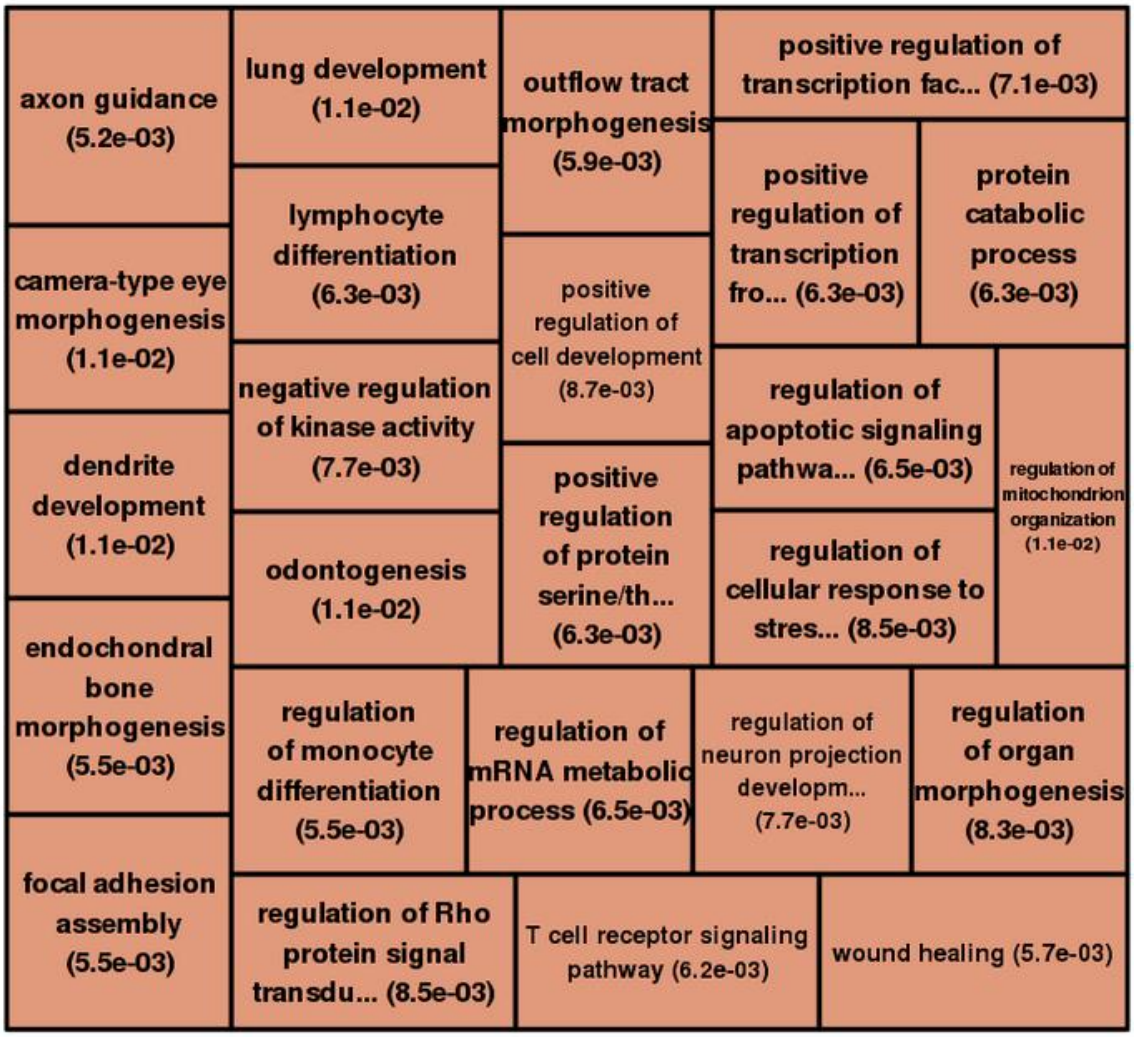


Figure 57. Go enrichments in Biological Process for Genes bound by BAZ1B and expressed in neural crest.

To narrow down the role of BAZ1B in the context of gene regulation and especially of genes following its levels across genotypes, I defined NC-specific enhancer regions focusing on the distribution of H3K4me1, H3K4me3, H3K27ac and H3K27me3 marks. To annotate enhancers, I identified regions showing H3K4me1 and H3K27ac in at least two samples, excluded regions marked with H3K4me3 (to exclude promoters) and associated that region to the closest TSS (using the same hg38 genome used for RNA-seq quantification). Notably, BAZ1B binds 75% of its putative targets at their enhancer regions. More precisely, it binds 6747 genes at enhancers, and 2297 at promoters. The total number of putative enhancers was

much larger of the promoters one (~300 thousands vs ~15 thousand), so I considered gene numbers to make a more even comparison. Thus, globally, ~40% of genes expressed in NC are bound by BAZ1B at regulatory regions, 27.4% only at enhancers, 3.5% only at promoters, 9% at both (Figure 58).

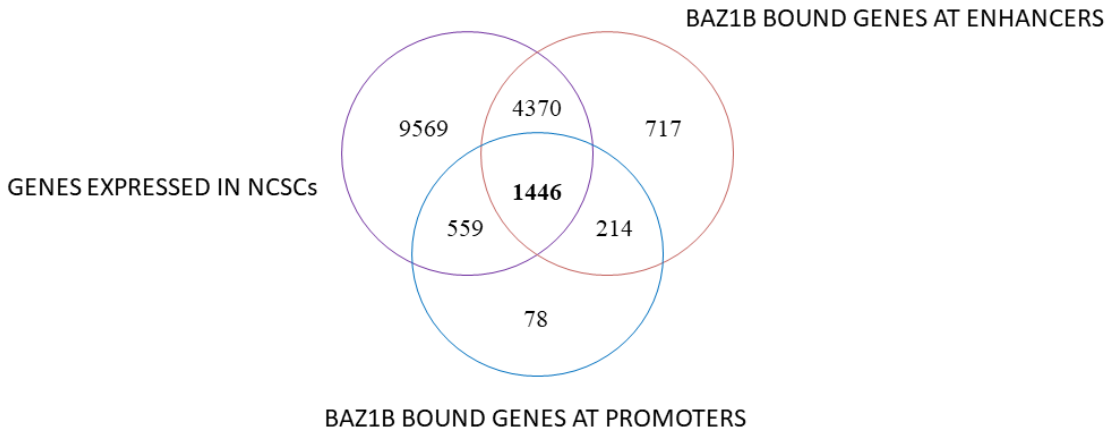


Figure 58. Venn diagram of BAZ1B bound genes at regulatory regions with respect to genes expressed in NCSCs

Moreover, I performed a motif enrichment on BAZ1B bound regions by means of HOMER (Heinz et al., 2010), which identified ~18% of them featuring one TFAP2A (AP2) like motif and another motif similar to the TFAP2A one, accounting for ~8% of BAZ1B bound regions. These similarities are relevant because AP2 has already been described as crucial for neural-crest specification and is associated with BOFS syndrome (Figure 59). Notably, a significant enrichment is found also for NGN2 motif (NEUROG2 in the figure text). Moreover, TAL1 motif (SCL), was the one aligning

exactly to BAZ1B one, but the same motif shows a strong alignment with Tcf12 and Ascl1 ones.

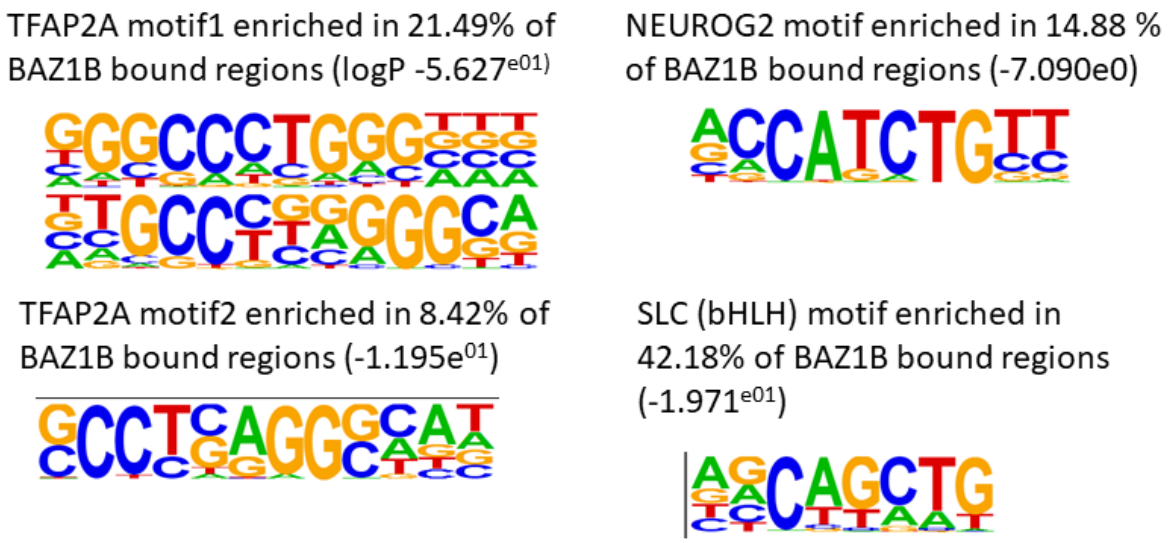


Figure 59. Relevant motif enrichment for BAZ1B bound regions identified with HOMER. In the case of TFAP2A "motif1", the alignment between TFAP2A (above) motif and BAZ1B one (below) is reported

Among regions following BAZ1B genetic copy-number, I identified ~200 regions, differentially bound by BAZ1B which will require further investigation (data not shown).

To characterise the extent of enhancer remodelling induced by interfering with BAZ1B levels in neural crest stem cells, I performed a regression analysis on BAZ1B levels for the distribution of the three histone marks in the above identified enhancer regions (a separate ~individual+BAZ1B analysis for histone marks H3K4me1, H3K27ac and H3K4me3). Interestingly, a high number of genes (7254) are associated with putative enhancers differentially marked in terms of H3K27ac, followed by genes associated with differentially marked regions of H3K4me1 (4048) and H3K27me3 (2136).

Within the 1192 DEGs identified upon regression of BAZ1B levels in the RNA-seq analysis, 21.3% (257/1192) are both bound by BAZ1B at their enhancers and have a concordant differential distribution of the H3K27ac mark at enhancers. Among

them, I noticed a more solid overlap for the genes whose expression resulted inversely correlated with BAZ1B levels (160 vs 97). The same observation can be done for DEGs that have a concordant differential distribution of the H3K4me1 mark at enhancers (123 vs 55), further supporting a stronger role for BAZ1B in remodelling at the level of distal enhancers/regions. A lower number of genes (about 30) were instead showing concordant differential expression and distribution of H3K27me3 marks at enhancer while these were bound by BAZ1B.

Thus, considering the limits of sequencing coverage, at least 30 genes i) are differentially expressed following BAZ1B levels (13 direct + 17 indirect), ii) have their enhancers differentially marked concordantly by all three marks (H3K27ac, H3K4me1 and H3K27me3) and iii) are bound by BAZ1B at enhancers (Figure 60).

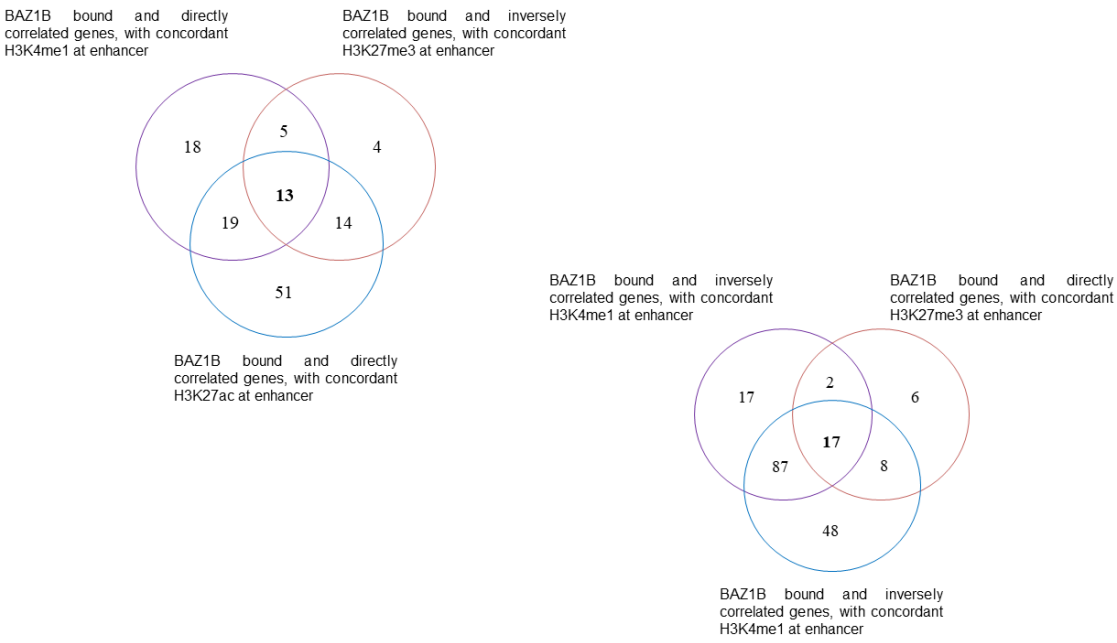


Figure 60. Venn diagrams of the intersections of genes bound by BAZ1B, differentially expressed following BAZ1B levels and differentially marked at enhancers

BAZ1B in the Neurocristic Axis

Having identified the transcriptional landscape of BAZ1B regulation, and recognized that ~9% of genes in the neural crest follow BAZ1B levels (~1200 out of ~13000 expressed in all samples with FDR < 0.25), I decided to do a linear regression analysis on BAZ1B levels along the Neurocristic Axis (~tissue+individual+BAZ1B), dropping tissue and individual, and using iPSCs as intercept. With this analysis I identified 454 genes following BAZ1B levels along NC Axis development with FDR < 0.05 and FC > 1.5. Intersecting those genes with the 12 modules of transcription previously reconstructed, I found several genes in each of them (Table 10). Considering that not all modules showed strong TF enrichments in the previous analysis (CL4 and CL7-11), this result poses BAZ1B up-stream most clusters down-regulated along the Neurocristic Axis (CL7,CL9,CL10 and CL11); and at the apex of Neurocristic Axis regulation.

Table 10. Number of genes following BAZ1B levels along Neurocristic Axis development in each of the 12 modules of developmental-wise expression identified previously for the same axis

CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CL9	CL10	CL11	CL12
15	19	8	10	19	13	20	11	26	35	18	23

Finally, given gathered evidences of BAZ1B potentially regulating several modules of transcription in the Neurocristic Axis, and having identified, among BAZ1B DEGs in NCSCs significant enrichments for GO categories associated with cranio-facial features and intellect, our lab developed a collaboration with Cedrick Boeckx. His lab is studying, among other genes and NDDs, BAZ1B and Williams Beuren Syndrome as a peculiar window onto the theory of self-domestication

(Theofanopoulou et al., 2017). This theory states that human facial traits have evolved along evolution in parallel with cognitive functions. This observation pairs with the fact that animal domestication inadvertently selects for deficits in animal neural crest induction (Wilkins et al., 2014). Thus, we crossed genes that have been deemed regulated by BAZ1B, and in particular those following its expression both at transcriptional level and with at least one enhancer mark, with genes modified by human evolution. A significant 5% overlap (p-adj < 0.01) between those genes and genes inversely correlated with BAZ1B levels in NCSCs was identified (Figure 61) and will be further characterised in the immediate future of this thesis writing. “Testa lists”, top-down refers to 7 lists of genes inversely correlated to BAZ1B levels (INV), and 7 of genes directly correlated to BAZ1B levels (DIR). More in details, 17 INV with all 3 enhancer marks concordant with expression changes; 25 with both H3K27ac and H3K27me3 concordant; 33 with only H3K27me3 concordant; 123 with concordant H3K4me1; all genes INV with FDR < 0.1, and FDR < 0.25. The same scheme is followed for DIR genes.

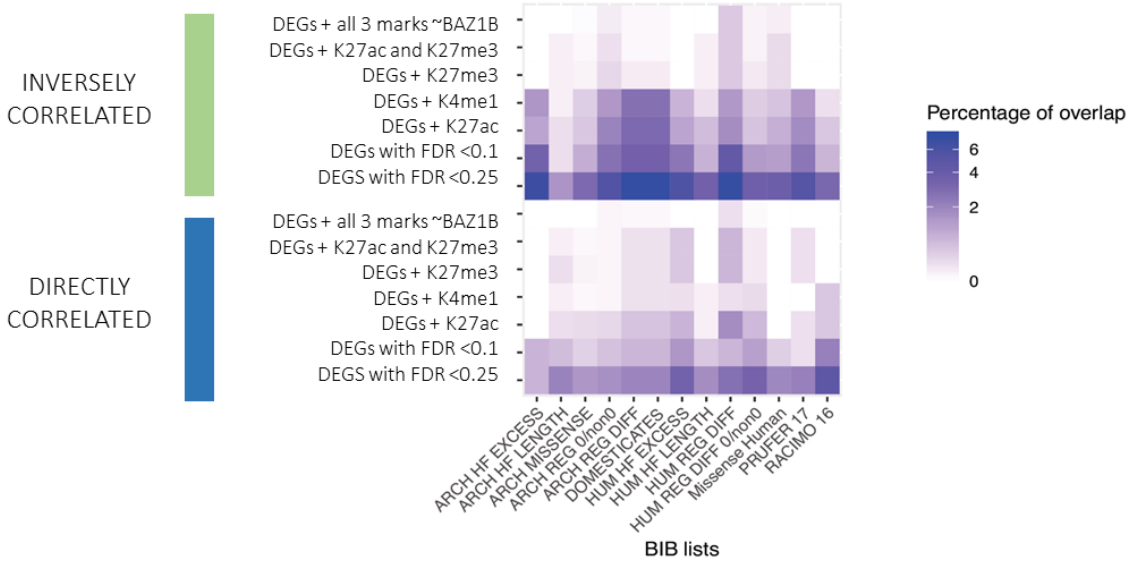


Figure 61. Heatmap of the intersection (made by Boeckx’s lab) of genes associated which mutations along evolution have been associated to face morphogenesis and human intellect (BIB lists) with Lists produced by my reconstruction of BAZ1B regulatory networks in NCSCs

Modelling of Cerebral Cortex Axis

In the introduction I described the cerebral cortex axis as the proxy to identify transcriptional deregulations associated with both mature and functional neurons and with early and mid-gestational stages of human cortical development. To do so I first characterized cortical neurons obtained with the NGN2 protocol, and analysed transcriptional deregulations in KS and WS, taking advantage of RNA-seq data of 8 control, 4 KS and 4 WS lines. In a similar manner to what was observed in NCSCs, upon data filtering for genes expressed in at least 3 samples (minimum number of read counts per gene = 20), Principal Component Analysis highlights a clear batch effect (Figure 62A). Focusing on each batch I observed, on the one hand, a sizeable distance between controls and KS samples, on the other hand a coarse overlap between controls and WS samples. The first batch (on the left in PCA) includes samples from the previously mentioned “orange” team while second batch comes from “purple” team (Figure 36).

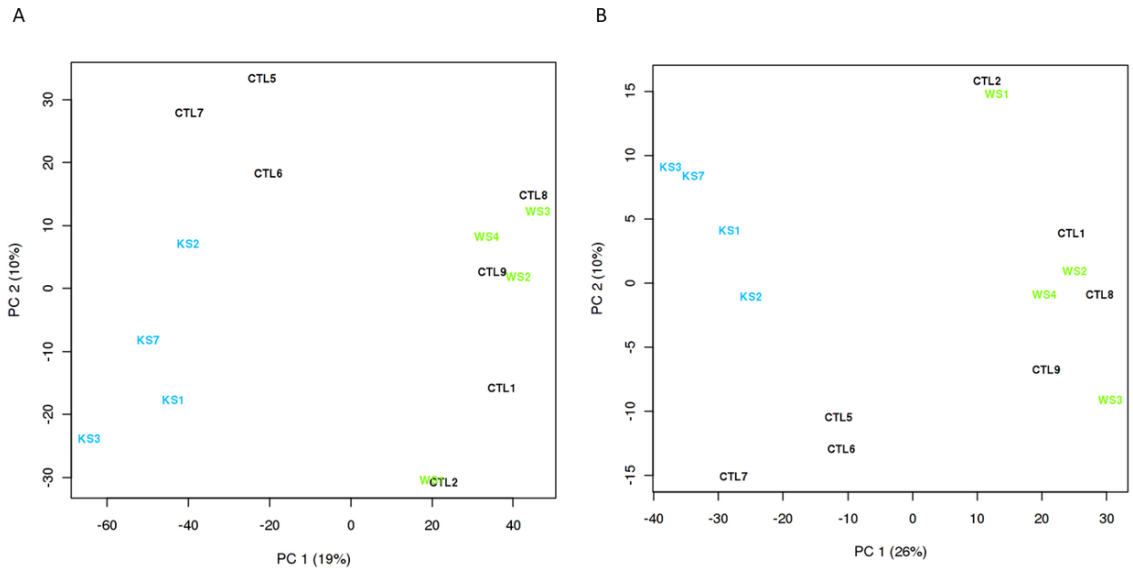


Figure 62. Principal component analysis of NGN2 RNA-seq data showing a tractable batch effect. A) PCA performed on data filtered for genes expressed in at least 3 samples with at least 20 read-counts; B) PCA after filtering for genes expressed in all samples.

Using stronger filtering (genes expressed in all samples, minimum read counts =35) I obtained an apparent reduction in distance between controls of the two batches and an even greater distance between KS and all the other samples (Figure 62B). Also in this case, to avoid filtering too many genes, and eventually overestimating differences which are more likely due to technical reasons than to effective genotype-dependent distances, I decided to work with the larger dataset (the less filtered), and to proceed with a multifactorial analysis with a design including batch as a dependent variable to discard, and an independent variable representing the different genotypes to be tested for (~batch+Genotype). As anticipated by the PCA, KS samples accounted for a higher difference than WS with respect to controls: 665 genes were differentially expressed in KS and 153 in WS (FC > 1.25, FDR < 0.05). Notably, considering separately the two batches, I could find 1338 DEGs in KS and 56 in WS (data not shown). Genes differentially expressed in KS in both analyses included most of the 665 identified here. The 56 genes deregulated in WS (not shown) were all present in the analysis here reported. I considerd this observation

as the minimum proof that the analysis including all samples was more robust, and further proceeded in the characterisation of disease-specific DEGs identified in the complete cohort. Selecting genes deregulated in both disorders I could detect 44 genes down-regulated in both disorders, 56 down- in KS and 46 up-in KS. No genes are consistently down-regulated in both disorders in NGN2 neurons. Crucially, among genes up-regulated in KS appears KDM6A (FC = 2, FDR < 0.0001), suggesting a potential compensatory mechanism within the Compass Complex to correct of KMT2D haploinsufficiency.

KS specific DEGs were enriched for GO biological process categories such as “behaviour”, “central nervous system neuron differentiation”, “sodium ion

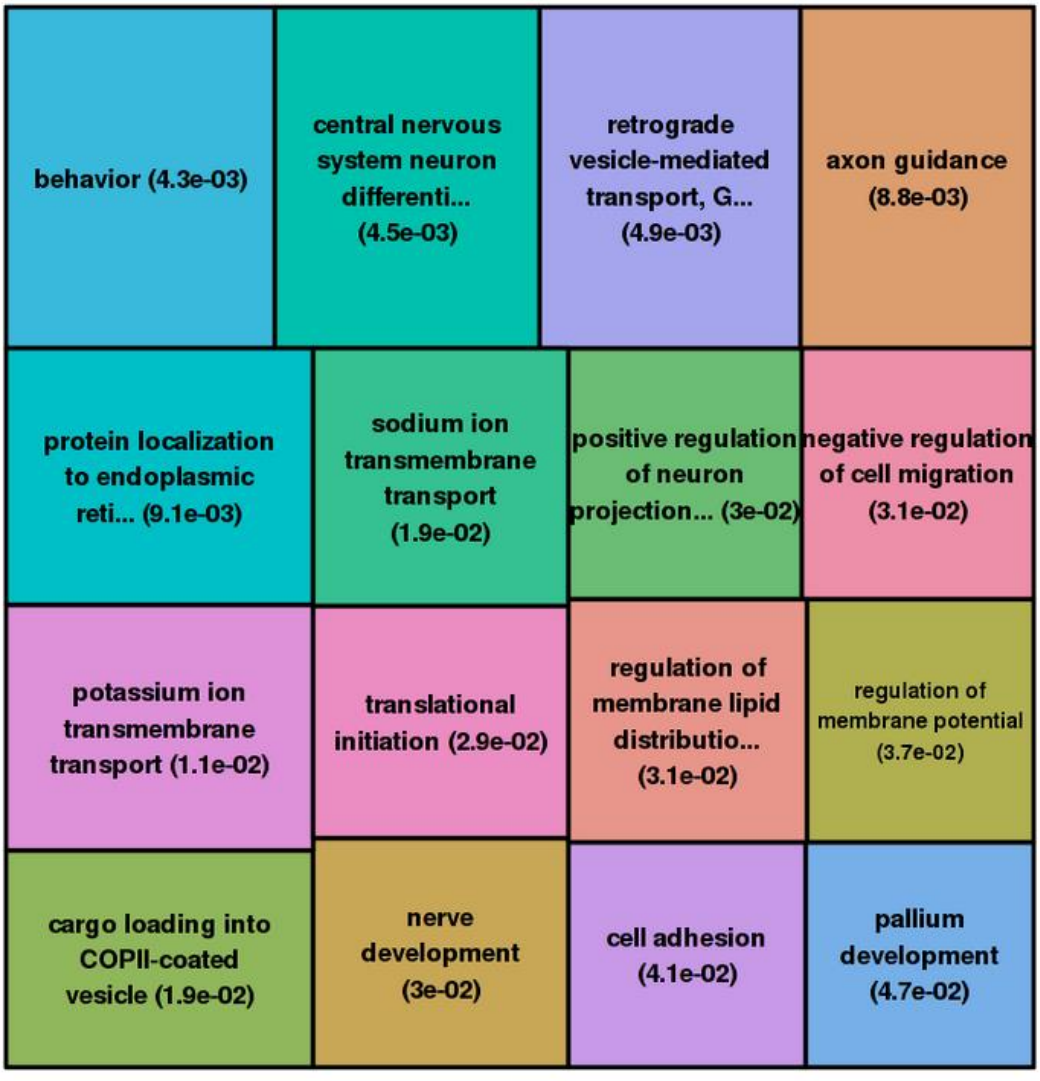


Figure 63. GO Biological process enrichments for Kabuki Syndrome DEGs in NGN2 neurons

transmembrane transport", "translational initiation", "nerve development", "pallium development", "regulation of cell migration", "Golgi vescicle transport" (FDR < 0.05) (Figure 63).

Together with a treemap of GO BP enriched categories (devoid of children categories) I report a heatmap of 154 genes with highest enrichments in non-redundant categories (Figure 64), which include "sodium ion transport", "behaviour", "axonogenesis", "central nervous system neuron differentiation", "central nervous system development" and "Golgi vescicle transport".

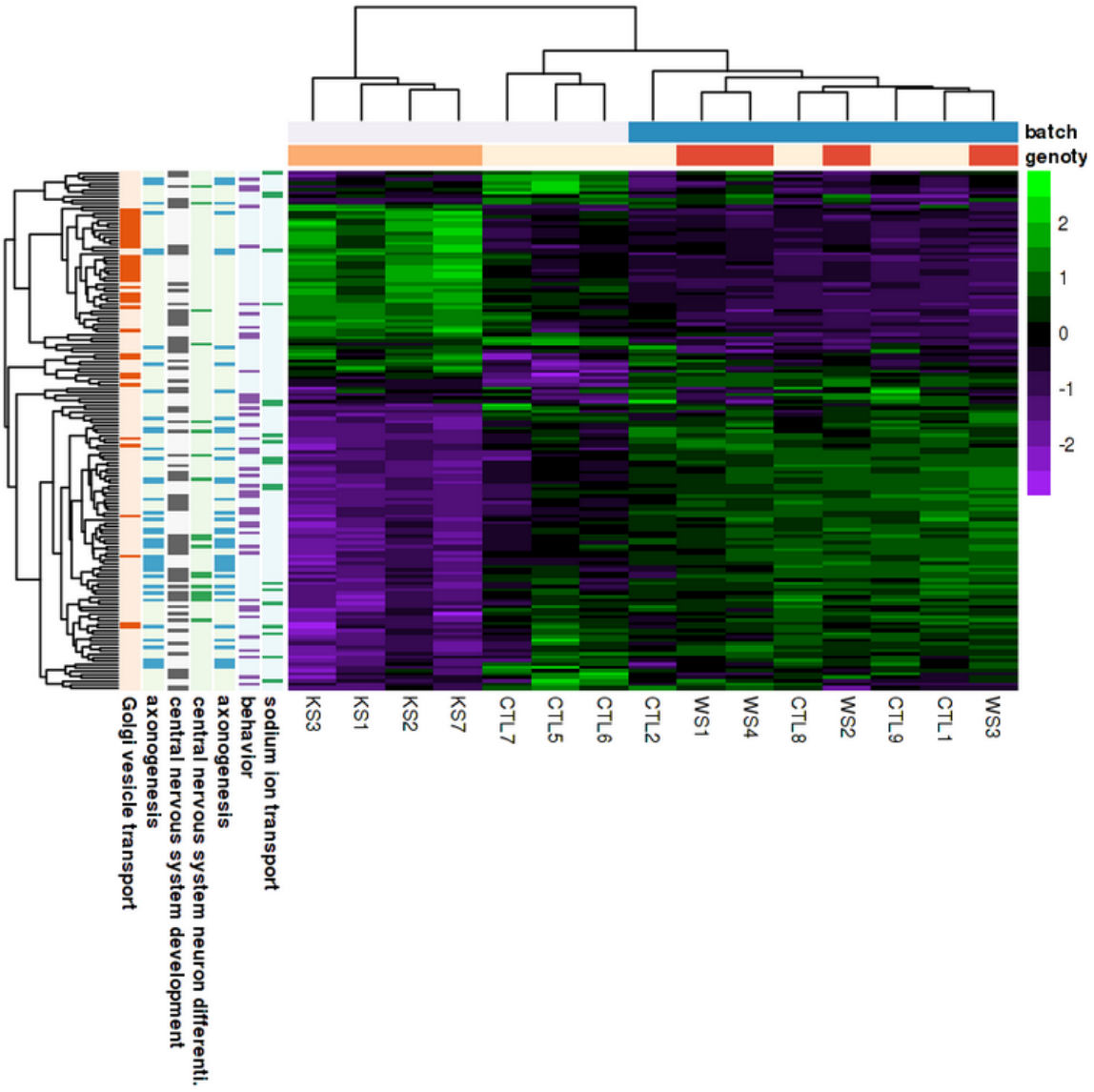


Figure 64.GO annotated heatmap of 154 KS differentially expressed genes in NGN2

These genes are mostly down-regulated, particularly those for “sodium ion transport”, “axonogenesis” and “behaviour”. “Golgi vescicle transport” appears to be mostly up-regulated.

Differentially expressed genes in WS show only two GO enrichments, most of them were down-regulated and characterised by opposite sign with respect to KS shared DEGs. A heatmap (of z-scores measured gene-wise on log-normalized read counts) shows major differences across batches but, with the help from row clustering, it is possible to recognise: *CST3, MYO7A, SFRP1, EGFR*, and *MCOLN3* down-regulated in WS and up- in KS, PAX8 up-regulated in WS and down- in KS, TBX2 and BMP7 down-regulated in WS. These genes enriched the GO terms “sensory organ development” and “G-protein coupled receptor signalling pathway” (FDR < 0.05) (Figure 65).

CST3 is a cysteine protease inhibitor, highly concentrated in body fluids, which mutation has been associated with cerebral amyloid angiopathy (OMIM #105150). *MYO7A* is a member of myosin family, associated with Usher Syndrome (OMIM #276900), which typical traits are reduced vestibular functions and hearing deficits but also visual impairment, depression and intellectual disability. *SFRP1* is associated with Wnt pathway signalling (Üren et al., 2000). In fact, it owns a cysteine-rich domain which is homologous of Frizzled proteins Wnt-binding site. *EGFR* is a well characterised growth factor receptor. *MCOLN3* mediates calcium release and its mutations (in mice) are associated with balance problems and hearing loss (OMIM *607400). TBX2 is a DNA binding domain associated with language impairments, developmental delay and growth retardation (OMIM *600747). PAX8, which expression is deregulated in both disorders in opposite direction is an important transcription factor, almost exclusively expressed during

development, which has been associated with cancer development and hypothyroidism (OMIM *167415).

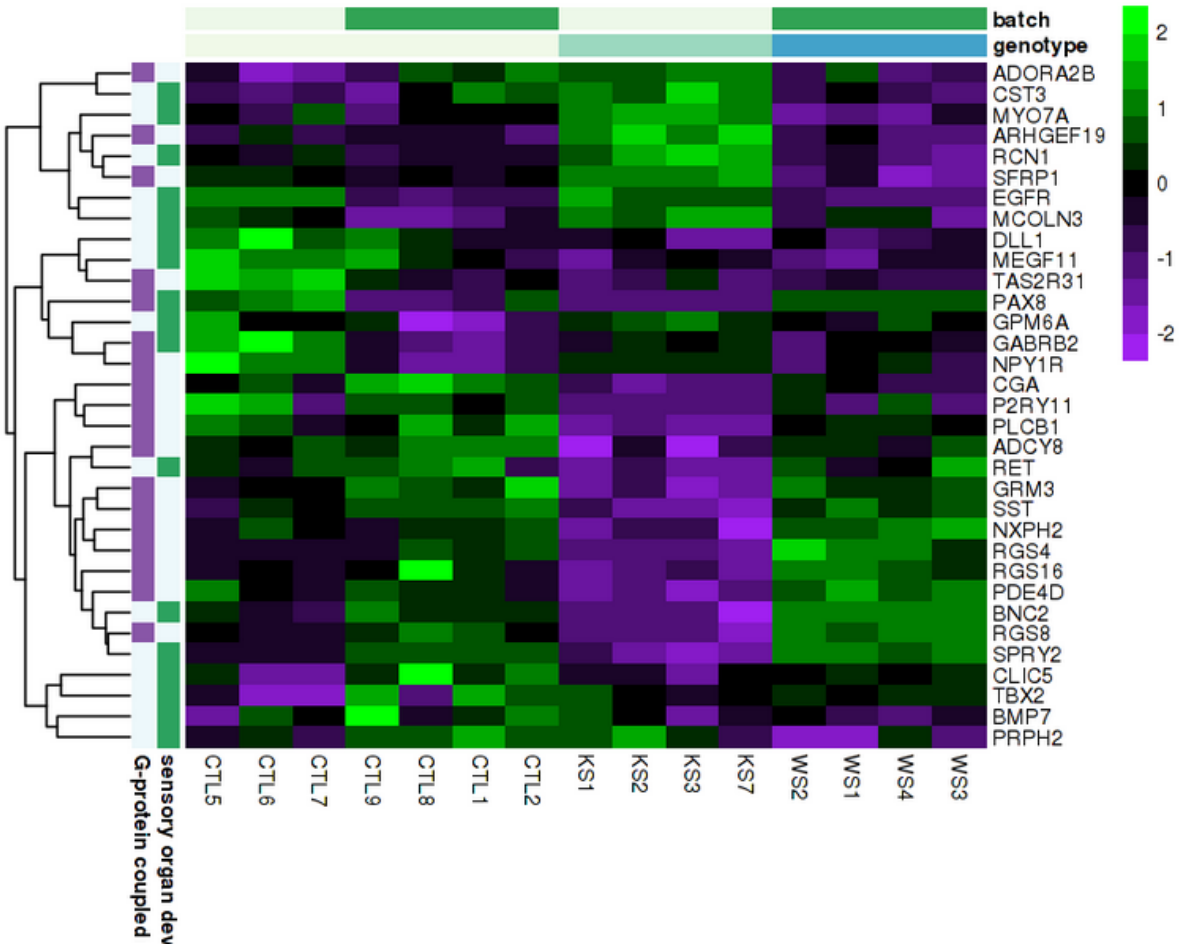


Figure 65. GO annotated heatmap of genes differentially expressed in Weaver syndrome in NGN2 neurons

Among transcription factors enriched by KS DEGs I found REST (accounting for 316 genes) and CRBP (342 genes), but also MAFK (264 genes). Among WS DEGs I could identify an enrichment for EZH2 (37 genes), REST (82), GTF2I (67), and MAFK (69).

Among genes up-regulated in WS and down-regulated in KS are mild enrichments for “transmembrane receptor activity”, “calcium binding”, and few analogous ones (see Figure 66).

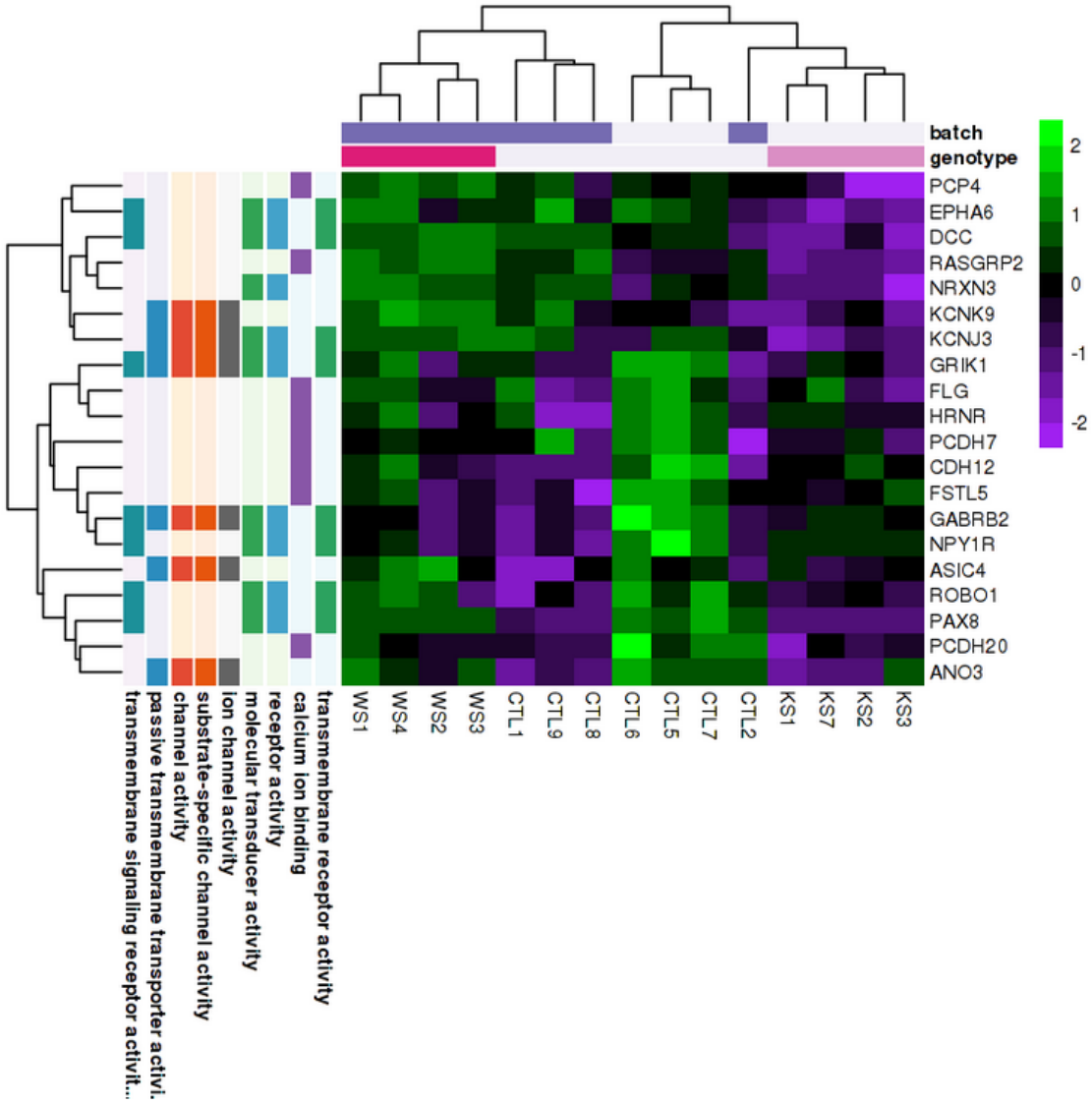


Figure 66. NGN2 genes up-regulated in Kabuki and down-regulated in Weaver Syndrome

Weaver Syndrome specific deregulation in brain organoids

While EZH2 has been largely demonstrated at the apex of timing and regulation of cortical development, Weaver Syndrome and its macrocephaly phenotype have not been specifically associated with it. Intellectual disability is highly variable among Weaver individuals, including mild cases of people able to complete high school studies. Brain Organoids give us the possibility to identify specific deregulations in WS patient-derived lineages that recapitulate early stages of this crucial developmental transition in a human-specific domain.

To dissect the transcriptional landscape of human cortical development I devised a strategy which included iPSCs as ground state, three stages of brain organoids culture (d25, d50 and d100), and NGN2 neurons as a progressive condition

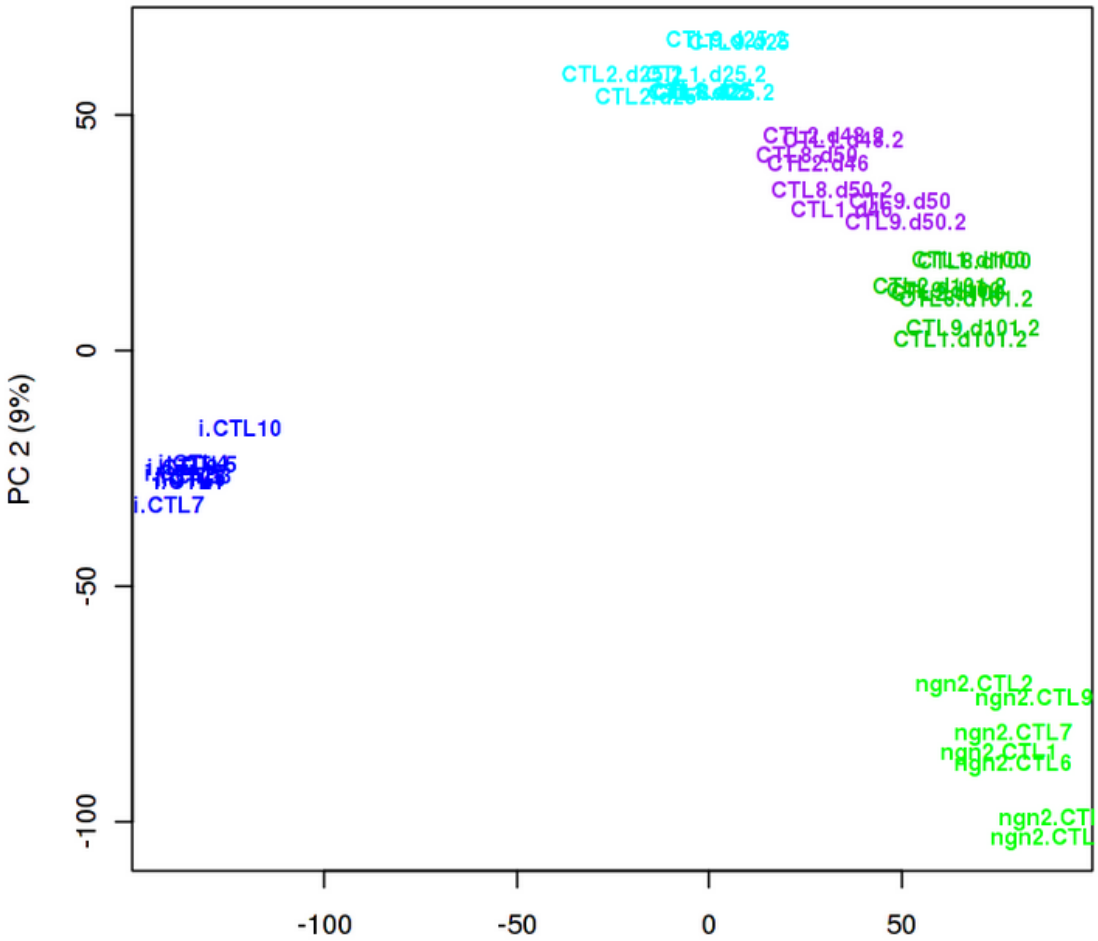


Figure 67. Principal Component Analysis conducted on iPSCs, NGN2 neurons and brain organoids. (performed on log-normalized read-counts)

(~individual+stage). Principal Component Analysis on RNA-seq data from control lines of iPSCs, brain organoids and NGN2 neurons show a similar distance between iPSCs and all the other tissues, but a clear temporal reproduction of development along PC2, going from d25 to NGN2 neurons, passing through d50 and d100 organoids (Figure 67).

To asses more clearly the position of these samples in the global context of development pictured by our inhouse bulk RNA-seq data, I calculated a Principal Component Analysis on all tissues, including all control fibroblasts at our disposal, from which some of our iPSCs were differentiated (Figure 68).

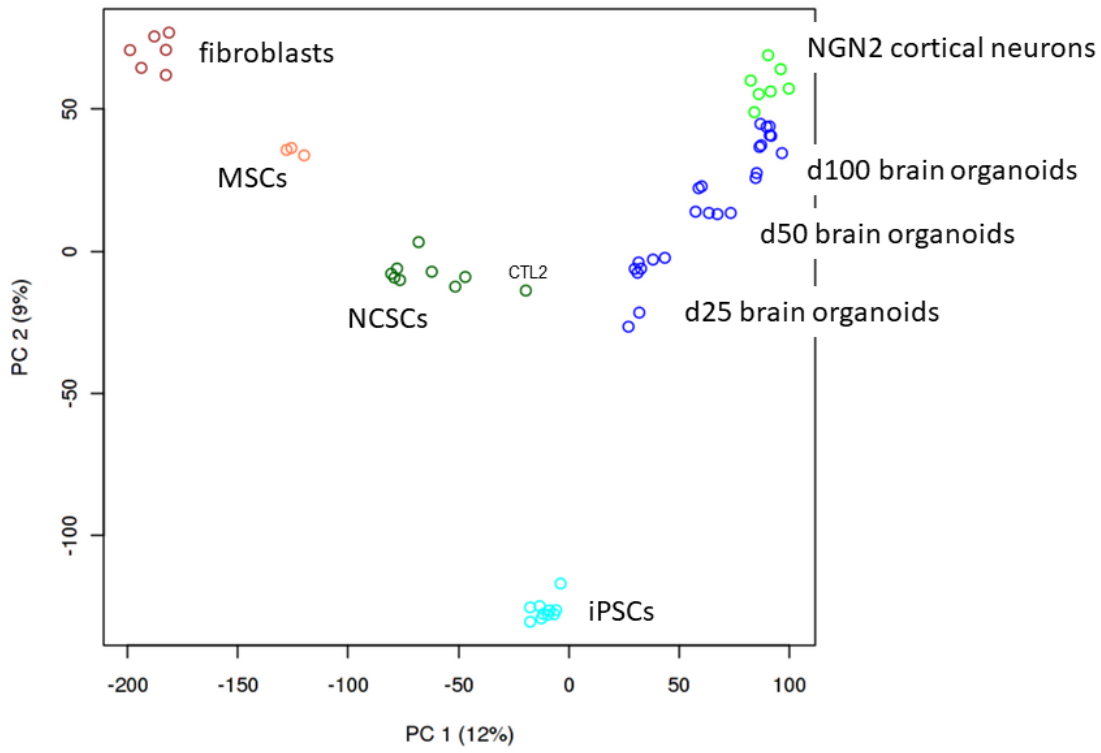


Figure 68. PCA of all control datasets available in the lab for fibroblasts, iPSCs, NCSCs, MSCs, brain organoids and NGN2 cortical neurons.

From this second analysis, it is clear that the distance between PSCs and all other tissues is probably too large to describe precisely, on a transcriptomic landscape, the developmental trajectory that goes from a pure pluripotent state (iPSCs) to cortical neurons (NGN2 in our hands), going through developmental stages represented by brain organoids. Nevertheless, this picture hints to an interesting and partially expressed similarity between neural crests and early stages of brain organoids culture, which is highly enriched for pluripotent lineages. Notably, the NCSCs line which is closer to brain organoids in this enlarged transcriptional landscape is CTL2, which had been suspected of possessing some outlier characteristic in our NCSCs based analyses.

To identify modules of genes changing expression along stages of organoids development, I built a comprehensive dataset including 4 control samples, and two technical replicates per sample. Indeed, each RNA-seq dataset has been obtained by sequencing an entire organoid. The high overlap between technical replicates and samples corroborates the high reproducibility of our culture protocol. A differential expression analysis along organoid stages (~individual+day) identify 6926 genes as differentially expressed across day (as factor), 3176 of which had a mean fold-change across stages > 2 (FDR < $1e^{-05}$) (Figure 69).

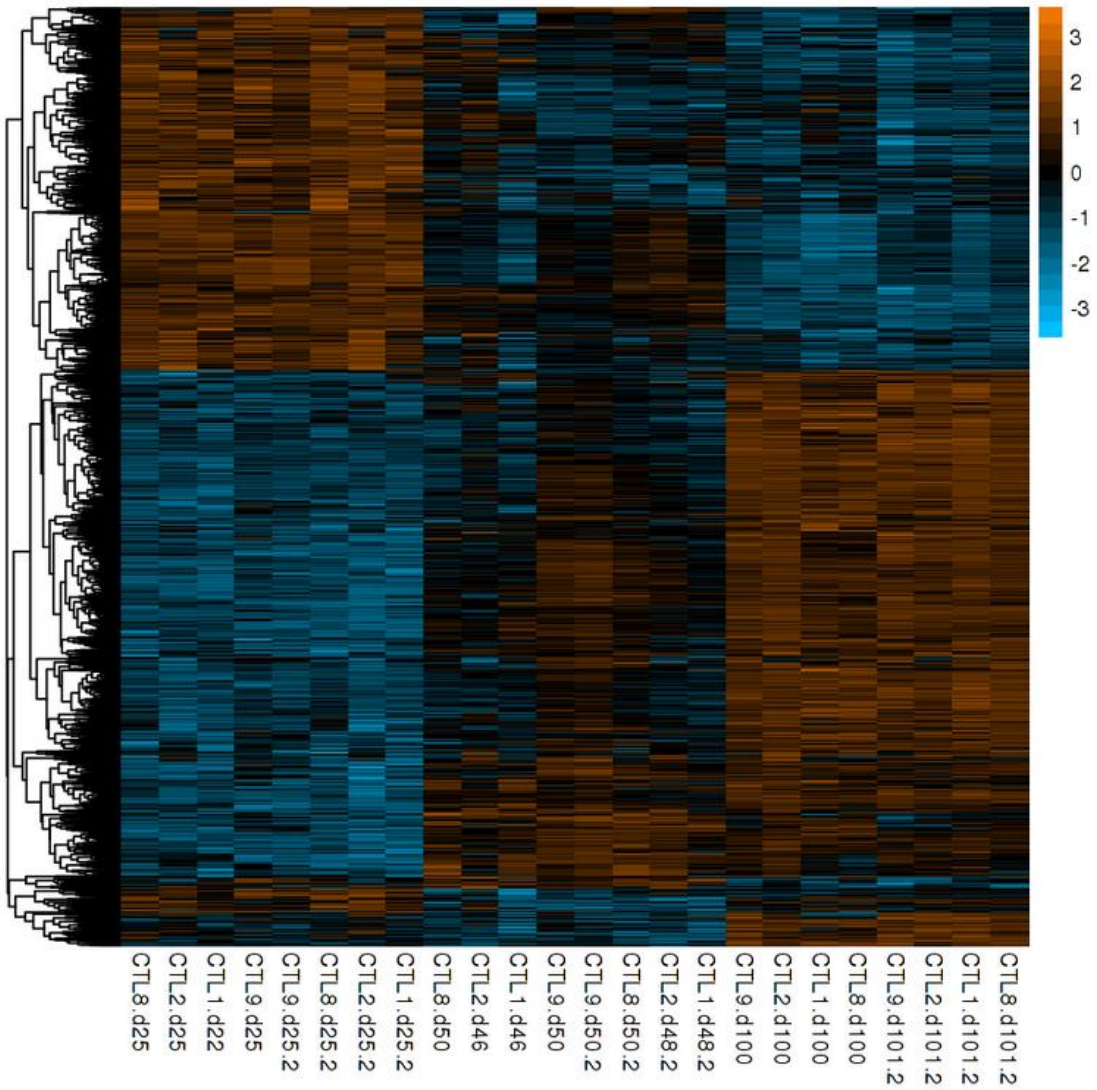


Figure 69. Genes differentially expressed across stages of brain organoids development in control samples (z-scores of log-normalized read-counts).

Most of these genes are up-regulated along development, the second largest group is down-regulated along development and two small subsets are specifically up- or down-regulated at day 50. Genes specifically down-regulated at day 50 (193, Figure 70) were enriched for several categories connected with brain and development in general, including lungs, while genes up-regulated exclusively at day 50 were fewer (43, Figure 71) and only 4 of them gave an enrichment for “sensory perception of pain” (FDR < 0.1).

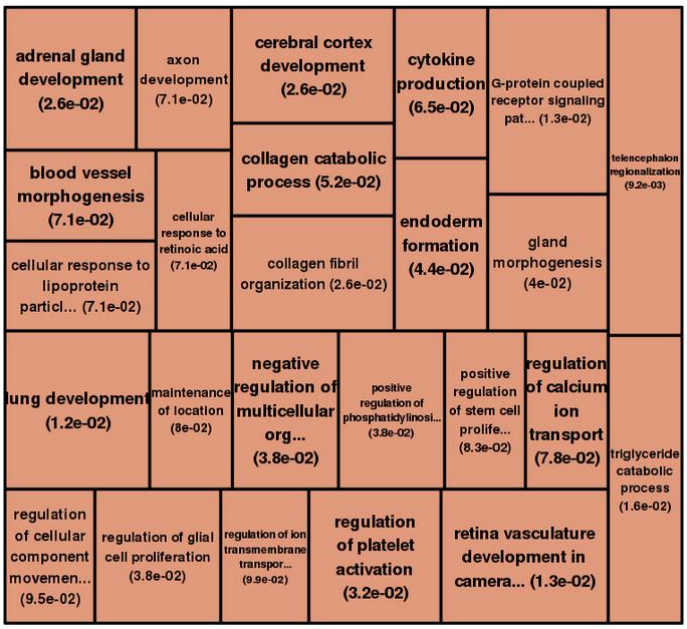


Figure 70. GO enrichments of genes down-regulated at day 50 in control organoids.

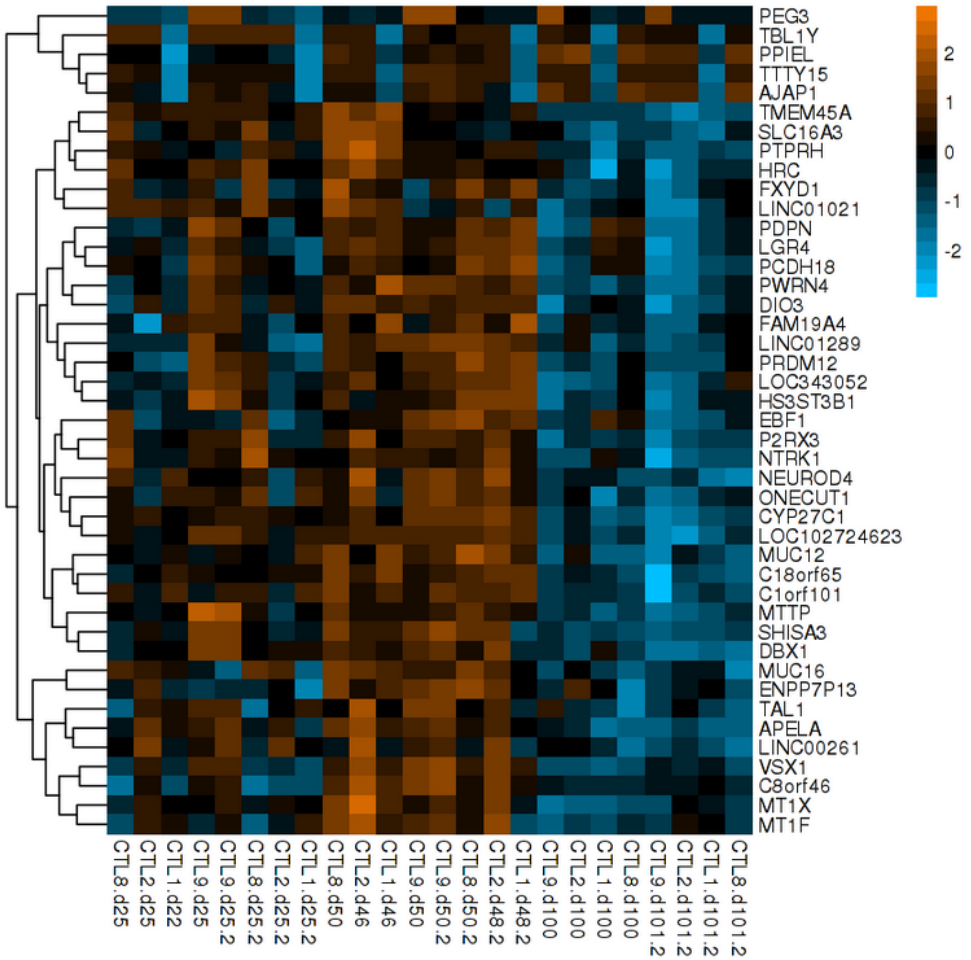


Figure 71. 43 Genes up-regulated in control brain organoids at day 50 (z-score)

These four genes are *NTRK1, P2RX3, PRDM12* and *FAM19A4*. Among them, PRDM12 is a kruppel-like zinc finger, which possess a SET domain and confers methyltransferase activity. The identified four modules of genes differentially expressed along our model of cortical development will be referred to as “Cortical Development DEGs”.

Once observed the largely reproducibility of organoid data, also at different stages, and the gradual transcriptional modulation across stages, I decided to focus on Weaver Syndrome-specific deregulations. In order to dissect the effect of EZH2 haploinsufficiency in WS in the context of brain development I performed stage-wise analyses (~Genotype) and multifactorial analysis across stages (~day+Genotype), considering this time day as numerical and Genotype as factorial variables. Genes differentially expressed in a stage specific fashion were 42 at day 25 and 39 at day 50 with FDR < 0.1 and FC > 1.5. Among DEGs at day 25 I focused my attention on *FAT4, WNT5A, LIMCH1* and *PCDH8*, which were all up-regulated at day 25, after hypothesizing a concerted dysregulation of a non-canonical Wnt pathway regulated by WNT5A in which all these genes could be involved. Moreover, genes differentially expressed at day 25 were enriched for GO categories such as “forebrain neuron differentiation”, “cell migration involved in gastrulation” and “bone development” (FDR < 0.1), while genes differentially expressed at day 50 were enriched for “axon extension” and “developmental growth involved in morphology”. Mining and interpretation of these lists is still ongoing.

In the context of the analysis of WS patients derived cortical organoids transcriptomic data along development (~day+genotype) I could identify, among the global differential expression, a large deregulation of genes expressed at day 25, which were down regulated gradually at day 50 and day 100 in control lines (significantly overlapping with genes down-regulated at day50 in controls, FDR 0.05,

Figure 70). These genes, which still need to be further characterized, appear to be clearly delayed in their down-regulation, and have a significant enrichment for EZH2 as master regulator (FDR < 0.0001). Here I report a kernel density distribution of logFC in controls (“black”) and WS samples (“red”), which shows a global shift of these fold-change towards 0, suggesting a failed down-regulation along cortex development (Figure 72).

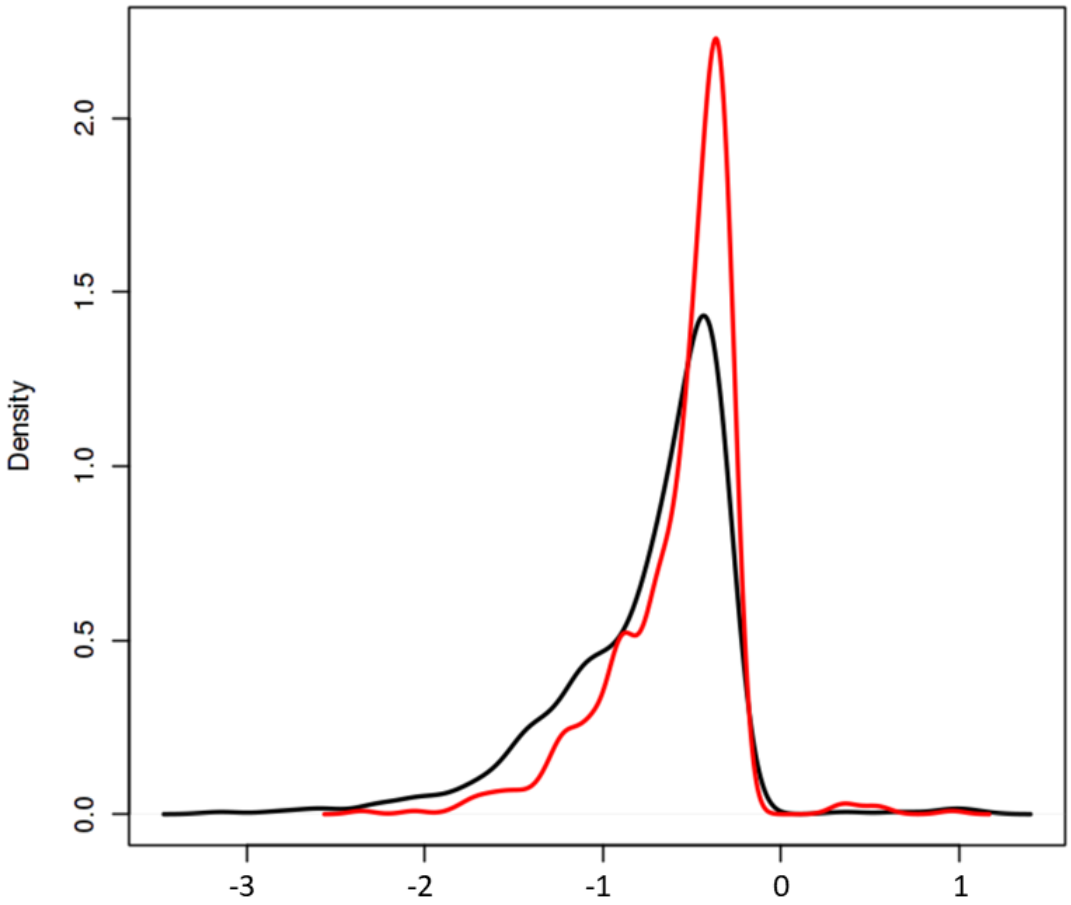


Figure 72. Kernel density distribution of log-FC in controls and Weaver samples for genes down-regulated in organoids after day25 along cortical development.

Dissection of Kabuki Syndrome epigenomic and transcriptional deregulations across the neocortical axis.

In order to further characterise the landscape of dysregulations induced by KMT2D mutations in cortical neurons, we profiled H3K4me1 and H3K27ac of KS and control NGN2 neurons for which RNA-seq analysis has already been described. Peak Calling and quantitative analyses have been performed with MACS2 and DeepTools as described for BAZ1B in the context of neural crest stem cells. I observed a global similarity between KS and control lines in terms of H3K27ac, with a slight global gain in KS. This global increase pairs with KDM6A overexpression in KS NGN2 (Figure 73).

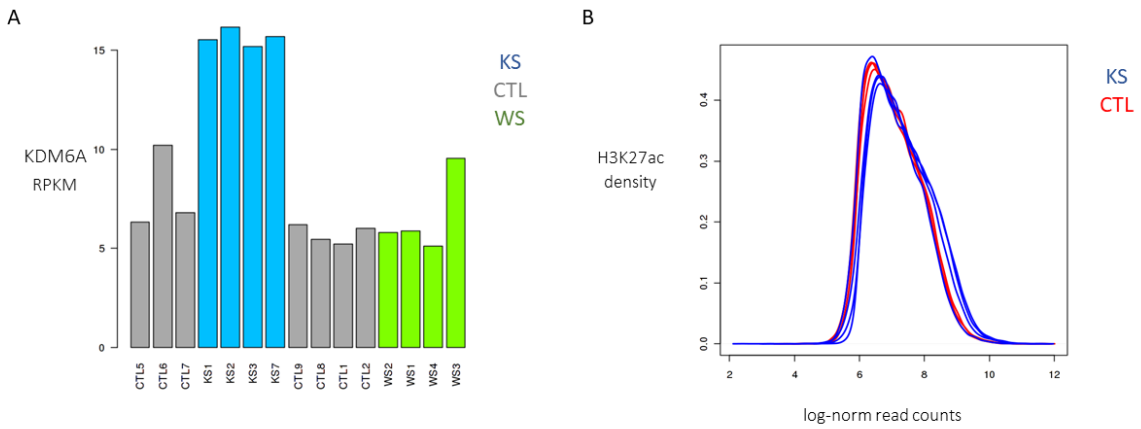


Figure 73. A) RPKM expression levels of KDM6A in cortical neurons; B) Kernel density function of log-normalized read counts (on library size) show a global increase in H3K27ac

In this context, being available data from 4DGenome initiative of glutamatergic neurons I used this database to associate genes with H3K27ac regions. 221 genes up-regulated in KS neurons show a significant increase in H3K27 acetylation at enhancers (FDR < 0.1). While 141 genes down-regulated show a loss in H3K27ac at enhancers. These 141 genes are clearly enriched for neuronal functions (Figure 74).

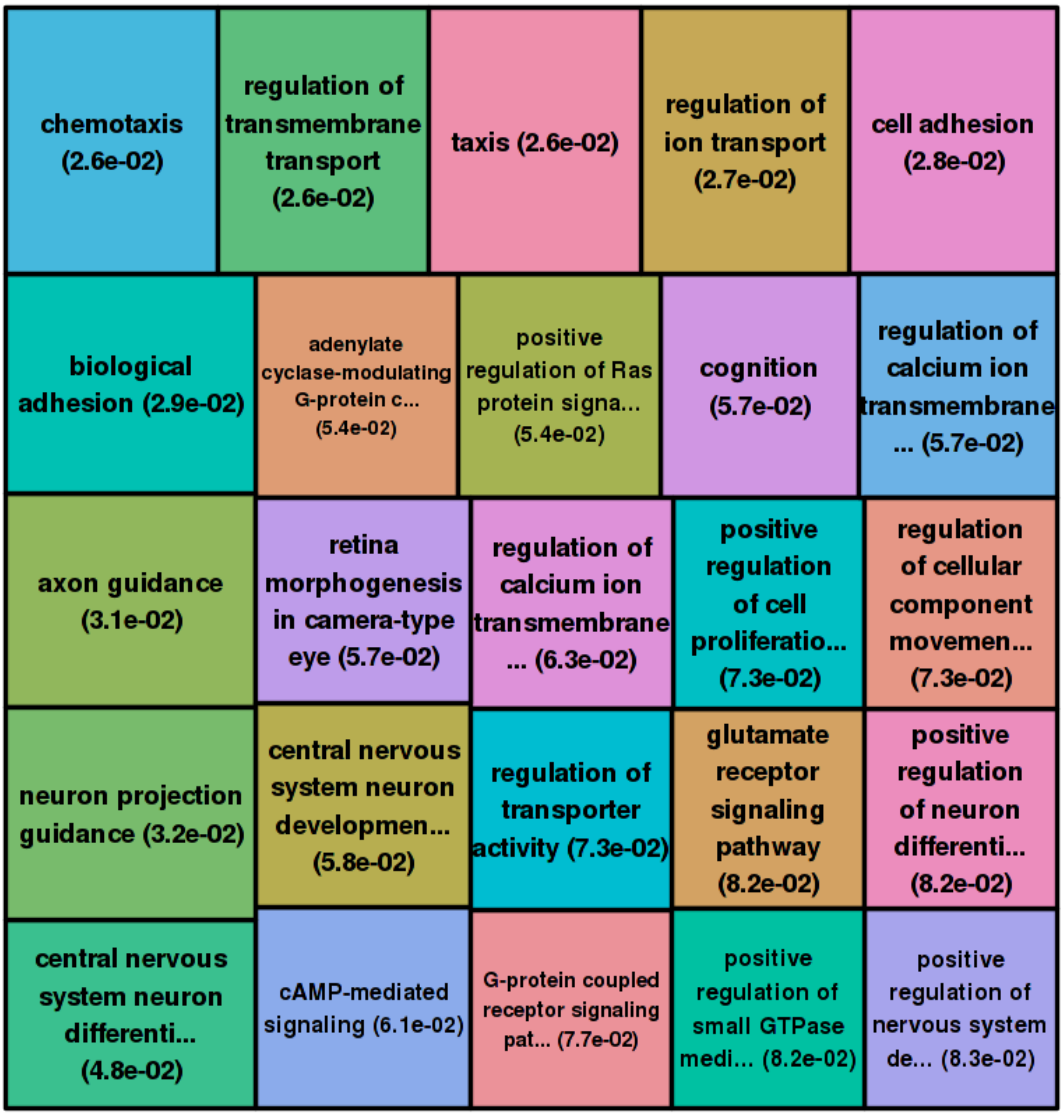


Figure 74. GO enrichments for genes hyper-acetylated at H3K27 regions in KS cortical neurons

Intriguingly, genes hypo-acetylated in KS that were not identified as targets of RBBP5 (a partner of KMT2D) do not enrich any category. Among BP terms enriched by genes down-regulated and hypo-acetylated I found “adenylate cyclase-modulating G-protein coupled receptors”, “positive regulation of Ras protein signal”, “cAMP-mediate signalling” and related ones, which clearly hint to signal transduction. Moreover, I could find categories such as “regulation of transmembrane transport”, “regulation of ion transport”, and related ones, which points towards neuronal electrical activity. Crossing this list of genes with RBBP5

targets keep genes enriched for “central nervous system neuron differentiation”, “dendrite development” and the likes. in the absence of ChIPseq data, among 496 genes hypomethylated at enhancers, 30 robust putative direct targets of KMT2D have been identified, by looking for genes which were down-regulated and enriched for RBBP5 but also hypo-methylated with respect to their H3K4 monomethylated enhancers in controls ($p < 3^{e-81}$). Crucially, among genes up-regulated in KS gaining H3K27ac marks at enhancers, I could identify categories unrelated to neuronal physiology but enriched for muscles function related categories such as “sarcoplasm” and “sarcomere” with FDR < 0.05 and “actomyosin” (FDR < 0.1). All these modifications appear in accordance with, as observed in other tissues where it was not significant, KDM6A being clearly up-regulated (FC > 2 and FDR < 0.0005) in KS patients-derived cortical neurons, providing evidence of the cellular epigenetic mechanisms activated to compensate for KMT2D haploinsufficiency.

Leveraging HipSci data to briefly verify false-positives proneness of model matrices and edgeR differential expression pipelines.

Human Induced Pluripotent Stem Cell Initiative (HipSci).

HipSci is a project based on the idea of gathering, in an open access easily available way, data and cell lines produced from healthy and rare disease individuals with their consent. iPSCs constituting HipSci cohort are all derived from fibroblasts in a standardized process and further characterised by means of genomics, proteomics and phenotypic assays (Streeter et al., 2017).

iPSCpoweR application *to test design matrices for differential expression analysis*

In the introduction I briefly described the crucial advances of iPSCs usage over hESCs in the contest of disease modelling. Here I try to describe how I took advantage of one of the latest publications of the lab (Germain et al., 2017) to optimize designs and models for my differential expression analyses at the pluripotent stage but also in the CNS- and neurocristic- axes.

Patient-derived lineages pave the way for reconstructing in vitro development, not simply because they can reproduce crucial, rare, and difficult to obtain cell-types, but mostly because they permit many repetitions of the first stages of a specific individual development, making it experimentally and statistically tractable (Colman, 2008; Germain et al., 2017). Yet we don’t have a crystallized idea of human

variability. Thus, we don't have a magic number or a set of rules to define how many replicates of each genetic background are enough to achieve a robust and sensitive experimental design. To answer this question, my co-supervisor resorted to two large datasets by HipSci (Kilpinen et al., 2017; Streeter et al., 2017) and the National Heart, Lung, Blood Institute (NHLBI) NextGen consortium (Carcamo-Orive et al., 2017). HipSci lines, in particular, come from fibroblasts and they were obtained via reprogramming with the same protocols used in my hosting lab, making them a fundamental reference to test our methods on a large - external - scale. Given their skin biopsies origin, one drawback of relying on fibroblasts is that they can harbour somatic mutations and epigenomic make-ups that could favour "selection in the dish" at reprogramming (Ji et al., 2012; Young et al., 2012); moreover, only a minor part of epigenetic marks is kept after the procedure (Kim et al., 2010); yet, iPSCs clones transcriptional heterogeneity appears to be mostly due to their epigenetic state (Kilpinen et al., 2017). Furthermore, iPSC clones transcriptomes can gain pure technical variability from clones harvesting to RNA extraction and RNA-seq library preparation. Finally, iPSCs derived by different people, in different labs, can cast further technical dissimilarities in the very same cell line, hampering the ability to see differences across genotypes or conditions, and the ability of making tests and hypotheses on further differentiated lineages (Volpato et al., 2018). In particular, Germain and Testa (Germain et al., 2017) demonstrated that:

- the use of more than one clone per individual in combination with current differential expression analytical practices can be detrimental to results robustness
- multiple clone settings can be leveraged with the application of mixed-models and/or aggregating data derived from multiple iPSC clones of the same individual

- it is possible to do power analysis with iPSCpower, using HipSci cohort data and a combination of i) number of samples and ii) a design matrix

Building on the iPSCpower package (Germain et al., 2017) I designed a short R script (Figure 75A) to verify empirically the tendency of model matrices (like those used in this thesis) to produce plainly false predictions of differentially expressed genes. Here I report a reproducible case. In Figure 75B I show the number of DEGs obtained with a model matrix considering 13 individuals and 2 clones per individuals. In this example I used a model matrix that is supposed to correct for a batch effect ("batch") and individuals background ("ind"), and test for expression levels of BAZ1B ("baz1b"): (~batch+ind+baz1b). A similar model matrix has been shown previously in the text, in the study of BAZ1B dependent regulatory networks observed in NCSCs, upon the gene knock-down. Batch effect corresponds to pure technical variability: RNA-extraction and libraries preparation happened in two different moments for two groups of samples. Each batch was represented both in each individual and in each group of the two short-hairpins used to knock down BAZ1B. The use of iPSCpower showed that a model matrix, with one third of the sample used in the real experiment (BAZ1B KD in NCSCs), poses the risk of obtaining high numbers of spurious DEGs only in one case out of 50, with most cases leading to less than 10 DEGs using the edg2 function. This function was thus chosen because it does not discard genes with high variability before the GLM fitting step.

A

```r
library(iPSCpoweR)
set.seed(123)
dat <- iPSCpoweR::aggByClone(getGeneExpr(),T)
data("hipsci_annotation")

indie <- dat$annotation$individual
indie <- unique(indie[which(duplicated(indie))])
cn <- t(combn(indie,11))
tes <- apply(cn[sample.int(1365,300),],1,an=dat$annotation,FUN=function(x,an){ dat$dat[,an$line[which(an$individual %in% x)]] })

mm <- model.matrix(~batch+ind+baz1b)

edg1 <- function(e,mm){
  row.names(mm) <- colnames(e)
  dds <- calcNormFactors(DGEList(e))
  dds <- estimateDisp(dds,mm,robust=T)
  dds <- glmFit(dds, mm)
  res <- as.data.frame(topTags(glmLRT(dds, "baz1b"), nrow(e)))
  res
}

edg2 <- function(e,mm){
  row.names(mm) <- colnames(e)
  dds <- calcNormFactors(DGEList(e))
  dds <- estimateGLMRobustDisp(dds,mm)
  dds <- glmFit(dds, mm)
  res <- as.data.frame(topTags(glmLRT(dds, "baz1b"), nrow(e)))
  res
}

edg3 <- function(e,mm){
  row.names(mm) <- colnames(e)
  dds <- calcNormFactors(DGEList(e))
  dds <- estimateDisp(dds,mm,robust=T)
  dds <- glmQLFit(dds,mm,robust=T)
  res <- as.data.frame(topTags(glmQLFTest(dds, "baz1b"), nrow(e)))
  res
}

edg4 <- function(e,mm){
  row.names(mm) <- colnames(e)
  dds <- calcNormFactors(DGEList(e))
  dds <- estimateGLMRobustDisp(dds,mm)
  dds <- glmQLFit(dds,mm,robust=T)
  res <- as.data.frame(topTags(glmQLFTest(dds, "baz1b"), nrow(e)))
  res
}

library(BiocParallel)
res1 <- bplapply(1:50, function(x){edg1(tes[[x]],mm)},BPPARAM = MulticoreParam(6))
res2 <- bplapply(1:50, function(x){edg2(tes[[x]],mm)},BPPARAM = MulticoreParam(6))
res3 <- bplapply(1:50, function(x){edg3(tes[[x]],mm)},BPPARAM = MulticoreParam(6))
res4 <- bplapply(1:50, function(x){edg4(tes[[x]],mm)},BPPARAM = MulticoreParam(6))

fdr025_FC125_1 <- unlist(lapply(1:50,function(x){sum(res1[[x]]$FDR < 0.25 & abs(res1[[x]]$logFC) > log2(1.25) )}))
fdr025_FC125_2 <- unlist(lapply(1:50,function(x){sum(res2[[x]]$FDR < 0.25 & abs(res2[[x]]$logFC) > log2(1.25) )}))
fdr025_FC125_3 <- unlist(lapply(1:50,function(x){sum(res3[[x]]$FDR < 0.25 & abs(res3[[x]]$logFC) > log2(1.25) )}))
fdr025_FC125_4 <- unlist(lapply(1:50,function(x){sum(res4[[x]]$FDR < 0.25 & abs(res4[[x]]$logFC) > log2(1.25) )}))

library(ggplot2)
df <- as.data.frame(cbind(c(fdr025_FC125_1,fdr025_FC125_2,fdr025_FC125_3,fdr025_FC125_4),c(rep(1,100),rep(2,100),rep(3,100),rep(4,100))))
names(df) <- c("DEGsNumber","DEAfunction")
ggplot(df,aes(x=as.factor(df$DEAfunction),y=df$DEGsNumber)) + geom_violin()
```

Load iPSCpower library, set a random number generation seed, load HipSci data

Build a dataframe with all possible combinations of 11 individuals

Build a model matrix

Introduce 4 differential-expression analysis pipelines (edg1, edg2, edg3, edg4)

Load BiocParallel
Do 50 differential-expression analyses with the first 50 datasets created previously

Save the number of DEGs found with a certain threshold in each analyses with a certain DEA function

Violin plot of DEGs found with each strategy

B

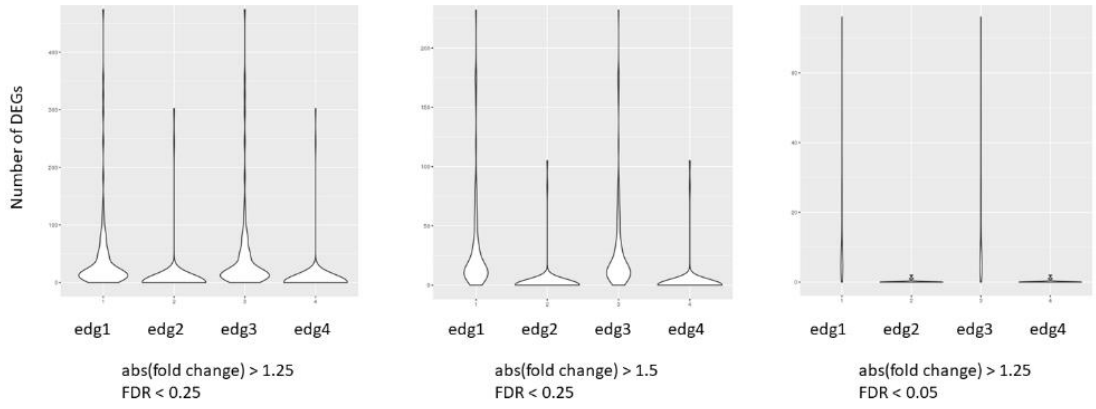


Figure 75. iPSCpower application; A) exemplary R code and B) Violin Plot of numbers of spurious DEGs generated by the given model matrix.

Geneset Characterization Tools

omimCrawler: from genesets to OMIM

In order to query the OMIM database in a reproducible and automatized fashion, I developed a small R script capable of searching for key words in OMIM phenotype and disease descriptions. While not being ready for publication I used it extensively to interpret the many lists produced during this thesis preparation. It works by receiving in input a “species” (e.g. “human”), a keyword to query (e.g.”intellectual disability”) and a list of genes (as character). Given the described input, it i) loads biomaRt libraries (Durinck et al., 2005); ii) build a copy of Ensembl database(Kinsella et al., 2011); iii) use the function getBM to retrieve unique entries including: gene symbols, ensemble ids, interpro, phenotype description, mim morbid accession, mim morbid description; and iv) it uses the given keyword to build a dataframe including HGNC symbols, phenotype description and OMIM disorder accession IDs.

HPO Enrichments

After defining the omimCrawler R script to query the OMIM database (Amberger et al., 2015), I decided to define a similar one for the Human Phenotype Ontology (HPO) (Köhler et al., 2017). As the acronym might recall, this is a database of relations between genes and phenotypic abnormalities that may arise from human diseases. This ontology stemmed from the Monarch Initiative (Mungall et al., 2017) and has been developed starting from clinical literature, OMIM itself, but also Orphanet and DECIPHER(Firth et al., 2009) . At the time of this thesis writing HPO accounts

for 13 thousand terms and 156 thousand annotations to hereditary disorders. Few tools have already been developed to query HPO and measure geneset enrichments against it, so mine has remained an incomplete work on which I will focus in the coming months/near future. Geneset enrichments on HPO have thus been performed using Enrichr (Chen et al., 2013; Kuleshov et al., 2016). Unfortunately, a major drawback - for which I started to build my own tool - is the impossibility to provide a background list of genes, which is instead instrumental to define the reference universe of expressed/considered genes in a non-trivial geneset enrichment analysis.

MINT-ChIP pipeline

As anticipated in the introduction, MINT-ChIP experiments permit to do many IPs at the same time, maximizing throughput via a pool-and-split multiplexing approach that reduces the well-known ChIPseq sensitivity both to the amount of chromatin input and to antibodies quality. This improves the accuracy with which chromatin landscapes can be quantitatively compared across samples. It also helps define global changes in histone modification levels due to different conditions such as cell type shifts or genetic mutations in epigenetic regulators. Given the double-barcoding strategy, reads coming from MINT-ChIP sequencing can be demultiplexed by “individual” after conversion of the raw binary base call data (BCL) into fastq file format: this part is performed by the Illumina Sequencing facility of my hosting institute. Common RNA-seq and ChIP-seq sample-specific raw fastq files are produced in the same way. Thus, fastq files coming from this demultiplexing step are devoid of sample-specific barcodes. As anticipated previously, MINT-ChIP libraries are pair-end, and one mate of each couple of reads contain the mark/IP specific barcode (MSB). Having these reads already been demultiplexed for patients’ specific barcodes, the first mate reads (R1) start with MSBs. The pipeline I

devised takes this in to account, so its first step is to couple all mate reads before MSB demultiplexing, which relies on fastx toolkit (Hannon Lab, hannonlab.cshl.edu/fastx_toolkit). After demultiplexing, part of the reads couples that do not contain any of the MSB are collected in an “unmatched” file that can be further mined to eventually recover miss-demultiplexed reads. R1 reads are then trimmed of the first nucleotides corresponding to the barcode and the resulting data is aligned to the reference genome in the same way used for ChIP-seq and CUT&RUN data. Following steps of analysis are the same enlisted above for classical ChIP-seq and CUT&RUN data.

DISCUSSION

This thesis leverages on transcriptomic and epigenomic data produced in my hosting lab. Other omic- data sources have been interrogated directly, such as Brainspan, Cortecon, 4DGenome, ENCODE and Roadmap, other indirectly, such as through GRID, Homer, StringDB and TFBS. The integration of all these datasets is incomplete and will require further investigations, but what I achieved so far clearly helped the deconvolution of regulatory networks underlying neurodevelopment both across relevant tissues, such as NCSCs, MSCs, NGN2 cortical neurons and brain organoids, and across human conditions such as neurodevelopmental disorders causing both intellectual disability and cranio-facial dysmorphisms. Moreover, I provided a first characterisation transcriptional deregulations in iPSCs separating non-autistic and autism spectrum neurodevelopmental disorders. To do so I designed, with strong support of my supervisors and colleagues, a multidimensional analysis for a new disease modelling paradigm. The multidimensionality is configured in terms of the number of layers of development being assessed, accounting for 6 tissue types (including fibroblasts) and recapitulating two main axes of development (The Cortex and the Neurocristic ones). The novelty of the paradigm, in which these analyses and studies have been carried on is due to the unprecedented intention of not only studying multiple neurodevelopmental disorders in concert - as primarily fostered by A.K. Smith and colleagues – but to confront directly, and sometimes without considering human control samples, neurodevelopmental disorders with utterly opposite genetic and phenotypic traits.

Among the analyses performed across disorders in iPSCs, data-driven identification of a selected, relatively small set of genes, having coarsely the same fold-change

across disorder, fosters the hypothesis of them working in a complex or in the same spatio-temporal phase of gene regulation (Figure 19, Figure 21).

The sixth and tenth cluster there identified highlight groups of genes almost exclusively deregulated in GaDeVS. Those up-regulated only in this disorder are many (289) and show very high fold-changes but no GO enrichments. The same genes are strongly enriched (FDR < $1.5e^{-04}$) in lists of TAF1, RBBP5, TBP and POL2 targets: all involved in active transcription. Given YY1's central role in active loop formation and transcription start, I hypothesis that these genes are direct targets of YY1 to be characterized and validated computationally and experimentally.

On the deconvolution of ASD features and NDDs features, in disorders having either both ASD and intellectual disability or only the latter, I could identify a large set of genes specifically dysregulated in iPSCs from aWBS individuals. These genes show a distinct transcriptional behaviour in aWBS with respect to 7DupASD and ADNP-ASD. Thus, they could suggest a potentially opposite transcriptional origin for the two groups of ASD patients that requires experimental validation and assessment along later stages of differentiation. On another end, having only WBS, as a genotype, both ASD and non-ASD individuals, my design could have forced the analysis towards the identification of genes with opposed expression in WBS samples (with or without ASD). Thus, I will have to verify the contribution of the design matrix to the identification of DEGs between non-ASD and ASD neurodevelopmental syndromes. The simplest action would be to consider aWBS and WBS as two separate conditions, losing the effective correction within genotype. In general, differential expression analysis on iPSCs has proven much more effective if conducted with a multifactorial design than by testing each disorder separately, discarding the others. Actually, the other setup, which identified few

genes differentially expressed in few disorders, highlighted the high level of similarity among them.

More in details, I could observe a smaller deregulation at the pluripotent stage for Kabuki and Weaver Syndromes with respect to the other disorders studied. Nevertheless, also in KS and WS lines, pluripotent stage specific dysregulations included several DEGs in the same direction in derived tissues in a disease-specific way. This hints to the definition, in early developmental lineages of an epigenetic scar that is kept in differentiated lineages. Notably, an inverse comorbidity has been identified for several neurodevelopmental and neurodegenerative disorders with respect to cancer (Tabarés-Seisdedos and Rubenstein, 2013). Indeed, several identical mutations occurring in *YY1, KMT2D* and *EZH2* are found as drivers in cancer and causing in NDDs. The hypothesis, developed along my whole PhD studies thanks to strong support and fruitful intellectual confrontation with my supervisor Giuseppe Testa and my colleague Michele Gabriele, is that when such mutations are gained during embryogenesis, a set of compensatory mechanisms are activated by the cellular machinery and translated into a protective epigenetic scar. This idea is supported by the sensitivity of cancer to the developmental context, demonstrated by nuclear transplantation experiments in mice in which egg-mediated reprogramming was able to suppress, through the blastocyst stage, the oncogenic potential of mutant genomes from leukemia, lymphoma, and breast cancer cells (Hochedlinger et al., 2004). In this outlook, the overexpression of KDM6A in KS post-mitotic neurons, coupled with the global (genome-wide) increase in H3K27ac in the very same tissue, suggests a starting point to dissect the molecular underpinnings of NDDs resistance to cancer.

As observed for KDM6A in KS cortical neurons, I observed in iPSCs of ASD patients the overexpression of two SFARI genes. These genes haploinsufficiency is known

to cause ASD, and their counter-intuitive up-regulation in disorders that are caused by other genes haploinsufficiency might hint to a compensatory mechanism.

Moving forward to the iPSCs derived axes described in this thesis, large parts of the work presented has been conducted by leveraging already published data from my hosting labs and from several publications and public databases. Particularly tools for Cortecon and Brainspan enrichments will be further developed to extend their application on both Neurocristic and Cerebral Cortex axes characterisation.

In the Neurocristic axis I identified 12 modules of genes, then collected into four subgroups: one gradually up- and one gradually down-regulated along development, one neural crest stem cells specific, and one mesenchymal stem cell specific. These four groups were further characterised by specific transcription factor enrichments, highlighting cross-talks across the four subgroups, or within each of them. The putative role of BAZ1B in orchestrating all these regulatory networks has been exposed, together with a precise reconstruction of its molecular role in neural crest stem cells.

For what concerns the characterisation of the Cerebral Cortex Axis, few experiments and analyses are required to gain insights into the network of regulation impacted by KMT2D and KDM6A mutations and their role in neuron specification. In the context of Kabuki Syndrome, the peculiar hyper-acetylation and up-regulation of genes involved in muscle development, coupled with the opposed down-regulation of neuron differentiation and brain development genes, strikes for its analogy with a paper recently published by my hosting lab, in which trans-differentiation of mouse fibroblasts to neuronal cells has been proved dependent on KMT2B (Barbagiovanni, 2018). In the human cortex, I observed iPSCs lineages deregulating several genes involved in brain development, neural crest lineages up-regulating genes enriching brain development relevant categories, cortical neurons down-regulating such

genes and up-regulating muscle development related ones. In most cases REST has been found as the main master regulator of genes differentially expressed in and across tissues and NDDs. Among targets of BAZ1B I identified CUL3, which is crucial for cranial neural crest and CNS branching in development, but most notably CoRest, which is recruited to chromatin by REST exactly to modulate neurogenesis. Among Weaver Syndrome deregulated genes in several tissues, I could identify *CDKN1A*, which is a partner of *EZH2* in neurogenesis regulation. All these genes work via a strict cross-talk with Reelin, which is up-regulated in WS and down-regulated in KS neural crest stem cells. The global picture I envision, which needs experimental validation and further proofs, is that all these mutations impinge on a very specific list of target genes and chromatin locations, on which BAZ1B, KMT2D, and EZH2 respective complexes compete for binding and regulation. Thus, the alternative haploinsufficiency, or loss of function, of one of these players causes the outbreak of the other ones.

The outcome I predict, on the cellular state and identity of NDDs relevant lineages, is of a general un-specification. What I imagine happening is the population of several mature tissues with meta-plastic non-tumorigenic cells, bearing hybrid epigenetic make-ups, and reduced presence of the “correct” cell types in each affected tissue. Indeed, the idea that an unbalance between proliferation and differentiation might be predominant cause of NDDs acquisition has already been hypothesised (Ernst, 2016). In this context, the final observation of a burst in expression at day 25 in our cortical organoids, of genes whose down-regulation is not completed, or delayed, in Weaver samples highlights a small set of genes to focus our future attention. In this outlook, *ADNP* and *YY1* working upstream *BAZ1B*, *EZH2*, and *KMT2D*, and *GTF2I* working downstream (given its pure and specific transcription factor function), should complete the picture.

Finally, for what concerns BAZ1B's role in chromatin regulation and transcription, while writing these conclusions, I could find by means of HHpred (https://toolkit.tuebingen.mpg.de/#/tools/hhpred), a specific homology between the BAZ1B DDT domain and domains binding H3K36 methylated regions. This histone modification is important to mark DNA damaged sites, and actively transcribed regions, to avoid spurious transcription and to regulate splicing (Li et al., 2013; Venkatesh and Workman, 2013). The presence of a kinase domain, whose role is to specifically phosphorylate H2AX upon DNA damage, immediately up-stream of the BAZ1B DDT domain, suggests a new role for the DDT domain itself, which was, up to now, only generally defined as "DNA-binding". The presence of LSM6 in both ~knockdown and ~BAZ1B analyses, in neural crest stem cells, defines it as one of the most robustly deregulated genes, and corroborates its primary role in splicing also with respect to the other differentially expressed genes enriching "splicing" categories. The supposed role in H3K36 methylated regions recognition of BAZ1B puts the two genes inside a hypothetical splicing regulating complex.

For what concerns pure computational and methodological concepts developed in this thesis, omimCrawler and HPO enrichment tools will be soon refined and made publicly available. All tools and types of analyses presented will be implemented into a publicly available queryable online framework.

The use of iPSCpower showed that using a model matrix with the same structure but with one third of the samples used in the real experiment (~batch+individual+BAZ1B, and BAZ1B KD in NCSCs), we risk obtaining high numbers of spurious DEGs only in one case out of 50, with most cases leading to less than 10 DEGs (using the edg2 function applied in this thesis). This function was used as a case study for the peculiar type of analysis conducted in NCSCs. Using

a numerical vector as an independent variable and data coming from different individuals with highly different genetic background can be dangerous without an internal reference. In this case we had 3 conditions per individual, reassuring the limit of spurious regression. Moreover, the edgeR function “estimateGLMRobustDisp” keeps all genes from the input dataframe to compute the negative binomial dispersion parameter, to later estimate the regression parameter. Usually, a more conservative “estimateDisp (dds,mm,robust=T)” configuration of edgeR pipeline cause the immediate removal of highly variable genes, to perform the linear regression only on the least variable ones. In this case, given the power of the design, the availability of two different short-hairpins – i.e. 3 measurements per individual per gene -, and multiple individuals per genotype, I decided to include all expressed genes in the differential expression analysis to ensure the recognition of genes following BAZ1B levels in the vast majority of samples. In addition, the possibility to compare the list of DEGs with other genesets coming from different analyses and different types of data allowed me to use loose threshold (FDR < 0.25) with the aim of not excluding potential false negatives and to leave the validation and exclusion of false positives to a second phase and or to an experimental validation (i.e. qRT-PCR, Western Blots, etc).

Last, but not least, from my humble chair, I’d like to place myself among those who acknowledge that seeds of self-domestication theory, barely touched in this thesis, were planted by Charles Darwin himself. While studying animal domestication in the context of evolution he first made a distinction between domestication and taming, because of the goal-oriented nature of the former, observing that variability of physical and intellectual traits is higher in domesticated species, as well as behavioural plasticity and educability. He further observed that the brain size of

domesticated animals is smaller than that of their wild ancestors, just like Homo sapiens, with respect to Neanderthals, and several NDDs individuals with respect to “healthy individuals”, just as opposite in the case of WS overgrowth. Moreover, he specifically considered dwarfism and gigantism as more present in domesticated species (Brüne, 2007).

He did not plainly theorise that a similar selection as “domestication” had happened in homo sapiens, accompanying the evolution of less aggressive traits with cognitive ones, thus binding these two aspects also in humans, but he observed that, just like dogs and other animals evolved less aggressive cranio-facial features along domestication, modern human traits had followed a similar path during their “final” evolution into sapiens. These hypotheses go far from embracing Lombroso’s theory, who thought he could prevent crimes by chasing people just because of their aspect, and supposing an impossibility of redeeming or “fighting” genetic inheritance, ignoring the power of environmental context, and luckily limited by his sample size. On an opposite track, I will try to characterize the cross-talk between molecular pathways modelling human face and skull morphogenesis and intellect, which are utterly orchestrated by common regulators. Indeed, we must observe that – again - Darwin proves himself largely ahead of his times, but now it is time to overcome “corroboration” by validating such hypotheses and dissecting this side of Nature at the molecular level.

REFERENCES

Acab, A., and Muotri, A.R. (2015). The Use of Induced Pluripotent Stem Cell Technology to Advance Autism Research and Treatment. Neurother. J. Am. Soc. Exp. Neurother. *12*, 534–545.

Adamo, A., Atashpaz, S., Germain, P.-L., Zanella, M., D'Agostino, G., Albertin, V., Chenoweth, J., Micale, L., Fusco, C., Unger, C., et al. (2015). 7q11.23 dosage-dependent dysregulation in human pluripotent stem cells affects transcriptional programs in disease-relevant lineages. Nat. Genet. *47*, 132–141.

Adolphs, R. (2015). The unsolved problems of neuroscience. Trends Cogn. Sci. *19*, 173–175.

Agger, K., Cloos, P.A.C., Christensen, J., Pasini, D., Rose, S., Rappsilber, J., Issaeva, I., Canaani, E., Salcini, A.E., and Helin, K. (2007). UTX and JMJD3 are histone H3K27 demethylases involved in *HOX* gene regulation and development. Nature *449*, 731–734.

Akizu, N., García, M.A., Estarás, C., Fueyo, R., Badosa, C., de la Cruz, X., and Martínez-Balbás, M.A. (2016). EZH2 regulates neuroepithelium structure and neuroblast proliferation by repressing p21. Open Biol. *6*, 150227–.

Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. *43*, D789–D798.

Antonell, A., Del Campo, M., Magano, L.F., Kaufmann, L., de la Iglesia, J.M., Gallastegui, F., Flores, R., Schweigmann, U., Fauth, C., Kotzot, D., et al. (2010).

Partial 7q11.23 deletions further implicate GTF2I and GTF2IRD1 as the main genes responsible for the Williams-Beuren syndrome neurocognitive profile. J. Med. Genet. *47*, 312–320.

Arrowsmith, C.H., Bountra, C., Fish, P.V., Lee, K., and Schapira, M. (2012). Epigenetic protein families: a new frontier for drug discovery. Nat. Rev. Drug Discov. *11*, 384–400.

Ashe, A., Morgan, D.K., Whitelaw, N.C., Bruxner, T.J., Vickaryous, N.K., Cox, L.L., Butterfield, N.C., Wicking, C., Blewitt, M.E., Wilkins, S.J., et al. (2008). A genome-wide screen for modifiers of transgene variegation identifies genes with critical roles in development. Genome Biol. *9*, R182.

Aydin, Ö.Z., Vermeulen, W., and Lans, H. (2014a). ISWI chromatin remodeling complexes in the DNA damage response. Cell Cycle *13*, 3016–3025.

Aydin, Ö.Z., Marteijn, J.A., Ribeiro-Silva, C., Rodríguez López, A., Wijgers, N., Smeenk, G., van Attikum, H., Poot, R.A., Vermeulen, W., and Lans, H. (2014b). Human ISWI complexes are targeted by SMARCA5 ATPase and SLIDE domains to help resolve lesion-stalled transcription. Nucleic Acids Res. *42*, 8473–8485.

Barak, B., and Feng, G. (2016). Neurobiology of social behavior abnormalities in autism and Williams syndrome. Nat. Neurosci. *19*, 647–655.

Barbagiovanni (2018). KMT2B is selectively required for neuronal transdifferentiation and its loss exposes dystonia candidate genes. Stem Cell Rep.

Barnett, C., and Krebs, J.E. (2011). WSTF does it all: a multifunctional protein in transcription, repair and replication. Biochem. Cell Biol. Biochim. Biol. Cell. *89*, 12–23.

Barnett, C., Yazgan, O., Kuo, H.-C., Malakar, S., Thomas, T., Fitzgerald, A., Harbour, W., Henry, J.J., and Krebs, J.E. (2012). Williams Syndrome Transcription Factor is critical for neural crest cell function in Xenopus laevis. Mech. Dev. *129*, 324–338.

Beagan, J.A., Duong, M.T., Titus, K.R., Zhou, L., Cao, Z., Ma, J., Lachanski, C.V., Gillis, D.R., and Phillips-Cremins, J.E. (2017). YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. Genome Res. *27*, 1139–1152.

Beby, F., and Lamonerie, T. (2013). The homeobox gene Otx2 in development and disease. Exp. Eye Res. *111*, 9–16.

Beck, D.B., Bonasio, R., Kaneko, S., Li, G., Li, G., Margueron, R., Oda, H., Sarma, K., Sims, R.J., Son, J., et al. (2010). Chromatin in the Nuclear Landscape. Cold Spring Harb. Symp. Quant. Biol. *75*, 11–22.

Belgrano, A., Rakicevic, L., Mittempergher, L., Campanaro, S., Martinelli, V.C., Mouly, V., Valle, G., Kojic, S., and Faulkner, G. (2011). Multi-tasking role of the mechanosensing protein Ankrd2 in the signaling network of striated muscle. PloS One *6*, e25519.

Bellugi, U., Lichtenberger, L., Jones, W., Lai, Z., and St George, M. (2000). I. The neurocognitive profile of Williams Syndrome: a complex pattern of strengths and weaknesses. J. Cogn. Neurosci. *12 Suppl 1*, 7–29.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53–59.

Berge, K.V.D., Hembach, K., Soneson, C., Tiberi, S., Clement, L., Love, M.I., Patro, R., and Robinson, M. (2018). RNA sequencing data: hitchhiker's guide to expression analysis (PeerJ Inc.).

Bhatt, S., Diaz, R., and Trainor, P.A. (2013). Signals and switches in Mammalian neural crest cell differentiation. Cold Spring Harb. Perspect. Biol. *5*.

Bonhoure, N., Bounova, G., Bernasconi, D., Praz, V., Lammers, F., Canella, D., Willis, I.M., Herr, W., Hernandez, N., Delorenzi, M., et al. (2014). Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. Genome Res. *24*, 1157–1168.

Boniolo, G., and Testa, G. (2012). The Identity of Living Beings, Epigenetics, and the Modesty of Philosophy. Erkenntnis *76*, 279–298.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. Cell *122*, 947–956.

Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K., et al. (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349–353.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

Brind'Amour, J., Liu, S., Hudson, M., Chen, C., Karimi, M.M., and Lorincz, M.C. (2015). An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. Nat. Commun. *6*, 6033.

Britsch, S., Li, L., Kirchhoff, S., Theuring, F., Brinkmann, V., Birchmeier, C., and Riethmacher, D. (1998). The ErbB2 and ErbB3 receptors and their ligand, neuregulin-1, are essential for development of the sympathetic nervous system. Genes Dev. *12*, 1825–1836.

Brown, J.L., Mucci, D., Whiteley, M., Dirksen, M.-L., and Kassis, J.A. (1998). The Drosophila Polycomb Group Gene pleiohomeotic Encodes a DNA Binding Protein with Homology to the Transcription Factor YY1. Mol. Cell *1*, 1057–1064.

Brüne, M. (2007). On human self-domestication, psychiatry, and eugenics. Philos. Ethics Humanit. Med. PEHM *2*, 21.

Burgess, P.W., and Stuss, D.T. (2017). Fifty Years of Prefrontal Cortex Research: Impact on Assessment. J. Int. Neuropsychol. Soc. JINS *23*, 755–767.

Cai, Y., Jin, J., Yao, T., Gottschalk, A.J., Swanson, S.K., Wu, S., Shi, Y., Washburn, M.P., Florens, L., Conaway, R.C., et al. (2007). YY1 functions with INO80 to activate transcription. Nat. Struct. Mol. Biol. *14*, 872–874.

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how and why? Mol. Cell *49*.

Carcamo-Orive, I., Hoffman, G.E., Cundiff, P., Beckmann, N.D., D'Souza, S.L., Knowles, J.W., Patel, A., Papatsenko, D., Abbasi, F., Reaven, G.M., et al. (2017). Analysis of Transcriptional Variability in a Large Human iPSC Library Reveals

Genetic and Non-genetic Determinants of Heterogeneity. Cell Stem Cell *20*, 518-532.e9.

Carelli, F.N., Sharma, G., and Ahringer, J. (2017). Broad Chromatin Domains: An Important Facet of Genome Regulation. BioEssays News Rev. Mol. Cell. Dev. Biol. *39*.

Caretti, G., Padova, M.D., Micales, B., Lyons, G.E., and Sartorelli, V. (2004). The Polycomb Ezh2 methyltransferase regulates muscle gene expression and skeletal muscle differentiation. Genes Dev. *18*, 2627–2638.

Carlén, M. (2017). What constitutes the prefrontal cortex? Science *358*, 478–482.

Casaletto, K.B., and Heaton, R.K. (2017). Neuropsychological Assessment: Past and Future. J. Int. Neuropsychol. Soc. JINS *23*, 778–790.

Cazals, Y., Bévengut, M., Zanella, S., Brocard, F., Barhanin, J., and Gestreau, C. (2015). KCNK5 channels mostly expressed in cochlear outer sulcus cells are indispensable for hearing. Nat. Commun. *6*, 8780.

de Chaldée, M., Brochier, C., Van de Vel, A., Caudy, N., Luthi-Carter, R., Gaillard, M.C., and Elalouf, J.M. (2006). Capucin: a novel striatal marker down-regulated in rodent models of Huntington disease. Genomics *87*, 200–207.

Chambers, S.M., Fasano, C.A., Papapetrou, E.P., Tomishima, M., Sadelain, M., and Studer, L. (2009). Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nat. Biotechnol. *27*, 275–280.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics *14*, 128.

Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J.B., and Gregor, T. (2018). Dynamic interplay between enhancer–promoter topology and gene activity. Nat. Genet. 1.

Christophersen, I.E., Olesen, M.S., Liang, B., Andersen, M.N., Larsen, A.P., Nielsen, J.B., Haunsø, S., Olesen, S.-P., Tveit, A., Svendsen, J.H., et al. (2013). Genetic variation in KCNA5: impact on the atrial-specific potassium current IKur in patients with lone atrial fibrillation. Eur. Heart J. *34*, 1517–1525.

Claes, P., Roosenboom, J., White, J.D., Swigut, T., Sero, D., Li, J., Lee, M.K., Zaidi, A., Mattern, B.C., Liebowitz, C., et al. (2018). Genome-wide mapping of global-to-local genetic effects on human facial shape. Nat. Genet. *50*, 414–423.

Colman, A. (2008). Induced Pluripotent Stem Cells and Human Disease. Cell Stem Cell *3*, 236–237.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.

Craft, A.M., and Johnson, M. (2017). From stem cells to human development: a distinctly human perspective on early embryology, cellular differentiation and translational research. Development *144*, 12–16.

Crespi, B.J., and Hurd, P.L. (2014). Cognitive-behavioral phenotypes of Williams syndrome are associated with genetic variation in the GTF2I gene, in a healthy population. BMC Neurosci. *15*, 127.

Croteau-Chonka, D.C., Rogers, A.J., Raj, T., McGeachie, M.J., Qiu, W., Ziniti, J.P., Stubbs, B.J., Liang, L., Martinez, F.D., Strunk, R.C., et al. (2015). Expression

Quantitative Trait Loci Information Improves Predictive Modeling of Disease Relevance of Non-Coding Genetic Variation. PloS One *10*, e0140758.

Culver-Cochran, A.E., and Chadwick, B.P. (2013). Loss of WSTF results in spontaneous fluctuations of heterochromatin formation and resolution, combined with substantial changes to gene expression. BMC Genomics *14*, 740.

Dai, L., Bellugi, U., Chen, X.-N., Pulst-Korenberg, A.M., Järvinen-Pasley, A., Tirosh-Wagner, T., Eis, P.S., Graham, J., Mills, D., Searcy, Y., et al. (2009). Is it Williams syndrome? GTF2IRD1 implicated in visual-spatial construction and GTF2I in sociability revealed by high resolution arrays. Am. J. Med. Genet. A. *149A*, 302–314.

Day, J.J., and Sweatt, J.D. (2011). Epigenetic Mechanisms in Cognition. Neuron *70*, 813–829.

De Mori, R., Romani, M., D'Arrigo, S., Zaki, M.S., Lorefice, E., Tardivo, S., Biagini, T., Stanley, V., Musaev, D., Fluss, J., et al. (2017). Hypomorphic Recessive Variants in SUFU Impair the Sonic Hedgehog Pathway and Cause Joubert Syndrome with Cranio-facial and Skeletal Defects. Am. J. Hum. Genet. *101*, 552–563.

Deciphering Developmental Disorders Study (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature *542*, 433–438.

Dekker, T.M., and Karmiloff-Smith, A. (2011). The dynamics of ontogeny: a neuroconstructivist perspective on genes, brains, cognition and behavior. Prog. Brain Res. *189*, 23–33.

Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J., and Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature *471*, 602–607.

DeLuca, D.S., Levin, J.Z., Sivachenko, A., Fennell, T., Nazaire, M.-D., Williams, C., Reich, M., Winckler, W., and Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics *28*, 1530–1532.

Dijk, E.L. van, Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. Trends Genet. *30*, 418–426.

Donohoe, M.E., Zhang, X., McGinnis, L., Biggers, J., Li, E., and Shi, Y. (1999). Targeted disruption of mouse Yin Yang 1 transcription factor results in peri-implantation lethality. Mol. Cell. Biol. *19*, 7237–7244.

Dorighi, K.M., Swigut, T., Henriques, T., Bhanu, N.V., Scruggs, B.S., Nady, N., Still, C.D., Garcia, B.A., Adelman, K., and Wysocka, J. (2017). Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. Mol. Cell *66*, 568-576.e4.

D'Souza, H., and Karmiloff-Smith, A. (2017). Neurodevelopmental disorders. Wiley Interdiscip. Rev. Cogn. Sci. *8*.

Dupin, E., Calloni, G.W., Coelho-Aguiar, J.M., and Le Douarin, N.M. (2018). The issue of the multipotency of the neural crest cells. Dev. Biol.

Durdiaková, J., Warrier, V., Banerjee-Basu, S., Baron-Cohen, S., and Chakrabarti, B. (2014). STX1A and Asperger syndrome: a replication study. Mol. Autism *5*, 14.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinforma. Oxf. Engl. *21*, 3439–3440.

Edelmann, L., Prosnitz, A., Pardo, S., Bhatt, J., Cohen, N., Lauriat, T., Ouchanov, L., González, P.J., Manghi, E.R., Bondy, P., et al. (2007). An atypical deletion of the Williams-Beuren syndrome interval implicates genes associated with defective visuospatial processing and autism. J. Med. Genet. *44*, 136–143.

Edgin, J.O., Clark, C.A.C., Massand, E., and Karmiloff-Smith, A. (2015). Building an adaptive brain across development: targets for neurorehabilitation must begin in infancy. Front. Behav. Neurosci. *9*, 232.

Eferl, R., Hoebertz, A., Schilling, A.F., Rath, M., Karreth, F., Kenner, L., Amling, M., and Wagner, E.F. (2004). The Fos-related antigen Fra-1 is an activator of bone matrix formation. EMBO J. *23*, 2789–2799.

Ernst, C. (2016). Proliferation and Differentiation Deficits are a Major Convergence Point for Neurodevelopmental Disorders. Trends Neurosci. *39*, 290–299.

Ferrero, G.B., Howald, C., Micale, L., Biamino, E., Augello, B., Fusco, C., Turturo, M.G., Forzano, S., Reymond, A., and Merla, G. (2010). An atypical 7q11.23 deletion in a normal IQ Williams-Beuren syndrome patient. Eur. J. Hum. Genet. EJHG *18*, 33–38.

Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Vooren, S.V., Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am. J. Hum. Genet. *84*, 524–533.

Flaherty, E.K., and Brennand, K.J. (2017). Using hiPSCs to model neuropsychiatric copy number variations (CNVs) has potential to reveal underlying disease mechanisms. Brain Res. *1655*, 283–293.

Fletcher, J.M., and Grigorenko, E.L. (2017). Neuropsychology of Learning Disabilities: The Past and the Future. J. Int. Neuropsychol. Soc. *23*, 930–940.

Froimchuk, E., Jang, Y., and Ge, K. (2017). Histone H3 lysine 4 methyltransferase KMT2D. Gene *627*, 337–342.

Fusaki, N., Ban, H., Nishiyama, A., Saeki, K., and Hasegawa, M. (2009). Efficient induction of transgene-free human pluripotent stem cells using a vector based on Sendai virus, an RNA virus that does not integrate into the host genome. Proc. Jpn. Acad. Ser. B Phys. Biol. Sci. *85*, 348–362.

Fyfe, J.C., Menotti-Raymond, M., David, V.A., Brichta, L., Schäffer, A.A., Agarwala, R., Murphy, W.J., Wedemeyer, W.J., Gregory, B.L., Buzzell, B.G., et al. (2006). An ~140-kb deletion associated with feline spinal muscular atrophy implies an essential LIX1 function for motor neuron survival. Genome Res. *16*, 1084–1090.

Gabriele, M., Silfhout, A.T.V., Germain, P.-L., Vitriolo, A., Kumar, R., Douglas, E., Haan, E., Kosaki, K., Takenouchi, T., Rauch, A., et al. (2017). YY1 Haploinsufficiency Causes an Intellectual Disability Syndrome Featuring Transcriptional and Chromatin Dysfunction. Am. J. Hum. Genet. *100*, 907–925.

Gabriele, M., Lopez Tobon, A., D'Agostino, G., and Testa, G. (2018a). The chromatin basis of neurodevelopmental disorders: Rethinking dysfunction along the molecular and temporal axes. Prog. Neuropsychopharmacol. Biol. Psychiatry.

Gabriele, M., Lopez Tobon, A., D'Agostino, G., and Testa, G. (2018b). The chromatin basis of neurodevelopmental disorders: Rethinking dysfunction along the molecular and temporal axes. Prog. Neuropsychopharmacol. Biol. Psychiatry.

van Galen, P., Viny, A.D., Ram, O., Ryan, R.J.H., Cotton, M.J., Donohue, L., Sievers, C., Drier, Y., Liau, B.B., Gillespie, S.M., et al. (2016). A Multiplexed System for Quantitative Comparisons of Chromatin Landscapes. Mol. Cell *61*, 170–180.

Gao, Z., Ure, K., Ding, P., Nashaat, M., Yuan, L., Ma, J., Hammer, R.E., and Hsieh, J. (2011). The Master Negative Regulator REST/NRSF Controls Adult Neurogenesis by Restraining the Neurogenic Program in Quiescent Stem Cells. J. Neurosci. *31*, 9772–9786.

Germain, P.-L., Ratti, E., and Boem, F. (2014). Junk or functional DNA? ENCODE and the function controversy. Biol. Philos. *29*, 807–831.

Germain, P.-L., Vitriolo, A., Adamo, A., Laise, P., Das, V., and Testa, G. (2016). RNAontheBENCH: computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. Nucleic Acids Res. *44*, 5054–5067.

Germain, P.-L., Testa, G., Sun, C.-W., George, D.R., Ding, L., Miller, C.A., Ley, T.J., Goldmann, J.M., Pervouchine, D.D., and Sullivan, T.J. (2017). Taming Human Genetic Variability: Transcriptomic Meta-Analysis Guides the Experimental Design and Interpretation of iPSC-Based Disease Modeling. Stem Cell Rep. *8*, 1784–1796.

Ghavi-Helm, Y., Klein, F.A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., and Furlong, E.E.M. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. Nature *512*, 96–100.

Gibson, W.T., Hood, R.L., Zhan, S.H., Bulman, D.E., Fejes, A.P., Moore, R., Mungall, A.J., Eydoux, P., Babul-Hirji, R., An, J., et al. (2012). Mutations in EZH2 Cause Weaver Syndrome. Am. J. Hum. Genet. *90*, 110–118.

Ginalski, K., Rychlewski, L., Baker, D., and Grishin, N.V. (2004). Protein structure prediction for the male-specific region of the human Y chromosome. Proc. Natl. Acad. Sci. *101*, 2305–2310.

Goodwin, L.R., and Picketts, D.J. (2018). The role of ISWI chromatin remodeling complexes in brain development and neurodevelopmental disorders. Mol. Cell. Neurosci. *87*, 55–64.

Greenfield, A., Carrel, L., Pennisi, D., Philippe, C., Quaderi, N., Siggers, P., Steiner, K., Tam, P.P.L., Monaco, A.P., Willard, H.F., et al. (1998). The UTX Gene Escapes X Inactivation in Mice and Humans. Hum. Mol. Genet. *7*, 737–742.

Grzybowski, A.T., Chen, Z., and Ruthenburg, A.J. (2015). Calibrating ChIP-Seq with Nucleosomal Internal Standards to Measure Histone Modification Density Genome Wide. Mol. Cell *58*, 886–899.

Hakre, S., Tussie-Luna, M.I., Ashworth, T., Novina, C.D., Settleman, J., Sharp, P.A., and Roy, A.L. (2006). Opposing functions of TFII-I spliced isoforms in growth factor-induced gene expression. Mol. Cell *24*, 301–308.

Hansen, P., Hecht, J., Ibrahim, D.M., Krannich, A., Truss, M., and Robinson, P.N. (2015). Saturation analysis of ChIP-seq data for reproducible identification of binding peaks. Genome Res. *25*, 1391–1400.

Hawrylycz, M.J., Lein, E.S., Guillozet-Bongaarts, A.L., Shen, E.H., Ng, L., Miller, J.A., van de Lagemaat, L.N., Smith, K.A., Ebbert, A., Riley, Z.L., et al. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. Nature *489*, 391–399.

He, Y., and Casaccia-Bonnefil, P. (2008). The Yin and Yang of YY1 in the nervous system. J. Neurochem. *106*, 1493–1502.

He, Y., Dupree, J., Wang, J., Sandoval, J., Li, J., Liu, H., Shi, Y., Nave, K.A., and Casaccia-Bonnefil, P. (2007). The Transcription Factor Yin Yang 1 Is Essential for Oligodendrocyte Progenitor Differentiation. Neuron *55*, 217–230.

He, Y., Kim, J.Y., Dupree, J., Tewari, A., Melendez-Vasquez, C., Svaren, J., and Casaccia, P. (2010). Yy1 as a molecular link between neuregulin and transcriptional modulation of peripheral myelination. Nat. Neurosci. *13*, 1472–1480.

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

Helsmoortel, C., Vulto-van Silfhout, A.T., Coe, B.P., Vandeweyer, G., Rooms, L., van den Ende, J., Schuurs-Hoeijmakers, J.H.M., Marcelis, C.L., Willemsen, M.H., Vissers, L.E.L.M., et al. (2014). A SWI/SNF-related autism syndrome caused by de novo mutations in ADNP. Nat. Genet. *46*, 380–384.

Henriquez, B., Bustos, F.J., Aguilar, R., Becerra, A., Simon, F., Montecino, M., and van Zundert, B. (2013). Ezh1 and Ezh2 differentially regulate PSD-95 gene transcription in developing hippocampal neurons. Mol. Cell. Neurosci. *57*, 130–143.

Hochedlinger, K., Blelloch, R., Brennan, C., Yamada, Y., Kim, M., Chin, L., and Jaenisch, R. (2004). Reprogramming of a melanoma genome by nuclear transplantation. Genes Dev. *18*, 1875–1885.

Hong, S., Cho, Y.-W., Yu, L.-R., Yu, H., Veenstra, T.D., and Ge, K. (2007). Identification of JmjC domain-containing UTX and JMJD3 as histone H3 lysine 27 demethylases. Proc. Natl. Acad. Sci. *104*, 18439–18444.

Hu, Z., and Tee, W.-W. (2017). Enhancers and chromatin structures: regulatory hubs in gene expression and diseases. Biosci. Rep. *37*.

Huang, C., and Zhu, B. (2018). Roles of H3K36-specific histone methyltransferases in transcription: antagonizing silencing and safeguarding transcription fidelity. Biophys. Rep. *4*, 170–177.

Ilieva, M., Fex Svenningsen, Å., Thorsen, M., and Michel, T.M. (2018). Psychiatry in a Dish: Stem Cells and Brain Organoids Modeling Autism Spectrum Disorders. Biol. Psychiatry *83*, 558–568.

Imagawa, E., Albuquerque, E.V.A., Isidor, B., Mitsuhashi, S., Mizuguchi, T., Miyatake, S., Takata, A., Miyake, N., Boguszewski, M.C.S., Boguszewski, C.L., et al. (2018). Novel SUZ12 mutations in Weaver-like syndrome. Clin. Genet. *94*, 461–466.

Imprialou, M., Petretto, E., and Bottolo, L. (2017). Expression QTLs Mapping and Analysis: A Bayesian Perspective. Methods Mol. Biol. Clifton NJ *1488*, 189–215.

Jeon, Y., and Lee, J.T. (2011). YY1 tethers Xist RNA to the inactive X nucleation center. Cell *146*, 119–133.

Jepsen, K., Solum, D., Zhou, T., McEvilly, R.J., Kim, H.-J., Glass, C.K., Hermanson, O., and Rosenfeld, M.G. (2007). SMRT-mediated repression of an H3K27 demethylase in progression from neural stem cell to neuron. Nature *450*, 415–419.

Ji, J., Ng, S.H., Sharma, V., Neculai, D., Hussein, S., Sam, M., Trinh, Q., Church, G.M., Mcpherson, J.D., Nagy, A., et al. (2012). Elevated Coding Mutation Rate During the Reprogramming of Human Somatic Cells into Induced Pluripotent Stem Cells. STEM CELLS *30*, 435–440.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science *337*, 816–821.

Johnson, M.H. (2011). Interactive Specialization: A domain-general framework for human functional brain development? Dev. Cogn. Neurosci. *1*, 7–21.

Jost, D., Carrivain, P., Cavalli, G., and Vaillant, C. (2014). Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. Nucleic Acids Res. *42*, 9553–9561.

Ju, J., Kim, D.H., Bi, L., Meng, Q., Bai, X., Li, Z., Li, X., Marma, M.S., Shi, S., Wu, J., et al. (2006). Four-color DNA sequencing by synthesis using cleavable

fluorescent nucleotide reversible terminators. Proc. Natl. Acad. Sci. U. S. A. *103*, 19635–19640.

Jung, Y.L., Luquette, L.J., Ho, J.W.K., Ferrari, F., Tolstorukov, M., Minoda, A., Issner, R., Epstein, C.B., Karpen, G.H., Kuroda, M.I., et al. (2014). Impact of sequencing depth in ChIP-seq experiments. Nucleic Acids Res. *42*, e74.

Kaang, B.-K., and Kim, S. (2017). Epigenetic regulation and chromatin remodeling in learning and memory. Exp. Mol. Med. *49*, e281.

Kamakaka, R.T., and Biggins, S. (2005). Histone variants: deviants? Genes Dev. *19*, 295–316.

Karmiloff-Smith, A. (2006). The tortuous route from genes to behavior: A neuroconstructivist approach. Cogn. Affect. Behav. Neurosci. *6*, 9–17.

Karmiloff-Smith, A. (2007). Atypical epigenesis. Dev. Sci. *10*, 84–88.

Karmiloff-Smith, A. (2009). Nativism versus neuroconstructivism: rethinking the study of developmental disorders. Dev. Psychol. *45*, 56–63.

Karmiloff-Smith, A. (2010). Neuroimaging of the developing brain: taking “developing” seriously. Hum. Brain Mapp. *31*, 934–941.

Karmiloff-Smith, A. (2012). Perspectives on the dynamic development of cognitive capacities: insights from Williams syndrome. Curr. Opin. Neurol. *25*, 106–111.

Karmiloff-Smith, A., D’Souza, D., Dekker, T.M., Van Herwegen, J., Xu, F., Rodic, M., and Ansari, D. (2012). Genetic and environmental vulnerabilities in children with neurodevelopmental disorders. Proc. Natl. Acad. Sci. U. S. A. *109 Suppl 2*, 17261–17265.

Khare, S.P., Habib, F., Sharma, R., Gadewal, N., Gupta, S., and Galande, S. (2012). HIstome—a relational knowledgebase of human histone proteins and histone modifying enzymes. Nucleic Acids Res. *40*, D337–D342.

Kilpinen, H., Goncalves, A., Leha, A., Afzal, V., Alasoo, K., Ashford, S., Bala, S., Bensaddek, D., Casale, F.P., Culley, O.J., et al. (2017). Common genetic variation drives molecular heterogeneity in human iPSCs. Nature *546*, 370–375.

Kim, K., Doi, A., Wen, B., Ng, K., Zhao, R., Cahan, P., Kim, J., Aryee, M.J., Ji, H., Ehrlich, L.I.R., et al. (2010). Epigenetic memory in induced pluripotent stem cells. Nature *467*, 285–290.

Kinsella, R.J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., et al. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database *2011*.

Kitagawa, H., Fujiki, R., Yoshimura, K., Oya, H., and Kato, S. (2011). Williams syndrome is an epigenome-regulator disease. Endocr. J. *58*, 77–85.

Ko, S.O., Chung, I.H., Xu, X., Oka, S., Zhao, H., Cho, E.S., Deng, C., and Chai, Y. (2007). Smad4 is required to regulate the fate of cranial neural crest cells. Dev. Biol. *312*, 435–447.

Köhler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S.M., Boerkoel, C.F., Boycott, K.M., et al. (2017). The Human Phenotype Ontology in 2017. Nucleic Acids Res. *45*, D865–D876.

Komori, T. (2018). Runx2, an inducer of osteoblast and chondrocyte differentiation. Histochem. Cell Biol. *149*, 313–323.

Kouzarides, T. (2007). Chromatin Modifications and Their Function. Cell *128*, 693–705.

Krum, S.A., Chang, J., Miranda-Carboni, G., and Wang, C.-Y. (2010). Novel functions for NFκB: inhibition of bone formation. Nat. Rev. Rheumatol. *6*, 607–611.

Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. *44*, W90-97.

Kuniba, H., Yoshiura, K., Kondoh, T., Ohashi, H., Kurosawa, K., Tonoki, H., Nagai, T., Okamoto, N., Kato, M., Fukushima, Y., et al. (2009). Molecular karyotyping in 17 patients and mutation screening in 41 patients with Kabuki syndrome. J. Hum. Genet. *54*, 304–309.

Kurosaki, T., and Maquat, L.E. (2016). Nonsense-mediated mRNA decay in humans at a glance. J. Cell Sci. *129*, 461–467.

Lan, F., Bayliss, P.E., Rinn, J.L., Whetstine, J.R., Wang, J.K., Chen, S., Iwase, S., Alpatov, R., Issaeva, I., Canaani, E., et al. (2007). A histone H3 lysine 27 demethylase regulates animal posterior development. Nature *449*, 689–694.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lathrop, M.J., Chakrabarti, L., Eng, J., Harker Rhodes, C., Lutz, T., Nieto, A., Denny Liggitt, H., Warner, S., Fields, J., Stöger, R., et al. (2010). Deletion of the Chd6 exon 12 affects motor coordination. Mamm. Genome *21*, 130–142.

Lawrence, M., Daujat, S., and Schneider, R. (2016). Lateral Thinking: How Histone Modifications Regulate Gene Expression. Trends Genet. *32*, 42–56.

Lee, C. (2018). Genome-Wide Expression Quantitative Trait Loci Analysis Using Mixed Models. Front. Genet. *9*.

Lee, E.-W., Lee, M.-S., Camus, S., Ghim, J., Yang, M.-R., Oh, W., Ha, N.-C., Lane, D.P., and Song, J. (2009). Differential regulation of p53 and p21 by MKRN1 E3 ligase controls cell cycle arrest and apoptosis. EMBO J. *28*, 2100–2113.

Lee, J.S., Galvin, K.M., See, R.H., Eckner, R., Livingston, D., Moran, E., and Shi, Y. (1995). Relief of YY1 transcriptional repression by adenovirus E1A is mediated by E1A-associated protein p300. Genes Dev. *9*, 1188–1198.

Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isono, K., et al. (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. Cell *125*, 301–313.

Li, M., and Belmonte, J.C.I. (2017). Ground rules of the pluripotency gene regulatory network. Nat. Rev. Genet. *18*, 180–191.

Li, F., Mao, G., Tong, D., Huang, J., Gu, L., Yang, W., and Li, G.-M. (2013). The Histone Mark H3K36me3 Regulates Human DNA Mismatch Repair through its Interaction with MutSα. Cell *153*, 590–600.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinforma. Oxf. Engl. *30*, 923–930.

Lloyd, J.P.B. (2018). The evolution and diversity of the nonsense-mediated mRNA decay pathway. F1000Research *7*, 1299.

Long, H.K., Prescott, S.L., and Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell *167*, 1170–1187.

Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251–260.

Lundqvist, J., Hansen, S.K., and Lykkesfeldt, A.E. (2013). Vitamin D analog EB1089 inhibits aromatase expression by dissociation of comodulator WSTF from the CYP19A1 promoter—a new regulatory pathway for aromatase. Biochim. Biophys. Acta BBA - Mol. Cell Res. *1833*, 40–47.

Ma, S.L., Ng, H.K., Baum, L., Pang, J.C.S., Chiu, H.F.K., Woo, J., Tang, N.L.S., and Lam, L.C.W. (2002). Low-density lipoprotein receptor-related protein 8 (apolipoprotein E receptor 2) gene polymorphisms in Alzheimer's disease. Neurosci. Lett. *332*, 216–218.

Maekawa, M., and Yamanaka, S. (2011). Glis1, a unique pro-reprogramming factor, may facilitate clinical applications of iPSC technology. Cell Cycle *10*, 3613–3614.

Makhlouf, M., Ouimette, J.-F., Oldfield, A., Navarro, P., Neuillet, D., and Rougeulle, C. (2014). A prominent and conserved role for YY1 in *Xist* transcriptional activation. Nat. Commun. *5*, 4878.

Malenfant, P., Liu, X., Hudson, M.L., Qiao, Y., Hrynchak, M., Riendeau, N., Hildebrand, M.J., Cohen, I.L., Chudley, A.E., Forster-Gibson, C., et al. (2012).

Association of GTF2i in the Williams-Beuren syndrome critical region with autism spectrum disorders. J. Autism Dev. Disord. *42*, 1459–1469.

Mandel, S., and Gozes, I. (2007). Activity-dependent neuroprotective protein constitutes a novel element in the SWI/SNF chromatin remodeling complex. J. Biol. Chem. *282*, 34448–34456.

Mandel, S., Rechavi, G., and Gozes, I. (2007). Activity-dependent neuroprotective protein (ADNP) differentially interacts with chromatin to regulate genes essential for embryogenesis. Dev. Biol. *303*, 814–824.

Margueron, R., and Reinberg, D. (2011). The Polycomb complex PRC2 and its mark in life. Nature *469*, 343–349.

Margueron, R., Justin, N., Ohno, K., Sharpe, M.L., Son, J., Drury Iii, W.J., Voigt, P., Martin, S.R., Taylor, W.R., De Marco, V., et al. (2009). Role of the polycomb protein EED in the propagation of repressive histone marks. Nature *461*, 762–767.

Martin, L.A., Iceberg, E., and Allaf, G. (2018). Consistent hypersocial behavior in mice carrying a deletion of Gtf2i but no evidence of hyposocial behavior with Gtf2i duplication: Implications for Williams-Beuren syndrome and autism spectrum disorder. Brain Behav. *8*, e00895.

Mas, G., Blanco, E., Ballaré, C., Sansó, M., Spill, Y.G., Hu, D., Aoi, Y., Dily, F.L., Shilatifard, A., Marti-Renom, M.A., et al. (2018). Promoter bivalency favors an open chromatin architecture in embryonic stem cells. Nat. Genet. *50*, 1452.

Matharu, N., and Ahituv, N. (2015). Minor Loops in Major Folds: Enhancer–Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease. PLOS Genet. *11*, e1005640.

Meloni, M., and Testa, G. (2014). Scrutinizing the epigenetics revolution. Biosocieties *9*, 431–456.

Menendez, L., Kulik, M.J., Page, A.T., Park, S.S., Lauderdale, J.D., Cunningham, M.L., and Dalton, S. (2013). Directed differentiation of human pluripotent cells to neural crest stem cells. Nat. Protoc. *8*, 203–212.

Meng, J., Zhang, X.-T., Liu, X.-L., Fan, L., Li, C., Sun, Y., Liang, X.-H., Wang, J.-B., Mei, Q.-B., Zhang, F., et al. (2016). WSTF promotes proliferation and invasion of lung cancer cells by inducing EMT via PI3K/Akt and IL-6/STAT3 signaling pathways. Cell. Signal. *28*, 1673–1682.

Mervis, C.B., Dida, J., Lam, E., Crawford-Zelli, N.A., Young, E.J., Henderson, D.R., Onay, T., Morris, C.A., Woodruff-Borden, J., Yeomans, J., et al. (2012). Duplication of GTF2I results in separation anxiety in mice and humans. Am. J. Hum. Genet. *90*, 1064–1070.

Meyer-Lindenberg, A., Kohn, P., Mervis, C.B., Kippenhan, J.S., Olsen, R.K., Morris, C.A., and Berman, K.F. (2004). Neural basis of genetically determined visuospatial construction deficit in Williams syndrome. Neuron *43*, 623–631.

Minoux, M., Holwerda, S., Vitobello, A., Kitazawa, T., Kohler, H., Stadler, M.B., and Rijli, F.M. (2017). Gene bivalency at Polycomb domains regulates cranial neural crest positional identity. Science *355*, eaal2913.

Mishina, Y., and Snider, T.N. (2014). Neural crest cell signaling pathways critical to cranial bone development and pathology. Exp. Cell Res. *325*, 138–147.

Moeller, C., Yaylaoglu, M.B., Alvarez-Bolado, G., Thaller, C., and Eichele, G. (2002). Murine Lix1, a novel marker for substantia nigra, cortical layer 5, and hindbrain structures. Gene Expr. Patterns *1*, 199–203.

Molnár, Z., and Pollen, A. (2014). How unique is the human neocortex? Development *141*, 11–16.

Morris, C.A., Mervis, C.B., Hobart, H.H., Gregg, R.G., Bertrand, J., Ensing, G.J., Sommer, A., Moore, C.A., Hopkin, R.J., Spallone, P.A., et al. (2003). GTF2I hemizygosity implicated in mental retardation in Williams syndrome: genotype-phenotype analysis of five families with deletions in the Williams syndrome region. Am. J. Med. Genet. A. *123A*, 45–59.

Motohashi, H., Igarashi, K., Onodera, K., Takahashi, S., Ohtani, H., Nakafuku, M., Nishizawa, M., Engel, J.D., and Yamamoto, M. (1996). **Mesodermal- vs. neuronal-specific expression of MafK is elicited by different promoters**. Genes Cells *1*, 223–238.

Mungall, C.J., McMurry, J.A., Köhler, S., Balhoff, J.P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., et al. (2017). The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res. *45*, D712–D722.

Muratore, C.R., Srikanth, P., Callahan, D.G., and Young-Pearse, T.L. (2014). Comparison and Optimization of hiPSC Forebrain Cortical Differentiation Protocols. PLOS ONE *9*, e105807.

Nagueh, S.F., Shah, G., Wu, Y., Torre-Amione, G., King, N.M.P., Lahmers, S., Witt, C.C., Becker, K., Labeit, S., and Granzier, H.L. (2004). Altered titin expression,

myocardial stiffness, and left ventricular function in patients with dilated cardiomyopathy. Circulation *110*, 155–162.

Nakagawa, M., Koyanagi, M., Tanabe, K., Takahashi, K., Ichisaka, T., Aoi, T., Okita, K., Mochiduki, Y., Takizawa, N., and Yamanaka, S. (2008). Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. Nat. Biotechnol. *26*, 101–106.

Nakato, R., and Shirahige, K. (2017). Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Brief. Bioinform. *18*, 279–290.

Noctor, S.C., Flint, A.C., Weissman, T.A., Dammerman, R.S., and Kriegstein, A.R. (2001). Neurons derived from radial glial cells establish radial units in neocortex. Nature *409*, 714–720.

Noctor, S.C., Martínez-Cerdeño, V., Ivic, L., and Kriegstein, A.R. (2004). Cortical neurons arise in symmetric and asymmetric division zones and migrate through specific phases. Nat. Neurosci. *7*, 136–144.

Okita, K., Ichisaka, T., and Yamanaka, S. (2007). Generation of germline-competent induced pluripotent stem cells. Nature *448*, 313–317.

Okita, Y., Kamoshida, A., Suzuki, H., Itoh, K., Motohashi, H., Igarashi, K., Yamamoto, M., Ogami, T., Koinuma, D., and Kato, M. (2013). Transforming Growth Factor-β Induces Transcription Factors MafK and Bach1 to Suppress Expression of the Heme Oxygenase-1 Gene. J. Biol. Chem. *288*, 20658–20667.

Oldre, A., Szafer, A., Jones, A.R., Stevens, A., Ebbert, A., Bernard, A., Sodt, A.J., Carey, A., Facer, B.A.C., Gregor, B.W., et al. (2014). Transcriptional landscape of the prenatal human brain. Nature *508*, 199.

Omwancha, J., Zhou, X.-F., Chen, S.-Y., Baslan, T., Fisher, C.J., Zheng, Z., Cai, C., and Shemshedini, L. (2006). Makorin RING finger protein 1 (MKRN1) has negative and positive effects on RNA polymerase II-dependent transcription. Endocrine *29*, 363–373.

Orlando, D.A., Chen, M.W., Brown, V.E., Solanki, S., Choi, Y.J., Olson, E.R., Fritz, C.C., Bradner, J.E., and Guenther, M.G. (2014). Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. Cell Rep. *9*, 1163–1170.

Osborne, L.R. (2010). Animal models of Williams syndrome. Am. J. Med. Genet. C Semin. Med. Genet. *154C*, 209–219.

Osborne, L.R., and Mervis, C.B. (2007). Rearrangements of the Williams-Beuren syndrome locus: molecular basis and implications for speech and language development. Expert Rev. Mol. Med. *9*, 1–16.

Ostapcuk, V., Mohn, F., Carl, S.H., Basters, A., Hess, D., Iesmantavicius, V., Lampersberger, L., Flemr, M., Pandey, A., Thomä, N.H., et al. (2018). Activity-dependent neuroprotective protein recruits HP1 and CHD4 to control lineage-specifying genes. Nature *557*, 739–743.

Osumi, N., Shinohara, H., Numayama-Tsuruta, K., and Maekawa, M. (2008). Concise review: Pax6 transcription factor contributes to both embryonic and adult neurogenesis as a multifunctional regulator. Stem Cells Dayt. Ohio *26*, 1663–1672.

Paşca, A.M., Sloan, S.A., Clarke, L.E., Tian, Y., Makinson, C.D., Huber, N., Kim, C.H., Park, J.-Y., O'Rourke, N.A., Nguyen, K.D., et al. (2015). Functional cortical

neurons and astrocytes from human pluripotent stem cells in 3D culture. Nat. Methods *12*, 671–678.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419.

Pereira, J.D., Sansom, S.N., Smith, J., Dobenecker, M.-W., Tarakhovsky, A., and Livesey, F.J. (2010). Ezh2, the histone methyltransferase of PRC2, regulates the balance between self-renewal and differentiation in the cerebral cortex. Proc. Natl. Acad. Sci. U. S. A. *107*, 15957–15962.

Pinhasov, A., Mandel, S., Torchinsky, A., Giladi, E., Pittel, Z., Goldsweig, A.M., Servoss, S.J., Brenneman, D.E., and Gozes, I. (2003). Activity-dependent neuroprotective protein: a novel gene essential for brain formation. Brain Res. Dev. Brain Res. *144*, 83–90.

Poirier, S., Mayer, G., Benjannet, S., Bergeron, E., Marcinkiewicz, J., Nassoury, N., Mayer, H., Nimpf, J., Prat, A., and Seidah, N.G. (2008). The proprotein convertase PCSK9 induces the degradation of low density lipoprotein receptor (LDLR) and its closest family members VLDLR and ApoER2. J. Biol. Chem. *283*, 2363–2372.

Pott, S., and Lieb, J.D. (2015). What are super-enhancers? Nat. Genet. *47*, 8–12.

Quadrato, G., Brown, J., and Arlotta, P. (2016). The promises and challenges of human brain organoids as models of neuropsychiatric disease. Nat. Med. *22*, 1220–1228.

Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr. Protoc. Bioinforma. *47*, 11.12.1-34.

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

Rada-Iglesias, A. (2018). Is H3K4me1 at enhancers correlative or causative? Nat. Genet. *50*, 4.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Rakic, P. (1972). Mode of cell migration to the superficial layers of fetal monkey neocortex. J. Comp. Neurol. *145*, 61–83.

Ralston, A., and Rossant, J. (2005). Genetic regulation of stem cell origins in the mouse embryo. Clin. Genet. *68*, 106–112.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165.

Rappaport, N., Nativ, N., Stelzer, G., Twik, M., Guan-Golan, Y., Iny Stein, T., Bahir, I., Belinky, F., Morrey, C.P., Safran, M., et al. (2013). MalaCards: an integrated compendium for diseases and their annotation. Database J. Biol. Databases Curation *2013*.

Ratan, Z.A., Son, Y.-J., Haidere, M.F., Uddin, B.M.M., Yusuf, M.A., Zaman, S.B., Kim, J.-H., Banu, L.A., and Cho, J.Y. (2018). CRISPR-Cas9: a promising genetic engineering approach in cancer research. Ther. Adv. Med. Oncol. *10*, 1758834018755089.

Reshef, Y.A., Finucane, H.K., Kelley, D.R., Gusev, A., Kotliar, D., Ulirsch, J.C., Hormozdiari, F., Nasser, J., O'Connor, L., Geijn, B. van de, et al. (2018). Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. Nat. Genet. 1.

Ribeyre, C., Zellweger, R., Chauvin, M., Bec, N., Larroque, C., Lopes, M., and Constantinou, A. (2016). Nascent DNA Proteomics Reveals a Chromatin Remodeler Required for Topoisomerase I Loading at Replication Forks. Cell Rep. *15*, 300–309.

Rickels, R., and Shilatifard, A. (2018). Enhancer Logic and Mechanics in Development and Disease. Trends Cell Biol. *28*, 608–630.

Rickels, R., Herz, H.-M., Sze, C.C., Cao, K., Morgan, M.A., Collings, C.K., Gause, M., Takahashi, Y., Wang, L., Rendleman, E.J., et al. (2017). Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. Nat. Genet. *49*, 1647–1653.

Riedel, B.C., Thompson, P.M., and Brinton, R.D. (2016). Age, APOE and sex: Triad of risk of Alzheimer's disease. J. Steroid Biochem. Mol. Biol. *160*, 134–147.

Ropers, H.H. (2010). Genetics of Early Onset Cognitive Impairment. Annu. Rev. Genomics Hum. Genet. *11*, 161–187.

Roy, A.L. (2001). Biochemistry and biology of the inducible multifunctional transcription factor TFII-I. Gene *274*, 1–13.

Ruthenburg, A.J., Allis, C.D., and Wysocka, J. (2007). Methylation of Lysine 4 on Histone H3: Intricacy of Writing and Reading a Single Epigenetic Mark. Mol. Cell *25*, 15–30.

Ryan, R.J.H., and Bernstein, B.E. (2012). Genetic Events That Shape the Cancer Epigenome. Science *336*, 1513–1514.

Sakurai, T., and Gamo, N.J. (2018). Cognitive functions associated with developing prefrontal cortex during adolescence and developmental neuropsychiatric disorders. Neurobiol. Dis.

Sanders, S.J., Ercan-Sencicek, A.G., Hus, V., Luo, R., Murtha, M.T., Moreno-De-Luca, D., Chu, S.H., Moreau, M.P., Gupta, A.R., Thomson, S.A., et al. (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron *70*, 863–885.

Santagati, F., and Rijli, F.M. (2003). Cranial neural crest and the building of the vertebrate head. Nat. Rev. Neurosci. *4*, 806–818.

Schmeisser, M.J., Ey, E., Wegener, S., Bockmann, J., Stempel, A.V., Kuebler, A., Janssen, A.-L., Udvardi, P.T., Shiban, E., Spilker, C., et al. (2012). Autistic-like behaviours and hyperactivity in mice lacking ProSAP1/Shank2. Nature *486*, 256–260.

Schmidt, C.K., and Jackson, S.P. (2013). On Your MARK, Get SET(D2), Go! H3K36me3 Primes DNA Mismatch Repair. Cell *153*, 513–515.

Schuurmans, C., Armant, O., Nieto, M., Stenman, J.M., Britz, O., Klenin, N., Brown, C., Langevin, L.-M., Seibt, J., Tang, H., et al. (2004). Sequential phases of

cortical specification involve Neurogenin-dependent and -independent pathways. EMBO J. *23*, 2892–2902.

Sebastian, A., Hum, N.R., Morfin, C., Murugesh, D.K., and Loots, G.G. (2018). Global gene expression analysis identifies Mef2c as a potential player in Wnt16-mediated transcriptional regulation. Gene *675*, 312–321.

Seidah, N.G., Benjannet, S., Wickham, L., Marcinkiewicz, J., Jasmin, S.B., Stifani, S., Basak, A., Prat, A., and Chrétien, M. (2003). The secretory proprotein convertase neural apoptosis-regulated convertase 1 (NARC-1): Liver regeneration and neuronal differentiation. Proc. Natl. Acad. Sci. *100*, 928–933.

Seifi, M., and Walter, M. a. (2018). Axenfeld-Rieger syndrome. Clin. Genet. *93*, 1123–1130.

Shi, Y., Seto, E., Chang, L.S., and Shenk, T. (1991). Transcriptional repression by YY1, a human GLI-Krüppel-related protein, and relief of repression by adenovirus E1A protein. Cell *67*, 377–388.

Short, P.J., McRae, J.F., Gallone, G., Sifrim, A., Won, H., Geschwind, D.H., Wright, C.F., Firth, H.V., FitzPatrick, D.R., Barrett, J.C., et al. (2018). *De novo* mutations in regulatory elements in neurodevelopmental disorders. Nature.

Sifrim, A., Hitz, M.-P., Wilsdon, A., Breckpot, J., Turki, S.H.A., Thienpont, B., McRae, J., Fitzgerald, T.W., Singh, T., Swaminathan, G.J., et al. (2016). Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. Nat. Genet. *48*, 1060–1065.

Sigova, A.A., Abraham, B.J., Ji, X., Molinie, B., Hannett, N.M., Guo, Y.E., Jangi, M., Giallourakis, C.C., Sharp, P.A., and Young, R.A. (2015). Transcription factor trapping by RNA in gene regulatory elements. Science *350*, 978–981.

Skene, P.J., and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. ELife *6*, e21856.

Smith, E., and Shilatifard, A. (2014). Enhancer biology and enhanceropathies. Nat. Struct. Mol. Biol. *21*, 210–219.

Smith, R.L., Redd, M.J., and Johnson, A.D. (1995). The tetratricopeptide repeats of Ssn6 interact with the homeo domain of alpha 2. Genes Dev. *9*, 2903–2910.

Sparmann, A., and van Lohuizen, M. (2006). Polycomb silencers control cell fate, development and cancer. Nat. Rev. Cancer *6*, 846–856.

Spokony, R.F., Aoki, Y., Saint-Germain, N., Magner-Fink, E., and Saint-Jeannet, J.-P. (2002). The transcription factor Sox9 is required for cranial neural crest development in Xenopus. Dev. Camb. Engl. *129*, 421–432.

van Steensel, B., and Belmont, A.S. (2017). Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. Cell *169*, 780–791.

Stiles, J., and Jernigan, T.L. (2010). The Basics of Brain Development. Neuropsychol. Rev. *20*, 327–348.

Streeter, I., Harrison, P.W., Faulconbridge, A., The HipSci Consortium, Flicek, P., Parkinson, H., and Clarke, L. (2017). The human-induced pluripotent stem cell initiative-data resources for cellular genetics. Nucleic Acids Res. *45*, D691–D697.

Sun, H., Martin, J.A., Werner, C.T., Wang, Z.-J., Damez-Werno, D.M., Scobie, K.N., Shao, N.-Y., Dias, C., Rabkin, J., Koo, J.W., et al. (2016). BAZ1B in Nucleus Accumbens Regulates Reward-Related Behaviors in Response to Distinct Emotional Stimuli. J. Neurosci. Off. J. Soc. Neurosci. *36*, 3954–3961.

Swartz, J.R., Waller, R., Bogdan, R., Knodt, A.R., Sabhlok, A., Hyde, L.W., and Hariri, A.R. (2017). A Common Polymorphism in a Williams Syndrome Gene Predicts Amygdala Reactivity and Extraversion in Healthy Adults. Biol. Psychiatry *81*, 203–210.

Tabarés-Seisdedos, R., and Rubenstein, J.L. (2013). Inverse cancer comorbidity: a serendipitous opportunity to gain insight into CNS disorders. Nat. Rev. Neurosci. *14*, 293–304.

Takahashi, K. (2014). Cellular Reprogramming. Cold Spring Harb. Perspect. Biol. *6*, a018606.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663–676.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861–872.

Talbert, P.B., and Henikoff, S. (2010). Histone variants — ancient wrap artists of the epigenome. Nat. Rev. Mol. Cell Biol. *11*, 264–275.

Tamming, R.J., Siu, J.R., Jiang, Y., Prado, M.A.M., Beier, F., and Bérubé, N.G. (2017). Mosaic expression of Atrx in the mouse central nervous system causes memory deficits. Dis. Model. Mech. *10*, 119–126.

Tan, G., and Lenhard, B. (2016). TFBSTools: an R/bioconductor package for transcription factor binding site analysis. Bioinforma. Oxf. Engl. *32*, 1555–1556.

Tatton-Brown, K., Loveday, C., Yost, S., Clarke, M., Ramsay, E., Zachariou, A., Elliott, A., Wylie, H., Ardissone, A., Rittinger, O., et al. (2017). Mutations in Epigenetic Regulation Genes Are a Major Cause of Overgrowth with Intellectual Disability. Am. J. Hum. Genet. *100*, 725–736.

Testa, G. (2011). The time of timing: how Polycomb proteins regulate neurogenesis. BioEssays News Rev. Mol. Cell. Dev. Biol. *33*, 519–528.

Theofanopoulou, C., Gastaldon, S., O'Rourke, T., Samuels, B.D., Martins, P.T., Delogu, F., Alamri, S., and Boeckx, C. (2017). Self-domestication in Homo sapiens: Insights from comparative genomics. PloS One *12*, e0185306.

Thomas, M., and Karmiloff-Smith, A. (2002). Are developmental disorders like cases of adult brain damage? Implications from connectionist modelling. Behav. Brain Sci. *25*, 727–750; discussion 750-787.

Tippens, N.D., Vihervaara, A., and Lis, J.T. (2018). Enhancer transcription: what, where, when, and why? Genes Dev. *32*, 1–3.

Tischfield, M.A., Robson, C.D., Gilette, N.M., Chim, S.M., Sofela, F.A., DeLisle, M.M., Gelber, A., Barry, B.J., MacKinnon, S., Dagi, L.R., et al. (2017). Cerebral Vein Malformations Result from Loss of Twist1 Expression and BMP Signaling from Skull Progenitor Cells and Dura. Dev. Cell *42*, 445-461.e5.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. *31*.

Tropberger, P., Pott, S., Keller, C., Kamieniarz-Gdula, K., Caron, M., Richter, F., Li, G., Mittler, G., Liu, E.T., Bühler, M., et al. (2013). Regulation of Transcription through Acetylation of H3K122 on the Lateral Surface of the Histone Octamer. Cell *152*, 859–872.

Ünsal-Kaçmaz, K., Mullen, T.E., Kaufmann, W.K., and Sancar, A. (2005). Coupling of Human Circadian and Cell Cycles by the Timeless Protein. Mol. Cell. Biol. *25*, 3109–3116.

Üren, A., Reichsman, F., Anest, V., Taylor, W.G., Muraiso, K., Bottaro, D.P., Cumberledge, S., and Rubin, J.S. (2000). Secreted Frizzled-related Protein-1 Binds Directly to Wingless and Is a Biphasic Modulator of Wnt Signaling. J. Biol. Chem. *275*, 4374–4382.

Van der Aa, N., Rooms, L., Vandeweyer, G., van den Ende, J., Reyniers, E., Fichera, M., Romano, C., Delle Chiaie, B., Mortier, G., Menten, B., et al. (2009). Fourteen new cases contribute to the characterization of the 7q11.23 microduplication syndrome. Eur. J. Med. Genet. *52*, 94–100.

Van Dijck, A., Vulto-van Silfhout, A.T., Cappuyns, E., van der Werf, I.M., Mancini, G.M., Tzschach, A., Bernier, R., Gozes, I., Eichler, E.E., Romano, C., et al. (2018). Clinical Presentation of a Complex Neurodevelopmental Disorder Caused by Mutations in ADNP. Biol. Psychiatry.

Van Laarhoven, P.M., Neitzel, L.R., Quintana, A.M., Geiger, E.A., Zackai, E.H., Clouthier, D.E., Artinger, K.B., Ming, J.E., and Shaikh, T.H. (2015). Kabuki

syndrome genes KMT2D and KDM6A: functional analyses demonstrate critical roles in craniofacial, heart and brain development. Hum. Mol. Genet. *24*, 4443–4453.

van de Leemput, J., Boles, N.C., Kiehl, T.R., Corneo, B., Lederman, P., Menon, V., Lee, C., Martinez, R.A., Levi, B.P., Thompson, C.L., et al. (2014). CORTECON: A Temporal Transcriptome Analysis of In Vitro Human Cerebral Cortex Development from Human Embryonic Stem Cells. Neuron *83*, 51–68.

Vandeweyer, G., Helsmoortel, C., Van Dijck, A., Vulto-van Silfhout, A.T., Coe, B.P., Bernier, R., Gerdts, J., Rooms, L., van den Ende, J., Bakshi, M., et al. (2014). The transcriptional regulator ADNP links the BAF (SWI/SNF) complexes with autism. Am. J. Med. Genet. C Semin. Med. Genet. *166C*, 315–326.

Vega-Lopez, G.A., Cerrizuela, S., Tribulo, C., and Aybar, M.J. (2018). Neurocristopathies: New insights 150 years after the neural crest discovery. Dev. Biol.

Vella, P., Barozzi, I., Cuomo, A., Bonaldi, T., and Pasini, D. (2012). Yin Yang 1 extends the Myc-related transcription factors network in embryonic stem cells. Nucleic Acids Res. *40*, 3403–3418.

Venkatesh, S., and Workman, J.L. (2013). Set2 mediated H3 lysine 36 methylation: Regulation of transcription elongation and implications in organismal development. Wiley Interdiscip. Rev. Dev. Biol. *2*, 685–700.

Visnapuu, V., Peltonen, S., Alivuotila, L., Happonen, R.-P., and Peltonen, J. (2018). Craniofacial and oral alterations in patients with Neurofibromatosis 1. Orphanet J. Rare Dis. *13*, 131.

Volpato, V., Smith, J., Sandor, C., Ried, J.S., Baud, A., Handel, A., Newey, S.E., Wessely, F., Attar, M., Whiteley, E., et al. (2018). Reproducibility of Molecular Phenotypes after Long-Term Differentiation to Human iPSC-Derived Neurons: A Multi-Site Omics Study. Stem Cell Rep. *11*, 897–911.

vonHoldt, B.M., Ji, S.S., Aardema, M.L., Stahler, D.R., Udell, M.A.R., and Sinsheimer, J.S. (2018). Activity of Genes with Functions in Human Williams–Beuren Syndrome Is Impacted by Mobile Element Insertions in the Gray Wolf Genome. Genome Biol. Evol. *10*, 1546–1553.

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., et al. (2018). Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. BioRxiv 447367.

Vulih-Shultzman, I., Pinhasov, A., Mandel, S., Grigoriadis, N., Touloumi, O., Pittel, Z., and Gozes, I. (2007). Activity-dependent neuroprotective protein snippet NAP reduces tau hyperphosphorylation and enhances learning in a novel transgenic mouse model. J. Pharmacol. Exp. Ther. *323*, 438–449.

Wai, D.C.C., Shihab, M., Low, J.K.K., and Mackay, J.P. (2016). The zinc fingers of YY1 bind single-stranded RNA with low sequence specificity. Nucleic Acids Res. *44*, 9153–9165.

Wallace, S.S. (2014). Base excision repair: a critical player in many games. DNA Repair *19*, 14–26.

Weaver, D.D., Graham, C.B., Thomas, I.T., and Smith, D.W. (1974). A new overgrowth syndrome with accelerated skeletal maturation, unusual facies, and camptodactyly. J. Pediatr. *84*, 547–552.

Weintraub, A.S., Li, C.H., Zamudio, A.V., Sigova, A.A., Hannett, N.M., Day, D.S., Abraham, B.J., Cohen, M.A., Nabet, B., Buckley, D.L., et al. (2017a). YY1 Is a Structural Regulator of Enhancer-Promoter Loops. Cell *171*, 1573-1588.e28.

Weintraub, A.S., Li, C.H., Zamudio, A.V., Sigova, A.A., Hannett, N.M., Day, D.S., Abraham, B.J., Cohen, M.A., Nabet, B., Buckley, D.L., et al. (2017b). YY1 Is a Structural Regulator of Enhancer-Promoter Loops. Cell *171*, 1573-1588.e28.

Weiss, K., Terhal, P.A., Cohen, L., Bruccoleri, M., Irving, M., Martinez, A.F., Rosenfeld, J.A., Machol, K., Yang, Y., Liu, P., et al. (2016). De Novo Mutations in CHD4, an ATP-Dependent Chromatin Remodeler Gene, Cause an Intellectual Disability Syndrome with Distinctive Dysmorphisms. Am. J. Hum. Genet. *99*, 934–941.

Werner, A., Iwasaki, S., McGourty, C., Medina-Ruiz, S., Teerikorpi, N., Fedrigo, I., Ingolia, N.T., and Rape, M. (2015). Cell fate determination by ubiquitin-dependent regulation of translation. Nature *525*, 523–527.

Wilkins, A.S., Wrangham, R.W., and Fitch, W.T. (2014). The “Domestication Syndrome” in Mammals: A Unified Explanation Based on Neural Crest Cell Behavior and Genetics. Genetics *197*, 795–808.

Wilson, S.W., and Rubenstein, J.L.R. (2000). Induction and Dorsoventral Patterning of the Telencephalon. Neuron *28*, 641–651.

Wiseman, F.K., Al-Janabi, T., Hardy, J., Karmiloff-Smith, A., Nizetic, D., Tybulewicz, V.L.J., Fisher, E.M.C., and Strydom, A. (2015). A genetic cause of Alzheimer disease: mechanistic insights from Down syndrome. Nat. Rev. Neurosci. *16*, 564–574.

Won, H., Lee, H.-R., Gee, H.Y., Mah, W., Kim, J.-I., Lee, J., Ha, S., Chung, C., Jung, E.S., Cho, Y.S., et al. (2012). Autistic-like social behaviour in *Shank2*-mutant mice improved by restoring NMDA receptor function. Nature *486*, 261–265.

Wu, S., Shi, Y., Mulligan, P., Gay, F., Landry, J., Liu, H., Lu, J., Qi, H.H., Wang, W., Nickoloff, J.A., et al. (2007). A YY1-INO80 complex regulates genomic stability through homologous recombination-based repair. Nat. Struct. Mol. Biol. *14*, 1165–1172.

Yabut, O.R., Fernandez, G., Huynh, T., Yoon, K., and Pleasure, S.J. (2015). Suppressor of Fused Is Critical for Maintenance of Neuronal Progenitor Identity during Corticogenesis. Cell Rep. *12*, 2021–2034.

Yang, S., Quaresma, A.J.C., Nickerson, J.A., Green, K.M., Shaffer, S.A., Imbalzano, A.N., Martin-Buley, L.A., Lian, J.B., Stein, J.L., Wijnen, A.J. van, et al. (2015). Subnuclear domain proteins in cancer cells support the functions of RUNX2 in the DNA damage response. J Cell Sci *128*, 728–740.

Yohe, S., and Thyagarajan, B. (2017). Review of Clinical Next-Generation Sequencing. Arch. Pathol. Lab. Med. *141*, 1544–1557.

Yoshioka, N., Gros, E., Li, H.-R., Kumar, S., Deacon, D.C., Maron, C., Muotri, A.R., Chi, N.C., Fu, X.-D., Yu, B.D., et al. (2013). Efficient generation of human iPSCs by a synthetic self-replicative RNA. Cell Stem Cell *13*, 246–254.

Young, M.A., Larson, D.E., Sun, C.-W., George, D.R., Ding, L., Miller, C.A., Lin, L., Pawlik, K.M., Chen, K., Fan, X., et al. (2012). Background mutations in parental cells account for most of the genetic heterogeneity of induced pluripotent stem cells. Cell Stem Cell *10*, 570–582.

Zhang, X.-B. (2013). Cellular Reprogramming of Human Peripheral Blood Cells. Genomics Proteomics Bioinformatics *11*, 264–274.

Zhang, S., and Cui, W. (2014). Sox2, a key factor in the regulation of pluripotency and neural differentiation. World J. Stem Cells *6*, 305–311.

Zhang, Y., Pak, C., Han, Y., Ahlenius, H., Zhang, Z., Chanda, S., Marro, S., Patzke, C., Acuna, C., Covy, J., et al. (2013). Rapid single-step induction of functional neurons from human pluripotent stem cells. Neuron *78*, 785–798.

Zhao, L., Li, J., Ma, Y., Wang, J., Pan, W., Gao, K., Zhang, Z., Lu, T., Ruan, Y., Yue, W., et al. (2015). Ezh2 is involved in radial neuronal migration through regulating Reelin expression in cerebral cortex. Sci. Rep. *5*, 15484.

Zhao, W., He, X., Hoadley, K.A., Parker, J.S., Hayes, D.N., and Perou, C.M. (2014). Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. BMC Genomics *15*, 419.

(2007). On the use of the word ‘epigenetic.’ Curr. Biol. *17*, R233–R236.