NEURODEVELOPMENTAL DISORDERS

FROM GENES TO REGULATORY ELEMENTS

ELENA PERENTHALER

NEURODEVELOPMENTAL DISORDERS: FROM GENES TO REGULATORY ELEMENTS | Elena Perenthaler

NEURODEVELOPMENTAL DISORDERS

FROM **GENES** TO **REGULATORY ELEMENTS**

ELENA PERENTHALER

The research described in this thesis was performed in the Barakat lab at the Department of Clinical Genetics, Erasmus MC University Medical Center, Rotterdam, the Netherlands.

Work in the Barakat lab has been supported amongst others by: Netherlands Organisation for Scientific Research (ZonMW Veni, grant 91617021), NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation, Erasmus MC Fellowship 2017, Erasmus MC Human Disease Model Award 2018, and grants from EpilepsieNL and CURE Epilepsy. Funding bodies did not have any influence on study design, results, data interpretation and final manuscript.

We are grateful to all patients and families that contributed to our research.

ISBN: 978-94-6419-563-7
Author: Elena Perenthaler
Cover design and layout: Elena Perenthaler and Mattia Fugatti
About the cover: the grey line represents DNA, the colourful spheres represent proteins and the dotted sphere represents degraded or absent proteins. In white is a section of the brain.
Chromatin is organized into topologically associating domains delimited by CTCF (yellow sphere). Within these loops, many proteins interact with DNA to regulate enhancer-promoter interactions and transcription.

Printed by: Gildeprint
Printing of this thesis has been kindly supported by the department of Clinical Genetics

Neurodevelopmental Disorders: from genes to regulatory elements

Neurologische ontwikkelingsstoornissen:

van genen tot regulerende elementen

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. A.L. Bredenoord

and in accordance with the decision of the Doctorate Board.
The public defence shall be held on

Tuesday 20th September 2022 at 10.30 hrs.

by

Elena Perenthaler
born in Trento, Italy.

Erasmus University Rotterdam

Doctoral Committee

Promotor:	Prof.dr. Y. Elgersma
Other members:	Prof.dr. D.F.E. Huylebroeck Dr. G.M.S. Mancini Prof.dr. N. Geijsen
Copromotor:	Dr. T.S. Barakat

To Paola

Contents

1

Introduction

Parts of this *Introduction* have been published in:

Perenthaler E*, Yousefi S*, Niggl E* and Barakat TS (2019). Beyond the Exome: The Non-coding Genome and Enhancers in Neurodevelopmental Disorders and Malformations of Cortical Development. Front Cell Neurosci 3, 352.

Verheul TCJ*, van Hijfte L*, Perenthaler E* and Barakat TS (2020). The Why of YY1: Mechanisms of Transcriptional Regulation by Yin Yang. Front Cell Dev Biol 8, 59264.

* These authors contributed equally

Brain development and disease

The brain lies at the foundation of what makes us human, as it not only regulates most of our body functions, but it is also central to our cognition and thoughts, defining our personalities, behaviour, and social interactions. How such a complex organ is formed during development has fascinated biologists for centuries. We are now living in a technology driven era where knowledge gained through various disciplines, such as medicine, developmental biology, biotechnology, computational biology, and neuroscience, enables us to get a glimpse on how these intricate processes are genetically regulated. Understanding the developmental biology of the human brain promises improvements and therapeutic options for various disorders, including neurodevelopmental disorders, which not only severely affect the quality of lives of patients and their families but also represent an economic burden on society.

Brain development

Animal models, among which frog, chick, fish and mouse, have been instrumental in understanding the basic principles of embryonic brain development[1], but in this section I will focus on human. The development of the human brain is a complex and tightly regulated process. It starts during early embryonic development, when neural progenitor cells are specified in the ectoderm by signals from the notochord, in a region known as neural plate. By the third gestation week (GW3), the neural plate folds to form the first neural structure, the neural tube, whose inner cavity will develop into the ventricles. During GW5, differences in the speed of proliferation of cells in the anterior part of the neural tube induce the formation of the primary brain vesicles (prosencephalon, mesencephalon and rhombencephalon), which will further divide into the secondary brain vesicles (telencephalon, diencephalon, mesencephalon, metencephalon and myelencephalon) establishing the basic organization of the brain (**Figure 1**). The posterior part of the neural tube, instead, gives rise to the spinal cord.

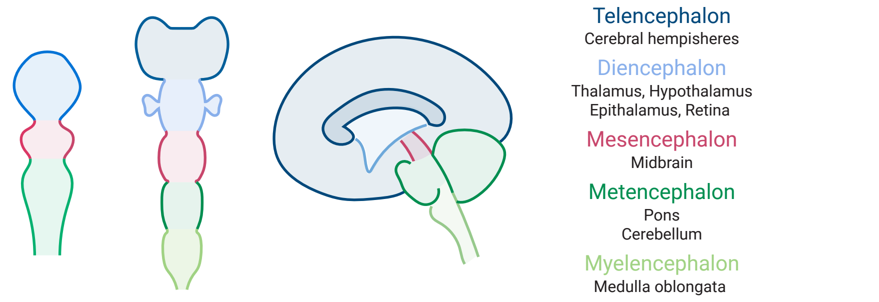


Figure 1 | Schematic diagram of brain vesicle development in the human brain.

The neural tube is composed of a single layer of **neuroepithelial cells** (NECs) that contact both the apical (towards the ventricle) and pial (or basal) surface and divide symmetrically to exponentially amplify their number. NECs represent the embryonic neural stem cell population and are organized in a pseudo-stratified monolayer, where, while proceeding through the cell cycle, the nucleus moves towards the pial surface during G1-phase and back to the ventricular surface during G2-phase in a process known as interkinetic nuclear movement[2]. Prior to neurogenesis, NECs lose the epithelial characteristics and differentiate to progenitor cells known as **apical radial glial cells** (aRGCs), that maintain the contact with both surfaces through the apical and the basal processes, at the ventricular and pial surface, respectively[3]. Until GW6, aRGCs self-renew by symmetric cell division, which is followed by asymmetric divisions generating one aRGC and either a post-mitotic excitatory neuron (direct neurogenesis), an intermediate progenitor (IP), or a **basal radial glial cell** (bRGCs) (indirect neurogenesis). bRGCs, highly present in brains of

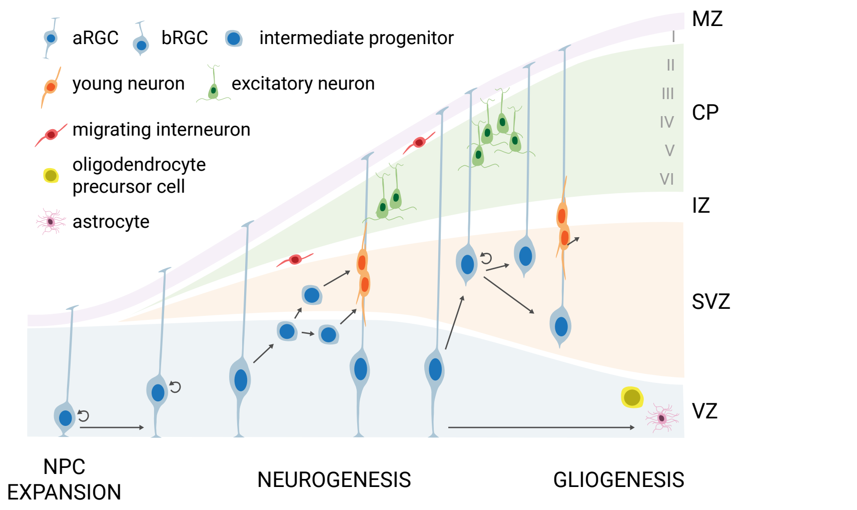


Figure 2 | Schematic diagram of cerebral cortex development. Neural progenitor cells (NPC) initially self renew (circular arrow) to amplify the progenitors pool. , Later, asymmetric divisions allow the generation of excitatory neurons. This process can be either direct or indirect, via the generation of intermediate progenitor cells or basal radial glial cells. Neurons migrate radially along the RGCs process and establish in the cortical plate in an inside out fashion, with early generated neurons occupying the deeper layers and late-born neurons in the superficial layers. Inhibitory interneurons migrate first tangentially along the intermediate and marginal zones and then radially to integrate into the cortical circuits. MZ: marginal zone; CP: cortical plate; IZ: intermediate zone; SVZ: subventricular zone; VZ: ventricular zone.

gyrencephalic species (such as ferrets and most primates), are located in the outer part of the subventricular zone and are in contact exclusively with the pial surface. These cells, in turn, can divide asymmetrically to generate either excitatory neurons or intermediate progenitors, and are thus thought to play a major role in the cortical surface expansion and folding in humans[4]. All neurons generated by asymmetric division of RGCs, IPs and bRGCs migrate radially along the basal process of RGCs towards the pial surface, to form, at first, the preplate. The preplate is a transient structure that includes, among others, reelin-secreting Cajal-Retzius cells (later becoming cortical layer I) that migrate tangentially from the medial ganglionic eminence, settle right below the pial surface and are responsible for the termination of cortical neuron migration[5]. The subsequent cortical **excitatory neurons** generated in the ventricular and subventricular zone migrate towards Cajal-Retzius cells. Here, they split the preplate in into two layers and settle in the middle forming the cortical plate in an inside-out fashion to form the other five layers of the cortex: early born neurons form layer VI while later born neurons form layer II[6]. The more superficial layer derived from the preplate is known as marginal zone and consists of Cajal-Retzius cells, while the deeper layer is known as subplate (**Figure 2**).

Unlike excitatory cortical neurons, **inhibitory interneurons** are generated in the ventral telencephalon, in different proliferative zones, known as ganglionic eminences (lateral, medial and caudal) and preoptic area. Once formed, the interneurons migrate tangentially to reach the cerebral cortex, similarly to Cajal-Retzius cells, and then radially to find their spot and integrate with the excitatory cortical neurons (**Figure 3**) (for a detailed *review* see[7]).

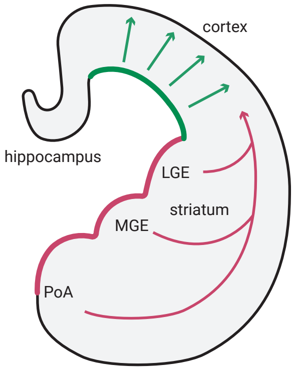


Figure 3 | Coronal section of the brain showing the tangential migration of interneurons generated in the ganglionic eminences (magenta) and the radial migration of excitatory neurons generated in the ventricular zone (green).

Besides neurons, the human brain is populated also by glial cells: oligodendrocytes, astrocytes, and microglia. The origin of forebrain **oligodendrocytes** is debated. The most accepted hypothesis is that they are formed in three distinct waves: (I) ventrally in the medial ganglionic eminence and anterior entopeduncular area, (II) in the lateral/caudal ganglionic eminence, and (III) postnatally, from the neural progenitors in the dorsal cortex, where later born cells replace the once generated in earlier waves[8]. Indeed, in the dorsal part of the telencephalon, neurogenesis is followed by gliogenesis, that continues into the postnatal period, where aRGCs finally differentiate into oligodendrocyte progenitors and **astrocytes**, that migrate into the cortex to establish connections with neurons[9] (**Figure 2**). Finally, **microglia**, the resident macrophages of the central nervous system (CNS), originate from the yolk sac primitive macrophages[10] and populate the human cortex around GW 10-12[11].

In human, the primary myelination of forebrain axons by oligodendrocytes starts during the third gestation trimester and continues for decades. Moreover, injuries or diseases such as multiple sclerosis can lead to loss of myelin, and re-myelination at these sites can occur throughout life, despite a decrease in efficiency with ageing. The involvement of myelination defects in neurodevelopmental disorders has long been understudied, however, myelination defects are increasingly reported in patients affected for instance by autism spectrum disorder[12,13] or epilepsy[14,15].

As a final step of human brain development, neurons, that are produced in excess to ensure appropriate connections, need to undergo a process of refinement. This happens via two routes: programmed cell death, where the entire neuron degenerates, and axon pruning, where only selected axon branches or synapses, and not the entire cell body, degenerate. Both processes are essential steps towards a properly functional brain. *In vivo* mouse models lacking apoptotic proteins display lethal neurodevelopmental phenotypes including an enlarged brain[16,17]. Likewise, aberrant axon pruning has been shown to be a hallmark of autism spectrum disorder, where patients have excessive synapses, likely due to reduced synaptic pruning[18], and schizophrenia, where, on the other hand, there is an increased elimination of synapses[19].

Taken together, the embryonic development of the cerebral cortex is a complex and dynamic process organized in three major steps: (I) stem cells proliferation, (II) differentiation and migration towards the cortical plate, and (III) post-migratory organization and circuitry formation. Alterations in any of these steps can be responsible for the development of a wide range of neurodevelopmental disorders, as I will discuss in the next section.

Neurodevelopmental disorders

Neurodevelopmental disorders (NDDs) are a group of clinically and genetically complex and heterogeneous disorders affecting more than 3% of children worldwide[20].

Disorders belonging to this class, as the name implies, share disturbances during CNS development as pathophysiological mechanism, and include various types of disorders and symptoms such as autism spectrum disorders (ASD), epilepsy, intellectual disability (ID) and malformations of cortical development. As the defects occur during embryogenesis, the symptoms often start during early childhood and the affected individuals fail to reach developmental or cognitive milestones[20]. Depending on the stage at which the defect in development arises, it is possible to identify several phenotypes. For example, impaired proliferation of the progenitor cells often results in microcephaly, defects in migration lead to diseases such as lissencephaly (characterized by a smooth brain surface with absent gyri), while defects in the connectivity and the post migratory organization of neurons are at the basis of ID and ASD. However, individuals with NDDs often present with more of these defects simultaneously. Such an example is represented by developmental and epileptic encephalopathies (DEEs), severe disorders where the affected children have early onset refractory seizures and psychomotor retardation often including ID, ASD and other behavioural issues. Seizures are driven by an abnormally excessive or synchronous neuronal firing, resulting from an imbalance of excitation/inhibition that can be caused by alterations in both neurons and glial cells. The seizures can be classified as either generalized, when they involve neuronal networks in both hemispheres, or focal, when they are restricted to anatomical brain regions in a single hemisphere.

There are many potential causes leading to NDDs that range from non-genetic and environmental insults to genetic alterations. Among the best recognized non-genetic origins are CNS injuries, including birth asphyxia or trauma, maternal infections during pregnancy ranging from viral infection by, for example, Zika virus or cytomegalovirus (CMV), to protozoan infection by *Toxoplasma gondii*[21], and environmental causes including maternal substance abuse throughout pregnancy (e.g., alcohol, tobacco, cocaine)[22]. Likewise, genetic causes have their roots in a wide spectrum of alterations. One of the first genetic causes recognized to be causative of NDDs are copy number variants (CNVs) that, in turn, range from aneuploidies, like trisomy 21 leading to Down's syndrome, or X monosomy in girls with Turner's syndrome, to a wide array of micro deletions and micro duplications affecting smaller parts of chromosomes. Also, polygenic and oligogenic causes are now recognized, where several risk factors together contribute to the onset of diseases such as ASD and epilepsy. Finally, NDDs can also be monogenic diseases. Some of the first genes identified to be causative of NDDs are located on the X chromosome and include *FMR1*, where a CGG expansion leads to Fragile X syndrome[23], and *MeCP2*, mutations of which are causative of Rett syndrome[24]. Besides X-linked disorders, autosomal variants are also found and dominant *de novo* mutations are currently considered the main genetic cause of NDDs. More recently other inheritance patterns, like autosomal recessive, are being recognized predominantly in inbred populations[25]. Disease causative variants can be missense variants, that induce an amino acid change in the encoded protein, non-sense variants, including start-loss or stop-gain variants, but also deletions or insertions that lead to a frameshift, altering the amino acid composition of the encoded protein or creating premature stop codons, which might result in truncated proteins or transcripts that are prone to nonsense-mediated decay.

Most of the identified variants in NDDs are located in genes that belong to specific biological pathways such as protein synthesis (for example the mTOR pathway), transcriptional/epigenetic regulation (chromatin remodelling factors and transcription factors), synaptic signalling, mitosis/microtubules and signalling pathways[20]. For example, DEEs are most frequently caused by mutations in ion channels such as *SCN1A*, a voltage-gated sodium channel, pathogenic variants of which are causative of Dravet syndrome[26], but also other sodium (e.g. *SCN8A*[27]), potassium (e.g. *KCNA2*[28], *KCNB1*[29]) or calcium channels (e.g. HCN1[30], CACNA1A/E[31,32]). A plethora of other pathways have also been identified, including neurotransmission, with variants affecting GABA or glutamate, both NMDA and AMPA, receptors (*GABA*$_A$[33], *GABA*$_B$[34], *GRIN2B*[35], *FRRS1L*[36]), membrane transporters (*SLC2A1*[37]), but also proteins involved in more general processes such as transcription, with variants in transcription factors (*CUX2*[38]) and chromatin remodelling factors (*ACTL6B*[39]), translation (*EEF1A2*[40]) and in post translational modifications (*UBA5*[41,42]).

Up to now more than 700 genes have been identified to have a role in intellectual disabilities[43] (more than 1200 genes are present in the gene panel "Intellectual disability, version 12" of the Clinical Genetics department at the Erasmus MC), more than 100 associated with DEE[44] and many more are expected to be identified in the near future. However, up to date many patients still do not have a molecular diagnosis and the identification of the genetic causes of NDDs is crucial not only for understanding the molecular mechanisms at the basis of the disease and to ultimately provide a specific treatment, but also for providing patients and their families with proper genetic counselling.

Molecular diagnosis of NDDs

Patients' diagnosis necessarily starts with a thorough clinical assessment by the physician, to determine which syndrome the patient might have, and the type of genetic testing required. Patients' DNA is usually isolated from blood and then subjected to a series of genetic testing to identify the possible causative variant(s). To exclude the presence of chromosomal aberrations or CNVs, the first step is to perform karyotyping or chromosomal microarrays, that, for patients with developmental delay or congenital abnormalities (excluding clear chromosomal syndromes such as Down syndrome) have a diagnostic yield of 3% and 15-20% respectively[45]. Among chromosomal microarrays are also SNP-microarrays that are pivotal to identify stretches of homozygosity in probands that indicate the consanguinity of the parents. This helps narrowing down the search for potential disease-causing variants. In case the clinical assessment of the patient suggests a monogenic disorder, the suspected gene can be sequenced by Sanger sequencing to identify the variant, while, if a clear candidate gene is not emerging from the differential diagnoses, or there is evidence of a run of homozygosity, a panel of genes often associated with the group of disorders or a specific genomic region can be targeted-sequenced. If all these genetic testing results are negative, DNA of the affected individual and the parents can be subjected to trio-whole exome sequencing (WES), a method to sequence all protein-coding exons in the genome. Analysis of these large-scale sequencing data then either focusses on selected panels of genes (e.g., using a bioinformatics filter to only investigate genes related to intellectual disability), all OMIM genes or all human genes.

The implementation of WES in the diagnostic process improved the diagnostic yield of Mendelian disorders to ~25-30%[46]. However, still many cases remain unexplained[47], even when focusing on NDD cases with a strong hint at genetic causes, for instance where multiple affected individuals are found in the same family, or other environmental causes have been excluded. Even though the diagnostic yield for some NDDs displaying defined features on brain imaging, such as lissencephaly, can reach up to 80%[48], for the majority of cases the yield is much lower[49]. This **missing heritability** is often reasoned to be caused by somatic mutations[50,51] or mosaicism[52-54]. However, as WES only interrogates the 2% of the human genome that encodes for proteins[55], it is tempting to speculate that at least some of this missing heritability might be caused or influenced by genetic variation in the non-coding genome. Whole genome sequencing (WGS) can improve the outcome of genetic testing, but WGS routine implementation requires a thorough understanding of the non-coding genome and the effect of its variation. As a result, most studies using WGS in a clinical setting, have limited their analysis to those nucleotides covering exons, deep intronic variants not covered in WES and copy number or structural variants[56-59]. Therefore, it remains crucial to gain more detailed information on the functional relevance of the non-coding genome and their variants from a basic science point of view.

The hypothesis that disease-causing variants might be located in non-coding regions of the genome is supported by several arguments. First, genome-wide association (GWAS) studies on multiple diseases have shown that more than 90% of disease-associated single nucleotide polymorphisms (SNPs) are located outside of coding genes[60], therefore potentially in regions involved in transcriptional regulation. Second, the last decade has witnessed an enormous progress in our understanding of the mechanisms involved in gene regulation, and it has become clear that aberrant gene regulation can cause a variety of genetic disorders[61-63]. Key elements in the non-coding genome such as promoters, insulators and enhancers ensure that genes are turned on or off at the right moment and place and ensure properly dosed levels of steady-state mRNA. When this tight spatio-temporal and/or dose regulation is disturbed, it can affect gene expression and could result in a genetic disorder. Although only very few large-scale genetic studies have investigated the role of the non-coding genome in genetic disorders[64-66] it is clear from a number of excellent studies that have recently been published[67-77], that the non-coding genome plays an important role in health and disease. Finally, one and the same mutation can show different degrees of severity in different patients, and it is likely that this phenotypic variability could be influenced by genetic variations outside of coding genes influencing gene expression[78,79].

Altogether, this strongly supports that alterations in the non-coding genome might play a role in disease and explain some missing heritability, but to properly investigate this hypothesis, we need a clearer understanding on regulatory elements and their location in the human genome, which is the focus of this *Thesis*.

Non-coding genome

According to the central dogma of molecular biology, there are three main processes taking place in a cell: replication of the genetic information, transcription of DNA into RNA, and translation of the RNA molecule into the final functional product, the protein[80]. As one of the main Human Genome Project surprises, it is now well established that more than 98% of the human genome does not encode proteins[55]. These non-protein-coding regions were initially considered junk DNA, which was assumed to be redundant and under no selective pressure, thus allowing for the accumulation of mutations without any harm to the organism[81,82]. However, by now, several structural elements of non-coding DNA have been described that regulate gene expression, by, for example, determining the 3D genomic organization. Regulation of gene transcription is particularly crucial during embryonic development, when a single cell needs to differentiate into distinct cell types and to establish diverse gene expression programs, while maintaining the same genotype. This is achieved by a tight spatio-temporal regulation of gene expression, that allows the transcription of the right gene, at the right level, in the right cell type, and is executed by the interplay between enhancers and gene promoters confined to the "playfield" established by the 3D organization of the genome. It is important to keep in mind in the following paragraphs, that gene regulation needs to be seen from a non-linear, 3D perspective where regulatory elements need to interact with target genes over long distances.

Chromatin organization

To grant efficient DNA packaging in the limited space of the nucleus while allowing for DNA replication and gene expression, DNA is wrapped around **nucleosomes**, histone octamers constituted of two copies of each histone protein H2A, H2B, H3 and H4. The genome is further organized in a hierarchical fashion. First, each chromosome is located in a different region of the nucleus in what is known as **chromosome territory**. Each territory is then organized into active and inactive compartments that are composed of either open (eu-) or condensed (hetero-) chromatin and which vary in size between 1 to 10 megabases (Mb). The **compartments** are often organized in a radial fashion, with inactive (or B) compartments close to the nuclear lamina, and active (or A) compartments more towards the centre of the nucleus[83,84]. At a sub-compartment level, chromatin is organized in **topologically associating domains** (TADs)[85] that are usually <1 Mb in size and delineate those regions of our chromosomes in which sequences interact preferentially with each-other. The prevailing model is that these TADs are formed by the dimerization of two CTCF molecules binding insulators at TAD boundaries, stabilized by the interaction with the ring-shaped cohesin complex through a process called loop extrusion[86-88]. Inside TADs, smaller DNA loops are formed to allow enhancer–promoter interactions and hence regulation of transcription[86,89]. These **enhancer-promoter loops**, similarly to the CTCF-mediated loops, are thought to be established by the binding and dimerization of the TF YY1 and its interaction with the cohesin complex (**Figure 4**)[90,91].

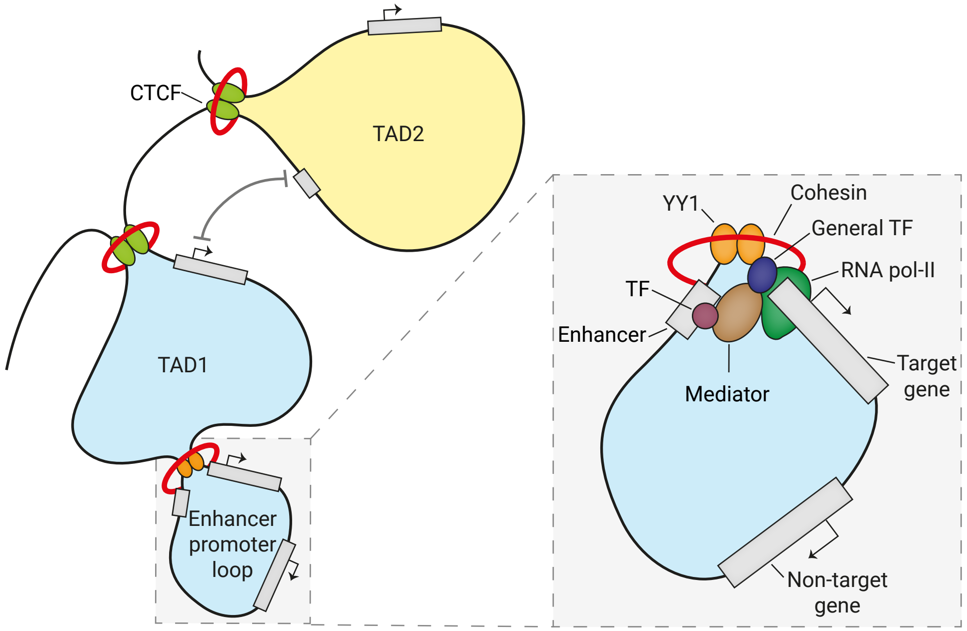


Figure 4 | Regulatory enhancer-promoter interactions are restricted within the TAD region. The genome (here represented as a black line) is tightly packaged and organized in topologically associating domains (TADs) established by the binding of CTCF to insulator elements, followed by dimerization and interaction with the cohesin complex. In order to establish the enhancer-promoter loops required for transcriptional regulation, enhancers and their target gene should reside in the same TAD. These regulatory loops are formed by the dimerization of YY1 and its interaction with cohesin. In the enlargement is a simplified scheme of transcription initiation (the size does not reflect the actual dimension of each component). Transcription factors (TFs) bind on the enhancer element while the pre-initiation complex formed by the RNA Pol II and the general TFs assembles at the promoter region. Mediator establishes the connection between enhancer and promoter via interactions with TF and pre-initiation complex components, without binding to DNA. Mediators regulates the phosphorylation of the RNA Pol II in order to release it from the promoter and start transcription.

Transcription initiation and its regulation

In eukaryotes, transcription is mediated via three large and multi-subunit DNA-dependent **RNA polymerases** that are responsible for the synthesis of different classes of RNA: (I) RNA polymerase I synthesizes the large 47S pre-ribosomal RNA (rRNA); (II) RNA pol II transcribes all the messenger RNAs (mRNA) and some non-coding RNAs; (III) and RNA pol III produces the 5S rRNA, all the transfer RNAs (tRNAs) and other short non-translated RNAs. In the next paragraphs, I will focus on RNA Pol II mediated transcription.

Two *cis*-regulatory elements play a major role in transcription, promoters, and enhancers. Promoters are located upstream of the gene transcriptional start site (TSS), while enhancers are distal *cis*-regulatory sequences that orchestrate the rate of transcription initiation. **Enhancers** are positive regulators of transcription[92], whose location relative to the TSS of the gene they control varies from adjacent to the promoter, to many kilobases (kb) upstream or downstream of it (also in introns, even of other genes). Besides acting in a position-independent manner, enhancers can regulate transcription irrespective of their orientation. A classic example of a long-range regulatory element is the limb *SHH* enhancer, which is located ~1 Mb away from its target gene[69]. Making the scenario even more complex, one enhancer can regulate several genes, and at the same time each gene can be regulated by multiple enhancers. This creates a redundancy in the system that results in phenotypic robustness, and probably gives advantages during evolution[93]. The position, identity, and arrangement of enhancers ultimately determines the time and place each gene is transcribed. On a mechanistic level, enhancers directly influence the recruitment of the transcriptional machinery to gene promoters[94,95]. Crucial for this long-range control of gene expression is the formation of enhancer-promoter loops which, as previously mentioned, preferentially occur within the neighbourhood of a TAD. Initiation of transcription requires the assembly of a **pre-initiation complex** (PIC) on promoters. This process is directed by several proteins such as the TATA box binding protein TBP, that binds to a specific sequence in the promoter, and a variety of general class II initiation factors, among which TFIIB that bridges the RNA Poll II and the promoter. The PIC is stabilized by Mediator, a large multi-subunit complex, that in turn bridges enhancers and promoters. Once the PIC is assembled, the DNA is unwound in an ATP hydrolysis-mediated fashion. Finally, Mediator stimulates a subunit of TFIIH, CDK7, to phosphorylate the C-terminal domain of RNA Pol II starting transcription elongation.

Both active promoters and enhancers are located in nucleosome-depleted regions of chromatin to allow RNA Pol II and the whole initiation complex to access DNA. However, not all genes, but only a fraction of them, are transcribed at a given time and place in the body, and this requires the combined action of chromatin remodelling factors, histone modifying enzymes and cell type-specific transcription factors (TFs).

ATP-dependent chromatin remodelling complexes

ATP-dependent chromatin remodelling complexes can modulate chromatin architecture and DNA accessibility by repositioning nucleosomes along DNA or by altering their subunit composition. Chromatin remodellers are classified into different families, including SWI/SNF (switch/sucrose-non-fermenting), INO80 (inositol requiring 80), and CHD (chromodomain-helicase-DNA binding). The SWI/SNF complex is characterized by a bromodomain that recognizes acetylated histones. Based on the core protein composition, in humans, it is possible to recognize three SWIN/SNF complexes known as BAF, PBAF (Polybromo-associated BAF complex) and ncBAF (non-canonical BAF complex), that further acquire tissue-specific roles during development due to a combinatorial assembly of different subunits. The INO80 complex, that includes YY1[96], exchanges the histone variant H2A.Z to H2A and slides nucleosomes at promoters[97,98], being involved in transcription but also in DNA repair and replication[99-102]. Finally, CHD is a large family of proteins that share a chromodomain and are recruited to chromatin via their interaction with different factors like transcription factors, histone modifications and methylated DNA. CHD family members can act both as monomers and as part of larger complexes such as NuRD that has been associated mainly with transcriptional repression by stimulating de-acetylation of chromatin via its interaction with histone deacetylases (HDACs).

Histone modifying enzymes

Histone modifying enzymes can alter chromatin accessibility by the deposition of various post-translational epigenetic modifications on histones tails. For example, histone acetylation by histone acetyltransferases (HATs) such as p300/CBP results in increased chromatin accessibility, which facilitates the binding of regulatory proteins like TFs. Many studies focused on a wide variety of histone modifications (for *review* [103,104]), and have led to a draft of a histone code, where various histone modifications are indicative of the functional role of the modified chromatin. For example, putative enhancers are enriched in chromatin regions surrounded by histone 3 lysine 4 monomethylation (H3K4me1) and lysine 27 acetylation (H3K27ac), while promoters are marked by histone 3 lysine 4 trimethylation (H3K4me3). On the opposite, tri-methylation of lysine 27, mediated by the Polycomb repressive complex 2 (PRC2), results in silencing.

Transcription factors

Transcription factors, as the name suggests, are another class of proteins controlling transcription. TFs generally bind to open chromatin regions, but some, known as pioneer TFs, can bind nucleosomal DNA and recruit chromatin remodelling complexes to displace nucleosomes and render chromatin accessible, both at promoters and at enhancers. TF binding sites (TFBS) consist of DNA motifs found at multiple sites in the genome, that are not necessarily all equally likely to be bound by the recognizing TF[105]. To provide higher than background activity, homotypic or heterotypic dimerization of TFs increases their DNA binding affinity and specificity[106]. Multiple TFs have been found to bind in a cooperative manner to TF binding site "hotspots"[107], later called stretch enhancers[108] or super-enhancers (SEs)[109]. The latter are described as long regions with a strong enrichment of H3K27ac, TFs and Mediator[109,110]. While on the one hand, a number of studies suggests that SEs represent a novel class of enhancers that maintain, define, and control mammalian cell identity and whose transcriptional regulatory output is larger than that of the individual enhancer constituents[109,111-113], on the other hand several other studies have challenged this view and consider super enhancers as a collection of normal enhancers that together do not have a larger activity than the sum of the individual parts[114,115]. Therefore, the debate on whether SE are a new class of regulatory elements or whether they simply reflect a clustering of normal enhancers remains to be settled.

The role of Ying Yang 1 in transcriptional regulation

One of the key TFs that has been already mentioned several times in this *Introduction* and is a central topic of investigation in this *Thesis*, is Yin Yang 1 (YY1). YY1 was first described in 1991 by three independent groups, who all named the protein differently based on the molecular mechanisms they associated it with. Park and Atchinson called it NF-E1, as it binds the μE1 intron enhancer at the immunoglobulin heavy chain (IgH) locus[116], Hariharan and colleagues δ, as it binds the delta motif in the promoter of ribosomal protein genes[117] and Shi et al. named it Yin Yang 1. The name YY1 was eventually broadly adopted, as it captures its dual activity as both a transcriptional activator and repressor[118].

YY1 is ubiquitously expressed in mammalian cells. It forms homodimers, that seem to be stabilized by low specificity RNA binding[119], and bind a relatively small DNA sequence motif [5'-CCGCCATNTT-3'], often found in enhancers and promoters, via the four C2H2 zinc fingers in its C-terminal domain[120,121] (**Figure 5**). The DNA binding zinc fingers partially overlap with protein sequences involved in transcriptional repression, while the transcriptional activator domain is located in the N-terminal region[118]. YY1 influences transcription by the recruitment of cofactors[122] that interact mainly with the REPO domain. For example, it interacts with polycomb group proteins, like the previously mentioned PRC2, to recruit repressive cofactors to specific genes[123,124] and with cohesin and condensin, key factors of 3D chromatin organization[125]. Another YY1 domain is a His tract consisting of eleven consecutive histidine residues, that is proposed to stimulate YY1 accumulation in nuclear speckles[126], that play an important role in RNA metabolism[127]. Interestingly, YY1 itself plays a role in pre-mRNA splicing[128] by binding intronic enhancer motives and promoting splicing while activating gene expression[129]. This RNA-binding capability of YY1 would be an interesting topic of further investigation in relation to transcriptional regulation, as it was shown that RNA binding stabilizes YY1 homodimers and that, compared to DNA binding, the RNA binding occurs with low sequence specificity[119,130].

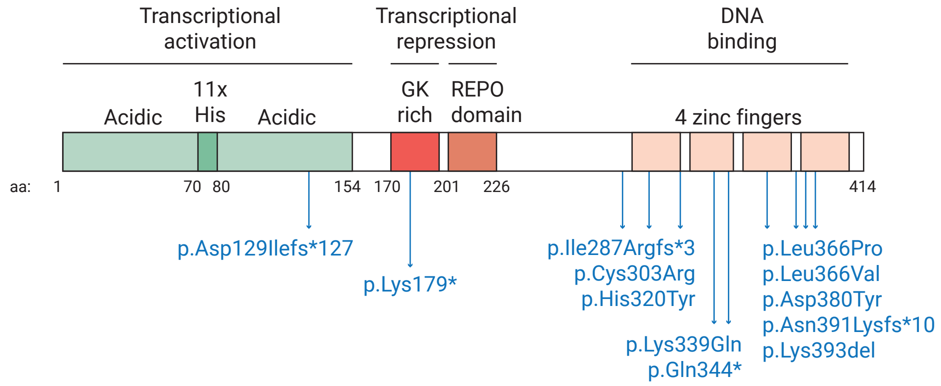


Figure 5 | Schematic diagram of human YY1 domains. Human YY1 is composed of 414 amino acids. It binds a small DNA sequence [5'-CCGCCATNTT-3'] through the four C2H2-type zinc fingers located at the C-terminal of the protein (amino acid 296-320, 325-347, 353-377, 383-407). The REPO domain (aa 201-226) and a glycine-lysine rich domain (GK-rich, aa 170-200) mediate transcriptional repression. The REPO domain is responsible for the interaction with polycomb group proteins while the GK rich domain with histone deacetylases (HDAC). The N-terminal region of the protein mediates transcriptional activation. It is composed mainly by acidic amino acids (aa 1-154) and by a stretch of 11 histidines (aa 70-80), that are thought to stimulate YY1 accumulation in nuclear speckles. In blue, the reported causative variants of Gabriele-de Vries syndrome are indicated[125,126].

YY1 is a crucial factor regulating cell proliferation and apoptosis[133,134] and it has been implicated in many regulatory mechanisms in different tissues. In brain, YY1 plays a well-established role in neuronal development[132,135,136]. In mice, a homozygous *Yy1* knock-out results in peri-implantation lethality while heterozygous mutations cause growth retardation and neurulation defects[137]. In humans, YY1 haploinsufficiency causes Gabriele-de Vries syndrome (OMIM #617557)[132,135], characterized by psychomotor delay and intellectual disability alongside many comorbidities, including craniofacial dysmorphisms, intra-uterine growth restriction and behavioural alterations.

The function of YY1 as a traditional DNA-binding transcription factor

A key question in the YY1 field is how a DNA-binding TF can act as both a transcriptional activator and repressor. Multiple mechanisms have been suggested, including post-translational modifications of YY1[138] and co-factor dependency[122]. Acetylation of the C-terminal domain of YY1, for example, reduces DNA-binding ability *in vitro,* whereas acetylation of the central domain is required for YY1 to act as a full repressor[138]. In mouse anterior neuroectoderm development, YY1 acetylation is required for *Otx2* activation[139]; indeed, only acetylated YY1 can bind an essential enhancer 5kb upstream of the homeobox gene *Otx2* to prompt its expression[139]. Post-translational modifications of YY1, thus, seem to influence whether YY1 can activate or repress transcription.

Another explanation for YY1s context-dependent transcriptional activation or repression is the interplay between YY1 and its cofactors[138] (**Figure 6A**), such as SpI[140] and c-Myc[141]. Another example is YY1 associated factor 1 (YY1AP), that is tethered directly to the promoters by YY1 and boosts transcription[142]. In light of its structural homology to YY1AP, also YARP, which is expressed in brain, heart and placenta, is suggested to act as another co-activator binding YY1[143]. *In vitro*, YY1 is able to bind promoter sequences and recruit polymerase by interacting with general TFs. Additionally, YY1 also recruits co-repressors to DNA, such as SMAD family members resulting in downregulation of activated SMAD-mediated TGF-β family signalling and therefore impacting on cell differentiation[144]. Moreover, YY1-mediated repression might be a result of the overlapping DNA binding sites of YY1 and some transcriptional activators. Since DNA-binding of YY1 and an activator at a given locus can be mutually exclusive, YY1 binding can block the activating factor. For example, in mammary epithelial cells, YY1 represses β-casein competing with the latent-state mammary gland specific factor (MGF) of the STAT family, STAT5A[145]. Upon lactation,

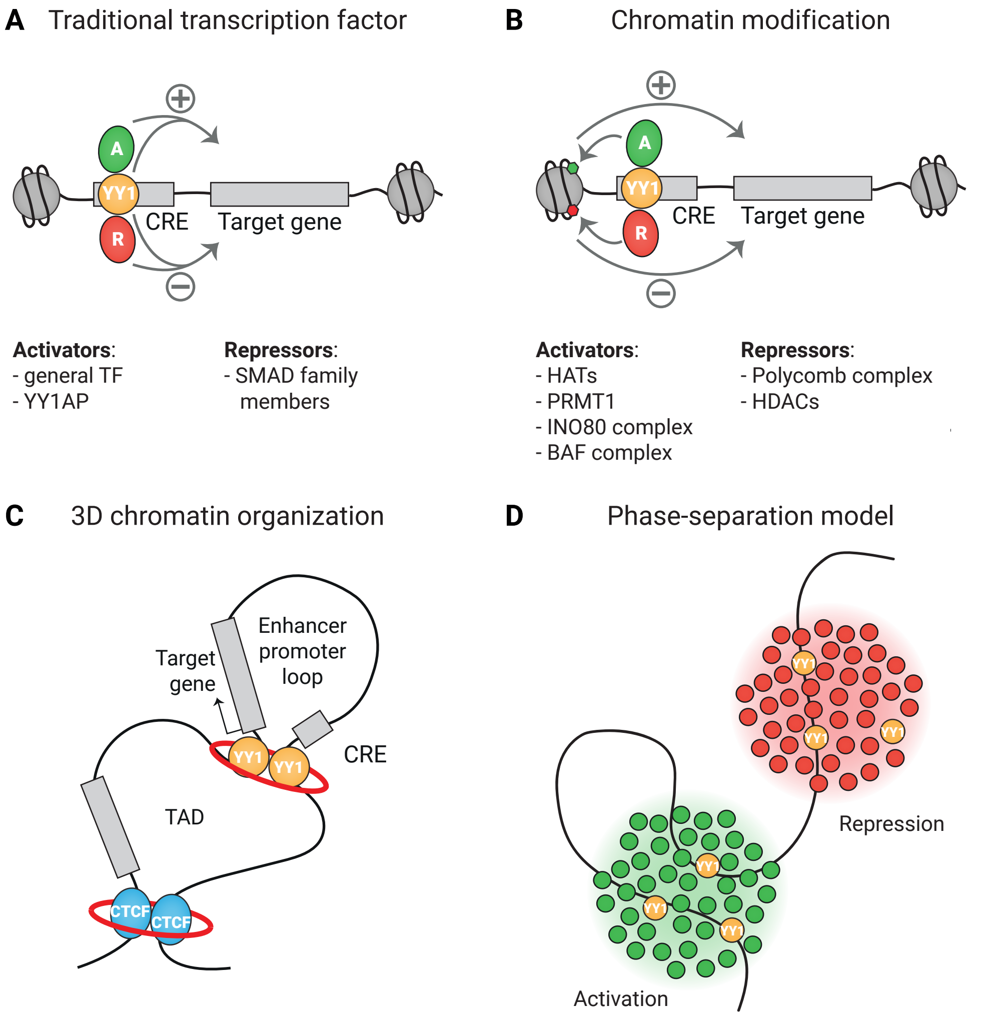


Figure 6 | Mechanism of transcriptional activation or repression by YY1. **A)** YY1 can act as a traditional DNA-binding transcription factor interacting with an extensive list of co-factors that mediate the activation or repression of transcription. Some are listed in the figure. **B)** YY1 can interact with chromatin remodelling complexes that regulate transcription by regulating chromatin accessibility. **C)** YY1 regulates transcription via the formation of enhancer-promoter loops within larger TADs. The red ring represents Cohesin. **D)** The phase separation model might also explain the activity of YY1 as a dynamic activator or repressor of transcription, based on the cocktail of co-factors present in the highly concentrated phase-separated droplets. CRE: cis-regulatory element.

MGF is activated, increasing its DNA-binding affinity and enabling it to replace YY1 and de-repress β-casein[146].

Finally, in mouse neurons, YY1 and its interacting partner BRD4 activate *Senp1*, an upstream regulator of glutamate signalling, which plays a pivotal role in neuronal plasticity[147]. De-phosphorylation of YY1 upon membrane depolarization depletes the *Senp1* promoter of YY1-BRD4, consequently repressing *Senp1* expression[147]. This example illustrates how post-translational modifications and co-factor binding act in concert and how a ubiquitously expressed TF can be an activator or repressor depending on the cellular context.

The interplay between YY1 and chromatin modifications

Among YY1 interactors, are multiple chromatin modifiers, suggesting that chromatin modifications might explain YY1 functioning as a transcriptional repressor or activator (**Figure 6B**). YY1 can direct the Polycomb complex to specific DNA loci, initiating deposition of H3K27me3[123,148]. Furthermore, YY1 interacts with histone deacetylases (HDACs) associated with gene silencing. Several members of this family, such as HDAC1, HDAC2 and HDAC3, interact with YY1 both *in vitro* and *in vivo*[149]. Interestingly, YY1 also interacts with histone acetyltransferases (HATs) like p300[150] and CREB binding protein (CBP), activating transcription[150,151]. Besides acetylating or de-acetylating histones, HATs and HDACs modify YY1 itself, regulating its DNA-binding affinity and activity as a transcriptional regulator[138]. In addition to histone 3 modifications, YY1 promotes transcription also via the methylation of histone 4 arginine 3 by the recruitment of the methyltransferase PRMT1[152].

Moreover, YY1 has been shown to activate transcription by interacting with chromatin remodelling complexes involved in the shifting and repositioning of nucleosomes, such as the INO80 complex[96,153] and, more recently, the BAF complex[154]. The interaction with the INO80 complex is also thought to play a role in facilitating the access of YY1 to its target genes[153]. Hence, through a plethora of molecular co-factor interactions, YY1 influences chromatin modifications and ultimately gene expression.

YY1 regulates transcription through the 3D chromatin organization

At first glance, YY1 does not seem essential for 3D chromatin organization, as the majority of its binding sites are close to transcription start sites (TSSs) and only a minority is located distal from regulated genes[135,136,155]. However, in YY1-haploinsufficient lymphoblastoid cell lines, the most differentially expressed genes are controlled by those distal YY1 binding sites[135] and in T-helper cells, YY1 seems to mainly influence gene expression through long-distance DNA interactions[156]. Further implicating a role in 3D chromatin organization, is the ability of YY1 to interact with proteins involved in chromatin organization, such as CTCF and cohesin[125,157].

As mentioned previously, loop extrusion allows cohesin to actively form DNA loops[158]. CTCF delimits the TAD loops[159-161], however it not crucial for the majority of enhancer-promoter interactions[162]. Recently, YY1 was identified as the structural factor that regulates the formation of enhancer-promoter loops within the larger CTCF-CTCF domains in a wide variety of mammalian cell types, indicating that this might be a general mechanism in mammalian cells[90] (**Figure 6C**). Like CTCF, YY1 forms homodimers and binds hypo-methylated DNA to facilitate long-distance DNA interactions. In contrast to CTCF, however, YY1 binds to a consensus sequence mainly present in promoters and enhancers and is only scarcely associates with insulators[90,91,163].

In neuronal differentiation specifically, TAD organization was found to be less conserved between cell types and differentiation stages than initially thought[164,165]. These findings triggered research into the possible role of dynamic chromatin organization during differentiation of neural progenitor cells (NPCs)[91]. Surprisingly, YY1 appeared to instigate DNA loop formation and NPC-specific promoter-enhancer interactions within TAD loops[91]. These findings introduce a new identity of YY1 as a structural protein in addition to its role as a traditional TF, providing an even broader understanding of the multitude of cellular mechanisms via which YY1 can regulate transcription.

As it is clear from the previous paragraphs, the whole process of transcription requires a massive number of factors, among which TFs like YY1, chromatin remodelling complexes, histone modifiers, and RNA Pol II, just to mention a few. One of the leading hypotheses arising in the latest years on how this can be carried out, is the formation of **condensates**, that, to a moderate extent, resemble the function of membrane-less organelles. These liquid-like droplets are suggested to be formed by the interactions of protein with intrinsically disordered regions, such as the transactivation domain of TFs, leading to the formation of hubs with a high concentration of proteins necessary for the activation or the repression of transcription[166]. For example, the TFs OCT4 and GCN4 can form such phase-separated droplets with a high concentration of many cofactors upon enhancer binding[167], while MeCP2 forms heterochromatin condensates that accumulate repressive factors[168]. The high concentration of proteins in phase-separated droplets is thought to allow rapid interactions of many factors that during *in vitro* assays would seem too weak[166]. In light of the extensive list of YY1 protein interactors, it is possible that also YY1 might act via a similar mechanism, that importantly could explain its ambivalence of being both a repressing and activating TF (**Figure 6D**).

The role of YY1 in neurodevelopment

YY1 has been shown to be essential for proper brain development, maintenance, and protection from degeneration. This holds true also in mouse[136,155,169], where *Yy1* conditional knock-out (cKO) induced at an early stage of cortical development, was shown to increase the apoptotic rate and to induce cell cycle arrest in neuroepithelium and NPCs[136,169]. This effect on NPCs however decreased markedly when cKO was induced at later stages[136]. Accordingly, another study by Varum and colleagues showed that *Yy1* cKO in mice affects neural crest (NC) development in a strict stage dependent manner. Early KO caused a reduction of multiple NC-derived lineages, whereas late KO (after embryonic day 11.5) resulted in no clear phenotypic difference compared to control[155], showing a decreased dependency on YY1 regulated processes at later stages of neuronal development. Surprisingly, early and late *Yy1* cKO, caused similar changes in gene expression, which indicates that the decreased importance of YY1 during neuronal development does not coincide with a shift in YY1 target genes[136]. As YY1-regulated genes seemed to be mainly implicated in metabolic pathways and protein translation, Zurkirchen and colleagues hypothesized that a decreasing dependency on YY1 during cortical development is due to a decreased biosynthetic demand and decreased proliferation rate of cells at later stages of corticogenesis, making these cells less vulnerable to defects in these pathways[136,155].

The importance of YY1 in early development is also attributed to apoptosis inhibition[136,169-171]. In mice, cKO of *Yy1* at an early embryonic stage caused an accumulation of p53 protein and increased apoptosis[136,169]. This effect could be partially reversed in *Yy1;Trp53* double cKO mice, indicating that YY1 downregulates p53, an important apoptosis regulator, to facilitate NPC survival[136]. Additionally, YY1 inhibits apoptosis by regulation of the planar cell polarity effector gene *FUZ*, an important apoptosis factor in neuronal development. Alterations of *FUZ/Fuz* expression cause neural tube defects in humans and are associated with an increased number and disorganization of NC cells in mice [172-174]. YY1 can induce hyper-methylation of the *FUZ* promoter, resulting in *FUZ* downregulation and inhibition of its apoptotic signal[170]. A reduction in soluble YY1 protein reverses this hyper-methylation at the *FUZ* promoter and is associated with increased apoptosis[170].

Recently, YY1 was also linked to NPC differentiation by downregulation of *Sox2* expression in mice[175]. SOX2 is a known pluripotency factor and is also involved in the maintenance of the undifferentiated state of NPCs[176]. YY1 was implicated in *Sox2* downregulation in mouse brain cortices during neuronal development by binding the *Sox2* locus and physically halting transcription. These results accentuate a pro-differentiation role for YY1[175] that contradicts previously described work, which shows that YY1 is vital for NPCs maintenance and proliferation in early development[136,155].

The role of YY1 in disease

In humans, YY1 haploinsufficiency causes Gabriele-de Vries Syndrome, which is characterized by cognitive impairment, behavioural alterations, intrauterine growth restriction, feeding problems and sometimes congenital malformations[132,135]. One of the consequences of YY1 haploinsufficiency in humans is a loss of H3K27 acetylation at enhancers bound by YY1[135]. H3K27 acetylation is tightly associated with active promoters and distal enhancers which indicates that downregulation of YY1 affects chromatin regulation and gene transcription[135,177]. Additionally, mutations in YY1 binding sites in specific non-coding regulatory regions cause neurodevelopmental disorders with a milder phenotype, since the effect of such mutations is limited to the expression of the gene associated with this regulatory region[178]. For example, a disrupted YY1-binding site in a brain-specific enhancer of *ADGRL3* leads to a predisposition to attention-deficit/hyperactivity disorder (ADHD)[179].

Throughout life, YY1 also regulates various neuroprotective pathways, playing a central role in ischemic damage, Parkinson's and Alzheimer's disease. For example, YY1 activates NRF2, which in turn initiates an antioxidant response to protect brain cells against ischemic damage following cerebrovascular accidents[180]. In Parkinson's Disease, YY1 is downregulated in microglia, along with other neuroprotective pathways like mTOR and TGF-β[181]. YY1 regulates the expression of NRF2-mediated antioxidant response and the transmembrane transporter SVCT2-dependent import of the protective drug ascorbate[182,183], which are key targets for Parkinson's disease treatment developments. A bioinformatics approach to uncover regulators of Alzheimer's Disease appointed YY1 as one of the master regulators[184]. Interestingly, in contrast to Parkinson's, in Alzheimer disease higher levels of YY1 mRNA were detected in human autopsy-derived whole brain samples and isolated neuron samples compared to controls[184]. It would be tempting to speculate that the protective function requires tightly regulated levels of YY1, while aberrant levels contribute to the onset and progression of neurodegenerative diseases.

YY1 has a highly context-dependent function also in cancer[185,186], where it can act both as a tumour suppressor or stimulator[185]. A detailed review of YY1 role in cancer, however, is beyond the scope of this *Introduction* (for a detailed overview see[187]).

Aberrations of non-coding elements in NDDs

As discussed in the previous sections, an increasing number of studies suggests that a high fraction of causative variants in neurodevelopmental disorders such as intellectual disability and autism, belong to pathways of transcriptional regulation and chromatin remodelling[135,188,189]. Besides mutations in *trans*-acting factors like TFs (e.g. YY1[135,188,189] and CTCF[135,188,189]) or chromatin modifiers (BICRA[190], SETD1B[191], *ACTL6B*[39], CHD8[192] just to mention some), also mutations of enhancers *in cis* have been proven to be causative of disease in an increasing number of cases. A classic example is pre-axial polydactyly caused by alterations of the ZRS, a long-distance enhancer that regulates Sonic hedgehog (*SHH*) expression in the embryonic limb[70,193]. Next to point mutations, also copy number variation (CNVs) such as duplications[194], and insertions[195] affecting this gene regulatory element have all been shown to cause polydactyly phenotypes, illustrating the wide range of alterations that can affect enhancer function resulting in a phenotype. In this section, I will discuss a number of other examples of enhancer's alterations, mainly in relation to disorders affecting the brain (**Table 1**).

Holoprosencephaly, a neurodevelopmental disorder characterized by craniofacial malformations, can be caused by coding mutations in the *SHH* gene. However, a point mutation in the *SHH* Brain Enhancer 2 (SBE2), located 460 kb upstream of the *SHH* gene, was identified in a patient with an identical phenotype[196]. This mutation was found to be disease-causing, as it disrupts the binding site of the TF SIX3, thereby leading to reduced forebrain SHH expression. In agreement, also mutations in *SIX3* can result in holoprosencephaly[197]. A disease-causing enhancer mutation is also found in the congenital eye malformation aniridia, that is often caused by haploinsufficiency of the TF PAX6. A point mutation in the *PAX6* eye-enhancer was found to disrupt PAX6 binding, affecting its own expression[198]. In another example, a 15-base pair deletion in a regulatory element upstream of an alternative transcript of *GPR56* was found in 5 individuals from 3 families[199]. *GPR56*, when mutant in its coding sequence, leads to widespread cobblestone malformation with cerebellar and white matter abnormalities. In the patients carrying the 15-base pair regulatory element deletion, polymicrogyria was bilaterally restricted to the Sylvian fissure, leading to a phenotype of speech delay, intellectual disability, and refractory seizures without further motor involvement. The authors could show that the deletion disrupts an RFX binding site, and thereby specifically alters the expression of *GPR56* in the perisylvian and lateral cortex, including the Broca area that is the primary language area.

Table 1 | Alterations of non-coding regulatory elements in diseases related to the central nervous system

Disease	Mutation	Affected gene	Ref
Holoprosencephaly	Point mutation	SHH	Jeong et al., 2009
Aniridia	Point mutation	PAX6	Bhatia et al., 2013
Polymicrogiria in the Sylvian fissure	Deletion	GPR56	Bae et al., 2014
Parkinson's disease	SNP	SNCA	Soldner et al., 2016
Schizophrenia	Tandem duplications	VIPR2	Vacic et al., 2011
Adult-onset demyelinating leukodystrophy	Deletion of TAD boundary and deletions	LMNB1	Nmezi et al., 2019; Giorgio et al., 2015
Intellectual disability	CNV	ARX	Ishibashi et al., 2015

Besides influencing disorders presenting early in life, also disease emerging later, such as neurodegenerative disorders and schizophrenia, are increasingly linked to enhancer variants. For example, a risk variant in an enhancer regulating α-synuclein expression was recently shown to affect gene expression by altering the binding of the TF EMX2 and NKX6-1[200]. In addition, tandem duplications of the non-coding upstream region of *VIPR2* have been observed in cases of schizophrenia and resulted in upregulated *VIPR2* expression[201]. Also, CNVs overlapping with enhancers in other schizophrenia related genes might be implicated in the disease pathogenesis, influencing the disease vulnerability[202].

Multiple CNVs have also been associated with periventricular nodular heterotopia (PNH), a brain malformation in which nodules of neurons are ectopically retained along the lateral ventricles[203]. Besides changing gene dosage, CNVs can also change the dosage and position of enhancers, as well as the higher-order chromatin organization of a locus[62,204]. Similarly, copy-number-neutral structural variants, such as inversions and translocations, can disrupt coding sequences or create fusion transcripts, but these types of variants can also disrupt or create new enhancer landscapes and chromatin domains, resulting in regulatory loss or gain of function. A clinical example of such a structural variant that changes the 3D architecture of the genome is the deletion of a TAD boundary at the *LMNB1* locus, which causes an enhancer to regulate a gene that is normally not regulated by that enhancer (so-called enhancer adoption). In this case, the enhancer adoption leads to adult-onset demyelinating leukodystrophy (ADLD), which is a progressive neurologic disorder affecting the myelination of the central nervous system[205]. More recently, deletions upstream of *LMNB1*, varying in size from 250 kb to 670 kb, occurring in repetitive elements, have revealed increased LMNB1 expression and an atypical ADLD phenotype[206]. Other rare inherited structural variants in *cis*-regulatory elements might influence the risk for children of developing autism spectrum disorders, depending on the parental origin of the structural variant[207]. Another study on autism using whole genome sequencing (WGS) on more than 2,000 individuals found that probands carry more gene disruptive CNVs and SNVs resulting in severe missense mutations and mapping to predicted foetal brain promoters and embryonic stem cell enhancers[208]. In addition, CNVs covering the regulatory elements of the *ARX* gene might cause an intellectual disability phenotype[209], and rare non-coding CNVs near previously known epilepsy genes were enriched in a cohort of 198 individuals affected with epilepsy compared to controls[210]. Similar findings are reported for multiple system atrophy[211] and non-coding variants might influence expression of *GLUT1* causing epilepsy[212].

Two large-scale analyses focusing on enhancers and their role in neurodevelopmental disorders have recently been performed. Using a targeted sequencing approach, Short and colleagues studied *de novo* occurring genomic variants in three classes of putative regulatory elements in 7,930 individuals suffering from developmental disorders recruited into the Deciphering Developmental Disorders (DDD) study and their parents[64]. The three classes of regulatory elements that they assessed consisted of 4,307 highly evolutionarily conserved non-coding elements[213], 595 experimentally validated enhancers[214], and 1,237 putative heart enhancers[215], together covering 4.2 Mb of genomic sequence. In the 6,239 individuals in which exome sequencing did not find a disease cause, they found that conserved non-coding elements were nominally significantly enriched for *de novo* variants (422 observed, 388 expected, P = 0.04), whereas in experimentally validated enhancers (153 observed, 156 expected, P = 0.605), heart enhancers (86 observed, 86 expected, P = 0.514), and intronic controls (901 observed, 919 expected, P = 0.728) *de novo* variants were not enriched. When focusing only on conserved non-coding elements that had evidence of activity in brain, they observed an even stronger enrichment. Based on their analysis, the authors estimate that only around 1-3% of exome-negative individuals will be explained by *de novo* variants in foetal brain-active regulatory elements. However, as in this study only *de novo* variants were assessed, and only a limited set of regulatory elements was used which were already defined in 2010, this is likely an underestimation of the possible impact of the non-coding genome alterations on neurodevelopmental disorders. Doan and colleagues performed a similar targeted sequencing approach assessing human accelerated regions (HARs)[65], conserved regions with elevated divergence in humans that might reflect potential roles in the evolution of human-specific traits. This study provides evidence that HARs can function as regulatory elements for dosage-sensitive genes expressed in the central nervous system. Using data from a large cohort study investigating 2100 sibling cases of autism spectrum disorder (ASD), they found that *de novo* CNV's affecting HARs, or HAR-containing genes, could be implicated in up to 1.9% of ASD cases in simplex families. Furthermore, they analysed consanguineous ASD cases using WGS from 30 affected and 5 unaffected individuals and designed a custom capture array to sequence HARs in another 188 affected and 172 unaffected individuals. Individuals with ASD exhibited an excess of rare (Allele Frequency <0.5%) bi-allelic HAR alleles (43% excess compared to unaffected, p=0.008), and this enrichment further increased when taking into consideration only HARs likely active as regulatory elements in brain. Using MPRA, 343 bi-allelic HAR variants were functionally tested, and 29% of these were shown to alter the regulatory activity of the reference sequence. Therefore, the enrichment of regulation-altering variants in HARs with predicted activity suggests that many may contribute to the pathogenesis and diversity of ASD. They further functionally validated their findings in three examples of bi-allelic variants in HARs identified in ASD families, regulating the genes *CUX1*, *PTBP2* and *GPC4*, further providing evidence that the investigation of regulatory elements such as HARs is promising to solve currently genetically unexplained disease cases.

What appears from the examples given above, is that a wide range of enhancer alterations can result in effects on gene transcription, leading to a disease phenotype. This can vary from point mutations affecting the binding of crucial TFs, to deletions, duplications, shuffling of the genomic location affecting their function (e.g., enhancer adoption) or alterations in the global chromatin landscape disrupting TAD borders, just to mention a few. Given the complexity of these disease mechanisms, and the current shortcomings in understanding the roles of the non-coding genome, it seems likely that in the next decade many more examples of enhance alterations in genetic diseases will be identified. The identification of the full repertoire of functional enhancers genome-wide in a given cell type or tissue is still challenging. However, several techniques and new technologies, that I will discuss in the next paragraphs, are leading the community towards a more thorough understanding of enhancers and their mechanism of action.

Genome-wide identification of putative enhancers

As introduced in the previous paragraphs, transcriptional enhancers were first described as DNA sequences that are able to enhance gene expression from an episomal plasmid (e.g. a non-integrating, extra chromosomal circular DNA), irrespective of their location and orientation relative to the TSS[92,216]; thus, enhancer identification was first limited to low-throughput reporter assays, where small fragments of DNA were tested for regulatory activity influencing reporter gene expression. Today, the most widely applied experimental techniques for genome-wide identification of putative enhancers at the endogenous genomic locus do not rely directly on this functional property, but rather on features that distinguish enhancers from non-regulatory regions at the chromatin level. Indeed, as aforementioned, enhancers are bound by TFs and transcription coactivators and are located in open chromatin regions depleted from nucleosomes. The surrounding nucleosomes have specific histone tail modifications, such as H3K4me1 and H3K27ac. Moreover, some enhancers are bi-directionally transcribed in so-called enhancer RNAs (eRNAs). However, even though these features correlate with enhancers, other genomic regions share the same chromatin characteristics, and more functional tests are required to prove that putative enhancers are indeed having a direct functional role in activation gene expression[217]. This led to the development of high-throughput functional screenings, overall known as massively parallel reporter assays (MPRAs) that quantify the enhancer activity of millions of sequences in parallel. In the next paragraphs I will discuss the most widely used techniques to identify putative regulatory regions (**Figure 7, Table 2**).

Table 2 | Methods for enhancers identification

Method	Description	Advantages	Disadvantages
ChIP-seq	Chromatin immunoprecipitation of histone modifications or TFs coupled with NGS.	Determines genome-wide binding patterns of protein of interest	Not all enhancers are marked by H3K27ac, H3K4me1, or tested TFs. Requires availability of ChIP-grade antibodies. Cannot determine enhancer activity. Cannot identify target gene.
ATAC-seq	Identification of open chromatin regions by the transposon Tn5, that cuts the DNA and inserts sequencing adapters.	Fast. Requires a low number of cells. No need for any a priori knowledge.	Other elements are in open chromatin regions. Cannot determine enhancer activity. Cannot identify target gene.
eRNA detection	Detection of the bidirectionally transcribed eRNA by sequencing the nascent RNA through techniques such as GRO-seq or CAGE.	Identifies enhancer transcription	Not all active enhancers are transcribed.
Chromosome Conformation capture	Detection of topological interactions between wo loci (3C) or genome wide (4C, 5C, Hi-C).	Identifies enhancer-target gene interactions	Cannot determine enhancer activity.
STARR-seq	Identification of functional enhancers by a massively parallel reporter assay where active enhancers drive their own transcription.	Identifies functional enhancers. Quantitatively measures enhancer activity. High throughput.	Episomal. Highly complex plasmid libraries requiring substantial number of cells for transfection. Possible false negative results.
CRISPR-Cas9 screenings	Endogenous manipulation of enhancers to force their activation or inactivation.	Identifies functional enhancers. Can be high throughput. Determines the endogenous effect of enhancer manipulation.	Off-target activity. Possible false negative results.

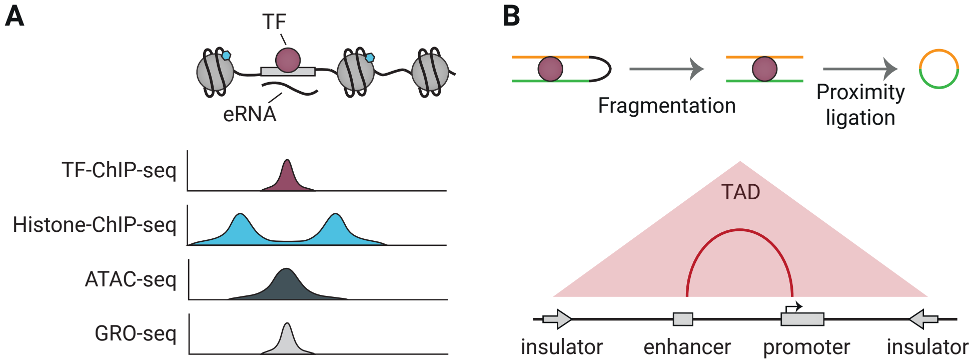


Figure 7 | Overview of the main techniques currently used to identify putative enhancer sequences and their interacting genes. **A**) Schematic drawing on an TF-bound enhancer, located in nucleosome depleted DNA from which eRNA is transcribed. Below are representative genome browser tracks shown, illustrating expected profiles for the same genetic region. Histone-ChIP-seq is illustrative for marks such as H3K27ac and H3K4me1. **B**) Cartoon representing the main steps of the workflow of Chromosome conformation capture technologies: nuclei are cross-linked, chromatin is then digested and re-ligated by proximity ligation. The two stretches of DNA that are normally located far away from each other (yellow and green), are now ligated together and can be tested by PCR or sequencing. In the bottom part is indicated the output of the experiment, with which TADs and enhancer-promoter interactions can be identified.

Chromatin Immunoprecipitation

Chromatin immunoprecipitation (ChIP) was introduced more than 30 years ago to study protein-DNA interactions[218] and it follows three basic steps. First, proteins are covalently cross-linked to their DNA binding-site by treating cells with formaldehyde. Chromatin is then sheared, and protein-DNA complexes are selectively co-immunoprecipitated with an antibody against the protein of interest. Finally, the cross-linking is reversed, and DNA is isolated and tested to identify the binding sites of the protein of interest. In more recent years, the emergence of Next Generation Sequencing (NGS) technologies, allowed genome-wide mapping of these protein-DNA binding sites (ChiP-seq)[219,220]. ChIP-seq is now primarily used to identify putative enhancers across the entire genome by immunoprecipitation of TFs, specific histone-tail post-translational modifications, including H3K4me1[221] and H3K27ac[222], and transcriptional coactivators, such as the histone acetyltransferase p300/CBP[223] and Mediator[109]. However, neither the binding of a TF nor the presence of histone modifications provide definitive evidence that a sequence acts as a transcriptional enhancer. For example, tissue specific enhancers can have a certain degree of H3K27ac enrichment in tissues where they are not active[224]. Several studies have used ChIP-seq for histone modifications to predict enhancers during human brain development[225,226] and in adult brain[227-229], and some have made direct comparisons to brains from other primates, providing important insights in human evolution[226,230]. Recent ChIP alternatives such as CUT&RUN or CUT&Tag, that are based on the tethering of nucleases to a specific antibody bound site on the genome of immobilized and permeabilized nuclei, require a reduced number of starting cells and reduced sequencing depth[231,232].

Identification of open chromatin regions

As abovementioned, *cis*-regulatory sequences like enhancers are enriched in chromatin regions depleted from nucleosomes[233], as nucleosomes would impede TF binding[234]. These accessible DNA regions can be identified in a genome-wide fashion thanks to several techniques such as DNase-seq, FAIRE-seq and ATAC-seq. DNase-seq takes advantage of the hypersensitivity of open chromatin to nuclease digestion. Briefly, cell nuclei are isolated, and DNA is digested with limiting concentrations of DNase-I. Fragments of about 500 bp are then selected and used for library preparation and sequencing[235]. FAIRE-seq (Formaldehyde-Assisted Isolation of Regulatory Elements) is based on the separation of free and nucleosome-bound DNA. Chromatin is cross-linked with formaldehyde to covalently bind nucleosomes to the DNA, and then sonicated and purified by phenol-chloroform extraction. Nucleosome-bound DNA is sequestered to the interphase, while accessible DNA can be recovered from the aqueous phase and sequenced[236]. Finally, the most recently developed method ATAC-seq (Assay for Transposase Accessible Chromatin with high-throughput sequencing) exploits the preference of transposons to land in open chromatin regions. Shortly, the transposon Tn5, loaded with sequencing adapters, is able to simultaneously cut the DNA and insert the adapters in a process known as tagmentation. The open chromatin regions where the transposon preferentially inserts are then amplified with primers binding to the adapters and sequenced. Compared to DNase-seq and FAIRE-seq, ATAC-seq is a simple and fast method that requires less starting material and does not require gel-purification or crosslinking reversal steps and is therefore less prone to loss of material[237]. However, as mentioned earlier, other regulatory elements such as insulators or promoters are also located in accessible chromatin[233]. Therefore, ATAC-seq should be used in combination with other techniques that are more selective for enhancers. However, also inactive enhancers can be in open-chromatin regions[217,238] and these techniques cannot discriminate functional from non-functional enhancers. A major advantage of all techniques assessing chromatin accessibility compared to ChIP-seq is that they screen for putative regulatory regions in an unbiased way, not requiring *a priori* knowledge of enhancer binding factors and not being restricted to the use of available ChIP-grade antibodies. A recent study has used ATAC-seq and RNA-seq to determine open chromatin regions and gene expression at different gestational weeks, and in different areas of the brain, i.e., the ventricular zone and the neuronal layers, providing a first glimpse on open chromatin dynamics during foetal brain development[239].

eRNA sequencing

Transcription of enhancer sequences was first reported in the early nineties in the Locus Control Region (LCR) of the β-globin gene cluster[240-242], where it was found that the expression of the LCR is restricted to the erythroid lineage. Later, transcription of regulatory elements into enhancer RNAs (eRNAs) was validated genome-wide with sequencing, at first, of total neuronal RNA[243], followed by sequencing of nascent RNA (GRO-seq, CAGE) in different cell types[244-247]. Enhancer RNAs are generally bi-directionally transcribed and not polyadenylated[243] but reports of unidirectional transcription and polyadenylation of eRNAs exist[248]. Enhancer transcription was shown to correlate with the presence of other enhancer marks such as histone tail post-translational modifications and p300/CBP and RNA pol II binding[244-246], but whether their expression is a cause, or a consequence of gene transcription is still debated[249]. If eRNA transcription has a direct functional role on gene regulation and is not just noise due to the recruitment of RNA pol II, the effect can either be mediated by the transcription process itself or by the transcript produced upon transcription, which might have direct *cis*-regulatory activity similar to other non-coding RNAs such as those involved in X chromosome inactivation[250,251]. However, even if eRNA presence correlates with enhancer activity at some loci, it seems that it is neither required nor sufficient in all instances[217]. For example, a recent study assessing eRNAs in brain only found that around 600 intergenic and intronic enhancers are transcribed in eRNAs, and this number even further decreased when considering only those eRNAs replicated in an independent data set or overlapping with enhancer associated histone modifications[252]. The FANTOM project has also found a small number of eRNAs in brain, although the majority is not overlapping with those from Yao and colleagues[247]. The number of predicted brain related enhancers based on other assays by far outnumbers this rather small set of transcribed enhancers, indicating that methods that just take eRNA transcription into account may oversimplify the identification of putative enhancers and may not catch the complete regulatory landscape.

Identification of long-distance chromatin interactions

All methods described until now identify putative enhancers but understanding which genes they regulate remains a challenge, as even though they often regulate nearby genes, they can also be found at long distances from the TSS of their target gene. Moreover, it is becoming more and more clear that chromatin organization plays an important role in transcription and enhancers and promoters need to be brought in close proximity for transcription to take place. In the past ~20 years several techniques have been developed to address this question (reviewed here[253,254]). The pioneering method, on which all the later developments are based, is known as chromosome conformation capture (3C) and relies on the formaldehyde cross-linking of chromatin within nuclei, followed by restriction digestion of chromatin and re-ligation by proximity ligation. The obtained fragments represent the junction of two chromatin regions that are normally located far away from each other on the linear genome, but are in close proximity in 3D space, and these junction products can be quantified by PCR[255]. 3C was developed to study whether two known regions are interacting with each other and is thus described as a “one vs one” method[254]. Further advances of 3C-based techniques allowed the identification of increasing numbers of contacts; for example, 4C, “one vs all”, allows the identification of all the regions interacting with a specific site of interest[256,257], 5C, “many vs many”, investigates all contacts that are happening in a specific locus[258], and targeted chromatin capture (T2C) improves the “many vs many” approach and allows a higher resolution analysis of all the interactions happening in a specific locus[259]. Finally, high-throughput contact identification became possible with Hi-C[260]. Hi-C allows the identification of genome-wide interactions thanks to the introduction of biotin-labelled nucleotides at the sites of restriction-digestion. The ends are then ligated, the chromatin is sheared, and the junctions are enriched by streptavidin pull-down and sequenced. By the application of an algorithm on Hi-C data, topologically associating domains (TADs) can be defined. A recent study has generated Hi-C maps from gestational week 15, 16 and 17 of human brain development, a critical period for cortex development[261], permitting the large-scale annotation of previously uncharacterized regulatory interactions relevant to the evolution of human brain. For example, the results of this study have linked several non-coding variants identified in GWAS to genes and pathways involved in schizophrenia, highlighting novel mechanisms underlying neuropsychiatric disorders. To investigate all the genome-

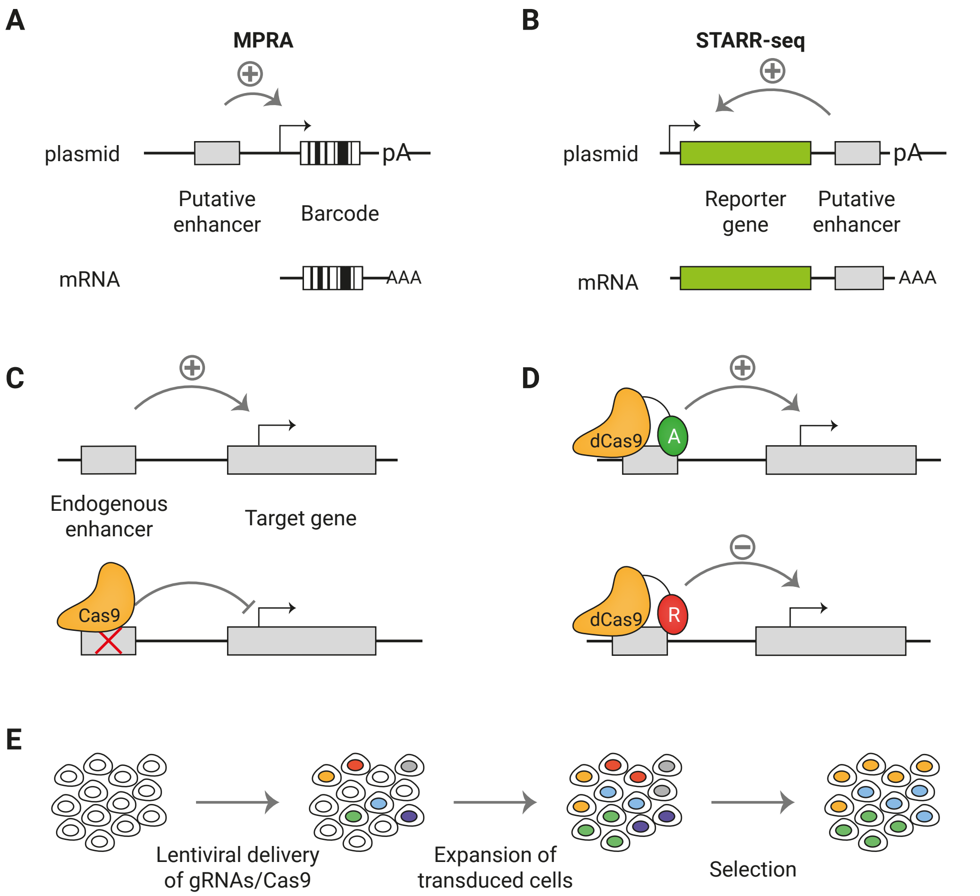


Figure 8 | Methods for functional identification of enhancers. **A**) Massively parallel reporter assays (MPRA) to test enhancer activity in an episomal setup. The putative enhancer sequence is cloned upstream a minimal promoter that drives the expression of a reporter gene and a unique barcode. **B**) With STARR-seq the putative enhancer sequence is cloned downstream the reporter gene and upstream of the polyA signal. When the enhancer sequence is active, is can drive the expression of the reporter (green) and of itself. In both MPRA and STARR-seq the mRNA is sequenced to identify the active enhancers. **C**) Cas9 can be used to knock out an enhancer at the endogenous genomic locus to assess its effect on the target gene transcription. **D**) A catalytically inactive Cas9 (dCas9) can be fused with activators (A: VP64; TET1; p300) or repressors (R: KRAB; SID4X; DNMT3A; KDM1A). **E**) Cas9 screens can be combined with high-throughput screenings by targeting Cas9 expressing cells with a lentiviral library of gRNA at a low MOI. By doing so, each cell will express a single gRNA and by different selections, such as drug resistance or reporter gene expression, it is possible to investigate the effect of the ablation of a large number of putative enhancers on gene expression in parallel.

wide interactions involving a specific protein of interest, techniques such as HiChIP[262] or PLAC-seq[263] were developed, by introducing a chromatin immunoprecipitation step. This method has the advantage of requiring less input material and less sequencing reads. For example, Nott and colleagues identified cell-type specific enhancer-promoter interactions by H3K4me3 PLAC-seq in microglia, neurons and oligodendrocytes sorted from human brain, and could identify Alzheimer's disease associated GWAS-hits in cell-type specific enhancers[264].

Despite their capacity to identify enhancer-promoter interactions, chromatin conformation techniques have the disadvantage of not directly measuring regulatory activity. Moreover, in most cases, interactions are determined on a population level on a high number of cells, which might only provide a snapshot of dynamic regulatory interactions. The spatial resolution at which interactions can be determined is heavily influenced by the sequencing depth of Hi-C experiments, hence, there is still a need for more functional tests to validate the regulatory activity of the identified interactions.

High-throughput functional identification of enhancers

ChIP-seq, open chromatin mapping and expression analysis have been of tremendous help to globally characterize the gene regulatory landscape of the non-coding genome, but in many instances, the identified putative enhancer sequences fail to perform as enhancers in functional validation experiments, giving rise to false positive predictions[265] (see[266] for an excellent *review*). Moreover, the resolution of commonly used techniques usually allows the identification of regions in the range of 500-1000 bp as potentially including an enhancer. But this makes it difficult to pinpoint those nucleotides that are of real functional relevance within a given predicted enhancer sequence, and this complicates the assignment of functional roles to nucleotide variants found in the human population. Finally, many of the currently used techniques are based on associations between epigenetic marks and putative enhancers, excluding other regions of the genome, that may nevertheless be functionally relevant, from being assessed[267]. Direct high-throughput functional tests of enhancer activity, such as massively parallel reporter assays (MRPAs) and CRISPR-Cas9 based screens have the potential to address these shortcomings (**Figure 8**).

Most traditional functional tests for enhancer activity are based on reporter assays, in which a putative enhancer sequence is cloned into a vector with a reporter gene driven by a minimal promoter that alone is not sufficient to induce reporter gene expression. The vectors are then transfected into a cell line or organism of interest, and the reporter gene expression is determined[92]. MPRAs are high throughput reporter assays where DNA sequences are inserted before the minimal promoter of a vector with a specific barcode sequence downstream of the open reading frame, which allows the simultaneous assessment of thousands of sequences for enhancer activity in parallel[109,238,268-273]. After cell transfection, RNA can be purified and sequenced. If the sequence cloned into the vector is a functional enhancer it drives the expression of the corresponding barcode. An adapted approach is Self-Transcribing Active Regulatory Region (STARR) sequencing[238]. STARR-seq takes advantage of the position-independent activity of enhancers. Indeed, differently from other MPRAs, STARR-seq does not rely on barcodes, but the candidate sequences are cloned downstream of the TSS and, when active, drive their own transcription. With this assay, millions of sequences can be tested in a single experiment. In both cases, the activity of the enhancer can be measured by the relative abundance of the barcode/sequence transcript from RNA-seq, in comparison to sequencing of the input plasmids. Similar episomal high-throughput approaches have recently been developed to also measure promoter responsiveness to enhancers[274] and autonomous promoter activity[275].

The major advantage of these tests is that they are unbiased since they are not based on any *a priori* hypothesis about TF binding or histone modifications. Nevertheless, the size of the human genome requires the construction and transfection of large plasmid libraries, and thus substantial numbers of cells and deeper sequencing and might therefore lead to a lower resolution. To overcome this limitation, it is possible to focus STARR-seq only on putative enhancers, testing only the sequences identified with ATAC[276] or other techniques[277,278].

Despite being incredibly useful to test millions of sequences for enhancer activity in a high-throughput manner, reporter gene assays may have several limitations. First, enhancer activity is tested most often on an episomal background, which might not completely reflect endogenous gene regulation in its native genomic context[279]. Interestingly, recent studies suggest that the effects of this might be less strong than initially suggested, as there is a high correlation between episomal enhancer activity and endogenous gene regulation when a set of enhancers is assessed on both plasmids and integrated at multiple genomic locations[280]. Second, MPRAs may potentially give false negative results. Indeed, if a sequence is found inactive in a reporter assay, this does not exclude that it is active as an enhancer in a different cell type, in a different moment in time or has another, but still biologically relevant, role independent on enhancer activity[281].

One way to overcome these possible limitations of transgenic reporter assays, is to use a CRISPR-Cas9 system (**BOX1**) to manipulate enhancers at the endogenous chromatin context. The regulatory element sequence can be deleted, allowing to test the effect of the enhancer ablation on gene expression in the endogenous chromatin environment. This approach can be used to study a selected enhancer of interest, but also in high-throughput screenings with large libraries of gRNAs that are introduced in Cas9-expressing cells. Lentiviral transduction of gRNAs at a low multiplicity of infection can result in a single gRNA integration per cell, and in combination with various means of positive or negative selection, such as drug selection or assessment of reporter gene expression, this can be used to investigate in parallel and on a large scale the effect of multiple putative enhancer ablations on gene expression. To this end, large populations of cells are transduced, and the quantitative presence of gRNAs is determined by next generation sequencing of isolated DNA prior and after a selection. If a sequence has an important role in gene regulation, the ablation of that sequence is expected to result in a disadvantage for the cells, and therefore gRNAs targeting relevant functional enhancers will be depleted over time. By comparing sequencing reads before and after the selection, it is possible to determine which gRNAs are lost over time, and as the targets of the gRNAs are known, the relevant enhancers can be identified. In one of the first applications, DNA regions around the *TP53* and *ESR1* gene loci were investigated, and it was shown that this approach was feasible to identify functional enhancers and, furthermore, using a dense CRISPR-Cas9 gRNA tiling screen, functional domain within these enhancer sequences were precisely mapped[282]. Using a similar approach, more than 18.000 gRNAs were used to test around 700 kb of sequence flanking genes involved in BRAF inhibitor resistance in melanoma, finding non-coding regions involved in gene regulation and chemotherapeutic resistance[283]. Other studies investigated putative enhancers involved in oncogene induced senescence[284], regulation of the *HPRT* gene involved in Lesch-Nyhan syndrome[285] and regulation of the *POU5F1* gene in embryonic stem cells[286,287], amongst others[288-291]. Besides genome-engineering, CRISPR-Cas9 can also be applied to edit the epigenome, even in high-throughput screenings. Indeed, by fusing a catalytically dead Cas9 (dCas9), that lacks endonuclease activity, to various functional domains it is possible to alter the status of an enhancer forcing its activation or inactivation, referred to as CRISPRa and CRISPRi, respectively. Functional additions to dCas9 leading to enhancer activation include transcription activating domains such as multiple repeats of the herpes simplex VP16 activation domain (VP64)[292,293], the nuclear factor-κB (NF-κB) trans-activating subunit activation domain (p65) and human heat-shock factor 1 (HSF1)[294], the ten-eleven translocation methylcytosine dioxygenase 1 (TET1)[295], and the p300 acetyltransferase[296]. In contrast, transcription repressive domains that can be used to silence enhancers include Krüppel-associated box (KRAB) domain[297,298], four concatenated mSin3 domains (SID4X)[299], cytosine-5-methyltransferase 3A (DNMT3A)[300], Histone deacetylase 3 (HDAC3)[301], and the lysine-specific histone demethylase 1A (KDM1A), called dCas9-LSD1[302]. Several of these dCas9 fusion have been used to activate or repress regulatory elements, and a number of studies have used them in high-throughput screening approaches, most of which focused on enhancer repression[303-306] but some included also activation[307,308]. It seems only a matter of time until more studies editing enhancers in various cell types using the full CRISPR-Cas9 toolbox will be published. Obviously, as all experimental approaches, also CRISPR-Cas9 has its pitfalls and is still far from perfect. For example, reduced on-target activity and off-target effects of gRNAs can introduce experimental noise, and it remains essential that screening results are validated independently. Also, it remains to be seen whether subtle enhancer effects on gene expression, that might still be of biological relevance, can be detected using CRISPR-based screens.

What is clear from the above, is that our knowledge on complex gene regulatory mechanisms and the tools to investigate them have increased dramatically over the last decade, providing insights into many sophisticated processes that need to occur correctly for development to proceed. Besides improvements in the molecular technologies to investigate gene expression, also the systems to model human brain development, as I will discuss in the following part of this *Introduction*, have seen a rapid evolution, promising a bright future for the investigation of the functional effects of non-coding variation during the development of such a complex organ as the human brain.

Neurodevelopment and disease modelling

The study of the human brain and related diseases has always been complicated by the paucity of available material for research, and the few available tissues can only be studied post-mortem, reflecting a snapshot of the endpoint of the disease, and not allowing the study of development or disease progression. Primary human neuronal culture can also be obtained from aborted foetuses[309] or neurosurgical samples[310], but the yield is low, and it incurs in various ethical problems. As a consequence, the focus moved onto the development of animal models carrying gene mutations similar to those found in patients, providing the advantage that different tissues can be examined at different time points, allowing the study of disease progression, but also of behavioural phenotypes.

Historically, mice have been the most used model, and they have proven extremely useful to deepen our understanding of diseases such as Rett syndrome[311] or Angelmann syndrome[312], but despite modelling some human NDD symptoms, they are far from recapitulating the full phenotype. The rodent brain has major physiological, anatomical, but also cognitive and behavioural differences with the human one. Just to mention a few, the mouse is lissencephalic (it has a smooth brain surface lacking the gyri and sulci typical of the human brain) and has significantly less bRGCs. The use of non-human primate models can partially overcome these issues, despite raising major ethical problems due to their similarity to humans. However, several diseases such as Rett syndrome and ASD have been successfully modelled in non-human primate models[313].

Other model organisms such as zebrafish (*D. rerio*) and fruit flies (*D. melanogaster*) are emerging as new valuable tools in the field of neurodevelopmental disorders. An important advantage of these models is their high-throughput potential, the large number of offspring, the short life cycle and the ease of genetic manipulation along with the easier compliance with ethical regulations and lower costs for experimentation and housing. Zebrafish are small vertebrates that share 70% of genetic identity to human[314]. As one of their major advantages, zebrafish larvae develop very fast (complete embryos are formed within 24 hours post fertilization), ex-utero and are transparent, rendering them very suitable for imaging experiments. *Drosophila*, despite being an invertebrate and thus quite far on the evolutionary scale from human, share many pathways at the molecular, physiological and behavioural level, and 75% of the genes currently known to be involved in intellectual disability have a functional orthologous gene in this organism[315]. Both these models contributed much to the understanding of diseases including ID/ASD[316,317] and microcephaly[318].

Undoubtedly, animal models will remain a key step in understanding the disease pathogenesis on a systemic, multi-organ level, and from a behavioural perspective, however there is the need to use in parallel alternative models to study disease mechanism in human cells. The establishment of human pluripotent cells revolutionized the field, as large number of progenitor cells could finally be obtained and potentially be differentiated towards any cell type of interest, allowing the investigations of the processes governing human brain development and disease in a relevant cell-type.

Human pluripotent stem cells

The term human pluripotent stem cell (hPSC) is used to describe cells that can indefinitely self-renew in culture and have the potential to be differentiated in cell-types belonging to all the three germ layers: endoderm, ectoderm and mesoderm, a capability that is referred to as pluripotency. hPSCs include human embryonic stem cells (hESCs), which were the first ones to be isolated, and human induced pluripotent stem cells (iPSCs).

Embryonic stem cells

The human body is composed of more than 200 different cell types that all arise from a single totipotent cell, the zygote, formed upon fertilization of the egg. About five days post-fertilization the cells organize to form the blastocyst, that consists of trophectoderm surrounding an inner cavity filled with fluid and an inner cell mass (ICM). The cells of the ICM give rise to all cells of the developing individual and can be isolated and expanded in culture, generating the so-called **embryonic stem cells**.

The first ESC lines were isolated from preimplantation mouse blastocyst in 1981[319] followed by several other mammals including cow[320], pig[321] and, almost two decades later, human[322]. Thomson and colleagues isolated and expanded single colonies of cells (although none of them originating from a single cell) derived from the ICM of human embryos obtained from in vitro fertilization (IVF), thereby establishing 5 hESCs lines characterized, morphologically, by a high nucleus to cytoplasm ratio and by clear nucleoli. Among these cell lines is H9, a female hESC line with a normal XX karyotype, which is used in the experiments described in this thesis. All the established cell lines showed high telomerase activity, expression of undifferentiated cells' surface markers, the ability to differentiate into all three germ layers and to produce teratomas when injected into SCID mice (severe combined immunodeficient: mice that lack mature B and T lymphocytes), the latter being the gold standard to assess pluripotency.

Two of the defining characteristics of ESCs are the ability to self-renew, i.e., divide creating copies of themselves, and to differentiate towards all the cell types of the human body. These two properties depend on the cells' ability to have an undifferentiated gene expression profile, while maintaining the potential to acquire different gene expression programs upon differentiation signals. Key mechanisms allowing this plasticity are epigenetic and transcriptional regulation. Three transcription factors involved in the intrinsic maintenance of the core transcriptional regulatory circuitry of hESCs are OCT4, NANOG and SOX2, that co-bind and regulate a large number of target genes. Moreover, they act in an auto-regulatory loop to maintain their own expression[323]. On the other hand, Polycomb group proteins silence genes involved in differentiation, such as key developmental transcription factors, that are often in a poised state marked by a bivalent histone code, involving both H3K27me3 (a repressive histone mark) and H3K4me3 (an active histone mark). Upon differentiation, only one of the two marks is retained, allowing the gene to stay silent or to become activated, respectively.

Human ESCs can be maintained in an undifferentiated state in well-defined culture conditions either on feeders (mitotically inactivated fibroblasts mostly of mouse origin (mouse embryonic fibroblasts, MEF) but also from other species including human) or on extracellular matrix coated dishes. Culturing in the absence of these factors leads to spontaneous differentiation, as does letting cells grow to confluency or in multiple layers. The culture medium generally consists of a basal medium supplemented with several factors among which are often present serum, amino acids, glucose and β-mercaptoethanol. In the absence of serum, the medium needs to be supplemented with FGF, that has been shown to be crucial for the maintenance of self-renewal by regulating TGFβ signalling in sustaining high *NANOG* expression[324]. Several commercial mediums, such as the widely used *mTESR1* (Stem Cell Technologies), are available for culturing hESCs in serum-free and feeder-free conditions.

Undifferentiated hESC cells can be kept in culture mainly in two different states: primed and naive. H9 hESCs maintained in culture as just described are considered **primed**, as they resemble the post-implantation epiblast, and can be converted to the earlier **naive** stage by either transient transgene expression of pluripotency factors[325,326] or by changing culture conditions[327,328]. Mouse ESCs instead resemble the preimplantation embryo, thus are considered naive. As one of their major characteristics, naive cells have a shorter doubling time, a rounded 3D morphology and exhibit a good survival as single cells. On the other hand, primed ESCs have a flat morphology and are vulnerable as single cells. Indeed, to improve their survival and reduce their differentiation upon passaging, the medium needs to be supplemented with inhibitors of Rho kinase (ROCK)[329]. Human naive cells are further characterized by an unrestricted lineage potential, being able to generate trophectoderm[330]. However, mESC even though naive do not have this potential[330].

hESCs are a valuable research tool but have also clinical potential (applications will be discussed in more detail in the next *Section*). However, they are not patient specific, so in transplantation therapy might induce immune rejection. A breakthrough in the field dealing with this concern happened in 2006 when an outstanding idea by Kazutoshi Takahashi and Shinya Yamanaka, granted Yamanaka (along with Sir John B. Gurdon) the Nobel prize in Physiology or Medicine 2012 "*for the discovery that mature cells can be reprogrammed to become pluripotent*".

Induced pluripotent stem cells (iPSC)

In 2006 Yamanaka and co-workers first described the induction of pluripotent stem cells (**iPSCs**) from mouse adult somatic cells (**reprogramming**) by the overexpression and subsequent endogenous re-activation of the key ESCs transcription factors Oct4, Klf4, Sox2, and c-Myc, now known as OKSM or Yamanaka factors. The first evidence of reprogramming was obtained in mouse, but subsequent experiments showed that the same cocktail of transcription factors could also reprogram human cells[331]. Later on, plenty of other factors' combinations have been shown to efficiently reprogram cells, for example KLF4 and cMYC can be replaced by NANOG and LIN28[332]. iPSCs can be obtained starting from a variety of adult somatic cells, including fibroblasts[331], peripheral blood cells[333,334], and keratinocytes[335]. Reprogrammed cells are very similar to ESCs of the corresponding species (mouse iPSCs resemble naïve mESCs while human iPSCs resemble primed hESCs) as they share the same morphology and surface markers expression, they can self-renew, differentiate into all three germ layers and they form teratomas when injected *in vivo*[336]. Moreover, they require identical culture conditions, both in terms of the feeder or the matrix on which cells can be grown and of medium.

The original strategy for the delivery of the transgenes was based on retroviral vectors[331,336], among which lentiviral vectors have a higher reprogramming efficiency[332,337], even though the generation of iPSCs remains a highly inefficient process. However, these vectors integrate in the genome, making the cells not suitable for therapy as stable integration of cMYC would increase tumorigenicity. The safest options that grant easy removal of the exogenous factors are represented by Sendai viruses, single-stranded RNA viruses that replicate in the cytoplasm[338,339], and synthetic RNA[340] or recombinant protein[341] delivery, with the last two being the most technically challenging but the safest for clinical applications.

Human iPSCs generation is highly variable in its quality, even when considering clones derived from the same original culture, and this might complicate the interpretation of results from downstream functional experiments. For this reason, it is suggested to proceed with more than one independent clone generated from each starting cell line.

Human iPSCs and ESCs share similar applications, as they can be used for disease modelling, drug testing, or cell replacement therapy. To model disease, hESCs can been obtained from embryos that undergo pre-implantation genetic diagnosis (i.e., a screening to identify embryos with a disease-causing variant and avoid their implantation into the mother). For instance, some studies obtained hESCs carrying variants responsible for Huntington's disease [342] and Fragile X syndrome[343]. However, the use of hESCs still incurs in ethical concerns because of the need to form and dissociate human embryos, that iPSCs do not have. To circumvent this problem, when no cells from patients are available for reprogramming, it is possible to introduce the desired genetic alteration in established hESCs lines, such as H9. These genome editing strategies became increasingly popular in the past decade primarily thanks to the introduction of CRISPR/Cas9, a ground-breaking discovery that led to the award of the Nobel prize in chemistry 2020 to Emmanuelle Charpentier and Jennifer Doudna "*for the development of a method for genome editing*" (**BOX1**).

One of the major advantages of iPSCs over genetically modified hESCs to model disease is that being derived from affected patients, they allow to study the variant of interest in the genetic background of the patient itself. This allows to exclude any confounding effect derived from a different genetic background (as can be the one of the control-hESCs) and to directly link *in vitro* phenotypes to the clinical presentation. The first disease-iPSCs were generated already in 2008, from patients carrying a wide variety of genetic diseases including Parkinson's disease, Huntington's disease, Duchenne muscular dystrophy and amyotrophic lateral sclerosis (ALS)[344,345]. In the case of ALS, patient-derived iPSCs could also be differentiated towards the affected cell type, motor neurons, showing the potential use of these cells to model disease[345]. In 2009, Ebert and colleagues demonstrated defects in motor neurons induced from iPSCs derived from an SMA-type 1 affected individual compared to the unaffected mother, proving that iPSCs can indeed recapitulate the disease

BOX1: Genome editing

The first successful attempts at targeted genome editing in mammals date back to the 80s when Oliver Smithies and Mario R Capecchi demonstrated that homologous recombination with a donor plasmid DNA could be used to target a desired gene in the human[351] and mouse[352] genome, respectively. The efficiency of recombination was very low (1 in 1000 cells), nevertheless, editing the genome of mESC, previously established by Sir Martin J. Evans[319], allowed the generation of the first knock out mouse models[353,354]. These three scientists were ultimately granted the 2007 Nobel prize in Physiology or Medicine "*for their discoveries of principles for introducing specific gene modifications in mice by the use of embryonic stem cells*".

Genome editing became increasingly popular in the last 10 years when several nucleases have been developed to target a specific genomic locus and induce a double-stranded break (DSB). These DSBs can then be repaired via either non-homologous-end-joining (NHEJ) that results in the formation of indels, or homology directed repair (HDR), that, upon co-delivery of a repair template, allows the integration of a desired sequence alteration of interest with an increased efficiency.

The first approach developed involved the engineering of different **zinc-finger nucleases** (ZFNs) units recognizing each three different base pairs of DNA. These ZFNs units are then assembled to target the specific DNA sequence of interest and fused to the non-target specific nuclease FokI. To induce a DSB, FokI needs to dimerize, thus two ZFN molecule need to be delivered to the cells and bind the target site[355].

Later, a second type of engineered nuclease was developed and called **TALENs** (transcription activator-like effector nuclease). This followed the same design as ZFNs, comprising the nuclease FokI and a customizable sequence-specific DNA-binding domain[356]. The DNA-binding domain used here is a highly conserved repeat sequence, derived from a phytopathogenic bacterium, naturally secreted to alter gene transcription in the host plant cells and promote bacteria survival. TALENs can easily be engineered to target a desired sequence by changing two residues of each repeat, using a simple "protein-DNA code" where, for instance, ND mediates the binding to the nucleotide C, NI to A, NG to T and NN to A or G[357]. Despite the basically limitless targeting range, the main challenge in the use of TALENs is the complex assembly of highly repeated sequences.

As previously mentioned, the major breakthrough came with the development of the **CRISPR** (clustered regularly interspaced short palindromic repeat)**/Cas** system. CRISPR was first described in *E. coli* as an adaptive prokaryotic immune system protecting bacteria from exogenous DNA by inducing a DSB[358]. The CRISPR/Cas follows once again a two-component design, with an endonuclease targeted to a specific DNA locus by 20 nt of a chimeric single guide RNA (gRNA). In order for the Cas protein to bind, the 20 nucleotides must bind the target sequence via base pairing and must be followed, on the target genome, by a protospacer adjacent motif (PAM), a 2-6 nucleotide sequence that is specific for each Cas protein, widening the applications of this tool. For example, the *Streptococcus pyogenes* Cas9, the most widely used, recognizes the protospacer sequence NGG at the 3' of the target sequence and cuts 3 nucleotides upstream of it. CRISPR/Cas is the most versatile and easy-to-use tool as it requires only the synthesis of a 20nt gRNA. Moreover, a catalytically "dead" variant of Cas9 (dCas9) was engineered to be delivered on the target locus without cleaving the DNA, further broadening the applications of this tool[359]. For instance, dCas9 can be fused to fluorophores to visualize the region(s) of interest[360], to biotinylating enzymes to identify the chromatin binding factors present in the targeted locus[361], or to activator/repressor domains to modulate gene expression[298].

phenotype. Moreover, they could also demonstrate the suitability of iPSCs as drug screening platforms for the development of new therapies[346]. With the advent of genome editing technologies, it became increasingly clear that, to demonstrate that a specific mutation is the one responsible for the phenotype, there is the need to have a genetically corrected control in the same genetic background[347].

One of the greatest potentials of hESCs and iPSCs besides disease modelling is the use in cell replacement therapy. For this purpose, iPSCs and hESCs need to be maintained in feeder-free and in xeno-free culture conditions, as xeno-derived molecules, severely hamper their use in transplantations. Still, no transplantation therapy involving hESCs or hiPSCs has been approved, but several proof-of-principle studies demonstrated this will likely change, in the near future. The use of hESCs in transplantation has been proven beneficial to treat spinal cord injury[348] and diabetes[349] in rat and mouse, respectively. However, hESCs have a high risk of inducing immune rejection and the use of autologous (i.e., from the same individual) iPSCs can circumvent this risk. iPSCs have been suggested to be ideal for therapy as correcting patient mutation ameliorates the phenotype, for example in Fanconi anaemia cells[350].

All the studies reported here, show that to understand the pathophysiological mechanisms of disease it is essential to study the variants' effect in a disease-relevant cell type.

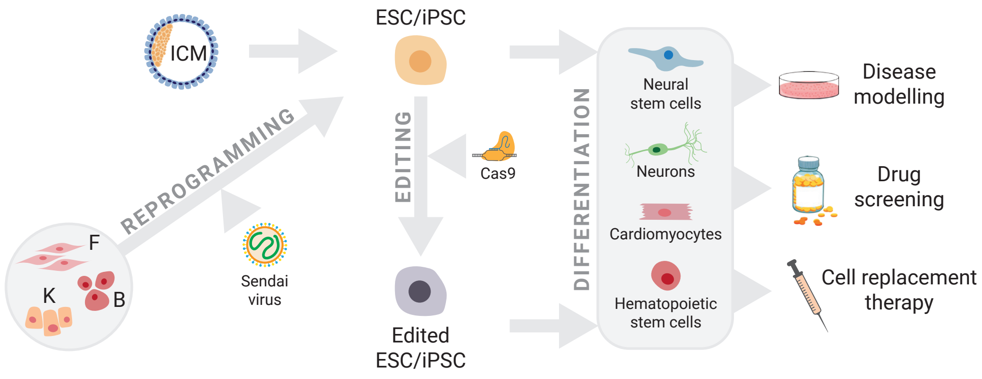


Figure 9 | Applications of pluripotent stem cells. ESC can be isolated from the ICM of human embryos obtained from IVF. Patient-specific iPSC can be obtined upon reprogramming of somatic cells, such a fibroblasts (F), peipheral blood cells (B), or keratinocytes (K), by transdution with viral vectors delivering Yamanaka factors. ESC and iPSC genome can be edited in order to introduce a patient mutation or to correct the patient mutation. Pluripotent cells can then be differentiated towards the cell type of interest and used for in vitro disease modelling, drug screening or cell transplantation therapies.

Neural induction

As mentioned in Part I of this introduction, all the cells of the central nervous system (with the exception of microglia) originate from the multipotent **neural stem cells** (NSCs). Thus, to study diseases affecting the CNS, it is important to first induce neural differentiation of pluripotent stem cells. Most of the published protocols for the generation of self-renewing NSCs rely on the generation of cell aggregates known as embryoid bodies (EBs) followed by monolayer cell culture[362,363]. However, knowledge gained via *in vivo* developmental studies[364-366], allowed the development of a simplified protocol for *in vitro* neural induction independently of labour-intensive embryoid bodies formation[367]. This protocol is based on the dual inhibition of the SMAD signalling pathway by blocking both TGFβ and BMP signalling with molecules such as SB431542 or A8301 and Noggin or Dorsomorphin, respectively. This differentiation protocol allows to obtain a large number of NSCs that can be further differentiated towards different mature and physiologically active neuronal subtype of interest such as glutamatergic neurons[368], GABAergic neurons[369], dopaminergic neurons, motor neurons[367], oligodendrocytes[370] and astrocytes[371]. However, the purity and differentiation potential of the obtained cells varies significantly from experiment to experiment.

Neural stem cells are self-renewing stem cells characterized by the expression of specific factors including NESTIN, SOX2 and PAX6. To maintain their multipotent potential in culture, NSC need to be grown in the presence of two growth factors: bFGF and EGF. Withdrawal of these factors leads to spontaneous differentiation towards neurons, astrocytes or oligodendrocytes. As pluripotent stem cells, to grow in culture NSC require to be plated onto matrix-coated vessels.

Conventional monolayer cultures of either NSCs or derived cell types have been an invaluable tool to explore developmental and disease-related processes, however one of their major limitations is the lack of complex interactions that characterize the human brain. As a consequence, co-culture systems and even three-dimensional culturing systems have recently been shown to better mimic the environment of the human brain and allow both cell-cell and cell-ECM communication.

Cerebral organoids are 3D aggregates of cells, derived from pluripotent stem cells embedded in a matrix droplet, that differentiate and self-organize to remarkably resemble the structure of the developing human brain, with ventricles, ventricular and sub-ventricular zones and a cortical plate[372]. The development of protocols to obtain these cerebral organoids built upon the knowledge gained via neural differentiation protocols in 2D, starting with neural induction via dual SMAD inhibition. Later, the organoid can be patterned by exposure to different factors. For instance, WNT and BMP act as dorsalizing factors, while SHH, antagonizing WNT and BMP is a ventralizing factor[373], retinoic acid promotes neurogenesis and migration[372], and BDNF and NT3 are needed to promote astrocyte formation[374]. Combining all this knowledge, cerebral organoids can be obtained that resemble various parts of the developing brain such as forebrain[375], hindbrain[376], retina[377], hippocampus[378] and even CSF-secreting choroid plexus[379], and they can even be fused to form even more complex structures[380,381]. Organoids have proven to be a valid model for several brain diseases such as microcephaly[372], lissencephaly[382], autism[383] but also Zika virus infection[375].

Taken together, the past decade witnessed an explosion of new technologies, starting from diagnostics, like whole genome sequencing, moving to molecular and bioinformatical methods to explore gene regulation and genome folding, and to biotechnological improvements of *in vitro* human disease modelling and genome editing. We are now in the position to combine all the knowledge gained though these various fields to shed light on the intricate gene regulation processes happening during development and disease of human neural tissue.

References

1 Harrington, M. J., Hong, E. & Brewster, R. Comparative analysis of neurulation: first impressions do not count. *Mol Reprod Dev* **76**, 954-965 (2009).

2 Sauer, F. C. Mitosis in the neural tube. *Journal of Comparative Neurology* **62**, 377-405 (1935).

3 Fernandez, V., Llinares-Benadero, C. & Borrell, V. Cerebral cortex expansion and folding: what have we learned? *Embo J* **35**, 1021-1044 (2016).

4 Sun, T. & Hevner, R. F. Growth and folding of the mammalian cerebral cortex: from molecules to malformations. *Nat Rev Neurosci* **15**, 217-232 (2014).

5 Bystron, I., Blakemore, C. & Rakic, P. Development of the human cerebral cortex: Boulder Committee revisited. *Nat Rev Neurosci* **9**, 110-122 (2008).

6 Rakic, P. Neurons in rhesus monkey visual cortex: systematic relation between time of origin and eventual disposition. *Science* **183**, 425-427 (1974).

7 Lim, L., Mi, D., Llorca, A. & Marín, O. Development and Functional Diversification of Cortical Interneurons. *Neuron* **100**, 294-313 (2018).

8 Kessaris, N. *et al.* Competing waves of oligodendrocytes in the forebrain and postnatal elimination of an embryonic lineage. *Nat Neurosci* **9**, 173-179 (2006).

9 Miller, F. D. & Gauthier, A. S. Timing is everything: making neurons versus glia in the developing cortex. *Neuron* **54**, 357-369 (2007).

10 Ginhoux, F. *et al.* Fate mapping analysis reveals that adult microglia derive from primitive macrophages. *Science* **330**, 841-845 (2010).

11 Monier, A., Evrard, P., Gressens, P. & Verney, C. Distribution and differentiation of microglia in the human encephalon during the first two trimesters of gestation. *J Comp Neurol* **499**, 565-582 (2006).

12 Graciarena, M., Seiffe, A., Nait-Oumesmar, B. & Depino, A. M. Hypomyelination and Oligodendroglial Alterations in a Mouse Model of Autism Spectrum Disorder. *Front Cell Neurosci* **12**, 517 (2018).

13 Malara, M. *et al.* SHANK3 deficiency leads to myelin defects in the central and peripheral nervous system. *Cell Mol Life Sci* **79**, 371 (2022).

14 Drenthen, G. S. *et al.* On the merits of non-invasive myelin imaging in epilepsy, a literature review. *J Neurosci Methods* **338**, 108687 (2020).

15 Knowles, J. K. *et al.* Maladaptive myelination promotes generalized epilepsy progression. *Nat Neurosci* **25**, 596-606 (2022).

16 Kuida, K. *et al.* Reduced apoptosis and cytochrome c-mediated caspase activation in mice lacking caspase 9. *Cell* **94**, 325-337 (1998).

17 Kuida, K. *et al.* Decreased apoptosis in the brain and premature lethality in CPP32-deficient mice. *Nature* **384**, 368-372 (1996).

18 Tang, G. *et al.* Loss of mTOR-dependent macroautophagy causes autistic-like synaptic pruning deficits. *Neuron* **83**, 1131-1143 (2014).

19 Sellgren, C. M. *et al.* Increased synapse elimination by microglia in schizophrenia patient-derived models of synaptic pruning. *Nat Neurosci* **22**, 374-385 (2019).

20 Parenti, I., Rabaneda, L. G., Schoen, H. & Novarino, G. Neurodevelopmental Disorders: From Genetics to Functional Pathways. *Trends Neurosci* **43**, 608-621 (2020).

21 Curcio, A. M., Shekhawat, P., Reynolds, A. S. & Thakur, K. T. Neurologic infections during pregnancy. *Handb Clin Neurol* **172**, 79-104 (2020).

22 van Loo, K. M. & Martens, G. J. Genetic and environmental factors in complex neurodevelopmental disorders. *Curr Genomics* **8**, 429-444 (2007).

23 Verkerk, A. J. *et al.* Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905-914 (1991).

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

25 Happ, H. C. & Carvill, G. L. A 2020 View on the Genetics of Developmental and Epileptic

Encephalopathies. *Epilepsy Curr* **20**, 90-96 (2020).

26 Depienne, C. *et al.* Spectrum of SCN1A gene mutations associated with Dravet syndrome: analysis of 333 patients. *J Med Genet* **46**, 183-191 (2009).

27 Veeramah, K. R. *et al.* De novo pathogenic SCN8A mutation identified by whole-genome sequencing of a family quartet affected by infantile epileptic encephalopathy and SUDEP. *Am J Hum Genet* **90**, 502-510 (2012).

28 Syrbe, S. *et al.* De novo loss- or gain-of-function mutations in KCNA2 cause epileptic encephalopathy. *Nat Genet* **47**, 393-399 (2015).

29 Calhoun, J. D., Vanoye, C. G., Kok, F., George, A. L., Jr. & Kearney, J. A. Characterization of a KCNB1 variant associated with autism, intellectual disability, and epilepsy. *Neurol Genet* **3**, e198 (2017).

30 Marini, C. *et al.* HCN1 mutation spectrum: from neonatal epileptic encephalopathy to benign generalized epilepsy and beyond. *Brain* **141**, 3160-3178 (2018).

31 Epi, K. C. De Novo Mutations in SLC1A2 and CACNA1A Are Important Causes of Epileptic Encephalopathies. *Am J Hum Genet* **99**, 287-298 (2016).

32 Helbig, K. L. *et al.* De Novo Pathogenic Variants in CACNA1E Cause Developmental and Epileptic Encephalopathy with Contractures, Macrocephaly, and Dyskinesias. *Am J Hum Genet* **104**, 562 (2019).

33 Maljevic, S. *et al.* Spectrum of GABAA receptor variants in epilepsy. *Curr Opin Neurol* **32**, 183-190 (2019).

34 Yoo, Y. *et al.* GABBR2 mutations determine phenotype in rett syndrome and epileptic encephalopathy. *Ann Neurol* **82**, 466-478 (2017).

35 Lemke, J. R. *et al.* GRIN2B mutations in West syndrome and intellectual disability with focal epilepsy. *Ann Neurol* **75**, 147-154 (2014).

36 Madeo, M. *et al.* Loss-of-Function Mutations in FRRS1L Lead to an Epileptic-Dyskinetic Encephalopathy. *Am J Hum Genet* **98**, 1249-1255 (2016).

37 Koch, H. & Weber, Y. G. The glucose transporter type 1 (Glut1) syndromes. *Epilepsy Behav* **91**, 90-93 (2019).

38 Chatron, N. *et al.* The epilepsy phenotypic spectrum associated with a recurrent CUX2 variant. *Ann Neurol* **83**, 926-934 (2018).

39 Bell, S. *et al.* Mutations in ACTL6B Cause Neurodevelopmental Deficits and Epilepsy and Lead to Loss of Dendrites in Human Neurons. *Am J Hum Genet* **104**, 815-834 (2019).

40 Cao, S. *et al.* Homozygous EEF1A2 mutation causes dilated cardiomyopathy, failure to thrive, global developmental delay, epilepsy and early death. *Hum Mol Genet* **26**, 3545-3552 (2017).

41 Colin, E. *et al.* Biallelic Variants in UBA5 Reveal that Disruption of the UFM1 Cascade Can Result in Early-Onset Encephalopathy. *Am J Hum Genet* **99**, 695-703 (2016).

42 Muona, M. *et al.* Biallelic Variants in UBA5 Link Dysfunctional UFM1 Ubiquitin-like Modifier Pathway to Severe Infantile-Onset Encephalopathy. *Am J Hum Genet* **99**, 683-694 (2016).

43 Vissers, L. E., Gilissen, C. & Veltman, J. A. Genetic studies in intellectual disability and related disorders. *Nat Rev Genet* **17**, 9-18 (2016).

44 Jeffrey, J. S. *et al.* Developmental and epileptic encephalopathy: Personal utility of a genetic diagnosis for families. *Epilepsia Open* **6**, 149-159 (2021).

45 Miller, D. T. *et al.* Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet* **86**, 749-764 (2010).

46 Yang, Y. *et al.* Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* **369**, 1502-1511 (2013).

47 de Wit, M. C. *et al.* Cortical brain malformations: effect of clinical, neuroradiological, and modern genetic classification. *Arch Neurol* **65**, 358-366 (2008).

48 Di Donato, N. *et al.* Analysis of 17 genes detects mutations in 81% of 811 patients with lissencephaly. *Genet Med* **20**, 1354-1364 (2018).

49 Wiszniewski, W. *et al.* Comprehensive genomic analysis of patients with disorders of cerebral cortical development. *Eur J Hum Genet* **26**, 1121-1131 (2018).

50 Gonzalez-Moron, D. *et al.* Germline and somatic mutations in cortical malformations: Molecular defects in Argentinean patients with neuronal migration disorders. *PLoS One* **12**, e0185103 (2017).

51 Jamuar, S. S. *et al.* Somatic mutations in cerebral cortical malformations. *N Engl J Med* **371**, 733-743 (2014).

52 Zillhardt, J. L. *et al.* Mosaic parental germline mutations causing recurrent forms of malformations of cortical development. *Eur J Hum Genet* **24**, 611-614 (2016).

53 Mirzaa, G. *et al.* PIK3CA-associated developmental disorders exhibit distinct classes of mutations with variable expression and tissue distribution. *JCI Insight* **1** (2016).

54 McMahon, K. Q. *et al.* Familial recurrences of FOXG1-related disorder: Evidence for mosaicism. *Am J Med Genet A* **167A**, 3096-3102 (2015).

55 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

56 Clark, M. M. *et al.* Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Sci Transl Med* **11** (2019).

57 Lionel, A. C. *et al.* Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genet Med* **20**, 435-443 (2018).

58 Stavropoulos, D. J. *et al.* Whole Genome Sequencing Expands Diagnostic Utility and Improves Clinical Management in Pediatric Medicine. *NPJ Genom Med* **1** (2016).

59 Scocchia, A. *et al.* Clinical whole genome sequencing as a first-tier test at a resource-limited dysmorphology clinic in Mexico. *NPJ Genom Med* **4**, 5 (2019).

60 Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195 (2012).

61 Smith, M. & Flodman, P. L. Expanded Insights Into Mechanisms of Gene Expression and Disease Related Disruptions. *Front Mol Biosci* **5**, 101 (2018).

62 Spielmann, M., Lupianez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat Rev Genet* **19**, 453-467 (2018).

63 Zeng, Y. *et al.* Aberrant gene expression in humans. *PLoS Genet* **11**, e1004942 (2015).

64 Short, P. J. *et al.* De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611-616 (2018).

65 Doan, R. N. *et al.* Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. *Cell* **167**, 341-354 e312 (2016).

66 Devanna, P., van de Vorst, M., Pfundt, R., Gilissen, C. & Vernes, S. C. Genome-wide investigation of an ID cohort reveals de novo 3'UTR variants affecting gene expression. *Hum Genet* **137**, 717-721 (2018).

67 Gloss, B. S. & Dinger, M. E. Realizing the significance of noncoding functionality in clinical genomics. *Exp Mol Med* **50**, 97 (2018).

68 Benko, S. *et al.* Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nat Genet* **41**, 359-364 (2009).

69 Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12**, 1725-1735 (2003).

70 Lettice, L. A. *et al.* Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc Natl Acad Sci U S A* **99**, 7548-7553 (2002).

71 Smemo, S. *et al.* Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum Mol Genet* **21**, 3255-3263 (2012).

72 Weedon, M. N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat Genet* **46**, 61-64 (2014).

73 Potuijt, J. W. P. *et al.* A point mutation in the pre-ZRS disrupts sonic hedgehog expression in the

limb bud and results in triphalangeal thumb-polysyndactyly syndrome. *Genet Med* **20**, 1405-1413 (2018).

74 Protas, M. E. *et al.* Mutations of conserved non-coding elements of PITX2 in patients with ocular dysgenesis and developmental glaucoma. *Hum Mol Genet* **26**, 3630-3638 (2017).

75 Ngcungcu, T. *et al.* Duplicated Enhancer Region Increases Expression of CTSB and Segregates with Keratolytic Winter Erythema in South African and Norwegian Families. *Am J Hum Genet* **100**, 737-750 (2017).

76 Mehrjouy, M. M. *et al.* Regulatory variants of FOXG1 in the context of its topological domain organisation. *Eur J Hum Genet* **26**, 186-196 (2018).

77 Bouman, A., van Haelst, M. & van Spaendonk, R. Blepharophimosis-ptosis-epicanthus inversus syndrome caused by a 54-kb microdeletion in a FOXL2 cis-regulatory element. *Clin Dysmorphol* **27**, 58-62 (2018).

78 Castel, S. E. *et al.* Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat Genet* **50**, 1327-1334 (2018).

79 Niemi, M. E. K. *et al.* Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature* **562**, 268-271 (2018).

80 Crick, F. H. On protein synthesis. *Symp Soc Exp Biol* **12**, 138-163 (1958).

81 Kuska, B. Should scientists scrap the notion of junk DNA? *J Natl Cancer Inst* **90**, 1032-1033 (1998).

82 Ohno, S. So much "junk" DNA in our genome. *Brookhaven Symp Biol* **23**, 366-370 (1972).

83 Girelli, G. *et al.* GPSeq reveals the radial organization of chromatin in the cell nucleus. *Nat Biotechnol* **38**, 1184-1193 (2020).

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

85 Bell, A. C., West, A. G. & Felsenfeld, G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98**, 387-396 (1999).

86 Dowen, J. M. *et al.* Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374-387 (2014).

87 Ong, C. T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet* **15**, 234-246 (2014).

88 Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat Rev Genet* **19**, 789-800 (2018).

89 Hnisz, D., Day, D. S. & Young, R. A. Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell* **167**, 1188-1200 (2016).

90 Weintraub, A. S. *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* **171**, 1573-1588 e1528 (2017).

91 Beagan, J. A. *et al.* YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res* **27**, 1139-1152 (2017).

92 Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299-308 (1981).

93 Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239-243 (2018).

94 Stadhouders, R. *et al.* Transcription regulation by distal enhancers: who's in the loop? *Transcription* **3**, 181-186 (2012).

95 Coulon, A., Chow, C. C., Singer, R. H. & Larson, D. R. Eukaryotic transcriptional dynamics: from single molecules to cell populations. *Nat Rev Genet* **14**, 572-584 (2013).

96 Cai, Y. *et al.* YY1 functions with INO80 to activate transcription. *Nat Struct Mol Biol* **14**, 872-874 (2007).

97 Brahma, S. *et al.* INO80 exchanges H2A.Z for H2A by translocating on DNA proximal to histone dimers. *Nat Commun* **8**, 15616 (2017).

98 Papamichos-Chronakis, M., Watanabe, S., Rando, O. J. & Peterson, C. L. Global regulation of H2A.Z localization by the INO80 chromatin-remodeling enzyme is essential for genome integrity. *Cell* **144**, 200-213 (2011).

99 Shen, X., Mizuguchi, G., Hamiche, A. & Wu, C. A chromatin remodelling complex involved in transcription and DNA processing. *Nature* **406**, 541-544 (2000).

100 Papamichos-Chronakis, M. & Peterson, C. L. The Ino80 chromatin-remodeling enzyme regulates replisome function and stability. *Nat Struct Mol Biol* **15**, 338-345 (2008).

101 Vincent, J. A., Kwong, T. J. & Tsukiyama, T. ATP-dependent chromatin remodeling shapes the DNA replication landscape. *Nat Struct Mol Biol* **15**, 477-484 (2008).

102 Eustermann, S. *et al.* Structural basis for ATP-dependent chromatin remodelling by the INO80 complex. *Nature* **556**, 386-390 (2018).

103 Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res* **21**, 381-395 (2011).

104 Hyun, K., Jeon, J., Park, K. & Kim, J. Writing, erasing and reading histone lysine methylations. *Exp Mol Med* **49**, e324 (2017).

105 Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650-665 (2018).

106 Funnell, A. P. & Crossley, M. Homo- and heterodimerization in transcriptional regulation. *Adv Exp Med Biol* **747**, 105-121 (2012).

107 Siersbaek, R. *et al.* Extensive chromatin remodelling and establishment of transcription factor 'hotspots' during early adipogenesis. *Embo J* **30**, 1459-1472 (2011).

108 Parker, S. C. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci U S A* **110**, 17921-17926 (2013).

109 Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319 (2013).

110 Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947 (2013).

111 Loven, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320-334 (2013).

112 Whyte, W. A. *et al.* Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. *Nature* **482**, 221-225 (2012).

113 Young, R. A. Control of the embryonic stem cell state. *Cell* **144**, 940-954 (2011).

114 Hay, D. *et al.* Genetic dissection of the alpha-globin super-enhancer in vivo. *Nat Genet* **48**, 895-903 (2016).

115 Shin, H. Y. *et al.* Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nat Genet* **48**, 904-911 (2016).

116 Park, K. & Atchison, M. L. Isolation of a candidate repressor/activator, NF-E1 (YY-1, delta), that binds to the immunoglobulin kappa 3' enhancer and the immunoglobulin heavy-chain mu E1 site. *Proc Natl Acad Sci U S A* **88**, 9804-9808 (1991).

117 Hariharan, N., Kelley, D. E. & Perry, R. P. Delta, a transcription factor that binds to downstream elements in several polymerase II promoters, is a functionally versatile zinc finger protein. *Proc Natl Acad Sci U S A* **88**, 9799-9803 (1991).

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

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

120 Hyde-DeRuyscher, R. P., Jennings, E. & Shenk, T. DNA binding sites for the transcriptional activator/repressor YY1. *Nucleic Acids Res* **23**, 4457-4465 (1995).

121 Yant, S. R. *et al.* High affinity YY1 binding motifs: identification of two core types (ACAT and CCAT) and distribution of potential binding sites within the human beta globin cluster. *Nucleic Acids Res* **23**, 4353-4362 (1995).

122 Gordon, S., Akopyan, G., Garban, H. & Bonavida, B. Transcription factor YY1: structure, function, and therapeutic implications in cancer biology. *Oncogene* **25**, 1125-1142 (2006).

123 Wilkinson, F. H., Park, K. & Atchison, M. L. Polycomb recruitment to DNA in vivo by the YY1 REPO domain. *Proc Natl Acad Sci U S A* **103**, 19296-19301 (2006).

124 Wilkinson, F., Pratt, H. & Atchison, M. L. PcG recruitment by the YY1 REPO domain can be mediated by Yaf2. *J Cell Biochem* **109**, 478-486 (2010).

125 Pan, X. *et al.* YY1 controls Igkappa repertoire and B-cell development, and localizes with condensin on the Igkappa locus. *EMBO J* **32**, 1168-1182 (2013).

126 Salichs, E., Ledda, A., Mularoni, L., Alba, M. M. & de la Luna, S. Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment. *PLoS Genet* **5**, e1000397 (2009).

127 Galganski, L., Urbanek, M. O. & Krzyzosiak, W. J. Nuclear speckles: molecular organization, biological function and role in disease. *Nucleic Acids Research* **45**, 10350-10368 (2017).

128 Rambout, X., Dequiedt, F. & Maquat, L. E. Beyond Transcription: Roles of Transcription Factors in Pre-mRNA Splicing. *Chem Rev* **118**, 4339-4364 (2018).

129 Bianchi, M. *et al.* Yin Yang 1 intronic binding sequences and splicing elicit intron-mediated enhancement of ubiquitin C gene expression. *PLoS One* **8**, e65932 (2013).

130 Sigova, A. A. *et al.* Transcription factor trapping by RNA in gene regulatory elements. *Science* **350**, 978-981 (2015).

131 Nabais Sá, M. J., Gabriele, M., Testa, G. & de Vries, B. B. A. Gabriele-de Vries Syndrome. (1993).

132 Morales-Rosado, J. A., Kaiwar, C., Smith, B. E., Klee, E. W. & Dhamija, R. A case of YY1-associated syndromic learning disability or Gabriele-de Vries syndrome with myasthenia gravis. *Am J Med Genet A* **176**, 2846-2849 (2018).

133 Deng, Z., Cao, P., Wan, M. M. & Sui, G. Yin Yang 1: a multifaceted protein beyond a transcription factor. *Transcription* **1**, 81-84 (2010).

134 Nicholson, S., Whitehouse, H., Naidoo, K. & Byers, R. J. Yin Yang 1 in human cancer. *Crit Rev Oncog* **16**, 245-260 (2011).

135 Gabriele, M. *et al.* YY1 Haploinsufficiency Causes an Intellectual Disability Syndrome Featuring Transcriptional and Chromatin Dysfunction. *Am J Hum Genet* **100**, 907-925 (2017).

136 Zurkirchen, L. *et al.* Yin Yang 1 sustains biosynthetic demands during brain development in a stage-specific manner. *Nat Commun* **10**, 2192 (2019).

137 Donohoe, M. E. *et al.* Targeted disruption of mouse Yin Yang 1 transcription factor results in peri-implantation lethality. *Mol Cell Biol* **19**, 7237-7244 (1999).

138 Yao, Y. L., Yang, W. M. & Seto, E. Regulation of transcription factor YY1 by acetylation and deacetylation. *Mol Cell Biol* **21**, 5979-5991 (2001).

139 Takasaki, N., Kurokawa, D., Nakayama, R., Nakayama, J. & Aizawa, S. Acetylated YY1 regulates Otx2 expression in anterior neuroectoderm at two cis-sites 90 kb apart. *EMBO J* **26**, 1649-1659 (2007).

140 Seto, E., Lewis, B. & Shenk, T. Interaction between transcription factors Sp1 and YY1. *Nature* **365**, 462-464 (1993).

141 Shrivastava, A. *et al.* Inhibition of transcriptional regulator Yin-Yang-1 by association with c-Myc. *Science* **262**, 1889-1892 (1993).

142 Wang, C. Y. *et al.* YY1AP, a novel co-activator of YY1. *J Biol Chem* **279**, 17750-17755 (2004).

143 Ohtomo, T., Horii, T., Nomizu, M., Suga, T. & Yamada, J. Molecular cloning of a structural homolog of YY1AP, a coactivator of the multifunctional transcription factor YY1. *Amino Acids* **33**, 645-652 (2007).

144 Kurisaki, K. *et al.* Nuclear factor YY1 inhibits transforming growth factor beta- and bone morphogenetic protein-induced cell differentiation. *Mol Cell Biol* **23**, 4494-4510 (2003).

145 Raught, B., Khursheed, B., Kazansky, A. & Rosen, J. YY1 represses beta-casein gene expression by preventing the formation of a lactation-associated complex. *Mol Cell Biol* **14**, 1752-1763 (1994).

146 Shi, Y., Lee, J. S. & Galvin, K. M. Everything you have ever wanted to know about Yin Yang 1. *Biochim Biophys Acta* **1332**, F49-66 (1997).

147 Wu, T. & Donohoe, M. E. Yy1 regulates Senp1 contributing to AMPA receptor GluR1 expression following neuronal depolarization. *J Biomed Sci* **26**, 79 (2019).

148 Atchison, L., Ghias, A., Wilkinson, F., Bonini, N. & Atchison, M. L. Transcription factor YY1 functions as a PcG protein in vivo. *EMBO J* **22**, 1347-1358 (2003).

149 Yang, W. M., Yao, Y. L., Sun, J. M., Davie, J. R. & Seto, E. Isolation and characterization of cDNAs corresponding to an additional member of the human histone deacetylase gene family. *J Biol Chem* **272**, 28001-28007 (1997).

150 Lee, J. S. *et al.* Relief of YY1 transcriptional repression by adenovirus E1A is mediated by E1A-associated protein p300. *Genes Dev* **9**, 1188-1198 (1995).

151 Austen, M., Luscher, B. & Luscher-Firzlaff, J. M. Characterization of the transcriptional regulator YY1. The bipartite transactivation domain is independent of interaction with the TATA box-binding protein, transcription factor IIB, TAFII55, or cAMP-responsive element-binding protein (CPB)-binding protein. *J Biol Chem* **272**, 1709-1717 (1997).

152 Rezai-Zadeh, N. *et al.* Targeted recruitment of a histone H4-specific methyltransferase by the transcription factor YY1. *Genes Dev* **17**, 1019-1029 (2003).

153 Wu, S. *et al.* A YY1-INO80 complex regulates genomic stability through homologous recombination-based repair. *Nat Struct Mol Biol* **14**, 1165-1172 (2007).

154 Wang, J. *et al.* YY1 Positively Regulates Transcription by Targeting Promoters and Super-Enhancers through the BAF Complex in Embryonic Stem Cells. *Stem Cell Reports* **10**, 1324-1339 (2018).

155 Varum, S. *et al.* Yin Yang 1 Orchestrates a Metabolic Program Required for Both Neural Crest Development and Melanoma Formation. *Cell Stem Cell* **24**, 637-653 e639 (2019).

156 Lee, G. R. Role of YY1 in long-range chromosomal interactions regulating Th2 cytokine expression. *Transcription* **5**, e27976 (2014).

157 Schwalie, P. C. *et al.* Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. *Genome Biol* **14**, R148 (2013).

158 Kim, Y., Shi, Z., Zhang, H., Finkelstein, I. J. & Yu, H. Human cohesin compacts DNA by loop extrusion. *Science* **366**, 1345-1349 (2019).

159 Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320 e324 (2017).

160 Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci U S A* **111**, 996-1001 (2014).

161 Merkenschlager, M. & Nora, E. P. CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. *Annu Rev Genomics Hum Genet* **17**, 17-43 (2016).

162 Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137**, 1194-1211 (2009).

163 López-Perrote, A. *et al.* Structure of Yin Yang 1 oligomers that cooperate with RuvBL1-RuvBL2 ATPases. *The Journal of biological chemistry* **289**, 22614-22629 (2014).

164 Maurano, M. T. *et al.* Role of DNA Methylation in Modulating Transcription Factor Occupancy. *Cell Rep* **12**, 1184-1195 (2015).

165 Wang, H. *et al.* Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome research* **22**, 1680-1688 (2012).

166 Hahn, S. Phase Separation, Protein Disorder, and Enhancer Function. *Cell* **175**, 1723-1725 (2018).

167 Boija, A. *et al.* Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* **175**, 1842-1855 e1816 (2018).

168 Li, C. H. *et al.* MeCP2 links heterochromatin condensates and neurodevelopmental disease. *Nature* **586**, 440-444 (2020).

169 Dong, X. & Kwan, K. M. Yin Yang 1 is critical for mid-hindbrain neuroepithelium development and involved in cerebellar agenesis. *Mol Brain* **13**, 104 (2020).

170 Chen, Z. S. *et al.* Planar cell polarity gene Fuz triggers apoptosis in neurodegenerative disease models. *EMBO Rep* **19** (2018).

171 Sui, G. *et al.* Yin Yang 1 is a negative regulator of p53. *Cell* **117**, 859-872 (2004).

172 Gray, R. S. *et al.* The planar cell polarity effector Fuz is essential for targeted membrane trafficking, ciliogenesis and mouse embryonic development. *Nat Cell Biol* **11**, 1225-1232 (2009).

173 Seo, J. H. *et al.* Mutations in the planar cell polarity gene, Fuzzy, are associated with neural tube defects in humans. *Hum Mol Genet* **24**, 3893 (2015).

174 Tabler, J. M. *et al.* Fuz mutant mice reveal shared mechanisms between ciliopathies and FGF-related syndromes. *Dev Cell* **25**, 623-635 (2013).

175 Knauss, J. L. *et al.* Long noncoding RNA Sox2ot and transcription factor YY1 co-regulate the differentiation of cortical neural progenitors by repressing Sox2. *Cell Death Dis* **9**, 799 (2018).

176 Graham, V., Khudyakov, J., Ellis, P. & Pevny, L. SOX2 functions to maintain neural progenitor identity. *Neuron* **39**, 749-765 (2003).

177 Hebbes, T. R., Thorne, A. W. & Crane-Robinson, C. A direct link between core histone acetylation and transcriptionally active chromatin. *EMBO J* **7**, 1395-1402 (1988).

178 Gabriele, M., Lopez Tobon, A., D'Agostino, G. & Testa, G. The chromatin basis of neurodevelopmental disorders: Rethinking dysfunction along the molecular and temporal axes. *Prog Neuropsychopharmacol Biol Psychiatry* **84**, 306-327 (2018).

179 Martinez, A. F. *et al.* An Ultraconserved Brain-Specific Enhancer Within ADGRL3 (LPHN3) Underpins Attention-Deficit/Hyperactivity Disorder Susceptibility. *Biol Psychiatry* **80**, 943-954 (2016).

180 Liu, W., Guo, Q. & Zhao, H. Oxidative stress-elicited YY1 potentiates antioxidative response via enhancement of NRF2-driven transcriptional activity: A potential neuronal defensive mechanism against ischemia/reperfusion cerebral injury. *Biomed Pharmacother* **108**, 698-706 (2018).

181 Pal, R., Tiwari, P. C., Nath, R. & Pant, K. K. Role of neuroinflammation and latent transcription factors in pathogenesis of Parkinson's disease. *Neurol Res* **38**, 1111-1122 (2016).

182 Qiao, H. & May, J. M. Interaction of the transcription start site core region and transcription factor YY1 determine ascorbate transporter SVCT2 exon 1a promoter activity. *PLoS One* **7**, e35746 (2012).

183 Gureev, A. P. & Popov, V. N. Nrf2/ARE Pathway as a Therapeutic Target for the Treatment of Parkinson Diseases. *Neurochem Res* (2019).

184 Aubry, S. *et al.* Assembly and interrogation of Alzheimer's disease genetic networks reveal novel regulators of progression. *PLoS One* **10**, e0120352 (2015).

185 Khachigian, L. M. The Yin and Yang of YY1 in tumor growth and suppression. *Int J Cancer* **143**, 460-465 (2018).

186 Sarvagalla, S., Kolapalli, S. P. & Vallabhapurapu, S. The Two Sides of YY1 in Cancer: A Friend and a Foe. *Front Oncol* **9**, 1230 (2019).

187 Verheul, T. C. J., van Hijfte, L., Perenthaler, E. & Barakat, T. S. The Why of YY1: Mechanisms of Transcriptional Regulation by Yin Yang 1. *Front Cell Dev Biol* **8**, 592164 (2020).

188 De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215 (2014).

189 Gregor, A. *et al.* De novo mutations in the genome organizer CTCF cause intellectual disability. *Am J Hum Genet* **93**, 124-131 (2013).

190 Barish, S. *et al.* BICRA, a SWI/SNF Complex Member, Is Associated with BAF-Disorder Related Phenotypes in Humans and Model Organisms. *Am J Hum Genet* **107**, 1096-1112 (2020).

191 Weerts, M. J. A. *et al.* Delineating the molecular and phenotypic spectrum of the SETD1B-related syndrome.

192 O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-250 (2012).

193 Gurnett, C. A. *et al.* Two novel point mutations in the long-range SHH enhancer in three families with triphalangeal thumb and preaxial polydactyly. *Am J Med Genet A* **143A**, 27-32 (2007).

194 Klopocki, E. *et al.* A microduplication of the long range SHH limb regulator (ZRS) is associated with triphalangeal thumb-polysyndactyly syndrome. *J Med Genet* **45**, 370-375 (2008).

195 Laurell, T. *et al.* A novel 13 base pair insertion in the sonic hedgehog ZRS limb enhancer (ZRS/LMBR1) causes preaxial polydactyly with triphalangeal thumb. *Hum Mutat* **33**, 1063-1066 (2012).

196 Jeong, Y. *et al.* Regulation of a remote Shh forebrain enhancer by the Six3 homeoprotein. *Nat Genet* **40**, 1348-1353 (2008).

197 Wallis, D. E. *et al.* Mutations in the homeodomain of the human SIX3 gene cause holoprosencephaly. *Nat Genet* **22**, 196-198 (1999).

198 Bhatia, S. *et al.* Disruption of autoregulatory feedback by a mutation in a remote, ultraconserved PAX6 enhancer causes aniridia. *Am J Hum Genet* **93**, 1126-1134 (2013).

199 Bae, B. I. *et al.* Evolutionarily dynamic alternative splicing of GPR56 regulates regional cerebral cortical patterning. *Science* **343**, 764-768 (2014).

200 Soldner, F. *et al.* Parkinson-associated risk variant in distal enhancer of alpha-synuclein modulates target gene expression. *Nature* **533**, 95-99 (2016).

201 Vacic, V. *et al.* Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. *Nature* **471**, 499-503 (2011).

202 Piluso, G. *et al.* Assessment of de novo copy-number variations in Italian patients with schizophrenia: Detection of putative mutations involving regulatory enhancer elements. *World J Biol Psychiatry* **20**, 126-136 (2019).

203 Cellini, E. *et al.* Multiple genomic copy number variants associated with periventricular nodular heterotopia indicate extreme genetic heterogeneity. *Eur J Hum Genet* **27**, 909-918 (2019).

204 Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat Rev Genet* **7**, 85-97 (2006).

205 Giorgio, E. *et al.* A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD). *Hum Mol Genet* **24**, 3143-3154 (2015).

206 Nmezi, B. *et al.* Genomic deletions upstream of lamin B1 lead to atypical autosomal dominant leukodystrophy. *Neurol Genet* **5**, e305 (2019).

207 Brandler, W. M. *et al.* Paternally inherited cis-regulatory structural variants are associated with autism. *Science* **360**, 327-331 (2018).

208 Turner, T. N. *et al.* Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* **171**, 710-722 e712 (2017).

209 Ishibashi, M. *et al.* Copy number variants in patients with intellectual disability affect the regulation of ARX transcription factor gene. *Hum Genet* **134**, 1163-1182 (2015).

210 Monlong, J. *et al.* Global characterization of copy number variants in epilepsy patients from whole genome sequencing. *PLoS Genet* **14**, e1007285 (2018).

211 Hama, Y. *et al.* Genomic copy number variation analysis in multiple system atrophy. *Mol Brain* **10**, 54 (2017).

212 Liu, Y. C. *et al.* Evaluation of non-coding variation in GLUT1 deficiency. *Dev Med Child Neurol* **58**, 1295-1302 (2016).

213 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050 (2005).

214 Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* **35**, D88-92 (2007).

215 May, D. *et al.* Large-scale discovery of enhancers from human heart tissue. *Nat Genet* **44**, 89-93 (2011).

216 Moreau, P. *et al.* The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Res* **9**, 6047-6068 (1981).

217 Catarino, R. R. & Stark, A. Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes Dev* **32**, 202-223 (2018).

218 Solomon, M. J., Larsen, P. L. & Varshavsky, A. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* **53**, 937-947 (1988).

219 Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497-1502 (2007).

220 Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**, 651-657 (2007).

221 Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-318 (2007).

222 Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936 (2010).

223 Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-858 (2009).

224 Cotney, J. *et al.* Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. *Genome Res* **22**, 1069-1080 (2012).

225 Amiri, A. *et al.* Transcriptome and epigenome landscape of human cortical development modeled in organoids. *Science* **362** (2018).

226 Reilly, S. K. *et al.* Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**, 1155-1159 (2015).

227 Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362** (2018).

228 Vermunt, M. W. *et al.* Large-scale identification of coregulated enhancer networks in the adult human brain. *Cell Rep* **9**, 767-779 (2014).

229 Sun, W. *et al.* Histone Acetylome-wide Association Study of Autism Spectrum Disorder. *Cell* **167**, 1385-1397 e1311 (2016).

230 Vermunt, M. W. *et al.* Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. *Nat Neurosci* **19**, 494-503 (2016).

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

232 Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun* **10**, 1930 (2019).

233 Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311-322 (2008).

234 Lidor Nili, E. *et al.* p53 binds preferentially to genomic regions with high DNA-encoded nucleosome occupancy. *Genome Res* **20**, 1361-1368 (2010).

235 John, S. *et al.* Genome-scale mapping of DNase I hypersensitivity. *Curr Protoc Mol Biol* **Chapter 27**, Unit 21 27 (2013).

236 Giresi, P. G. & Lieb, J. D. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). *Methods* **48**, 233-239 (2009).

237 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218 (2013).

238 Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074-1077 (2013).

239 de la Torre-Ubieta, L. *et al.* The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. *Cell* **172**, 289-304 e218 (2018).

240 Collis, P., Antoniou, M. & Grosveld, F. Definition of the minimal requirements within the human beta-globin gene and the dominant control region for high level expression. *Embo J* **9**, 233-240 (1990).

241 Tuan, D., Kong, S. & Hu, K. Transcription of the hypersensitive site HS2 enhancer in erythroid cells. *Proc Natl Acad Sci U S A* **89**, 11219-11223 (1992).

242 Ashe, H. L., Monks, J., Wijgerde, M., Fraser, P. & Proudfoot, N. J. Intergenic transcription and transinduction of the human beta-globin locus. *Genes Dev* **11**, 2494-2509 (1997).

243 Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182-187 (2010).

244 Hah, N., Murakami, S., Nagari, A., Danko, C. G. & Kraus, W. L. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res* **23**, 1210-1223 (2013).

245 Wang, D. *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474**, 390-394 (2011).

246 Kaikkonen, M. U. *et al.* Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Mol Cell* **51**, 310-325 (2013).

247 Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461 (2014).

248 Koch, F. *et al.* Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat Struct Mol Biol* **18**, 956-963 (2011).

249 Lam, M. T., Li, W., Rosenfeld, M. G. & Glass, C. K. Enhancer RNAs and regulated transcriptional programs. *Trends Biochem Sci* **39**, 170-182 (2014).

250 Natoli, G. & Andrau, J. C. Noncoding transcription at enhancers: general principles and functional models. *Annu Rev Genet* **46**, 1-19 (2012).

251 Barakat, T. S. & Gribnau, J. X chromosome inactivation and embryonic stem cells. *Adv Exp Med Biol* **695**, 132-154 (2010).

252 Yao, P. *et al.* Coexpression networks identify brain region-specific enhancer RNAs in the human brain. *Nat Neurosci* **18**, 1168-1174 (2015).

253 Davies, J. O., Oudelaar, A. M., Higgs, D. R. & Hughes, J. R. How best to identify chromosomal interactions: a comparison of approaches. *Nat Methods* **14**, 125-134 (2017).

254 de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev* **26**, 11-24 (2012).

255 Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-1311 (2002).

256 Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* **38**, 1341-1347 (2006).

257 Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* **38**, 1348-1354 (2006).

258 Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* **16**, 1299-1309 (2006).

259 Kolovos, P. *et al.* Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. *Epigenetics Chromatin* **7**, 10 (2014).

260 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).

261 Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523-527 (2016).

262 Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* **13**, 919-922 (2016).

263 Fang, R. *et al.* Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res* **26**, 1345-1348 (2016).

264 Nott, A. *et al.* Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. *Science* **366**, 1134-1139 (2019).

265 Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res* **24**, 1595-1602 (2014).

266 Halfon, M. S. Studying Transcriptional Enhancers: The Founder Fallacy, Validation Creep, and Other Biases. *Trends Genet* **35**, 93-103 (2019).

267 Pradeepa, M. M. *et al.* Histone H3 globular domain acetylation identifies a new class of enhancers. *Nat Genet* **48**, 681-686 (2016).

268 Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**, 265-270 (2012).

269 Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**, 271-277 (2012).

270 Dickel, D. E. *et al.* Function-based identification of mammalian enhancers using site-specific

integration. *Nat Methods* **11**, 566-571 (2014).

271 Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* **23**, 800-811 (2013).

272 Murtha, M. *et al.* FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. *Nat Methods* **11**, 559-565 (2014).

273 Ernst, J. *et al.* Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat Biotechnol* **34**, 1180-1190 (2016).

274 Arnold, C. D. *et al.* Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nat Biotechnol* **35**, 136-144 (2017).

275 van Arensbergen, J. *et al.* Genome-wide mapping of autonomous promoter activity in human cells. *Nat Biotechnol* **35**, 145-153 (2017).

276 Wang, X. *et al.* High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat Commun* **9**, 5380 (2018).

277 Shen, S. Q. *et al.* Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Res* **26**, 238-255 (2016).

278 Vanhille, L. *et al.* High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nat Commun* **6**, 6905 (2015).

279 Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res* **27**, 38-52 (2017).

280 Maricque, B. B., Chaudhari, H. G. & Cohen, B. A. A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. *Nat Biotechnol* (2018).

281 Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**, 272-286 (2014).

282 Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat Biotechnol* **34**, 192-198 (2016).

283 Sanjana, N. E. *et al.* High-resolution interrogation of functional elements in the noncoding genome. *Science* **353**, 1545-1549 (2016).

284 Han, R. *et al.* Functional CRISPR screen identifies AP1-associated enhancer regulating FOXF1 to modulate oncogene-induced senescence. *Genome Biol* **19**, 118 (2018).

285 Gasperini, M. *et al.* CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. *Am J Hum Genet* **101**, 192-205 (2017).

286 Diao, Y. *et al.* A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat Methods* **14**, 629-635 (2017).

287 Diao, Y. *et al.* A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. *Genome Res* **26**, 397-405 (2016).

288 Rajagopal, N. *et al.* High-throughput mapping of regulatory DNA. *Nat Biotechnol* **34**, 167-174 (2016).

289 Sen, D. R. *et al.* The epigenetic landscape of T cell exhaustion. *Science* **354**, 1165-1169 (2016).

290 Canver, M. C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192-197 (2015).

291 Canver, M. C. *et al.* Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. *Nat Genet* **49**, 625-634 (2017).

292 Maeder, M. L. *et al.* CRISPR RNA-guided activation of endogenous human genes. *Nat Methods* **10**, 977-979 (2013).

293 Mali, P. *et al.* CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol* **31**, 833-838 (2013).

294 Konermann, S. *et al.* Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583-588 (2015).

295 Liu, X. S. *et al.* Editing DNA Methylation in the Mammalian Genome. *Cell* **167**, 233-247 e217 (2016).

296 Hilton, I. B. *et al.* Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat Biotechnol* **33**, 510-517 (2015).

297 Thakore, P. I. *et al.* Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat Methods* **12**, 1143-1149 (2015).

298 Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442-451 (2013).

299 Konermann, S. *et al.* Optical control of mammalian endogenous transcription and epigenetic states. *Nature* **500**, 472-476 (2013).

300 Vojta, A. *et al.* Repurposing the CRISPR-Cas9 system for targeted DNA methylation. *Nucleic Acids Res* **44**, 5615-5628 (2016).

301 Kwon, D. Y., Zhao, Y. T., Lamonica, J. M. & Zhou, Z. Locus-specific histone deacetylation using a synthetic CRISPR-Cas9-based HDAC. *Nat Commun* **8**, 15315 (2017).

302 Kearns, N. A. *et al.* Functional annotation of native enhancers with a Cas9-histone demethylase fusion. *Nat Methods* **12**, 401-403 (2015).

303 Fulco, C. P. *et al.* Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769-773 (2016).

304 Xie, S., Duan, J., Li, B., Zhou, P. & Hon, G. C. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Mol Cell* **66**, 285-299 e285 (2017).

305 Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**, 377-390 e319 (2019).

306 Carleton, J. B., Berrett, K. C. & Gertz, J. Multiplex Enhancer Interference Reveals Collaborative Control of Gene Regulation by Estrogen Receptor alpha-Bound Enhancers. *Cell Syst* **5**, 333-344 e335 (2017).

307 Simeonov, D. R. *et al.* Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature* **549**, 111-115 (2017).

308 Klann, T. S. *et al.* CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat Biotechnol* **35**, 561-568 (2017).

309 Xu, J. *et al.* Dopamine-dependent neurotoxicity of alpha-synuclein: a mechanism for selective neurodegeneration in Parkinson disease. *Nat Med* **8**, 600-606 (2002).

310 Park, T. I. *et al.* Isolation and culture of functional adult human neurons from neurosurgical brain specimens. *Brain Commun* **2**, fcaa171 (2020).

311 Ricceri, L., De Filippis, B. & Laviola, G. Mouse models of Rett syndrome: from behavioural phenotyping to preclinical evaluation of new therapeutic approaches. *Behav Pharmacol* **19**, 501-517 (2008).

312 Rotaru, D. C., Mientjes, E. J. & Elgersma, Y. Angelman Syndrome: From Mouse Models to Therapy. *Neuroscience* **445**, 172-189 (2020).

313 Aida, T. & Feng, G. The dawn of non-human primate models for neurodevelopmental disorders. *Curr Opin Genet Dev* **65**, 160-168 (2020).

314 Davis, E. E., Frangakis, S. & Katsanis, N. Interpreting human genetic variation with in vivo zebrafish assays. *Biochim Biophys Acta* **1842**, 1960-1970 (2014).

315 Şentürk, M. & Bellen, H. J. Genetic strategies to tackle neurological diseases in fruit flies. *Curr Opin Neurobiol* **50**, 24-32 (2018).

316 Meshalkina, D. A. *et al.* Zebrafish models of autism spectrum disorder. *Exp Neurol* **299**, 207-216 (2018).

317 Bellosta, P. & Soldano, A. Dissecting the Genetics of Autism Spectrum Disorders: A Drosophila Perspective. *Front Physiol* **10**, 987 (2019).

318 Kim, H. T. *et al.* The microcephaly gene aspm is involved in brain development in zebrafish. *Biochem Biophys Res Commun* **409**, 640-644 (2011).

319 Evans, M. J. & Kaufman, M. H. Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292**, 154-156 (1981).

320 Cherny, R. A. *et al.* Strategies for the isolation and characterization of bovine embryonic stem cells.

Reprod Fertil Dev **6**, 569-575 (1994).

321 Li, M. *et al.* Isolation and culture of embryonic stem cells from porcine blastocysts. *Mol Reprod Dev* **65**, 429-434 (2003).

322 Thomson, J. A. *et al.* Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145-1147 (1998).

323 Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947-956 (2005).

324 Greber, B., Lehrach, H. & Adjaye, J. Fibroblast growth factor 2 modulates transforming growth factor beta signaling in mouse embryonic fibroblasts and human ESCs (hESCs) to support hESC self-renewal. *Stem Cells* **25**, 455-464 (2007).

325 Buecker, C. *et al.* A murine ESC-like state facilitates transgenesis and homologous recombination in human pluripotent stem cells. *Cell Stem Cell* **6**, 535-546 (2010).

326 Hanna, J. *et al.* Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs. *Proc Natl Acad Sci U S A* **107**, 9222-9227 (2010).

327 Gafni, O. *et al.* Derivation of novel human ground state naive pluripotent stem cells. *Nature* **504**, 282-286 (2013).

328 Theunissen, T. W. *et al.* Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. *Cell Stem Cell* **15**, 471-487 (2014).

329 Watanabe, K. *et al.* A ROCK inhibitor permits survival of dissociated human embryonic stem cells. *Nat Biotechnol* **25**, 681-686 (2007).

330 Guo, G. *et al.* Human naive epiblast cells possess unrestricted lineage potential. *Cell Stem Cell* **28**, 1040-1056 e1046 (2021).

331 Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861-872 (2007).

332 Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917-1920 (2007).

333 Staerk, J. *et al.* Reprogramming of human peripheral blood cells to induced pluripotent stem cells. *Cell Stem Cell* **7**, 20-24 (2010).

334 Loh, Y. H. *et al.* Reprogramming of T cells from human peripheral blood. *Cell Stem Cell* **7**, 15-19 (2010).

335 Aasen, T. *et al.* Efficient and rapid generation of induced pluripotent stem cells from human keratinocytes. *Nat Biotechnol* **26**, 1276-1284 (2008).

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

337 Wernig, M. *et al.* A drug-inducible transgenic system for direct reprogramming of multiple somatic cell types. *Nat Biotechnol* **26**, 916-924 (2008).

338 Fusaki, N., Ban, H., Nishiyama, A., Saeki, K. & Hasegawa, M. 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 (2009).

339 Ban, H. *et al.* Efficient generation of transgene-free human induced pluripotent stem cells (iPSCs) by temperature-sensitive Sendai virus vectors. *Proc Natl Acad Sci U S A* **108**, 14234-14239 (2011).

340 Warren, L. *et al.* Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. *Cell Stem Cell* **7**, 618-630 (2010).

341 Kim, D. *et al.* Generation of human induced pluripotent stem cells by direct delivery of reprogramming proteins. *Cell Stem Cell* **4**, 472-476 (2009).

342 Niclis, J. C. *et al.* Characterization of forebrain neurons derived from late-onset Huntington's disease human embryonic stem cell lines. *Front Cell Neurosci* **7**, 37 (2013).

343 Eiges, R. *et al.* Developmental study of fragile X syndrome using human embryonic stem cells derived from preimplantation genetically diagnosed embryos. *Cell Stem Cell* **1**, 568-577 (2007).

344 Park, I. H. *et al.* Disease-specific induced pluripotent stem cells. *Cell* **134**, 877-886 (2008).

345 Dimos, J. T. *et al.* Induced pluripotent stem cells generated from patients with ALS can be differentiated into motor neurons. *Science* **321**, 1218-1221 (2008).

346 Ebert, A. D. *et al.* Induced pluripotent stem cells from a spinal muscular atrophy patient. *Nature* **457**, 277-280 (2009).

347 Reinhardt, P. *et al.* Genetic correction of a LRRK2 mutation in human iPSCs links parkinsonian neurodegeneration to ERK-dependent changes in gene expression. *Cell Stem Cell* **12**, 354-367 (2013).

348 Kerr, C. L. *et al.* Efficient differentiation of human embryonic stem cells into oligodendrocyte progenitors for application in a rat contusion model of spinal cord injury. *Int J Neurosci* **120**, 305-313 (2010).

349 Kroon, E. *et al.* Pancreatic endoderm derived from human embryonic stem cells generates glucose-responsive insulin-secreting cells in vivo. *Nat Biotechnol* **26**, 443-452 (2008).

350 Raya, A. *et al.* Disease-corrected haematopoietic progenitors from Fanconi anaemia induced pluripotent stem cells. *Nature* **460**, 53-59 (2009).

351 Smithies, O., Gregg, R. G., Boggs, S. S., Koralewski, M. A. & Kucherlapati, R. S. Insertion of DNA sequences into the human chromosomal beta-globin locus by homologous recombination. *Nature* **317**, 230-234 (1985).

352 Thomas, K. R., Folger, K. R. & Capecchi, M. R. High frequency targeting of genes to specific sites in the mammalian genome. *Cell* **44**, 419-428 (1986).

353 Snouwaert, J. N. *et al.* An animal model for cystic fibrosis made by gene targeting. *Science* **257**, 1083-1088 (1992).

354 Ratcliff, R. *et al.* Production of a severe cystic fibrosis mutation in mice by gene targeting. *Nat Genet* **4**, 35-41 (1993).

355 Kim, Y. G., Cha, J. & Chandrasegaran, S. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proc Natl Acad Sci U S A* **93**, 1156-1160 (1996).

356 Mussolino, C. *et al.* A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic Acids Res* **39**, 9283-9293 (2011).

357 Zhang, F. *et al.* Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nat Biotechnol* **29**, 149-153 (2011).

358 Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012).

359 Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173-1183 (2013).

360 Ma, H. *et al.* Multiplexed labeling of genomic loci with dCas9 and engineered sgRNAs using CRISPRainbow. *Nat Biotechnol* **34**, 528-530 (2016).

361 Schmidtmann, E., Anton, T., Rombaut, P., Herzog, F. & Leonhardt, H. Determination of local chromatin composition by CasID. *Nucleus* **7**, 476-484 (2016).

362 Zhang, S. C., Wernig, M., Duncan, I. D., Brüstle, O. & Thomson, J. A. In vitro differentiation of transplantable neural precursors from human embryonic stem cells. *Nat Biotechnol* **19**, 1129-1133 (2001).

363 Koch, P., Opitz, T., Steinbeck, J. A., Ladewig, J. & Brüstle, O. A rosette-type, self-renewing human ES cell-derived neural stem cell with potential for in vitro instruction and synaptic integration. *Proc Natl Acad Sci U S A* **106**, 3225-3230 (2009).

364 Khokha, M. K., Yeh, J., Grammer, T. C. & Harland, R. M. Depletion of three BMP antagonists from Spemann's organizer leads to a catastrophic loss of dorsal structures. *Dev Cell* **8**, 401-411 (2005).

365 Dal-Pra, S., Fürthauer, M., Van-Celst, J., Thisse, B. & Thisse, C. Noggin1 and Follistatin-like2 function redundantly to Chordin to antagonize BMP activity. *Dev Biol* **298**, 514-526 (2006).

366 Bachiller, D. *et al.* The organizer factors Chordin and Noggin are required for mouse forebrain development. *Nature* **403**, 658-661 (2000).

367 Chambers, S. M. *et al.* Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nat Biotechnol* **27**, 275-280 (2009).

368 Shi, Y., Kirwan, P., Smith, J., Robinson, H. P. & Livesey, F. J. Human cerebral cortex development from pluripotent stem cells to functional excitatory synapses. *Nat Neurosci* **15**, 477-486, S471 (2012).

369 DeRosa, B. A. *et al.* Derivation of autism spectrum disorder-specific induced pluripotent stem cells from peripheral blood mononuclear cells. *Neurosci Lett* **516**, 9-14 (2012).

370 Douvaras, P. *et al.* Efficient generation of myelinating oligodendrocytes from primary progressive multiple sclerosis patients by induced pluripotent stem cells. *Stem Cell Reports* **3**, 250-259 (2014).

371 Roybon, L. *et al.* Human stem cell-derived spinal cord astrocytes with defined mature or reactive phenotypes. *Cell Rep* **4**, 1035-1048 (2013).

372 Lancaster, M. A. *et al.* Cerebral organoids model human brain development and microcephaly. *Nature* **501**, 373-379 (2013).

373 Kadoshima, T. *et al.* Self-organization of axial polarity, inside-out layer pattern, and species-specific progenitor dynamics in human ES cell-derived neocortex. *Proc Natl Acad Sci U S A* **110**, 20284-20289 (2013).

374 Paşca, A. M. *et al.* Functional cortical neurons and astrocytes from human pluripotent stem cells in 3D culture. *Nat Methods* **12**, 671-678 (2015).

375 Qian, X. *et al.* Brain-Region-Specific Organoids Using Mini-bioreactors for Modeling ZIKV Exposure. *Cell* **165**, 1238-1254 (2016).

376 Muguruma, K., Nishiyama, A., Kawakami, H., Hashimoto, K. & Sasai, Y. Self-organization of polarized cerebellar tissue in 3D culture of human pluripotent stem cells. *Cell Rep* **10**, 537-550 (2015).

377 Nakano, T. *et al.* Self-formation of optic cups and storable stratified neural retina from human ESCs. *Cell Stem Cell* **10**, 771-785 (2012).

378 Sakaguchi, H. *et al.* Generation of functional hippocampal neurons from self-organizing human embryonic stem cell-derived dorsomedial telencephalic tissue. *Nat Commun* **6**, 8896 (2015).

379 Pellegrini, L. *et al.* Human CNS barrier-forming organoids with cerebrospinal fluid production. *Science* **369** (2020).

380 Xiang, Y. *et al.* Fusion of Regionally Specified hPSC-Derived Organoids Models Human Brain Development and Interneuron Migration. *Cell Stem Cell* **21**, 383-398 e387 (2017).

381 Bagley, J. A., Reumann, D., Bian, S., Lévi-Strauss, J. & Knoblich, J. A. Fused cerebral organoids model interactions between brain regions. *Nat Methods* **14**, 743-751 (2017).

382 Bershteyn, M. *et al.* Human iPSC-Derived Cerebral Organoids Model Cellular Features of Lissencephaly and Reveal Prolonged Mitosis of Outer Radial Glia. *Cell Stem Cell* **20**, 435-449 e434 (2017).

383 Mariani, J. *et al.* FOXG1-Dependent Dysregulation of GABA/Glutamate Neuron Differentiation in Autism Spectrum Disorders. *Cell* **162**, 375-390 (2015).

Aim of the thesis

As outlined in the *Introduction*, about 50% of individuals affected by neurodevelopmental disorders still do not have a molecular diagnosis. In this thesis, I aimed at improving this by (I) identifying a novel gene involved in severe cases of developmental and epileptic encephalopathy, (II) investigating functional transcriptional enhancers, often neglected in the investigation of genetic forms of NDDs, in embryonic and neural stem cells, and finally (III) gaining more insight into the role of YY1, an important protein for enhancer activity and transcriptional regulation, which is also involved in a neurodevelopmental disorder when mutant.

PART I

In **chapter 2** we reported a recurrent homozygous mutation in the gene *UGP2*, that has never been associated to disease before, as causative of severe developmental and epileptic encephalopathy. In **chapter 3** the chromatin architecture of the *UGP2* locus is investigated.

PART II

We present a method that, combining chromatin immunoprecipitation with the massively parallel reporter assay STARR-seq, allows the genome wide identification of functional enhancers in human embryonic stem cells, in **chapter 4**, and neural stem cells in **chapter 5**.

PART III

In **chapter 6**, we identified the YY1 protein interactome in ESCs and in NSCs, while in **chapter 7** we investigated the role of YY1 in gene expression and enhancer activation in these two cell types.

Part I

Loss of UGP2 in brain leads to a severe DEE

Insights on *UGP2* regulation

2

Loss of UGP2 in brain leads to a severe DEE

- A founder mutation abolishing the short *UGP2* isoform expression arose in Balochistan 600 years ago
- The short isoform of *UGP2* is the predominant in brain
- The complete absence of UGP2 is likely lethal
- Bi-allelic isoform-specific start-loss mutations of essential genes can cause genetic diseases

This chapter has been published in:

E. Perenthaler *et al.* (2020). Loss of UGP2 in brain leads to a severe epileptic encephalopathy, emphasizing that bi-allelic isoform-specific start-loss mutations of essential genes can cause genetic diseases. Acta Neuropathol 39, 45-442.

Loss of UGP2 in brain leads to a severe epileptic encephalopathy, emphasizing that bi-allelic isoform specific start-loss mutations of essential genes can cause genetic diseases

Elena Perenthaler[1], Anita Nikoncuk[1*], Soheil Yousefi[1*], Woutje M. Berdowski[1*], Maysoon Alsagob[2*], Ivan Capo[3], Herma C. van der Linde[1], Paul van den Berg[1], Edwin H. Jacobs[1], Darija Putar[1], Mehrnaz Ghazvini[4], Eleonora Aronica[5,6], Wilfred F.J. van IJcken[7], Walter G. de Valk[1], Evita Medici–van den Herik[8], Marjon van Slegtenhorst[1], Lauren Brick[9], Mariya Kozenko[9], Jennefer N. Kohler[10], Jonathan A. Bernstein[11], Kristin G. Monaghan[12], Amber Begtrup[12], Rebecca Torene[12], Amna Al Futaisi[13], Fathiya Al Murshedi[14], Renjith Mani[13], Faisal Al Azri15, Erik-Jan Kamsteeg[16], Majid Mojarrad[17,18,19], Atieh Eslahi[17,20], Zaynab Khazaei[21], Fateme Massinaei Darmiyan[21], Mohammad Doosti[22], Ehsan Ghayoor Karimiani[23], Jana Vandrovcova[24], Faisal Zafar[25], Nuzhat Rana[25], Krishna K. Kandaswamy[26], Jozef Hertecant[27], Peter Bauer[26], Mohammed A. AlMuhaizea[28], Mustafa Salih[29], Mazhor Aldosary[2], Rawan Almass[2], Laila Al-Quait[2], Wafa Qubbaj[30], Serdar Coskun[30], Khaled O. Alahmadi[31], Muddathir H.A. Hamad[29], Salem Alwadaee[30], Khalid Awartani[32], Anas M. Dababo[30], Futwan Almohanna[33], Dilek Colak[34], Stephanie Efthymiou[24], Henry Houlden[24], Aida M. Bertoli-Avella[26], Reza Maroofian[24], Kyle Retterer[12], Alice S. Brooks[1], Namik Kaya[2], Tjakko J. van Ham[1] and Tahsin Stefan Barakat[1, #]

[1]Department of Clinical Genetics, Erasmus MC University Medical Center, Rotterdam, The Netherlands
[2]Department of Genetics, King Faisal Specialist Hospital and Research Centre, Riyadh 11211, Kingdom of Saudi Arabia
[3]Department for Histology and Embryology, Faculty of Medicine Novi Sad, University of Novi Sad, Serbia
[4]iPS cell core facility, Erasmus MC University Medical Center, Rotterdam, The Netherlands
[5]Amsterdam UMC, University of Amsterdam, Department of (Neuro)pathology, Amsterdam, Amsterdam Neuroscience, The Netherlands
[6]Stichting Epilepsie Instellingen Nederland (SEIN), The Netherlands
[7]Center for Biomics, Department of Cell Biology, Erasmus MC University Medical Center, Rotterdam, The Netherlands
[8]Department of Neurology, Erasmus MC University Medical Center, Rotterdam, The Netherlands
[9]Division of Genetics, McMaster Children's Hospital, Hamilton, Ontario, L8S 4J9, Canada.
[10]Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA, 94035, USA
[11]Department of Pediatrics, Division of Medical Genetics, Stanford University School of Medicine, Stanford, CA 94035, USA
[12]GeneDx, Gaithersburg, MD, 20877, USA
[13]Department of Child health, college of Medicine and Health Sciences, Sultan Qaboos University, Muscat, Oman
[14]Genetic and Developmental Medicine Clinic, Sultan Qaboos University Hospital, Muscat, Oman
[15]Department of Radiology and molecular imaging, Sultan Qaboos University Hospital, Muscat, Oman
[16]Department of Clinical Genetics, Radboud University, Nijmegen, The Netherlands
[17]Department of Medical Genetics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran
[18]Medical Genetics Research Center, Mashhad University of Medical Sciences, Mashhad, Iran
[19]Genetic Center of Khorasan Razavi, Mashhad, Iran
[20]Student Research Committee, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran
[21]Genetic Counseling Center, Welfare Organization of Sistan and Baluchestan, Zahedan, Iran
[22]Department of Modern Sciences and Technologies, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.
[23]Genetics Research Centre, Molecular and Clinical Sciences Institute, St. George's, University, London, SW17 ORE, United Kingdom
[24]Department of Neuromuscular Disorders, UCL Queen Square Institute of Neurology, London, WC1N 3BG, United Kingdom.
[25]Department of Paediatric Neurology, Children's hospital and institute of Child health, Multan 60000, Pakistan
[26]CENTOGENE AG, Rostock, Germany
[27]Department of Pediatrics, Tawam Hospital, and College of Medicine and Health Sciences, UAE University, Al-Ain, UAE
[28]Department of Neurosciences, King Faisal Specialist Hospital and Research Centre, Riyadh 11211, Kingdom of Saudi Arabia
[29]Department of Pediatric Neurology, College of Medicine, King Saud University, Riyadh 11211, Kingdom of Saudi Arabia
[30]Department of Pathology and Laboratory Medicine, King Faisal Specialist Hospital and Research Centre, Riyadh 11211, Kingdom of Saudi Arabia
[31]Radiology Department, King Faisal Specialist Hospital and Research Centre, Riyadh 11211, Kingdom of Saudi Arabia
[32]Obstetrics/Gynecology Department, King Faisal Specialist Hospital and Research Centre, Riyadh 11211, Kingdom of Saudi Arabia
[33]Department of Cell Biology, King Faisal Specialist Hospital and Research Centre, Riyadh 11211, Kingdom of Saudi Arabia
[34]Department of Biostatistics, Epidemiology and Scientific Computing, King Faisal Specialist Hospital and Research Centre, Riyadh 11211, Kingdom of Saudi Arabia

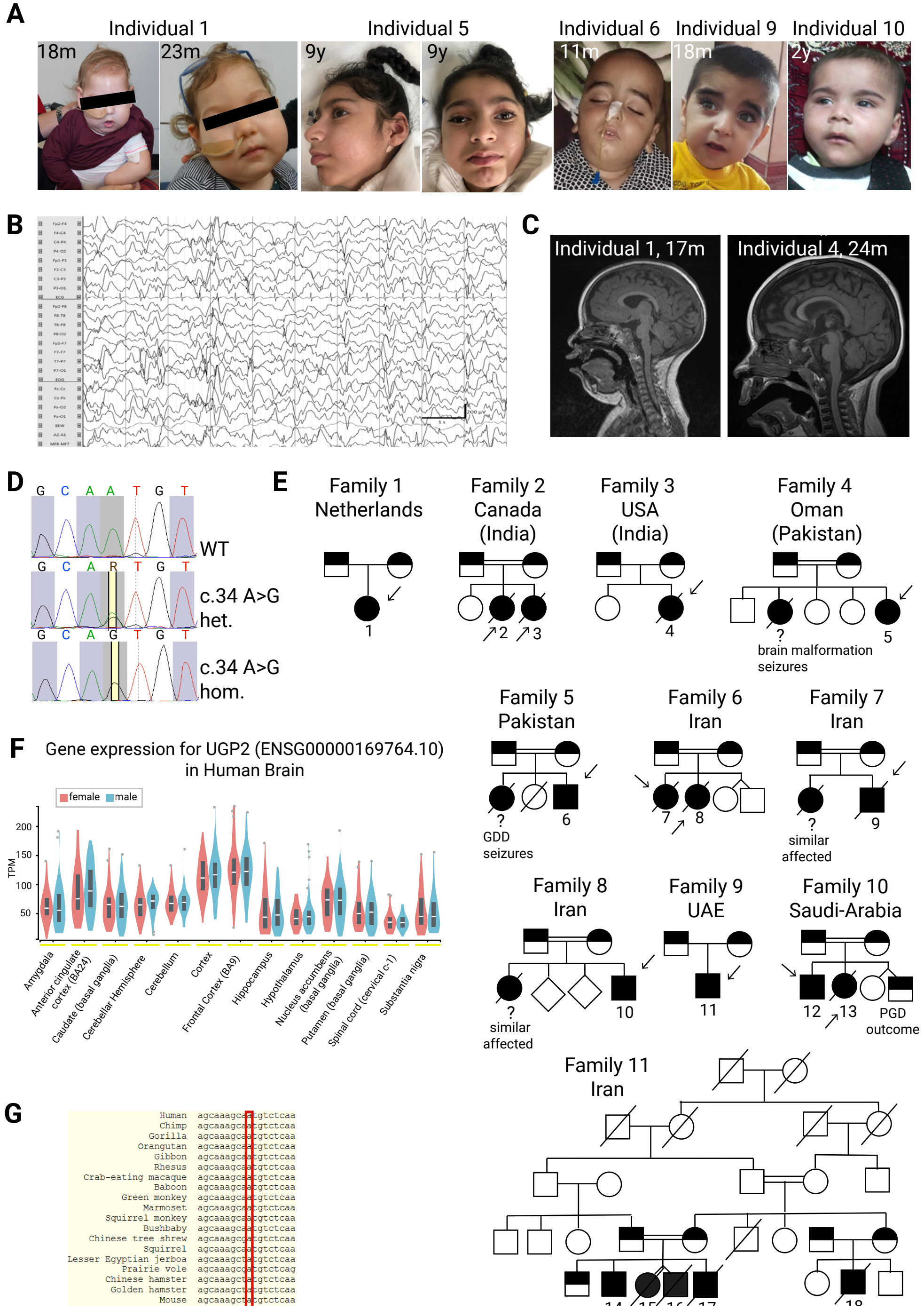
Developmental and/or epileptic encephalopathies (DEEs) are a group of devastating genetic disorders, resulting in early onset, therapy resistant seizures and developmental delay. Here we report on 19 individuals from 12 families presenting with a severe form of intractable epilepsy, severe developmental delay, progressive microcephaly and visual disturbance. Whole exome sequencing identified a recurrent, homozygous variant (chr2:64083454A>G) in the essential UDP-glucose pyrophosphorylase (UGP2) gene in all probands. This rare variant results in a tolerable Met12Val missense change of the longer UGP2 protein isoform but causes a disruption of the start codon of the shorter isoform. We show that the absence of the shorter isoform leads to a reduction of functional UGP2 enzyme in brain cell types, leading to altered glycogen metabolism, upregulated unfolded protein response and premature neuronal differentiation, as modelled during pluripotent stem cell differentiation in vitro. In contrast, the complete lack of all UGP2 isoforms leads to differentiation defects in multiple lineages in human cells. Reduced expression of Ugp2a/Ugp2b in vivo in zebrafish mimics visual disturbance and mutant animals show a behavioral phenotype. Our study identifies a recurrent start codon mutation in UGP2 as a cause of a novel autosomal recessive

DEE. Importantly, it also shows that isoform specific start-loss mutations causing expression loss of a tissue relevant isoform of an essential protein can cause a genetic disease, even when an organism-wide protein absence is incompatible with life. We provide additional examples where a similar disease mechanism applies.

Introduction

Developmental and/or epileptic encephalopathies (DEEs) are a heterogeneous group of genetic disorders, characterized by severe epileptic seizures in combination with developmental delay or regression [1]. Genes involved in multiple pathophysiological pathways have been implicated in DEEs, including synaptic impairment, ion channel alterations, transporter defects and metabolic processes such as disorders of glycosylation[2]. Mostly, dominant acting, de novo mutations have been identified in children suffering from DEEs[3], and only a limited number of genes with a recessive mode of inheritance are known so far, with a higher occurrence rate in consanguineous populations[4]. A recent cohort study on DEEs employing whole exome sequencing (WES) and copy-number analysis, however, found that up to 38% of diagnosed cases might be caused by recessive genes, indicating that the importance of this mode of inheritance in DEEs has been underestimated[5].

The human genome contains ~20,000 genes of which more than 5,000 have been implicated in genetic disorders. Wide-scale population genomics studies and CRISPR-Cas9 based loss-of-function (LoF) screens have identified around 3000-7000 genes that are essential for the viability of the human organism or result in profound loss of fitness when mutated. In agreement with that they are depleted for LoF variants in the human population[6]. For some of these essential genes it is believed that LoF variants are incompatible with life and are therefore unlikely to be implicated in genetic disorders presenting in postnatal life[7]. One such example is the UDP-glucose pyrophosphorylase (UGP2) gene at chromosome 2. UGP2 is an essential octameric enzyme in nucleotide-sugar metabolism[8-10], as it is the only known enzyme capable of catalyzing the conversion of glucose-1-phosphate to UDP-glucose[11,12]. UDP-glucose is a crucial precursor for the production of glycogen by glycogen synthase (GYS)[13,14], and also serves as a substrate for UDP-glucose:glycoprotein transferases (UGGT) and UDP-glucose-6-dehydrogenase (UGDH), thereby playing important roles in glycoprotein folding control, glycoconjugation and UDP-glucuronic acid synthesis. The latter is an obligate precursor for the synthesis of glycosaminoglycans and proteoglycans of the extracellular matrix[15,16], of which aberrations have been associated with DEEs and neurological disorders[17-20]. UGP2 has previously been identified as a marker protein in various types of malignancies including gliomas where its upregulation is correlated with a poor disease outcome[21-28], but has so far not been implicated in genetic diseases and it has been speculated that this is given its essential role in metabolism[8].



Many genes are differentially expressed amongst tissues, regulated by non-coding regulatory elements[29]. In addition, it has become clear that there are more than 40,000 protein isoforms encoded in the human genome, whose expression levels vary amongst tissues. Although there are examples of genetic disorders caused by the loss of tissue specific protein isoforms[30-33], it is unknown whether a tissue-relevant loss of an essential gene can be involved in human disease. Here, we report on such a scenario, providing evidence that a novel form of a severe DEE is caused by the brain relevant loss of the essential gene UGP2 due to an isoform specific and germ line transmitted start codon mutation. We present data that this is likely a more frequent disease mechanism in human genetics, illustrating that essential genes for which organism-wide loss is lethal can still be implicated in genetic disease when only absent in certain tissues due to expression misregulation.

Results

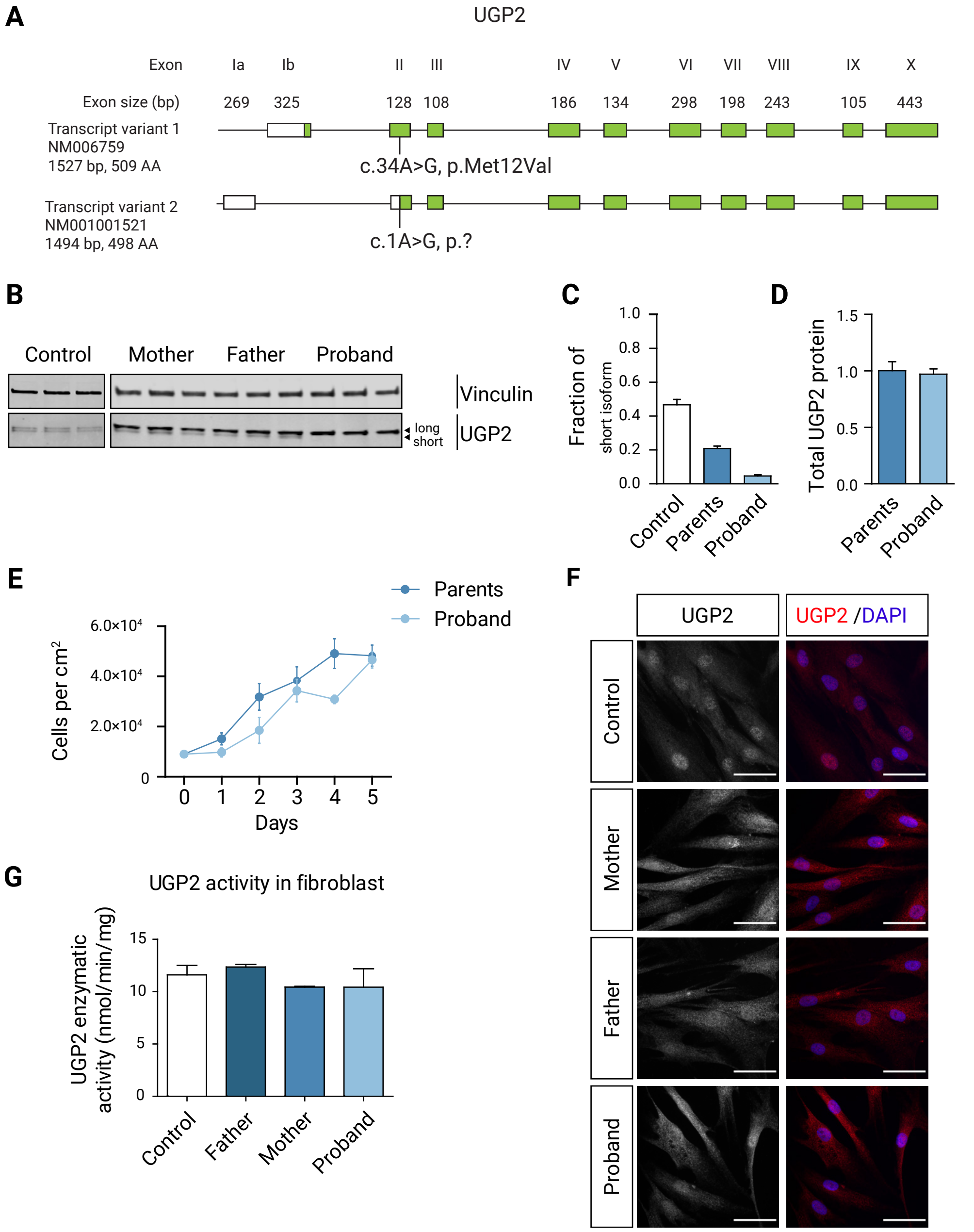
A recurrent ATG mutation in UGP2 in 19 individuals presenting with a severe DEE

We encountered a three-month old girl (**Figure 1A**, family 1, individual 1), that was born as the first child to healthy non-consanguineous Dutch parents, by normal vaginal delivery after an uneventful pregnancy conceived by ICSI. She presented

Figure 1 | UGP2 homozygous variants in 13 individuals with severe epileptic encephalopathy. A) Facial pictures of individual 1 (at 3, 18 and 23 month), individual 5 (at 9 years), individual 6 (at 11 month) and individual 10 (at 2 years). Note the progressive microcephaly with sloping forehead, suture ridging, bitemporal narrowing, high hairline, arched eyebrows, pronounced philtrum, a relatively small mouth and large ears. **B)** Electroencephalogram of individual 1 at the age of 8 month showing a highly disorganized pattern with high voltage irregular slow waves intermixed with multifocal spikes and polyspikes. **C)** T1-weighted mid sagittal brain MRI of individual 1 (age 17 month) and individual 4 (age 24 month) illustrating global atrophy and microcephaly but no major structural anomalies. **D)** Sanger sequencing traces of family 1, confirming the chr2:64083454A>G variant in UGP2 in a heterozygous and homozygous state in parents and affected individual 1, respectively. **E)** Family pedigrees of ascertained patients. Affected individuals and heterozygous parents are indicated in black and half black, respectively. Affected individuals with confirmed genotype are indicated with an arrow, and numbers. Other affected siblings presenting with similar phenotypes are indicated with a question mark. Consanguineous parents are indicated with a double connection line. Male are squares, females circles; unknown sex indicated with rotated squares; deceased individuals are marked with a line. **F)** Violin plots showing distribution of gene expression (in TPM) amongst male and female samples from the GTEx portal43 for various brain regions. Outliers are indicated by dots. **G)** Multiple species sequence alignment from the UCSC browser, showing that the ATG start site is highly conserved.

in the first weeks of life with irritability and jitteriness, that developed into infantile spasms and severe epileptic activity on multiple electroencephalograms, giving rise to a clinical diagnosis of West syndrome (**Figure 1B**). Despite the use of multiple anti-epileptic drugs, including ACTH and a ketogenic diet, seizures remained intractable and occurred daily. Severe developmental delay was evident without acquisition of any noticeable developmental milestones, causing the need for gastrointestinal tube feeding. Visual tracking was absent, and foveal hypopigmentation, hypermetropia and mild nystagmus were noticed upon ophthalmological investigation. MRI brain imaging showed no gross structural abnormalities or migration disorders at the age of 4 months, but displayed reduced white matter, that further developed into global atrophy with wide sulci and wide pericerebral liquor spaces at the age of 17 months (**Figure 1C**, **Figure S1B**). At that time, she had become progressively microcephalic, with a head circumference of -2.96 SD at the last investigation at 23 months of age (**Figure S1A**). She showed a number of minor dysmorphisms, including a sloping forehead, elongated head with suture ridging, bitemporal narrowing, a relatively small mouth and large ears (**Figure 1A**). Neurological examination showed brisk, symmetric deep tendon reflexes, more pronounced at the upper limbs. Routine investigations, including metabolic screening in urine, plasma and cerebrospinal fluid were normal. A SNP-array showed a normal female chromosomal profile, with a large, ~30 Mb run of homozygosity (ROH) at chromosome 2, and a few smaller ROH regions, adding up to 50 Mb ROH regions in total, pointing to an unrecognized common ancestor of both parents (coefficient of inbreeding 1/64). Subsequent trio WES did not show any disease-causing variants in known DEE genes, but identified a homozygous variant (chr2:64083454A>G) in *UGP2*, located in the large ROH region (**Figure 1D**), with no other disease implicated variants observed in that region. Both parents were heterozygous carriers of the same variant. Via Genematcher[34] and our network of collaborators, we identified 18 additional individuals from 11 unrelated families (of which 9 were consanguineous), harboring the exact same homozygous variant and presenting with an almost identical clinical phenotype of intractable seizures, severe developmental delay, visual disturbance, microcephaly and similar minor dysmorphisms (**Figure 1A, C, E, Figure S1B, Supplementary Case reports, Supplementary Table 1** for detailed information on 13 cases). Seven of these individuals passed away before the age of 3.5 years. In 4 families, at least 4 already deceased siblings had a similar phenotype but could not be investigated. Two families were of Indian descent (both with ancestors from regions currently belonging to Pakistan), living in Canada (family 2) and the USA (family 3), with the remaining families from Oman (family 4, originally from Pakistan), Pakistan (family 5), Iran (family 6, 7, and 8), UAE (family 9) and Saudi-Arabia (family 10). One additional case in a family from Oman, and 5 additional cases in a family from Iran were identified presenting with intractable seizures and microcephaly, but no detailed medical information could be obtained at this point.

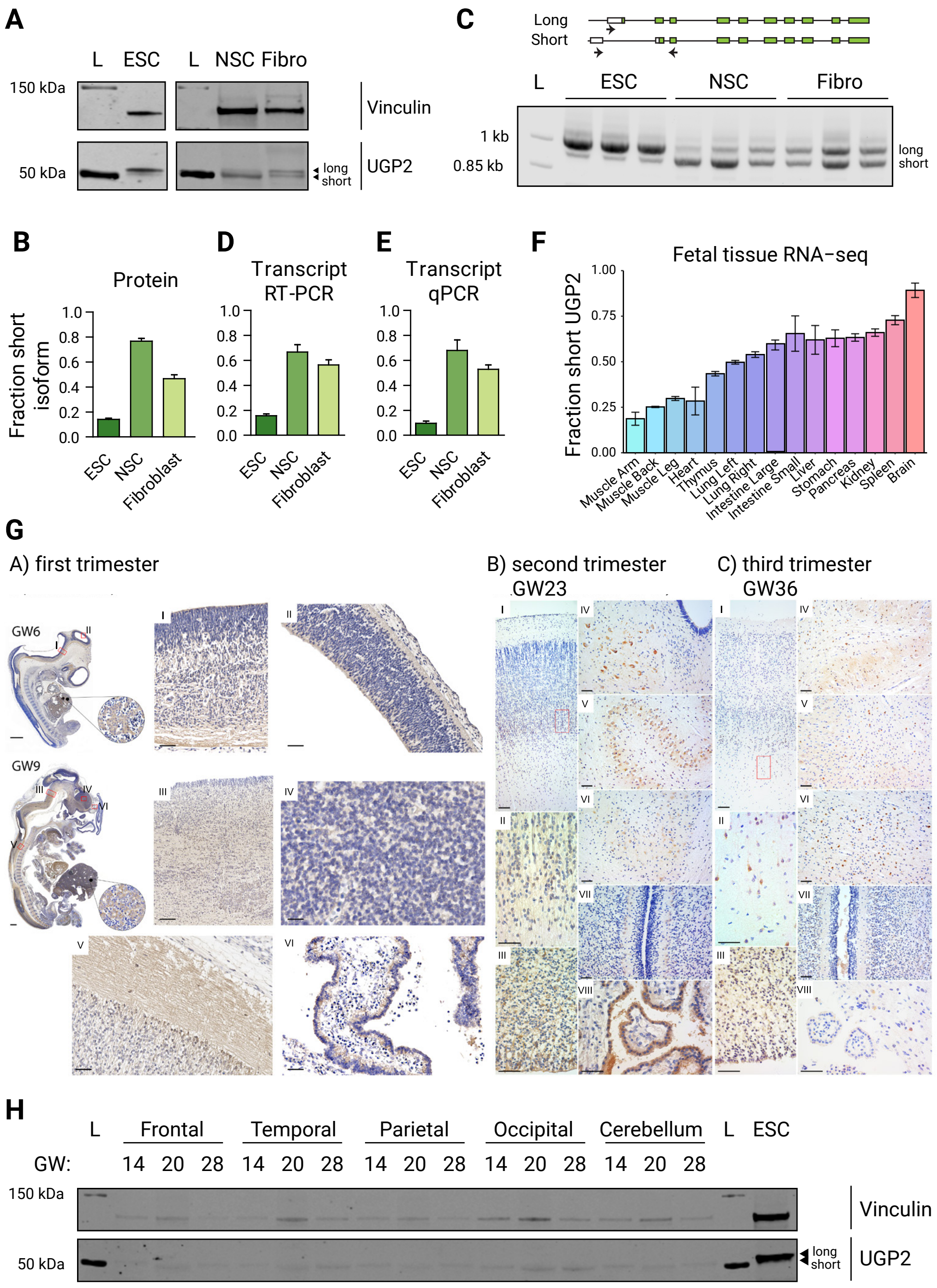
Having identified at least 19 individuals with an almost identical clinical phenotype and an identical homozygous variant in the same gene, led us to pursue *UGP2* as a candidate gene for a new genetic form of DEE. *UGP2* is highly expressed in various brain regions (**Figure 1F**), and also widely expressed amongst other tissues, including liver and muscle according to the data from the GTEx portal[35] (**Figure S1D**). The (chr2:64083454A>G) variant is predicted to cause a missense variant (c.34A>G, p.Met12Val) in *UGP2* isoform 1 (NM_006759), and to cause a translation start loss (c.1A>G, p.?.) of *UGP2* isoform 2 (NM_001001521), referred to as long and short isoform, respectively. The variant has not been reported in the Epi25 web browser[36], ClinVar[37], LOVD[38], Exome Variant Server[39], DECIPHER[40], GENESIS[41] , GME variome[42] or Iranome databases[43], is absent from our in-house data bases and is found only 15 times in a heterozygous, but not homozygous, state in the 280,902 alleles present in *gnomAD* (MAF: 0.00005340)[44]. In the *GeneDx* unaffected adult cohort, the variant was found heterozygous 10 times out of 173,502 alleles (MAF: 0.00005764), in the ~10,000 exomes of the Queen Square Genomic Center database two heterozygous individuals were identified, and out of 45,921 individuals in the *Centogene* cohort, 10 individuals are heterozygous for this variant. The identified variant has a CADD score (v1.4) of 19.22[45] and Mutation Taster[46] predicted this variant as disease causing. The nucleotide is strongly conserved over multiple species (**Figure 1G**). Analysis of WES data from 6 patients did provide evidence of a shared ROH between patients from different families, indicating that this same variant might represent an ancient mutation that originated some 26 generations ago (**Figure S1C**). Interestingly, since most families originally came from regions of India, Pakistan and Iran, overlapping with an area called Balochistan, this could indicate that the mutation has originated there around 600 years ago. As Dutch traders settled in that area in the 17th century, it is tempting to speculate that this could explain the co-occurrence of the variant in these distant places[47].



Short UGP2 isoform is predominantly expressed in brain and absent in patients with ATG mutations

Both UGP2 isoforms only differ by 11 amino acids at the N-terminal (Figure 2A) and are expected to be functionally equivalent[8]. To investigate how the A>G variant may cause DEE, we first obtained fibroblasts from individual 1 (homozygous for the A>G variant) and her heterozygous parents and analyzed the isoform expression by Western blotting (**Figure 2B**). Whereas the two isoforms were equally expressed in wild type fibroblasts, the expression of the shorter isoform was diminished to ~25% of total UGP2 in heterozygous parents, both of individual 1 (**Figure 2B, C**) and of individual 2 and 3 (**Figure S2A, B**), and was absent in cells from the affected individual 1 (**Figure 2B, C**; fibroblasts of the affected children in family 2 or other families were not available). Total UGP2 levels were not significantly different between the affected child and her parents, or between parents and wild type controls (**Figure 2D, Figure S2C**). This indicates that the long isoform harboring the Met12Val missense variant is upregulated in fibroblast when the short isoform is missing. Moreover, this indicates that Met12Val does not affect the stability of the long isoform at the protein or transcript level (**Figure S2D, E, F**). RNA-seq on peripheral blood samples of family 1 did not identify altered splicing events of *UGP2* and the

Figure 2 | UGP2 homozygous variant leads to a loss of the shorter protein isoform in patient fibroblasts. **A)** Schematic drawing of the human UGP2 locus, with both long and short transcript isoforms. Boxes represent exons, with coding sequences indicated in green. The location of the recurrent mutation is indicated in both transcripts. **B)** Western blotting of cellular extracts derived from control fibroblasts or fibroblasts obtained from family 1, detecting the housekeeping control vinculin or UGP2. Note the two separated isoforms of UGP2 that have a similar intensity in wild type cells. The shorter isoform is less expressed in fibroblasts from heterozygous parents and absent in fibroblasts from the affected proband. **C)** Western blotting quantification of the fraction of the short UGP2 protein isoform compared to total UGP2 expression in control, parental heterozygous and proband homozygous fibroblasts, as determined in three independent experiments. Error bars represent SEM. **D)** Western blotting quantification of total UGP2 protein levels, as determined by the relative expression to the housekeeping control vinculin. Bar plot showing the results from three independent experiments. Error bars represent SEM; no significant differences were found between parents and proband, t-test, two-tailed. **E)** Cell proliferation experiment of fibroblast from heterozygous parents and homozygous proband from family 1, during a 5 days period, determined in three independent experiments. Error bars represent SEM. **F)** Immunocytochemistry on cultured control and UGP2 heterozygous and homozygous mutant fibroblast derived from family 1, detecting UGP2 (red). Nuclei are stained with DAPI. Scale bar = 50 μm. **G)** Enzymatic activity of UGP2 as measured in control and UGP2 heterozygous and homozygous mutant fibroblast derived from family 1. Shown is the mean of two independent experiments. Error bars represent SEM; no significant differences were found, unpaired t-test, two-tailed.



global transcriptome of the proband was not different from her parents, although only a limited analysis could be performed as only a single sample was available for each individual (**Figure S2G, H**). Both homozygous and heterozygous fibroblasts had a similar proliferation rate compared to wild type fibroblasts (**Figure 2E, Figure S2I**), and immunocytochemistry confirmed a similar subcellular localization of UGP2 in mutant and wild type cells (**Figure 2F**). We then measured the enzymatic activity of UGP2 in wild type, heterozygous and homozygous fibroblasts, and found that mutant fibroblast had a similar capacity to produce UDP-glucose in the presence of exogenously supplied glucose-1-phosphate and UTP (**Figure 2G**). Altogether, this indicates that the long UGP2 isoform harboring the Met12Val missense change is functional and is therefore unlikely to contribute to the patient phenotype.

As the A>G variant results in a functional long UGP2 isoform but abolishes the translation of the shorter UGP2 isoform, we next investigated whether the ratio between short and long isoform differs amongst tissues. If so, the homozygous A>G variant would lead to depletion of UGP2 in tissues where mainly the short isoform is expressed, possibly below a threshold that is required for normal development or function. Western blotting on cellular extracts derived from wild type H9 human embryonic stem cells (ESCs), commercially acquired H9-derived neural stem cells (NSCs) and fibroblasts (**Figure 3A**) showed that, whereas the ratio between short

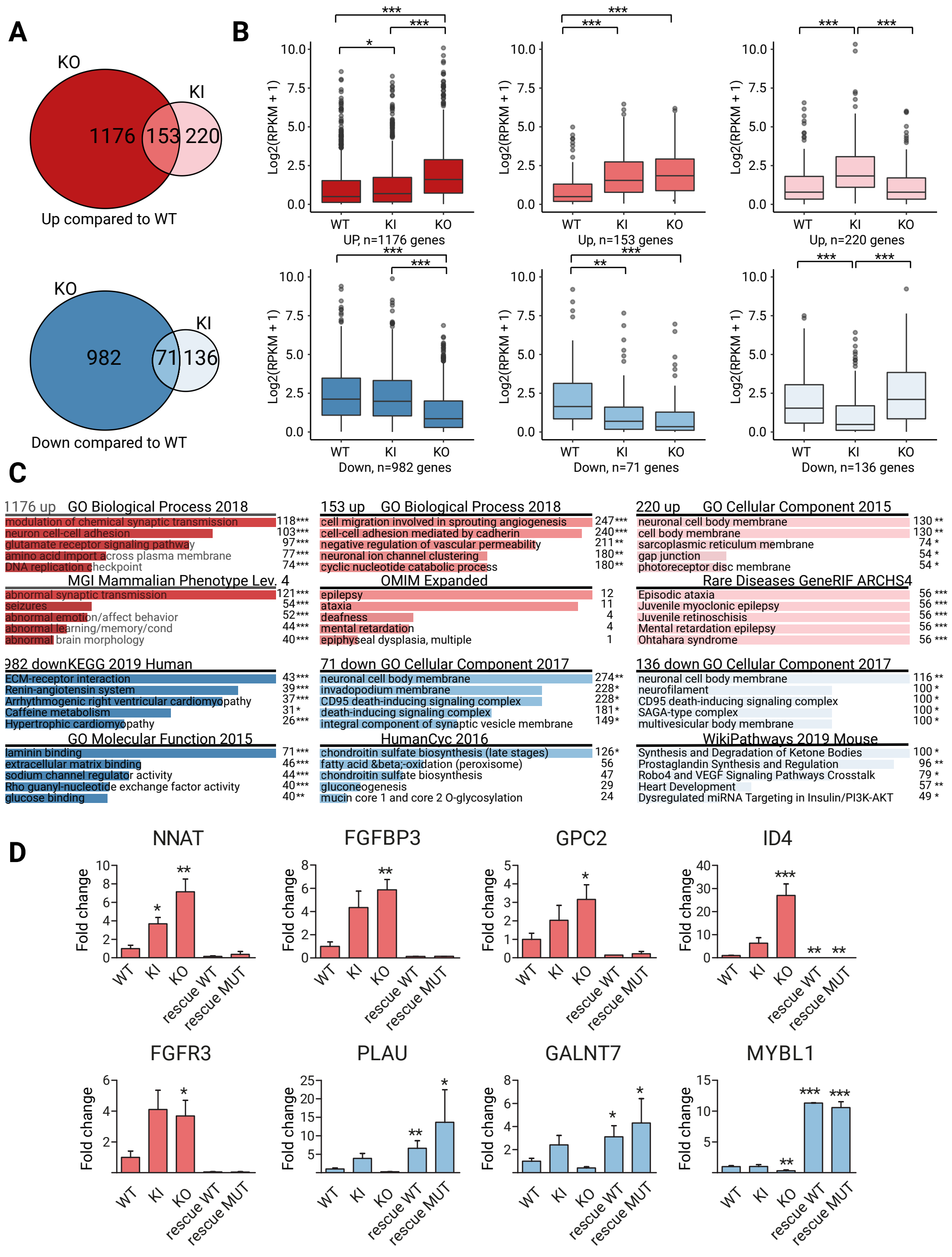
Figure 3 | UGP2 short isoform is predominant in brain related cell types. A) Western blotting showing UGP2 expression in H9 human embryonic stem cells (ESCs), H9 derived neural stem cells (NSCs) and fibroblasts (Fibro). Vinculin is used as a housekeeping control. Note the changes in relative expression between the two UGP2 isoforms in the different cell types. L, ladder. **B)** Western blotting quantification of the fraction of the short UGP2 protein isoform compared to total UGP2 expression, as determined in three independent experiments. Error bars represent SEM. **C)** Multiplex RT-PCR of ESCs, NSCs and fibroblasts, showing a similar variability in isoform expression at the transcript as at the protein level. Each cell line was tested in triplicates. **D)** Quantification of the fraction of the short UGP2 transcript isoform compared to total UGP2 expression, from the multiplex RT-PCR from C). Error bars represent SEM. **E)** Quantification of the fraction of the short UGP2 transcript isoform compared to total UGP2 expression by qRT-PCR in three independent experiments. Error bars represent SEM. **F)** Ratio of RNA-seq reads covering the short transcript isoform compared to the total reads (covering both short and long isoforms), in multiple fetal tissues. In RNA-seq samples derived from brain, virtually all UGP2 expression comes from the short isoform. Error bars represent SD. **G)** Immunohistochemistry detecting UGP2 in human fetal brains from the first, second and third trimester (gestational week (GW) 6, 9, 23 and 36). See text for details. **H)** Western blotting detecting UGP2 in various human brain regions at week 14, 20 and 28 of gestation, showing the virtual absence of the long isoform expression in fetal brain. Vinculin is used as a housekeeping control. L, ladder.

and long isoform in fibroblasts was around 0.5, in ESCs it was 0.14 and in NSCs 0.77, indicating that the shorter UGP2 isoform is the predominant one in NSCs (**Figure 3B**). A similar trend was observed when assessing the transcript level, both by multiplex RT-PCR and RT-qPCR, using primers detecting specifically the short and long transcript isoform (**Figure 3C, D, E**). This indicates that differential isoform expression between cell types is regulated at the transcriptional level, possibly hinting at tissue-specific regulatory elements driving isoform expression. We next analyzed RNA-seq data from human fetal tissues[48-51] to determine the fraction of reads covering short versus total *UGP2* transcripts (**Figure 3F**). This showed that in human fetal brain the short transcript isoform is predominantly expressed. To gain more insight into the cell type-specific expression of UGP2, we performed immunohistochemistry on human fetal brain tissues from the first to third trimester of pregnancy (**Figure 3G**). In the first trimester we found pale labeling of neuropil in the proliferative neuroepithelium of the hypothalamic, cortical, mesencephalic and thalamic regions (**Figure 3G-A/I, II, III, IV**), as well as the marginal zone of the spinal cord (**Figure 3G-A/V**) and cuboidal epithelial cells of choroid plexus (**Figure 3G-A/VI**). During the second trimester, UGP2 positivity was detected in neurons from the subplate region of the cerebral cortex (**Figure 3G-B/I, II**) and still in some of the cells in the neuroepithelium and subventricular zone (**Figure 3G-B/III**). Almost the same pattern of UGP2 distribution was found in the cerebral cortex of fetuses from the 3rd trimester. Also, we found clear cytoplasmatic UGP2 expression in neurons from mesencephalic, inferior olivary and cerebellar nuclei during the second (**Figure 3G-B/IV, V, and VI**) and third trimester, respectively (**Figure 3G-C/IV, V**). In the white matter of the cerebellum in the third trimester, we identified single positive glial cells (**Figure 3G-C/VI**). In the cerebellar cortex we did not find specific positivity of cells on UGP2 (**Figure 3G-B, C/VII**). Cuboidal epithelial cells of choroid plexus preserved UGP2 positivity during the second trimester (**Figure 3G-B/VIII**) but lost it in the third trimester (**Figure 3G-C/VIII**). Together this indicates that UGP2 can be detected in a broad variety of cell types during brain development. On Western blotting, we noticed preferential expression of the shorter UGP2 isoform in the developing cortex and cerebellum from gestational weeks 14, 20 and 28 (**Figure 3H**) and in the frontal cortex of brains from weeks 21 and 23 (**Figure S2J**). Together, this supports the hypothesis that the DEE phenotype in patients is caused by a major loss of functional UGP2 in the brain, as the short isoform represents virtually all UGP2 produced in this tissue.

Lack of the short UGP2 isoform leads to transcriptome changes upon differentiation into neural stem cells

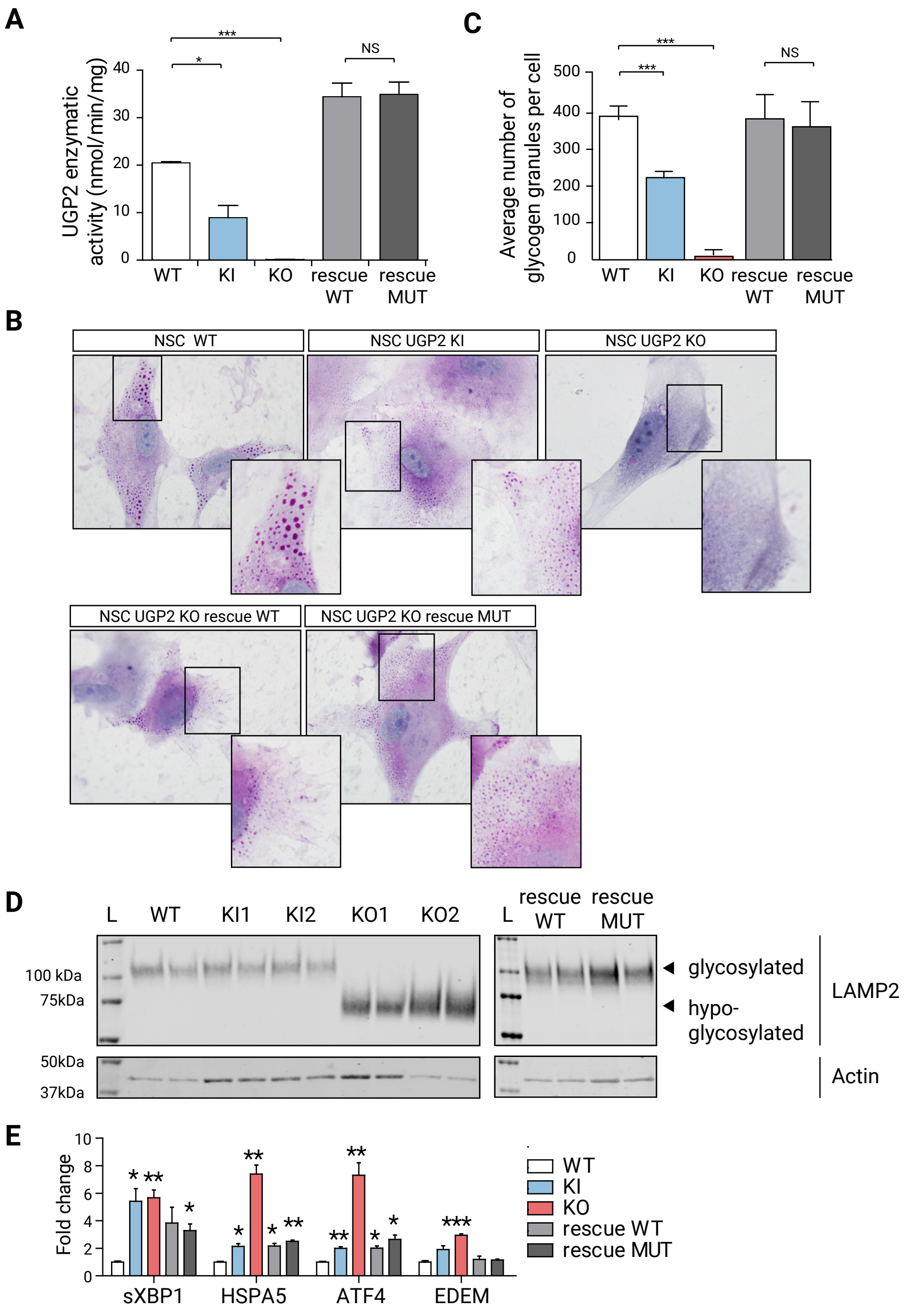
To model the disease *in vitro*, we first engineered the homozygous A>G mutation in H9 ESCs to study the mutation in a patient independent genetic background and compare it to isogenic parental cells. We obtained two independent clones harboring the homozygous A>G change (referred to as knock-in, KI, mutant) and two cell lines harboring an insertion of an additional A after nucleotide position 42 of *UGP2* transcript 1 (chr2:64083462_64083463insA) (**Figure S3A, B**) (referred to as knockout, KO). This causes a premature stop codon at amino acid position 47 (D15Rfs*33), leading to nonsense mediated mRNA decay and complete absence of UGP2 protein (**Figure S3C**). All derived ESCs had a normal morphology and remained pluripotent as assessed by marker expression (**Figure S3D, E**), indicating that the absence of UGP2 in ESCs is tolerated, in agreement with genome-wide LoF CRISPR screens which did not identify *UGP2* as an essential gene in ESCs[52,53]. We differentiated wild type, KI and KO ESCs into NSCs, using dual SMAD inhibition (**Figure S4A-C**). Wild type cells could readily differentiate into NSCs, having a normal morphology and marker expression, whereas differentiation of KI and KO cells was more variable and not all differentiations resulted in viable, proliferating NSCs. KO cells could not be propagated for more than 5 passages under NSC culture conditions (data not shown), which could indicate that the total absence of UGP2 protein is not tolerated in NSCs. When assessed by Western blotting, total UGP2 protein levels were reduced in KI cells and depleted in KO cells compared to wild type (**Figure S4D, E**).

Next, we performed RNA-seq of wild type, KI and KO ESCs and NSCs to assess how depletion of UGP2 upon NSCs differentiation would impact on the global transcriptome (**Figure 4, Figure S5, Supplementary Table 2**). In agreement with normal proliferation and morphology of KI and KO ESCs, all ESCs shared a similar expression profile of pluripotency associated genes and only few genes were differentially expressed between the three genotypes (**Figure S5C, Supplementary Table 3**). This indicates that the absence of UGP2 in ESCs does not lead to major transcriptome alterations despite the central role of this enzyme in metabolism. Upon differentiation, cells from all genotypes expressed NSC markers (**Figure S5F**), but when comparing wild type and KO cells, we observed noticeable changes, that were less pronounced in KI NSCs but still followed a similar trend (**Figure 4A, B, Figure S5D, E**). Gene enrichment analysis showed that genes downregulated in KO and KI cells were implicated in processes related to the extra-cellular matrix, cell-cell interactions and metabolism, while genes upregulated in KO and KI cells were enriched



for synaptic processes and genes implicated in epilepsy (**Figure 4C, Supplementary Table 4**). Both KO and KI cells showed an upregulation of neuronal expressed genes, indicating a tendency to differentiate prematurely. To validate RNA-seq findings, we tested several genes by RT-qPCR in wild type, KI and KO cells (**Figure 4D**). We also included KO rescue cells, in which we had restored the expression of either the wild type or the mutant UGP2 long isoform, leading each to an approximately 4-fold UGP2 overexpression at the NSC state compared to WT (**Figure S4F**). Amongst the tested genes was *NNAT*, which showed a significant upregulation in KI and KO cells, which was rescued by restoration of UGP2 expression in KO NSCs. *NNAT* encodes neuronatin that stimulates glycogen synthesis by upregulating glycogen synthase and was previously found to be upregulated in Lafora disease. This lethal teen-age onset neurodegenerative disorder presenting with myoclonic epilepsy is caused by mutations in the ubiquitin ligase malin, leading to accumulation of altered polyglucosans[54]. Malin can ubiquitinate neuronatin leading to its degradation. As reduced UGP2 expression might impact on glycogen production, it seems plausible that this results in compensatory *NNAT* upregulation and in downstream aberrations contributing to the patient phenotypes. Indeed, neuronatin upregulation was shown to cause increased intracellular Ca^{2+} signaling, ER stress, proteasomal dysfunction and cell death in Lafora disease[55,56], and was shown to be a stress responsive protein in the outer segment of retina photoreceptors[57,58]. Another interesting gene upregulated in KI and KO NSCs and downregulated in rescue cell lines was the autism candidate gene *FGFBP3*[59]. This secreted proteoglycan that enhances FGF signaling is broadly expressed in brain [60], and functions as an extracellular chaperone for locally

Figure 4 | RNA-seq of UGP2 mutant H9 derived cell lines. A) Venn diagram showing the overlap between differentially expressed genes in UGP2 KO or KI NSCs that are upregulated (upper panel genes with FDR<0.05 and LogFC>1) or downregulated (lower panel, genes with FDR<0.05 and LogFC<-1) compared to wild type NSCs. **B)** Box plot showing the distribution of gene expression levels (in Log2(RPKM+1)) from RNA-seq for the groups of genes displayed in A), in wild type, UGP2 KI or KO NSCs. Boxes are IQR; line is median; and whiskers extend to 1.5x the IQR (*=p<0.05; **=p<0.01,***=p<0.001, unpaired t-test, two-tailed). **C)** Enrichment analysis using Enrichr119 of up- or downregulated genes in NSCs from A) for selected gene ontology sets, showing the 5 most enriched terms per set. Combined score and p-value calculated by Enrichr are depicted (*=p<0.05; **=p<0.01; ***=p<0.001). **D)** qRT-PCR validation of differentially expressed genes from RNA-seq in wild type, UGP2 KI, UGP2 KO NSCs and KO NSCs that were rescued with either WT or MUT (Met12Val) transcript isoform 1, at p5 of NSC differentiation. Bar plot showing the mean fold change for the indicated genes compared to wild type, normalized for the housekeeping gene TBP. Results of two biological and two independent technical replicates are plotted. Colors match the Venn diagram group to which the tested genes belong to from A). Error bars represent SEM; (*=p<0.05; **=p<0.01,***=p<0.001, unpaired t-test, one-tailed).



stored FGFs in the ECM, thereby influencing glucose metabolism by regulating rate-limiting enzymes in gluconeogenesis[61]. Other potentially relevant genes displaying the same expression trend were the heparan sulphate proteoglycan GPC2 (a marker of immature neurons[62,63]), the helix-loop-helix transcription factor ID4 (a marker of postmitotic neurons[64]), and the signaling molecule FGFR3 that has been implicated in epilepsy[65]. Genes downregulated in KO cells and upregulated in rescue cells included urokinase-type plasminogen activator PLAU (deficiency in mouse models increases seizure susceptibility[66]), the glycoprotein GALNT7 (upregulation of which has been found to promote glioma cell invasion[67]) and the brain tumor gene MYBL1 (that has been shown to be regulated by O-linked N-acetylglucosamine[68]. Similar expression changes were observed in NSCs differentiated from induced pluripotent stem cells (iPSCs) that we had generated from family 1 (**Figure S6**). Together, RNA-seq showed that whereas the absence of UGP2 is tolerated in ESCs, its complete absence or reduced expression results in global transcriptome changes in NSCs, with many affected genes implicated in DEE relevant pathways.

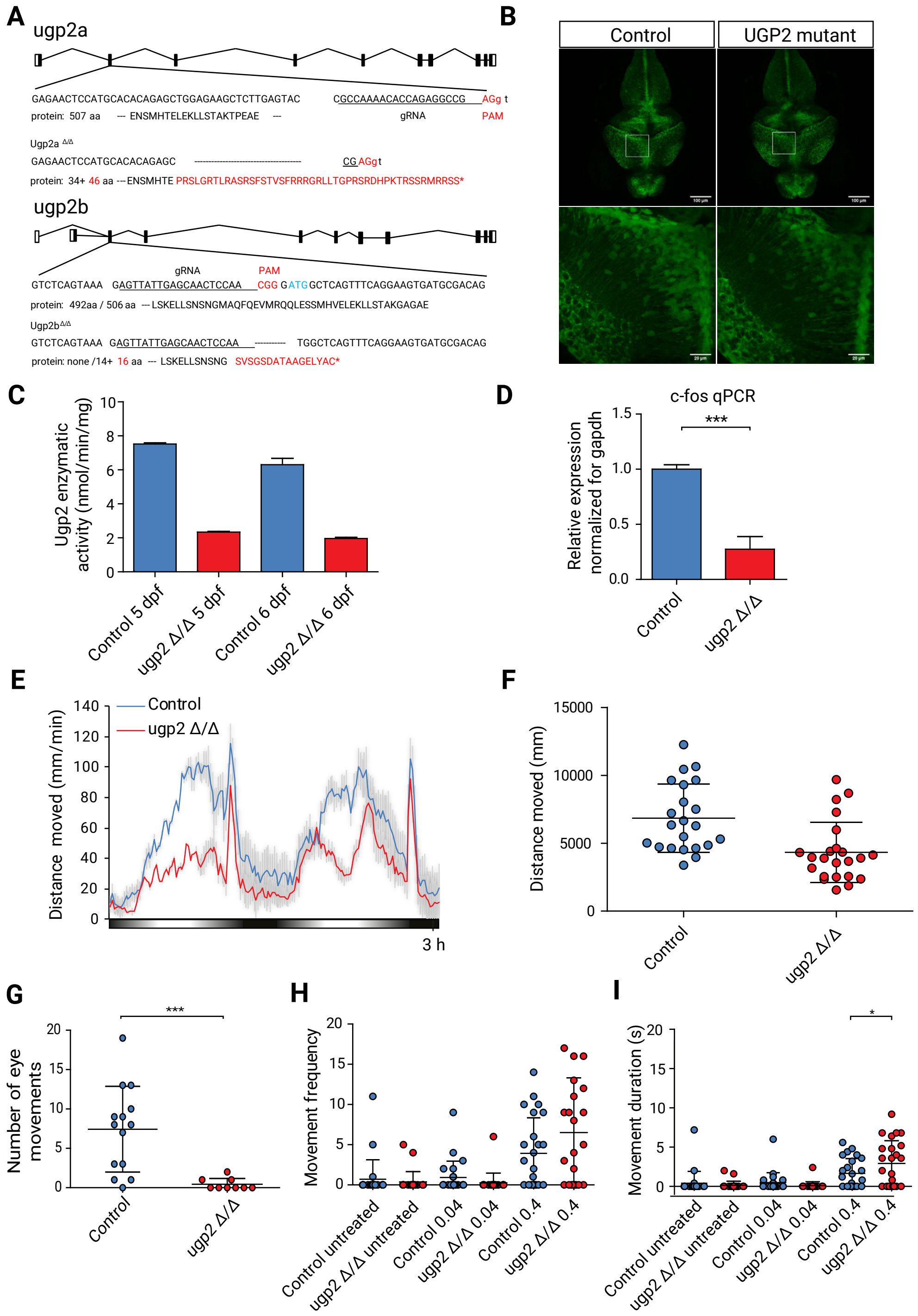
Figure 5 | Metabolic changes upon UGP2 loss. A) UGP2 enzymatic activity in WT, UGP2 KI, KO and KO NSCs rescued with wildtype or mutant Met12Val isoform 1 of UGP2. Bar plot showing the mean of two replicate experiments, error bar is SEM. *=p<0.05; ***=p<0.001, unpaired t-test, two-tailed. **B)** Representative pictures of PAS staining in WT, KI, KO and rescue NSCs. Nuclei are counterstained with hematoxylin (blue). Inserts show zoom-in of part of the cytoplasm. Note the presence of glycogen granules in WT NSCs, their diminished number in KI NSCs, their absence in KO NSCs and their reappearance upon rescue with both wild type long UGP2 as with Met12Val long UGP2. **C)** Quantification of the number of glycogen granules per cell in WT, UGP2 KI, KO and rescue NSCs, after 48 hours culture under low-oxygen conditions. Shown is the average number of glycogen granules per cell, n=80-100 cells per genotype. Error bars represent the SD. ***=p<0.001, unpaired t-test, two-tailed. **D)** Western blotting detecting LAMP2 (upper panel) and the housekeeping control ACTIN (lower panel) in cellular extracts from ESC-derived NSCs, that are wt, UGP2 KI, KO and KO cells rescued with either the long wildtype isoform 1 or the mutant Met12Val isoform 1. Glycosylated LAMP2 runs at ~110 kDa, whereas hypo-glycosylated LAMP2 is detected around 75 kDa. The absence of detectable changes in LAMP2 glycosylation in KI cells is likely explained by a non-complete isoform switch upon in vitro NSC differentiation, resulting in residual UGP2 levels (c.p. Supplementary Figure 5D). **E)** qRT-PCR expression analysis for UPR marker genes (spliced XBP1, HSPA5, ATF4 and EDEM) in WT, KI, KO and rescue NSCs. Shown is the mean fold change for the indicated genes compared to wild type, normalized for the housekeeping gene TBP. Results of two biological and two independent technical replicates are plotted, from two experiments. Error bars represent SEM; *=p<0.05; **=p<0.01,***=p<0.001, unpaired t-test, two-tailed.

Absence of short UGP2 isoform leads to metabolic defects in neural stem cells

To investigate how reduced UGP2 expression levels in KO and KI cells would impact on NSC metabolism, we investigated the capacity to produce UDP-glucose in the presence of exogenously supplied glucose-1-phospate and UTP. KO NSCs showed a severely reduced ability to produce UDP-glucose (**Figure 5A)**. This reduction was rescued by ectopic overexpression of both long wild type and long mutant UGP2. KI cells showed a slightly reduced activity in ESCs (**Supplementary Figure 7A**), but a more strongly reduced activity in NSCs compared to wild type (**Figure 5A**), correlating with total UGP2 expression levels (**Figure S4D, E**). Surprisingly, contrary to KO NSCs, KO ESC showed some residual capacity to produce UDP-glucose despite the complete absence of UGP2 (**Figure S7A**). This could indicate that a yet to be identified enzyme can partially take over the function of UGP2 in ESCs but not NSCs, which might explain the lack of expression changes in this cell type upon UGP2 loss. iPSCs showed similar results (**Figure S7B**). We next assessed the capacity to synthesize glycogen under low oxygen conditions by PAS staining, as it was previously shown that hypoxia triggers increased glycogen synthesis[69]. As expected, wild type ESCs cultured for 48 hours under hypoxia showed an intense cytoplasmic PAS staining in most cells (**Figure S7C, D**), while KO ESCs showed a severely reduced staining intensity. This indicates that under hypoxia conditions, the residual capacity of ESC to produce UDP-glucose in the absence of UGP2 is insufficient to produce glycogen. KI ESCs were indistinguishable from wild type (**Figure S7D)**. At the NSC state, many KO cells kept at low oxygen conditions for 48 hours died (data not shown) and those KO cells that did survive were completely depleted from glycogen granules (**Figure 5B, C**). This could be rescued by overexpression of both wild type or mutant long UGP2 isoform. KI NSCs showed a more severe reduction in PAS staining compared to the ESC state (**Figure 5B, C**), and we observed similar findings in patient iPSC derived NSCs (**Figure S7E**). Together, this further indicates that upon neural differentiation the isoform expression switch renders patient cells depleted of UGP2, leading to a reduced capacity to synthesize glycogen. This can directly be involved in the DEE phenotype, as, besides affecting energy metabolism, reduction of glycogen in brain has been shown to result in I) impairment of synaptic plasticity[70]; II) reduced clearance of extracellular potassium ions leading to neuronal hypersynchronization and seizures[71-73]; and III) altered glutamate metabolism[74]. To investigate how reduced UDP-glucose levels would impact on glycosylation, we next, investigated glycosylation levels by means of LAMP2, a lysosomal protein known to be extensively glycosylated both by N-linked and O-linked glycosylation[75]. We found that KO NSCs show hypoglycosylation of LAMP2 that is rescued by the over expression of both WT and mutant long isoform (**Figure 5D**). In contrast, in ESCs no glycosylation defects were noticed (**Figure S7F**). Finally, we investigated whether the absence of UGP2, affecting protein glycosylation, could induce ER stress and thus unfolded protein response (UPR). Whereas in ESCs, the absence of UGP2 did not result in a detectable effect on UPR markers (**Figure S7G**), in NSCs we noticed an increased expression of these genes both in KO and in KI cells (**Figure 5E**). This indicates that NSCs having UGP2 levels under a certain threshold are more prone to ER-stress and UPR. In agreement with this, we did not observe upregulation of UPR markers in patient derived fibroblast, which have similar total UGP2 expression levels compared to controls (**Figure S7H**). Together this indicates that upon differentiation to NSCs, KI cells become sufficiently depleted of UGP2 to have reduced synthesis of UDP-glucose, leading to defects in glycogen synthesis and protein glycosylation and to the activation of UPR response. Alterations of these crucial processes are likely to be implicated in the pathogenesis leading to increased seizure susceptibility, altered brain microstructure and progressive microcephaly.

Ugp2a and Ugp2b double mutant zebrafish recapitulate metabolic changes during brain development, have an abnormal behavioral phenotype, visual disturbance, and increased seizure susceptibility

Finally, to model the consequences of the lack of UGP2 *in vivo*, we generated zebrafish mutants for both *ugp2a* and *ugp2b*, the zebrafish homologs of *UGP2*, using CRISPR-Cas9 injections in fertilized oocytes in a background of a radial glia/neural stem cell reporter[76]. Double homozygous mutant lines having frameshift deletions for both genes confirmed by Sanger sequencing could be generated but the only viable combination, obtained with *ugp2a* loss, created a novel ATG in exon 2 of ugp2b, leading to a hypomorphic allele (**Figure 6A**). Homozygous *ugp2a/b* mutant zebrafish had a normal gross morphology of brain and radial glial cells (**Figure 6B**), showed a largely diminished activity to produce UDP-glucose in the presence of exogenously supplied glucose-1-phospate and UTP (**Figure 6C**), and showed a reduction in c-FOS expression levels, indicating reduced global neuronal activity (**Figure 6D**). To monitor possible spontaneous seizures, we performed video tracking experiments of developing larvae under light-dark cycling conditions at 5 days post fertilization (dpf). Control larvae show increased locomotor activity under light conditions, and although *ugp2* double mutant larvae still responded to increasing light conditions, they showed a strongly reduced activity (**Figure 6E, F**). This could indicate that

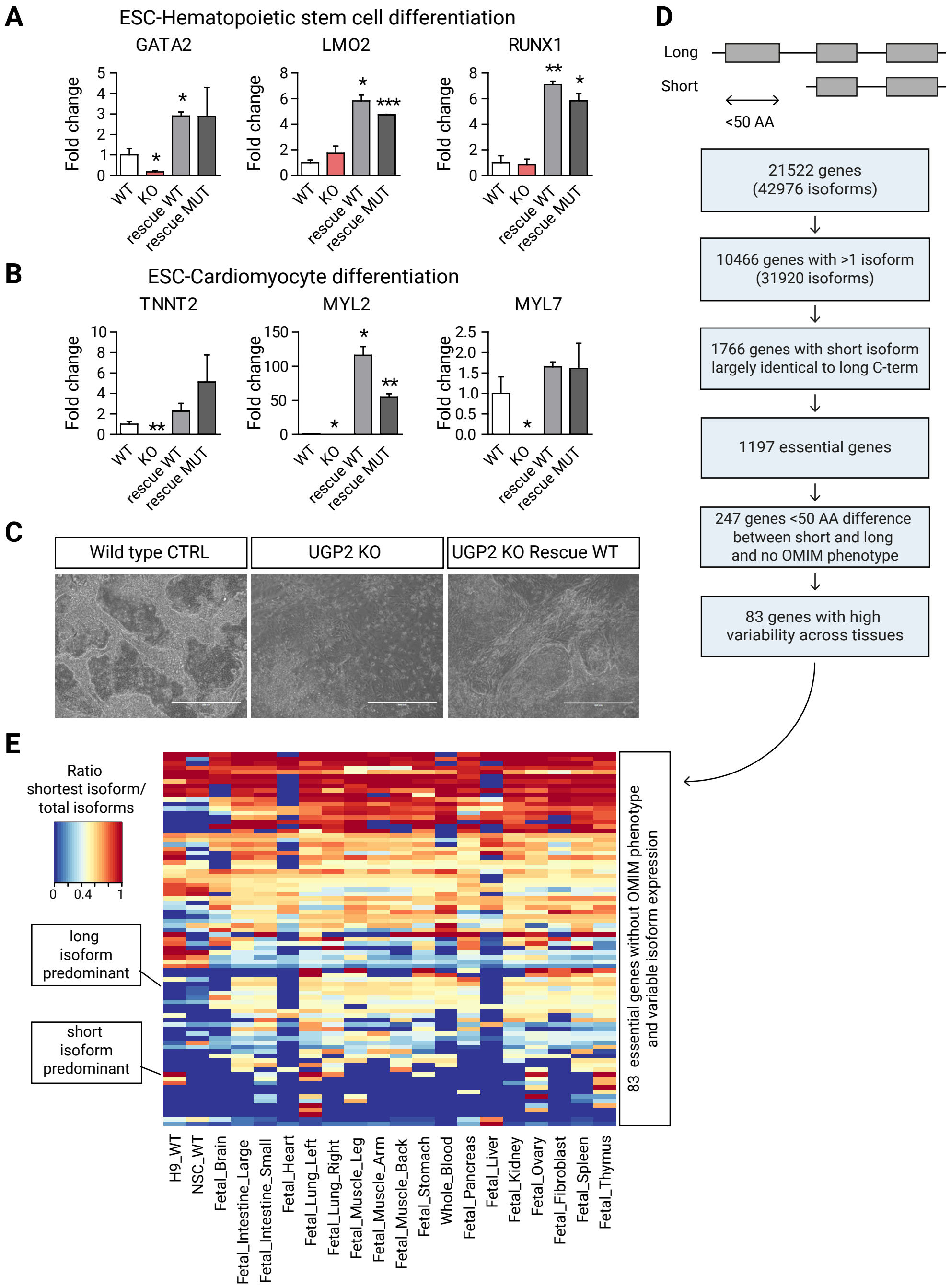


their capability to sense visual cues is diminished, or that their tectal processing of visual input is delayed, resulting in reduced movements. Strikingly, upon careful inspection, we noticed that *ugp2* double mutant larvae did not show spontaneous eye movements, in contrast to age-matched control larvae (**Figure 6G, Supplemental Movie 1 and 2**). Whereas we did not observe an obvious spontaneous epilepsy phenotype in these double mutant larvae, upon stimulation with 4-aminopyridine (4-AP), a potent convulsant, double mutant larvae showed an increased frequency and duration of movements at high velocity compared to controls, which might indicate an increased seizure susceptibility (**Figure 6H, I**). Taken together, severely reduced Ugp2a/Ugp2b levels result in a behavior defect with reduced eye movements, indicating that also in zebrafish Ugp2 plays an important role in brain function.

UGP2 is an essential gene in humans and ATG mutations of tissue specific isoforms of essential genes potentially cause more rare genetic diseases

Several lines of evidence argue that UGP2 is essential in humans. First, no homozygous LoF variants or homozygous exon-covering deletions for *UGP2* are

Figure 6 | Zebrafish disease modelling. A) Schematic drawing of the ugp2a and ugp2b loci in zebrafish and the generated mutations indicated. **B)** Confocal images (Maximum projection of confocal Z-stacks) of the brain of wild type (left) and ugp2aΔ/Δ; ugp2bΔ/Δ mutant zebrafish larvae (right), both in an slc1a2b-citrine reporter background, at 4 days post fertilization (dpf). IThe lower panels are higher magnifications of the boxed regions indicated in the upper panels. Scale bar in upper panel is 100 µm, in lower panel 20 µm. In upper panel, Z = 45 with step size 4 µm; In lower panel, Z = 30 with step size 2 µm. **C)** Enzymatic activity in ugp2 double mutant zebrafish larvae at 4 and 5 dpf, compared to wild type age matched controls, showing reduced Ugp2 enzyme activity in double mutant zebrafish. **D)** qRT-PCR for the neuronal activity marker c-FOS in wild type and ugp2 double mutant larvae at 3 dpf. For each group, 2 batches of 12 larvae were pooled. Shown is the mean fold change for the indicated genes compared to wild type, normalized for the housekeeping gene gapdh. Error bars represent SEM; ***= p<0.001, unpaired t-test, two-tailed. **E)** Representative graph of a locomotion assay showing the total distance moved by larvae during the dusk-dawn routine (total time: 3 hr 12 min), n = 24 larvae per genotype. Grey shading shows the standard error of the mean. **F)** Quantification of the total distance moved throughout the experiment from E) excluding the dark period. **G)** Quantification of the number of observed spontaneous eye movements during a 2 minutes observation in wild type and ugp2 double mutant larvae at 4 dpf. Each dot represents one larva; shown is the average and SD; ***p<0.001, t-test, two tailed. **H)** Quantification of the frequency of movements at a speed of > 15 mm/s, for wild type control and ugp2 double mutant zebrafish larvae at 4 dpf, treated with mock control or with 0.04 nM or 0.4 nM 4-AP during a 35 minutes observation. Each dot represents a single larva; results of two experiments are shown, with in total 24 larvae per condition. **I)** As H, but now assessing movement duration at a speed of > 15 mm/s. * =p<0.05, two way ANOVA with Bonferoni post-test.



present in *gnomAD* or *GeneDx* controls, and homozygous variants in this gene are limited to non-coding changes, synonymous variants and 5 missense variants, together occurring only 7 times homozygous (**Supplementary Table 5**). Also, no homozygous or compound heterozygous *UGP2* LoF variants were found in published studies on dispensable genes in human knockouts[77-79], or in the *Centogene* (*CentoMD*) or *GeneDx* patient cohorts, encompassing together many thousands of individuals, further indicating that this gene is intolerant to loss-of-function in a bi-allelic state. In addition, no homozygous deletions of the region encompassing *UGP2* are present in DECIPHER[40] or ClinVar[37]. Second, *UGP2* has been identified as an essential gene using gene-trap integrations[80] and in CRISPR-Cas9 LoF screens in several human cell types[81-85]. Finally, studies in yeast [86,87], fungus[88] and plants[89-91] consider the orthologs of *UGP2* as essential, and the absence of *Ugp2* in mice is predicted to be lethal[92]. In flies, homozygous UGP knock-outs are lethal while only hypomorphic compound heterozygous alleles are viable but have a severe movement defect with altered neuromuscular synaptogenesis due to glycosylation defects[93]. To further investigate the essentiality of UGP2, we performed differentiation experiments of our WT, KO and rescue ESCs. Differentiation of KO ESCs into hematopoietic stem cells (HSCs) resulted in severe downregulation of *GATA2* compared to wild type cells, and this was restored in rescue cell lines (**Figure 7A**). GATA2 is a key transcription factor in the developing blood system, and knockout of *Gata2* is embryonic lethal in mice due to defects in HSC generation and maintenance[94,95]. Differentiation of ESCs into cardiomyocytes similarly affected key marker gene expression in KO cells, and these changes were restored upon UGP2 rescue (**Figure 7B, C**). Whereas WT ESCs

Figure 7 | Essentiality of UGP2 and other disease candidate genes with a similar mutation mechanism. **A)** qRT-PCR analysis of the hematopoietic stem cell markers GATA2, LMO2 and RUNX1, after 12 days of differentiation of wild type, UGP2 KO and UGP2 KO rescue ESCs. Shown is the mean fold change for the indicated genes compared to wild type, normalized for the housekeeping gene TBP. Results of two biological and two technical replicates are plotted. Error bars represent SEM; $*=p<0.05$; $**=p<0.01$,$***=p<0.001$, unpaired t-test, two-tailed. **B)** As A), but now for cardiomyocyte differentiation at day 15, assessing expression of the cardiomyocyte markers TNNT2, MYL2 and MYL7. **C)** Bright-field image of cardiomyocyte cultures of wild type, UGP2 KO and rescue cells. Note the elongated organized monolayer structures cardiomyocytes capable of beating in wild type and rescue cells, that are absent in KO cultures. Scale bar is 400 µm. **D)** Scheme showing the homology search to identify genes with a similar structure as UGP2, where ATG altering mutations could affect a tissue specific isoform causing genetic disease. **E)** Heat map showing the ratio of short isoform expression over total isoform expression from published RNA-seq data amongst 20 tissues for 83 out 247 essential genes that are not yet implicated in disease and in which the short and longer protein isoform differ by less than 50 amino acids at the N-terminal.

could generate beating cardiomyocytes after 10 days, these were not seen in KO ESCs. Taken together this argues that the complete absence of UGP2 in humans is probably incompatible with life, a hypothesis that cannot be tested directly. However, if true, this could well explain the occurrence of the unique recurrent mutation in all cases presented herein. Given the structure of the *UGP2* locus (**Figure 2A**), every LoF variant would affect either the long isoform, when located in the first 33 nucleotides of the cDNA sequence, or both the short and long isoform when downstream to the ATG of the short isoform. Therefore, the short isoform start codon is the only mutational target that can disrupt specifically the short isoform. In this case, the Met12Val change introduced into the long isoform does not seem to disrupt UGP2 function to such an extent that this is intolerable and therefore allows development to proceed for most tissues. However, the lack of the short UGP2 isoform caused by the start codon mutation results in a depletion of functional UGP2 in tissues where normally the short isoform is predominantly expressed. In brain this reduction diminishes total UGP2 levels below a threshold for normal development, causing a severe epileptic encephalopathy syndrome. Given the complexity of the human genome with 42,976 transcripts with RefSeq peptide IDs, perhaps also other genetic disorders might be caused by such tissue restricted depletion of essential proteins. Using a computational homology search of human proteins encoded by different isoforms, we have identified 1,766 genes that share a similar structure to the *UGP2* locus (e.g. a shorter protein isoform that is largely identical to the longer protein isoform, translated from an ATG that is contained within the coding sequence of the long isoform) (**Figure 7D**). When filtering these genes for 1) those previously shown to be essential[6], 2) not associated with disease (e.g. no OMIM phenotype) and 3) those proteins where the shorter isoform is no more than 50 amino acids truncated at the N-terminal compared to the longer isoform, we identified 247 genes (**Supplementary Table 6**). When comparing the ratios of isoform specific reads obtained from different fetal RNA-seq data[48-51] we noticed that many of these genes show differential isoform expression amongst multiple tissues, with many genes showing either expression of the long or the short isoform in a particular tissue (**Figure 7E**). Homozygous LoF variants or start codon altering mutations in these genes are rare in *gnomAD* (**Supplementary Table 7**), and it is tempting to speculate that mutations in start codons of these genes could be associated with human genetic diseases, as is the case for *UGP2*. Using mining of data from undiagnosed patients from our own exome data base, the Queen Square Genomic Center database and those from *Centogene* and *GeneDx*, we found evidence for several genes out of the 247 having rare, bi-allelic variants affecting the start codon of one of the isoforms that could be implicated in novel disorders (*unpublished observations*) and give one such example in the **Supplementary Note**. Together, these findings highlight the relevance of mutations resulting in tissue-specific protein loss of essential genes for genetic disorders.

Discussion

Here we describe a recurrent variant in 19 individuals from 12 families, affecting the start codon of the shorter isoform of the essential gene *UGP2* as a novel cause of a severe DEE. Using *in vitro* and *in vivo* disease modeling, we provide evidence that the reduction of *UGP2* expression in brain cells leads to global transcriptome changes, a reduced ability to produce glycogen, alterations in glycosylation and increased sensitivity to ER stress, which together can explain the phenotype observed in the patients. Most likely our findings *in vitro* underestimate the downstream effects in patient cells, as in fetal brain the longer isoform expression is almost completely silenced and virtually all UGP2 comes from the shorter isoform, which in patient cells cannot be translated. During our *in vitro* NSC differentiation this isoform switch is less complete, leaving cells with the patient mutation with some residual UGP2. Strikingly, the clinical phenotype seems to be very similar in all cases, including intractable seizures, absence of developmental milestones, progressive microcephaly and a disturbance of vision, with retinal pigment changes observed in all patients who had undergone ophthalmological examination. Also, all patients seem to share similar, although mild, dysmorphisms, possibly making this condition a recognizable syndrome.

The involvement of *UGP2* in genetic disease is surprising. Given its central role in nucleotide-sugar metabolism it is expected that loss of this essential protein would be incompatible with life, and therefore loss-of-function should not be found in association with postnatal disease. Our data argue that indeed a total absence of UGP2 in all cells is lethal, but that tissue-specific loss, as caused here by the start codon alteration of an isoform important for brain, can be compatible with postnatal development but still results in a severe phenotype. Given that any other LoF variant across this gene would most likely affect both protein isoforms, this could also explain why only a single mutation is found in all individuals. The fact that the Met12Val long isoform was able to rescue the full KO phenotype indicates that the missense change introduced to the long protein isoform does not affect UGP2 function. As other variants at this start codon, even heterozygous, are not found, possibly missense variants encoding for leucine, lysine, threonine, arginine or isoleucine (e.g. amino acids that would be encoded by alternative changes affecting the ATG codon) at this amino acid location in the long isoform could not produce a functional protein and are therefore not tolerated. Although start codon mutations have previously been implicated in disease[96,97], there are no reports, to our knowledge, on disorders describing start codon alterations of other essential genes, leading to alterations of tissue specific isoforms. Using a genome-wide homology search, we have identified a large list of other essential genes with a similar locus structure and variable isoform expression amongst tissues, where similar ATG altering variants could affect tissue-relevant expression. An intriguing question is why evolution has resulted in a large number of genes encoding almost identical protein isoforms. It will be interesting to further explore the mutational landscape of these genes in cohorts of currently unexplained patients.

Experimental procedure

Patient recruitment

All affected probands were investigated by their referring physicians and all genetic analysis was performed in a diagnostic setting. Legal guardians of affected probands gave informed consent for genomic investigations and publication of their anonymized data.

Next generation sequencing of index patients

Individual 1: Genomic DNA was isolated from peripheral blood leukocytes of proband and both parents and exome-coding DNA was captured with the Agilent Sure Select Clinical Research Exome (CRE) kit (v2). Sequencing was performed on an Illumina HiSeq 4000 with 150 bp paired end reads. Reads were aligned to hg19 using BWA (BWA-MEM v0.7.13) and variants were called using the GATK haplotype caller (v3.7 (reference: http://www.broadinstitute.org/gatk/)[98]. Detected variants were annotated, filtered and prioritized using the Bench lab NGS v5.0.2 platform (Agilent technologies). Initially, only genes known to be involved in epilepsy were analyzed, followed by a full exome analysis revealing the homozygous UGP2 variant

Individuals 2, 3 and 4: Using genomic DNA from the proband and parents (individual 4) or the proband, parents, and affected sibling (individual 2 and 3), the exonic regions and flanking splice junctions of the genome were captured using the SureSelect Human All Exon V4 (50 Mb) (individual 4) or the IDT xGen Exome Research Panel v1.0 (individual 2 and 3). Massively parallel (NextGen) sequencing was done on an Illumina system with 100 bp or greater paired-end reads. Reads were aligned to human genome build GRCh37/UCSC hg19, and analyzed for sequence variants using a custom-developed analysis tool. Additional sequencing technology and variant interpretation protocol has been previously described[99]. The general assertion criteria for variant classification are publicly available on the GeneDx ClinVar submission page (http://www.ncbi.nlm.nih.gov/clinvar/submitters/26957/)

Individual 5: Diagnostic exome sequencing was done at the Departments of Human Genetics of the Radboud University Medical Center Nijmegen, The Netherlands and performed essentially as described previously[100].

Individual 6, 7, 8, 9, 10, 15, 16, 17, 18 and 19: After informed consent, we collected blood samples from the probands, their parents and unaffected siblings, and extracted DNA using standard procedures. To investigate the genetic cause of the disease, WES was performed in the affected proband. Nextera Rapid Capture Enrichment kit (Illumina) was used according to the manufacturer instructions. Libraries were sequenced in an Illumina HiSeq3000 using a 100-bp paired-end reads protocol. Sequence alignment to the human reference genome (UCSC hg19), and variants calling, and annotation were performed as described elsewhere[101]. After removing all synonymous changes, we filtered single nucleotide variants (SNVs) and indels, only considering exonic and donor/acceptor splicing variants. In accordance with the pedigree and phenotype, priority was given to rare variants [<1% in public databases, including 1000 Genomes project, NHLBI Exome Variant Server, Complete Genomics 69, and Exome Aggregation Consortium (ExAC v0.2)] that were fitting a recessive or a de novo model.

Individual 11 and 14: Whole exome sequencing was performed at CENTOGENE AG, as previously described[102].

Individual 12 and 13: High quality DNA was used to capture exomic sequences using the SureSelect kit (Agilent, Santa Clara, CA, US). Then genomic libraries were created according to manufacturer's protocols. Sequences were read on Proton (Life Technologies Inc., Carlsbad, CA, US). Downstream analyses such as sequence alignment, indexing, raw variant calling were done using publicly and commercially available tools such as Ion Reporter, SAMTools, and Genomic Analysis ToolKit. Moreover, variant interrogations were done using sequence-variant databases, such as dbSNP, Ensembl, and the National Heart, Lung, and Blood Institute (NHLBI) Exome Variant Server (EVS), 1000 genome project.

Human brain samples

Tissue was obtained, upon informed consent, and used in a manner compliant with the Declaration of Helsinki and the Research Code provided by the local ethical committees. Fetal brains were preserved after spontaneous or induced abortions with appropriate maternal written consent for brain autopsy and use of rest material for research. We performed a careful histological and immunohistochemical analysis and evaluation of clinical data (including genetic data, when available). We only included specimens displaying a normal cortical structure for the corresponding age and without any significant brain pathology.

Brain tissue immunohistochemistry

For immunohistochemical analysis, we used 2 cases from the first trimester (GW6 and GW9), 4 cases from the second trimester (GW21, GW23, GW24 and GW26) and 2 cases from the third trimester (GW33 and GW36). Anatomical regions were determined according to the atlas of human brain development[103-106]. We cut 4 µm sections from formalin-fixed, paraffin embedded whole fetuses (GW6 and GW9) and brain tissue from cerebral, mesencephalic, cerebellar and brain stem region (from GW21 to GW36). Slides were stained with mouse anti-UGP2 (C-6) in a 1:150 dilution (Santa Cruz) and visualized using Mouse and Rabbit Specific HRP/DAB (ABC) Detection IHC kit (Abcam). Mayer's hematoxylin was used as a counterstain for immunohistochemistry followed by mounting and coverslipping (Bio-Optica) for slides. Prepared slides were analyzed and scanned under a VisionTek® Live Digital Microscope (Sakura).

Cloning of UGP2 cDNA

RNA was isolated using TRI reagent (Sigma) from whole peripheral blood of index patient 1 and her parents, after red blood cell depletion with RBC lysis buffer (168mM NH_4Cl, 10mM $KHCO_3$, 0.1mM EDTA). cDNA was synthesized following the iSCRIPT cDNA Synthesis Kit (Bio-Rad) protocol, and the coding sequence of the long and short UGP2 isoform (wild type or mutant) was PCR-amplified together with homology arms for Gibson assembly (see **Supplementary Table 8** for primer sequences) using Phusion High-Fidelity DNA polymerase (NEB). PCR amplified DNA was then cloned by Gibson assembly as previously described[107] in a pPyCAG-IRES-puro plasmid (a kind gift of Ian Chambers, Edinburgh) opened with EcoRI for experiments in mammalian cells. All obtained plasmids were sequenced verified by Sanger sequencing (complete plasmid sequences available upon request).

Fibroblast cell culture

Fibroblasts from index patient 1 and her parents were obtained using a punch biopsy according to standard procedures, upon informed consent (IRB approval MEC-2017-341). Fibroblasts from the parents of index patient 2 and 3 were also obtained upon informed consent at McMaster Children's Hospital. All fibroblasts were cultured in standard DMEM medium supplemented with 15% Fetal calf serum, MEM Non-Essential amino acids (Sigma), 100 U/ml penicillin and 100 µg/ml streptomycin, as done previously[108], in routine humidified cell culture incubators at 20% O2. Fibroblast cell lines were transfected using Lipofectamine 3000 (Invitrogen) with the indicated plasmid constructs. All the cell lines used in this report were regularly checked for the presence of mycoplasma and were negative during all experiments.

Genome engineering in human embryonic stem cells

H9 human embryonic stem cells were cultured as previously described[107,109]. In short, cells were maintained on feeder free conditions in mTeSR-1 medium (STEMCELL technologies) on Matrigel (Corning) coated culture dishes. To engineer the patient specific UGP2 mutation by homologous recombination[110], ESC were transfected using Lipofectamine 3000 with a plasmid expressing eSpCas9-t2a-GFP (a kind gift of Feng Zhang) and a gRNA targeting the UGP2 gene (see **Supplementary Table 8** for the sequence), together with a 60 bp single stranded oligonucleotide (ssODN) homology template encoding the patient mutation (synthesized at IDT). To increase the stability of the ssODN and therefore homologous recombination efficiency, the first two 5' and 3' nucleotides were synthesized using phosphorothioate bonds[111]. 48 hours post transfection, GFP expressing cells were sorted, and 6000 single GFP-positive cells were plated on a Matrigel coated 6-well plate in the presence of 10µM ROCK-inhibitor (Y27632, Millipore). After approximately 10 days, single colonies where manually picked, expanded and genotyped using Sanger sequencing (see **Supplementary Table 8** for primer sequences). As a by-product of non-homologous end joining, knock-out clones were identified which showed a single nucleotide A insertion at position 42 of UGP2 transcript 1 (chr2:64083462_64083463insA), leading to an out of frame transcript and a premature termination of the protein at amino acid position 47 (D15Rfs*33). Western blotting confirmed the absence of all UGP2 protein in knock-out clones and the loss of the short UGP2 isoform in clones with the patient mutation. To produce a stable rescue cell line, ESC cells were transfected as previously described with the pPyCAG-IRES-puro plasmid expressing either the long WT or mutant UGP2 isoform. After 48 hours, the population of cells with the transgene integration was selected with 1µg/ml puromycin. Engineered ESC clones had a normal colony morphology and pluripotency factor expression.

Patient specific Induced pluripotent stem cell generation

Patient fibroblast cell lines were reprogrammed using the CytoTune™-iPS 2.0 Sendai Reprogramming Kit (Thermo Scientific, A16517) expressing the reprogramming factors OCT4, SOX2, KLF4 and C-MYC on matrigel coated cell culture plates, upon informed consent (IRB approval MEC-2017-341). After approximately 4-5 weeks, emerging colonies were manually picked and expanded. Multiple clones were assessed for their karyotype, pluripotency factor expression and three lineage differentiation potential (Stem Cell Technologies, #05230), following the routine procedures of the Erasmus MC iPS Cell facility, as previously described[108]. Sanger sequencing was used to verify the genotype of each obtained iPSC line. We used three validated clones for each individual in our experiments.

Neural stem cell differentiation

Pluripotent cells were differentiated in neural stem cells (NSCs), using a modified dual SMAD inhibition protocol[112]. In short, 18000 cells/cm^2 were plated on matrigel coated cell culture dishes in mTeSR-1 medium in the presence of 10µM Y27632. When cells reached 90% confluency, the medium was switched to differentiation medium (KnockOut DMEM (Gibco), 15% KnockOut serum replacement (Gibco), 2mM L-glutamine (Gibco), MEM Non-Essential amino acids (Sigma), 0.1 mM β-mercaptoethanol, 100U/ml penicillin and 100 µg/ml streptomycin) supplemented with 2µM A 83-01 (Tocris) and 2µM Dorsomorphin (Sigma-Aldrich). At day 6, medium was changed to an equal ratio of differentiation medium and NSC medium (KnockOut DMEM-F12 (Gibco), 2mM L-glutamine (Gibco), 20ng/ml bFGF (Peprotech), 20ng/ml EGF (Peprotech), 2% StemPro Neural supplement (Gibco), 100U/ml penicillin and 100µg/ml streptomycin) supplemented with 2µM A 83-01 (Tocris) and 2µM Dorsomorphin (Sigma-Aldrich). At day 10, cells were passaged (NSC p=0) using Accutase (Sigma) and maintained in NSC medium. We used commercially available H9-derived NSCs (Gibco) as a control (a kind gift of Raymond Poot, Rotterdam).

Other stem cell differentiation experiments

ESCs were differentiated into hematopoietic stem cells and cardiomyocyte using commercially available STEMCELL technologies kits (STEMdiff Hematopoietic kit #05310, STEMdiff Cardiomyocyte differentiation kit #05010) according to manufacturer's instructions. Cells were finally harvested and lysed with TRI reagent to isolate RNA for further RT-qPCR analysis.

RNA-sequencing and data analysis

For RNA-seq on blood derived patient RNA, peripheral blood was obtained from index patient 1 and her parents, collected in PAX tubes and RNA was isolated following standard diagnostic procedures in the diagnostics unit of the Erasmus MC Clinical Genetics department. RNA-seq occurred in a diagnostic setting, and sequencing was performed at GenomeScan (Leiden, The Netherlands). For RNA-seq of in vitro cultured cell lines, RNA was obtained from 6-well cultures using TRI reagent, and further purified using column purification (Qiagen, #74204). mRNA capture, library prep including barcoding and sequencing on an Illumina HiSeq2500 machine were performed according to standard procedures of the Erasmus MC Biomics facility. Approximately 20 million reads were obtained per sample. For the cell line experiments, two independent H9 wild type cultures, two independent knock-out clones harboring the same homozygous UGP2 genetic alteration and two independent clones harboring the patient homozygous UGP2 mutation were used. Each cell line was sequenced in two technical replicates at ESC state and differentiated NSC state (at passage 5). FASTQ files obtained after de-multiplexing of single-end, 50 bp sequencing reads were trimmed by removing possible adapters using Cutadapt after quality control checks on raw data using the FastQC tool. Trimmed reads were aligned to the human genome (hg38) using the HISAT2 aligner[113]. To produce Genome Browser Tracks, aligned reads were converted to bedgraph using bedtools genomecov, after which the bedGraphToBigWig tool from the UCSC Genome Browser was used to create a bigwig file. Aligned reads were counted for each gene using htseq-count[114] and GenomicFeatures[115] was used to determine the gene length by merging all non-overlapping exons per gene from the Homo_sapiens.GRCh38.92.gtf file (Ensemble). Differential gene expression and RPKM (Reads Per Kilobase per Million) values were calculated using edgeR[116] after removing low expressed genes and normalizing data. The threshold for significant differences in the gene expression was FDR < 0.05. To obtain a list of ESC and NSC reference genes used in Supplementary Figure 6F, we retrieved genes annotated in the following GO terms using GSEA/MSigDB web site v7.0: GO_FOREBRAIN_NEURON_DEVELOPMENT (GO:0021884), GO_CEREBRAL_CORTEX_DEVELOPMENT (GO:0021987), GO_NEURAL_TUBE_DEVELOPMENT (GO:0021915), BHATTACHARYA_EMBRYONIC_STEM_CELL (PMID: 15070671) and BENPORATH_NOS_TARGETS (PMID: 18443585).

Functional enrichment analysis

Metascape[117], g:profiler[118] and Enrichr[119] were used to assess functional enrichment of differential expressed genes. **Supplementary Table 4** reports all outputs in LogP, log(q-value) and Adjusted p-value (q-value) for Metascape and g:profiler, and in p-value, Adjusted p-value (q-value) and combined-score (which is the estimation of significance based on the combination of Fisher's exact test p-value and z-score deviation from the expected rank) for Enrichr. All tools were used with default parameters and whole genome set as background.

Genome-wide homology search

To make a genome-wide list of transcripts sharing a similar structure as UGP2 transcripts, 42976 transcripts from 21522 genes (Human genes GRCh38.p12) were extracted using BioMart of Ensembl (biomaRt R package). 11056 out of 21522 genes had only 1 transcript and the remaining 31920 transcripts from 10466 genes were selected, the protein sequences were obtained with biomaRt R package and homology analysis was performed using the NCBI`s blastp (formatting option: -outfmt=6) command line. We grouped longest and shorter transcript based on coding sequence length and only kept those that matched a pairwise homology comparison between the longest and the shorter transcript with the following criteria: complete 100 percent identity, without any gap and mismatch, and starting ATG codon of shortest transcript being part of the longest transcript(s). 1766 genes meet these criteria. We then filtered these genes for published essential genes[6], leaving us with 1197 genes. Using BioMart (Attributes: Phenotype description and Study external reference) of Ensembl we then evaluated the probability that these genes were implicated in disease and identified 850 genes that did not have an association with disease phenotype/OMIM number. Of those, 247 genes encoded proteins of which the shorter isoform differed less than 50 amino acids from the longer isoform. We chose this arbitrary threshold to exclude those genes where both isoforms could encode proteins differing largely in size and might therefore encode functionally completely differing proteins (although we cannot exclude that this will also hold true for some of the genes in our selection).

Differential isoform expression in fetal tissues

Publicly available RNA-seq data from various fetal tissue samples (**Supplementary Table 2**) were analyzed using the same workflow as described for the RNA-seq data analysis above. To determine differential isoform expression in these tissues, we calculated a ratio between the unique exon(s) of the shortest and longest transcript for each gene and assessed its variability across different fetal tissue samples. The number of reads for each unique exon of a transcript was calculated by mapping aligned RNA-seq reads against the unique exon coordinate using bedtools multicov. The longest and shortest transcripts were separated and the transcript ratio (number of counts of shortest transcript / (number of counts of shortest transcript + number of counts of longest transcript)) for each gene was obtained from the average reads of RNA-seq samples per tissue. 382 genes out of 1197 genes showed high variability across different samples (defined as a difference between highest and lowest ratio > 0.5), 277 of those high variable genes were not associated with a disease phenotype/OMIM number and of these 83 genes had a length less than 50 amino acids (a subset of the 247 genes with no OMIM and length less than 50 amino acids)

Haplotype Analysis

The 30 MB region surrounding UGP2 was extracted from exome sequencing VCF files to include both common and rare polymorphisms. Variants were filtered for a minimum depth of coverage of at least 10 reads and a genotype quality of at least 50. The filtered variants, were then used as input in PLINK (v1.07) with the following settings:

- homozyg-snp 5
- homozyg-kb 100
- homozyg-gap 10000
- homozyg-window-het 0

ROH around the UGP2 variant were identified in all 5 probands examined. The minimum ROH in common between all samples was a 5Mb region at chr2: 60679942-65667235. We note that targeted sequencing leads to uneven SNP density, so the shared ROH may, in fact, be larger or smaller. Next, we used recombination maps from deCODE to estimate the size of the region in centiMorgans (cM). We then used the region size in cM to estimate the time to event in generations using methods previously described[120].

qPCR analysis

RNA was obtained using TRI reagent, and cDNA prepared using iSCRIPT cDNA Synthesis Kit according to manufacturer's instructions. qPCR was performed using iTaq universal SYBR Green Supermix in a CFX96RTS thermal cycler (Bio-Rad). **Supplementary Table 8** summarizes all primers used in this study. Relative gene expression was determined following the ΔΔct method. To calculate the ratio of the short isoform, we performed absolute quantification as previously described[121]. Briefly, we performed qPCR on known copy numbers, ranging from 10^3 to 10^8 copies, of a plasmid containing the short UGP2 isoform (5' UTR included) using primers detecting specifically either the total or the short isoform. After plotting the log copy number versus the ct, we obtained a standard curve that we used to extrapolate the copy number of the unknown samples. To test for significance, we used Student's T-test and considered p<0.05 as significant.

Western blotting

Proteins were extracted with NE buffer (20mM Hepes pH 7.6, 1.5mM MgCl2, 350mM KCl, 0.2mM EDTA and 20% glycerol) supplemented with 0.5% NP40, 0.5mM DTT, cOmplete Protease Inhibitor Cocktail (Roche) and 150U/ml benzonase Protein concentration was determined by BCA (Pierce) and 20-50µg of proteins were loaded onto a 4–15% Criterion TGX gel (Bio-Rad). Proteins were then transferred to a nitrocellulose membrane using the Trans-Blot Turbo Transfer System (Bio-Rad). The membrane was blocked in 5% milk in PBST and subsequently incubated overnight at 4°C with primary antibody diluted in milk. After PBST washes, the membrane was incubated 1 hour at RT with the secondary antibody and imaged with an Odyssey CLX scanning system (Li-Cor). Band intensities were quantified using Image Studio (Li-cor). Antibodies used were: Ms-α-UGP2 (sc-514174) 1:250; Ms-α-Vinculin (sc-59803) 1:10000; Gt-α-actin (sc-1616) 1:500; Ms-α-LAMP2 (H4B4) 1:200; IRDye 800CW Goat anti-Mouse (926-32210) 1:5000; IRDye 680 Donkey anti-Goat (926-32224) 1:5000.

Zebrafish disease modelling

Animal experiments were approved by the Animal Experimentation Committee at Erasmus MC, Rotterdam. Zebrafish embryos and larvae were kept at 28°C on a 14–10-hour light–dark cycle in 1 M HEPES buffered (pH 7.2) E3 medium (34.8 g NaCl, 1.6 g KCl, 5.8 g CaCl$_2$ · 2H$_2$O, 9.78 g MgCl$_2$ · 6 H$_2$O). For live imaging, the medium was changed at 1 dpf to E3 + 0.003% 1-phenyl 2-thiourea (PTU) to prevent pigmentation. Ugp2a and ugp2b were targeted by Cas9/gRNA RNP-complex as we did before[76]. Briefly, fertilized oocytes from a tgBAC(slc1a2b:Citrine)re01tg reporter line[76] maintained on an TL background strain were obtained, and injected with Cas9 protein and crRNA and tracrRNA synthesized by IDT (Alt-R CRISPR-Cas9 System), targeting the open reading frame of zebrafish ugp2a and ugp2b. DNA was extracted from fin clips and used for genotyping using primers flanking the gRNA location (**Supplementary Table 8**) followed by sequencing. Mutants with a high level of out of frame indels in both genes were identified using TIDE[122] and intercrossed to obtain germ line transmission. Upon re-genotyping, mutant zebrafish with the following mutations as indicated in **Figure 6** were selected and further intercrossed. In this study, we describe two new mutant fish lines containing deletions in ugp2a (ugp2a$^{\Delta/\Delta}$) and ugp2b (ugp2b$^{\Delta/\Delta}$): ugp2a$^{re08/re08}$ containing a 37 bp deletion in exon 2 and ugp2b$^{re09/re09}$ containing a 5 bp deletion in exon 2. Intravital imaging, and analysis of eye movement, was performed as previously described[76]. Briefly, zebrafish larvae anesthetized in tricaine were mounted in low melting point agarose containing tricaine and imaged using a Leica SP5 intravital imaging setup with a 20×/1.0 NA water-dipping lens. To assess the locomotor activity of zebrafish larvae from 3 to 5 dpf, locomotor activity assays were performed using an infrared camera system (DanioVision™ Observation chamber, Noldus) and using EthoVision® XT software (Noldus) as described[76]. Briefly, control (n = 24) and ugp2a$^{\Delta/\Delta}$; ugp2b$^{\Delta/\Delta}$ (n = 24) zebrafish larvae, in 48 well plates, were subjected to gradually increasing (to bright light) and decreasing light conditions (darkness) as in Kuil et al[76]. Distance traveled (mm) per second was measured. For 4-AP (Sigma) stimulation animals were treated with 4-AP dissolved in DMSO 30 minutes before the onset of the experiments. For these experiments locomotor activity was measured over 35 minutes, with the first 5 minutes going from dark to light, followed by 30 minutes under constant light exposure.

Periodic acid- schiff (PAS) staining

ESCs or differentiated NSCs (wild type, KO, KI or rescue) were incubated under hypoxia conditions (3% O2) for 48 hours. Cells were fixed with 5.2% formaldehyde in ethanol, incubated 10 min with 1% Periodic acid, 15 min at 37°C with Schiff's reagent (Merck) and 5 min with Hematoxylin solution (Klinipath) prior to air drying and mounting. Every step of the protocol is followed by a 10 minutes wash with tap water. Imaging occurred on an Olympus BX40 microscope. Images were acquired at a 100x magnification, and ImageJ software was used for quantification. For ESCs, we used a minimum of 20 images per genotype for the quantification, containing on average 20 cells each, calculating the percentage of PAS positive area. For NSCs, we imaged between 80 to 100 cells per genotype, counting the number of glycogen granules in the cytoplasm. We report the average of two independent experiments at 48 hours low oxygen.

UGP2 enzymatic activity

The measurement of UGP2 enzyme activity was performed according to a modified GALT enzyme activity assay as described previously[123]. Frozen cell pellets were defrosted and homogenized on ice. 10 µl of each cell homogenate (around 0.5 mg protein/ml as established by BSA protein concentration determination) was pre-incubated with 10 µl of dithiothreitol (DDT) for 5 min at 25°C. 80 µl of a mixture of glucose-1-phosphate (final concentration 1 mM), UTP (0.2 mM), magnesium chloride (1 mM), glycine (125 mM) and Tris-HCl (pH8) (40 mM) was added and incubated for another 15 min at 25°C. The reaction was stopped by adding 150 µl of 3.3% perchloric acid. After 10 min on ice the mixture was centrifuged (10,000 rpm for 5 min at 4°C), the supernatant isolated and neutralized with ice cold 8 µl potassium carbonate for 10 min on ice. After centrifugation the supernatant was isolated and 1:1 diluted with eluent B (see below) after which the mixture was added to a MilliPore Amicon centrifugal filter unit. After centrifugation the supernatant was stored at -20°C until use. The separation was performed by injection of 10 µl of the defrosted supernatant onto a HPLC system with UV/VIS detector (wave length 262 nm) equipped with a reversed phase Supelcosil LC-18-S 150 mm x 4.6 mm, particle size 5 µm, analytical column and Supelguard LC18S guard column (Sigma-Aldrich). During the experiments the temperature of the column was maintained at 25°C. The mobile phase consisted of eluent A (100% methanol) and eluent B (50 mM ammonium phosphate buffer pH7.0 and 4 mM tetrabutylammonium bisulphate). A gradient of 99% eluent B (0-20 min), 75% eluent B (20-30 min) and 99% eluent B (30-45 min) at a flow rate of 0.5 m/min was used. The reaction product UDP-glucose was quantified using a calibration curve with known concentrations of UDP-glucose. UGP2 activity was expressed as the amount of UDP-glucose formed per mg protein per min. Experiments were performed in duplicate and for every cell line two independently grown cell pellets were used.

Immunostaining / Immunohistochemistry

For immuonofluorescence staining, cells were seeded on coverslips coated with 100µg/ml poly-D-lysine (Sigma) overnight. For ESC, coverslips were further coated with Matrigel (Corning) for one hour at 37°C. When cells reached about 70% confluency, they were fixed with 4% PFA for 15 min at RT. Cells were then permeabilized with 0.5% triton in PBS, incubated one hour in blocking solution (3% BSA in PBS) and then overnight at 4°C with the primary antibody diluted in blocking solution. The following day the coverslips were incubated one hour at room temperature in the dark with a Cy3-conjugated secondary antibody and mounted using ProLong Gold antifade reagent with DAPI (Invitrogen) to counterstain the nuclei. Images were acquired with a ZEISS Axio Imager M2 using a 63X objective.

Data availability

RNA-Seq of in vitro studies are publicly available through the National Center for Biotechnology

Information (NCBI) Gene Expression Omnibus (GEO) under accession number GSE137129. A token for reviewer access is present in the supplement. Due to privacy regulations and consent, raw RNA-seq data from patient blood cannot be made available. To retrieve tissue wide expression levels of UGP2, the GTEx Portal was accessed on 16/07/2019 (https://gtexportal.org/home/). RNA-seq data from various tissues were downloaded from various publications[48-51]. All publicly available data that were re-analyzed here are summarized in **Supplementary Table 2**.

AUTHOR CONTRIBUTIONS

EP performed molecular biology experiments, with help from AN and DP. HvdL, WB and TvH performed zebrafish work. PvdB and EHJ performed enzymatic analyses. IC performed brain immunohistochemistry and supplied tissues samples. EA supplied tissue samples. MG generated iPSCs. WvI and WGdV performed and SY analysed RNA-seq. SY performed gene homology search. Patient recruitment and diagnosing was performed in the different families as follows: Family 1: TSB, ASB, and EM phenotyped patient 1, MvS analysed WES; Family 2: LB and MK phenotyped patient 2 and 3, KGM, AB, KR analyzed WES; Family 3: JNK and JB phenotyped patient 4, KGM, AB, KR analyzed WES. Family 4: AaF, FaM, RM and FaA phenotyped patient 5, EJK analyzed WES; Family 5: FZ and NR phenotyped patient 6, SE, HH analyzed WES; Family 6, family 7 and family 8: MM, AE, ZK, FMD, MD, EGK phenotyped patients 7-10, JV, RM, HH analyzed WES; Family 9: JH phenotyped patient 11, KKK, ABA analysed WES; Family 10: MA, MAA, MS, MA, RA, LAQ, WQ, SC, KA, MHAH, SA, KA, AD, FA, DC, NK phenotyped patient 12 and 13, performed WES analysis and PGD; RT, KR, KKK, PB, ABA, RM, HH provided genetic data for population analysis. TSB identified patient 1, conceived the study, obtained funding, supervised the lab work and wrote the manuscript, with input from all main authors. All authors approved the final version of the manuscript.

ACKNOWLEDGEMENTS

We are indebted to the parents of the patients for their kind cooperation. We thank Virginie Verhoeven and Gerben Schaaf for critically reading our manuscript and Grazia Mancini for helpful discussions. We thank Gerben Schaaf for providing the LAMP2 antibody, and Eskeatnaf Mulugeta for bioinformatics advice. DP was supported by an Erasmus+ Traineeship Programme. HH is supported by the Rosetree Trust, Ataxia UK, MSA Trust, Brain Research UK, Muscular Dystrophy UK, Muscular Dystrophy Association (MDA USA), Higher Education Commission of Pakistan, The MRC (MR/S01165X/1, MR/S005021/1, G0601943), Wellcome Trust (WT093205MA, WT104033AIA and the Synaptopathies Strategic Award, 165908) and the National Institute for Health Research University College London Hospitals Biomedical Research Centre. Families 5, 6, 7 and 8 were collected as part of the SYNaPS Study Group collaboration funded by The Wellcome Trust and strategic award (Synaptopathies) funding. Research for these families was conducted as part of the Queen Square Genomics group at University College London, supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. NK is supported by intramural funds provided by King Faisal Specialist Hospital and Research Center (KFSHRC), the National Plan for Science, Technology and Innovation program under King Abdulaziz City for Science and Technology and the King Salman Center for Disability Research. TVH is supported by an Erasmus University Rotterdam (EUR) fellowship. TSB's lab is supported by the Netherlands Organisation for Scientific Research (ZonMW Veni, grant 91617021), a NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation, an Erasmus MC Fellowship 2017 and Erasmus MC Human Disease Model Award 2018. TSB, IC and EA acknowledge support from COST action CA16118 that facilitated this collaboration.

DISCLOSURE

KGM, AB, RT and KR are employees of GeneDx, Inc. KR holds stock in OPKO Health, Inc. KKK, PB and ABA are employees of CENTOGENE AG.

References

1 Kalser, J. & Cross, J. H. The epileptic encephalopathy jungle - from Dr West to the concepts of aetiology-related and developmental encephalopathies. Curr Opin Neurol 31, 216-222 (2018).

2 (!!! INVALID CITATION !!! (McTague et al., 2016)).

3 Epi, K. C. et al. De novo mutations in epileptic encephalopathies. Nature 501, 217-221 (2013).

4 Nashabat, M. et al. The landscape of early infantile epileptic encephalopathy in a consanguineous population. Seizure 69, 154-172 (2019).

5 Papuc, S. M. et al. The role of recessive inheritance in early-onset epileptic encephalopathies: a combined whole-exome sequencing and copy number study. Eur J Hum Genet 27, 408-421 (2019).

6 Bartha, I., di Iulio, J., Venter, J. C. & Telenti, A. Human gene essentiality. Nat Rev Genet 19, 51-62 (2018).

7 Robbins, S. M., Thimm, M. A., Valle, D. & Jelin, A. C. Genetic diagnosis in first or second trimester pregnancy loss using exome sequencing: a systematic review of human essential genes. J Assist Reprod Genet 36, 1539-1548 (2019).

8 Fuhring, J. et al. Octamerization is essential for enzymatic function of human UDP-glucose pyrophosphorylase. Glycobiology 23, 426-437 (2013).

9 Fuhring, J. I. et al. A quaternary mechanism enables the complex biological functions of octameric human UDP-glucose pyrophosphorylase, a key enzyme in cell metabolism. Sci Rep 5, 9618 (2015).

10 Yu, Q. & Zheng, X. The crystal structure of human UDP-glucose pyrophosphorylase reveals a latch effect that influences enzymatic activity. Biochem J 442, 283-291 (2012).

11 Turnquist, R. L., Gillett, T. A. & Hansen, R. G. Uridine diphosphate glucose pyrophosphorylase. Crystallization and properties of the enzyme from rabbit liver and species comparisons. J Biol Chem 249, 7695-7700 (1974).

12 Flores-Diaz, M. et al. Cellular UDP-glucose deficiency caused by a single point mutation in the UDP-glucose pyrophosphorylase gene. J Biol Chem 272, 23784-23791 (1997).

13 Higuita, J. C., Alape-Giron, A., Thelestam, M. & Katz, A. A point mutation in the UDP-glucose pyrophosphorylase gene results in decreases of UDP-glucose and inactivation of glycogen synthase. Biochem J 370, 995-1001 (2003).

14 Adeva-Andany, M. M., Gonzalez-Lucan, M., Donapetry-Garcia, C., Fernandez-Fernandez, C. & Ameneiros-Rodriguez, E. Glycogen metabolism in humans. BBA Clin 5, 85-100 (2016).

15 Magee, C., Nurminskaya, M. & Linsenmayer, T. F. UDP-glucose pyrophosphorylase: up-regulation in hypertrophic cartilage and role in hyaluronan synthesis. Biochem J 360, 667-674 (2001).

16 Vigetti, D., Viola, M., Karousou, E., De Luca, G. & Passi, A. Metabolic control of hyaluronan synthases. Matrix Biol 35, 8-13 (2014).

17 Perkins, K. L., Arranz, A. M., Yamaguchi, Y. & Hrabetova, S. Brain extracellular space, hyaluronan, and the prevention of epileptic seizures. Rev Neurosci 28, 869-892 (2017).

18 Soleman, S., Filippov, M. A., Dityatev, A. & Fawcett, J. W. Targeting the neural extracellular matrix in neurological disorders. Neuroscience 253, 194-213 (2013).

19 Cope, E. C. & Gould, E. Adult Neurogenesis, Glia, and the Extracellular Matrix. Cell Stem Cell 24, 690-705 (2019).

20 Arranz, A. M. et al. Hyaluronan deficiency due to Has3 knock-out causes altered neuronal activity and seizures via reduction in brain extracellular space. J Neurosci 34, 6164-6176 (2014).

21 Zeng, C., Xing, W. & Liu, Y. Identification of UGP2 as a progression marker that promotes cell growth and motility in human glioma. J Cell Biochem 120, 12489-12499 (2019).

22 Li, S., Hu, Z., Zhao, Y., Huang, S. & He, X. Transcriptome-Wide Analysis Reveals the Landscape of Aberrant Alternative Splicing Events in Liver Cancer. Hepatology 69, 359-375 (2019).

23 Li, Y. et al. Multiomics Integration Reveals the Landscape of Prometastasis Metabolism in Hepatocellular Carcinoma. Mol Cell Proteomics 17, 607-618 (2018).

24 Wang, L. et al. Expression of UGP2 and CFL1 expression levels in benign and malignant pancreatic lesions and their clinicopathological significance. World J Surg Oncol 16, 11 (2018).

25 Wang, Q. et al. SHP2 and UGP2 are Biomarkers for Progression and Poor Prognosis of Gallbladder Cancer. Cancer Invest 34, 255-264 (2016).

26 Tan, G. S. et al. Novel proteomic biomarker panel for prediction of aggressive metastatic hepatocellular carcinoma relapse in surgically resectable patients. J Proteome Res 13, 4833-4846 (2014).

27 Thorsen, K. et al. Tumor-specific usage of alternative transcription start sites in colorectal cancer identified by genome-wide exon array analysis. BMC Genomics 12, 505 (2011).

28 de Jonge, H. J. et al. Gene expression profiling in the leukemic stem cell-enriched CD34+ fraction identifies target genes that predict prognosis in normal karyotype AML. Leukemia 25, 1825-1833 (2011).

29 Perenthaler, E., Yousefi, S., Niggl, E. & Barakat, T. S. Beyond the Exome: The Non-coding Genome and Enhancers in Neurodevelopmental Disorders and Malformations of Cortical Development. Front Cell Neurosci 13, 352 (2019).

30 Jenkins, Z. A. et al. Differential regulation of two FLNA transcripts explains some of the phenotypic heterogeneity in the loss-of-function filaminopathies. Hum Mutat 39, 103-113 (2018).

31 Gostynska, K. B. et al. Mutation in exon 1a of PLEC, leading to disruption of plectin isoform 1a, causes autosomal-recessive skin-only epidermolysis bullosa simplex. Hum Mol Genet 24, 3155-3162 (2015).

32 Li, J. et al. Point Mutations in Exon 1B of APC Reveal Gastric Adenocarcinoma and Proximal Polyposis of the Stomach as a Familial Adenomatous Polyposis Variant. Am J Hum Genet 98, 830-842 (2016).

33 Ta-Shma, A. et al. Mutations in TMEM260 Cause a Pediatric Neurodevelopmental, Cardiac, and Renal Syndrome. Am J Hum Genet 100, 666-675 (2017).

34 Sobreira, N., Schiettecatte, F., Valle, D. & Hamosh, A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. Hum Mutat 36, 928-930 (2015).

35 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. Nat Genet 45, 580-585 (2013).

36 Epi25 Collaborative. Electronic address, s. b. u. e. a. & Epi, C. Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals. Am J Hum Genet (2019).

37 Landrum, M. J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42, D980-985 (2014).

38 Fokkema, I. F. et al. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat 32, 557-563 (2011).

39 Exome Variant Server NHLBI GO Exome Sequencing Project (ESP) Seattle WA. (accessed Juli 2019).

40 Firth, H. V. et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet 84, 524-533 (2009).

41 Gonzalez, M. et al. Innovative genomic collaboration using the GENESIS (GEM.app) platform. Hum Mutat 36, 950-956 (2015).

42 Scott, E. M. et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat Genet 48, 1071-1076 (2016).

43 Fattahi, Z. et al. Iranome: A catalog of genomic variations in the Iranian population. Hum Mutat (2019).

44 Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285-291 (2016).

45 Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 47, D886-D894 (2019).

46 Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods 11, 361-362 (2014).

47 in Encyclopædia Iranica Vol. III fasc. 6, pp 598-632 (2010).

48 Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317-330 (2015).

49 Yan, L. et al. Epigenomic Landscape of Human Fetal Brain, Heart, and Liver. J Biol Chem 291, 4386-4398 (2016).

50 Hwang, T. et al. Dynamic regulation of RNA editing in human brain development and disease. Nat Neurosci 19, 1093-1099 (2016).

51 Shih, H. P. et al. A Gene Regulatory Network Cooperatively Controlled by Pdx1 and Sox9 Governs Lineage Allocation of Foregut Progenitor Cells. Cell Rep 13, 326-336 (2015).

52 Mair, B. et al. Essential Gene Profiles for Human Pluripotent Stem Cells Identify Uncharacterized Genes and Substrate Dependencies. Cell Rep 27, 599-615 e512 (2019).

53 Yilmaz, A., Peretz, M., Aharony, A., Sagi, I. & Benvenisty, N. Defining essential genes for human pluripotent stem cells by CRISPR-Cas9 screening in haploid cells. Nat Cell Biol 20, 610-619 (2018).

54 Turnbull, J. et al. Lafora disease. Epileptic Disord 18, 38-62 (2016).

55 Sharma, J., Rao, S. N., Shankar, S. K., Satishchandra, P. & Jana, N. R. Lafora disease ubiquitin ligase malin promotes proteasomal degradation of neuronatin and regulates glycogen synthesis. Neurobiol Dis 44, 133-141 (2011).

56 Sharma, J. et al. Neuronatin-mediated aberrant calcium signaling and endoplasmic reticulum stress underlie neuropathology in Lafora disease. J Biol Chem 288, 9482-9490 (2013).

57 Shinde, V., Pitale, P. M., Howse, W., Gorbatyuk, O. & Gorbatyuk, M. Neuronatin is a stress-responsive protein of rod photoreceptors. Neuroscience 328, 1-8 (2016).

58 Sel, S. et al. Temporal and spatial expression pattern of Nnat during mouse eye development. Gene Expr Patterns 23-24, 7-12 (2017).

59 Salyakina, D. et al. Copy number variants in extended autism spectrum disorder families reveal candidates potentially involved in autism risk. PLoS One 6, e26049 (2011).

60 Li, Y. et al. Temporal and spatial expression of fgfbp genes in zebrafish. Gene 659, 128-136 (2018).

61 Tassi, E. et al. Fibroblast Growth Factor Binding Protein 3 (FGFBP3) impacts carbohydrate and lipid metabolism. Sci Rep 8, 15973 (2018).

62 Oikari, L. E. et al. Cell surface heparan sulfate proteoglycans as novel markers of human neural stem cell fate determination. Stem Cell Res 16, 92-104 (2016).

63 Lugert, S. et al. Glypican-2 levels in cerebrospinal fluid predict the status of adult hippocampal neurogenesis. Sci Rep 7, 46543 (2017).

64 Diotel, N., Beil, T., Strahle, U. & Rastegar, S. Differential expression of id genes and their potential regulator znf238 in zebrafish adult neural progenitor cells and neurons suggests distinct functions in adult neurogenesis. Gene Expr Patterns 19, 1-13 (2015).

65 Okazaki, T. et al. Epileptic phenotype of FGFR3-related bilateral medial temporal lobe dysgenesis. Brain Dev 39, 67-71 (2017).

66 Kyyriainen, J. et al. Deficiency of urokinase-type plasminogen activator and its receptor affects social behavior and increases seizure susceptibility. Epilepsy Res 151, 67-74 (2019).

67 Hua, S. et al. High expression of GALNT7 promotes invasion and proliferation of glioma cells. Oncol Lett 16, 6307-6314 (2018).

68 Guo, H. et al. O-Linked N-Acetylglucosamine (O-GlcNAc) Expression Levels Epigenetically Regulate Colon Cancer Tumorigenesis by Affecting the Cancer Stem Cell Compartment via Modulating Expression of Transcriptional Factor MYBL1. J Biol Chem 292, 4123-4137 (2017).

69 Pescador, N. et al. Hypoxia promotes glycogen accumulation through hypoxia inducible factor (HIF)-mediated induction of glycogen synthase 1. PLoS One 5, e9644 (2010).

70 Duran, J., Saez, I., Gruart, A., Guinovart, J. J. & Delgado-Garcia, J. M. Impairment in long-term memory formation and learning-dependent synaptic plasticity in mice lacking glycogen synthase in the brain. J Cereb Blood Flow Metab 33, 550-556 (2013).

71 Lopez-Ramos, J. C., Duran, J., Gruart, A., Guinovart, J. J. & Delgado-Garcia, J. M. Role of brain glycogen in the response to hypoxia and in susceptibility to epilepsy. Front Cell Neurosci 9, 431 (2015).

72 Choi, H. B. et al. Metabolic communication between astrocytes and neurons via bicarbonate-responsive soluble adenylyl cyclase. Neuron 75, 1094-1104 (2012).

73 Xu, J. et al. Requirement of glycogenolysis for uptake of increased extracellular K+ in astrocytes: potential implications for K+ homeostasis and glycogen usage in brain. Neurochem Res 38, 472-485 (2013).

74 Schousboe, A., Sickmann, H. M., Walls, A. B., Bak, L. K. & Waagepetersen, H. S. Functional importance of the astrocytic glycogen-shunt and glycolysis for maintenance of an intact intra/extracellular glutamate gradient. Neurotox Res 18, 94-99 (2010).

75 Wang, X. et al. Histone H3K4 methyltransferase Mll1 regulates protein glycosylation and tunicamycin-induced apoptosis through transcriptional regulation. Biochim Biophys Acta 1843, 2592-2602 (2014).

76 Kuil, L. E. et al. Hexb enzyme deficiency leads to lysosomal abnormalities in radial glia and microglia in zebrafish brain development. Glia 67, 1705-1718 (2019).

77 Narasimhan, V. M. et al. Health and population effects of rare gene knockouts in adult humans with related parents. Science 352, 474-477 (2016).

78 Saleheen, D. et al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. Nature 544, 235-239 (2017).

79 Sulem, P. et al. Identification of a large set of rare complete human knockouts. Nat Genet 47, 448-452 (2015).

80 Blomen, V. A. et al. Gene essentiality and synthetic lethality in haploid human cells. Science 350, 1092-1096 (2015).

81 Bakke, J. et al. Genome-wide CRISPR screen reveals PSMA6 to be an essential gene in pancreatic cancer cells. BMC Cancer 19, 253 (2019).

82 Bertomeu, T. et al. A High-Resolution Genome-Wide CRISPR/Cas9 Viability Screen Reveals Structural Features and Contextual Diversity of the Human Cell-Essential Proteome. Mol Cell Biol 38 (2018).

83 Wang, T. et al. Identification and characterization of essential genes in the human genome. Science 350, 1096-1101 (2015).

84 Wang, X. et al. BRD9 defines a SWI/SNF sub-complex and constitutes a specific vulnerability in malignant rhabdoid tumors. Nat Commun 10, 1881 (2019).

85 Hart, T. et al. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell 163, 1515-1526 (2015).

86 Daran, J. M., Bell, W. & Francois, J. Physiological and morphological effects of genetic alterations leading to a reduced synthesis of UDP-glucose in Saccharomyces cerevisiae. FEMS Microbiol Lett 153, 89-96 (1997).

87 Daran, J. M., Dallies, N., Thines-Sempoux, D., Paquet, V. & Francois, J. Genetic and biochemical characterization of the UGP1 gene encoding the UDP-glucose pyrophosphorylase from Saccharomyces cerevisiae. Eur J Biochem 233, 520-530 (1995).

88 Li, M. et al. UDP-glucose pyrophosphorylase influences polysaccharide synthesis, cell wall components, and hyphal branching in Ganoderma lucidum via regulation of the balance between glucose-1-phosphate and UDP-glucose. Fungal Genet Biol 82, 251-263 (2015).

89 Chen, R. et al. Rice UDP-glucose pyrophosphorylase1 is essential for pollen callose deposition and its cosuppression results in a new type of thermosensitive genic male sterility. Plant Cell 19, 847-861 (2007).

90 Park, J. I. et al. UDP-glucose pyrophosphorylase is rate limiting in vegetative and reproductive phases in Arabidopsis thaliana. Plant Cell Physiol 51, 981-996 (2010).

91 Woo, M. O. et al. Inactivation of the UGPase1 gene causes genic male sterility and endosperm chalkiness in rice (Oryza sativa L.). Plant J 54, 190-204 (2008).

92 Tian, D. et al. Identifying mouse developmental essential genes using machine learning. Dis Model Mech 11 (2018).

93 Jumbo-Lucioni, P. P., Parkinson, W. M., Kopke, D. L. & Broadie, K. Coordinated movement, neuromuscular synaptogenesis and trans-synaptic signaling defects in Drosophila galactosemia models. Hum Mol Genet 25, 3699-3714 (2016).

94 Tsai, F. Y. et al. An early haematopoietic defect in mice lacking the transcription factor GATA-2. Nature 371, 221-226 (1994).

95 de Pater, E. et al. Gata2 is required for HSC generation and survival. J Exp Med 210, 2843-2850 (2013).

96 Binder, J. et al. Clinical and molecular findings in a patient with a novel mutation in the deafness-dystonia peptide (DDP1) gene. Brain 126, 1814-1820 (2003).

97 Caridi, G. et al. A novel mutation in the albumin gene (c.1A>C) resulting in analbuminemia. Eur J Clin Invest 43, 72-78 (2013).

98 McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297-1303 (2010).

99 Retterer, K. et al. Clinical application of whole-exome sequencing across clinical indications. Genet Med 18, 696-704 (2016).

100 Snoeijen-Schouwenaars, F. M. et al. Diagnostic exome sequencing in 100 consecutive patients with both epilepsy and intellectual disability. Epilepsia 60, 155-164 (2019).

101 Mencacci, N. E. et al. De Novo Mutations in PDE10A Cause Childhood-Onset Chorea with Bilateral Striatal Lesions. Am J Hum Genet 98, 763-771 (2016).

102 Trujillano, D. et al. Clinical exome sequencing: results from 2819 samples reflecting 1000 families. Eur J Hum Genet 25, 176-182 (2017).

103 Bayer SA, A. J. Vol. volume 2 (CRC Press, Boca Raton, 2004).

104 Bayer SA, A. J. Vol. volume 3 (CRC Press, Boca Raton, 2005).

105 Bayer SA, A. J. Vol. volume 4 (CRC Press, Boca Raton, 2006).

106 Bayer SA, A. J. Vol. volume 5 (CRC Press, Boca Raton, 2008).

107 Barakat, T. S. et al. Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. Cell Stem Cell 23, 276-288 e278 (2018).

108 Barakat, T. S. et al. Stable X chromosome reactivation in female human induced pluripotent stem cells. Stem Cell Reports 4, 199-208 (2015).

109 Barakat, T. S. & Gribnau, J. X chromosome inactivation and embryonic stem cells. Adv Exp Med Biol 695, 132-154 (2010).

110 Barakat, T. S. & Gribnau, J. Generation of knockout alleles by RFLP based BAC targeting of polymorphic embryonic stem cells. Methods Mol Biol 1227, 143-180 (2015).

111 Renaud, J. B. et al. Improved Genome Editing Efficiency and Flexibility Using Modified Oligonucleotides with TALEN and CRISPR-Cas9 Nucleases. Cell Rep 14, 2263-2272 (2016).

112 Chambers, S. M. et al. Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nat Biotechnol 27, 275-280 (2009).

113 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12, 357-360 (2015).

114 Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166-169 (2015).

115 Lawrence, M. et al. Software for computing and annotating genomic ranges. PLoS computational biology 9, e1003118, doi:10.1371/journal.pcbi.1003118 (2013).

116 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140 (2010).

117 Zhou, Y. et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun 10, 1523 (2019).

118 Raudvere, U. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 47, W191-W198 (2019).

119 Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44, W90-97 (2016).

120 Ying, D. et al. HaploShare: identification of extended haplotypes shared by cases and evaluation against controls. Genome Biol 16, 92 (2015).

121 Turton, K. B., Esnault, S., Delain, L. P. & Mosher, D. F. Merging Absolute and Relative Quantitative PCR Data to Quantify STAT3 Splice Variant Transcripts. J Vis Exp (2016).

122 Brinkman, E. K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. Nucleic Acids Res 42, e168 (2014).

123 Lindhout, M., Rubio-Gozalbo, M. E., Bakker, J. A. & Bierau, J. Direct non-radioactive assay of galactose-1-phosphate:uridyltransferase activity using high performance liquid chromatography. Clin Chim Acta 411, 980-983 (2010).

Supplementary Material

Supplementary Case reports

Individual 4: The patient was born at 36+4 weeks after pregnancy complicated by maternal cholestasis. Her parents are of Indian ancestry. There is no recognized consanguinity. The patient was diagnosed with beta thalassemia in the newborn period which required regular transfusions. Feeding difficulties were also noted in the newborn period and persisted. Gastrostomy feeding was initiated at 7 months of age. Seizures were first observed at 3 months of age. The seizures were initially myoclonic and hypsarrhythmia was seen on EEG. The patient's epilepsy had been intractable and over time she has demonstrated a variety of seizure types including hemiclonic, focal motor, generalized tonic-clonic and tonic. A trial of the ketogenic diet was not effective. Multiple anti-epileptic drugs have been used with limited improvement of seizure frequency. Her primary regimen consisted of phenobarbital and clonazepam. Beginning at age of 10 months the patient began to have severe, dystonic episodes that featured posturing and variation in heart rate. The was also diagnosed with and treated for intussusception at this time. The dystonic episodes improved some with the administration of clonidine and propranolol. Benzodiazepines and opioids were not effective. MRI of the brain was performed at ages 1, 2 and 3 years. A thin corpus callosum was noted and over time there was cortical and striatal volume loss. She has been diagnosed with cortical visual impairment. Eye exam noted lagophthalmos and mild disc pallor. Her linear growth and weight were typical for age. She was able to vocalize but did not achieve other developmental milestones before she passed at age 3.5 years.

Individual 5: A 9-year-old female child from Oman, who presented at the age of 10 weeks with one day history of recurrent episodes of generalized tonic clonic seizures. She was born to first degree consanguineous parents at full-term, via spontaneous vaginal delivery with a birth weight of 2860 grams and an Apgar of 7 and 10 at 1 and 5 minutes respectively. She is the 5th child for the parents and one of her elder siblings died at 4years of age with some brain malformation (No documents available) and all other siblings are normal except a boy who reportedly has intellectual disability. Her clinical examination, on initial admission showed a head circumference of 37 cm (between 50th and 75th centiles) with weak Moro and sucking reflex, power of 4/5 on the limbs and exaggerated deep tendon reflexes. All the baseline investigations were within normal limits and she was loaded with phenobarbitone and continued with a maintenance dose. As the seizures were not well controlled, she required phenytoin, levetiracetam, topiramate and midazolam infusion during the first admission.

Her seizures got controlled after she was started on midazolam infusion. Her EEG at that time showed multifocal seizures with burst suppression and MRI brain showed cerebral atrophy with a thin corpus callosum and delayed myelination. An oral pyridoxine trial was started and she was referred to for further metabolic work-up. She was seen by a metabolic consultant, but her parents refused further investigations at that time and went against medical advice. During a second opinion in Pakistan she was started on ACTH for 6 weeks but this did not result in improvements. Parents stopped phenytoin treatment after hospital discharge and the child continued to get daily recurrent episodes of multiple types of seizures (generalized tonic clonic, tonic seizures, flexor spasms).

After 4 months parents visited again our outpatient clinic and at that time the girl had not attained any head control, did not visually track and had bilateral pyramidal signs. She was on phenobarbitone, topiramate and levetiracetam at that time. After adjustments of medication doses, clonazepam was added, which resulted in a slightly reduced seizure frequency. Her ophthalmic assessment showed generalized disc pallor with severe visual impairment. During her follow up as the seizures were not well controlled, she was started on trial of folinic acid and parents felt that the seizures improved after starting folinic acid. Parents noticed that seizure frequency had increased while they ran out of folinic acid for a week. During the follow-up, she was admitted twice to complete the detailed metabolic work ups. Relevant investigations:

- FBC - Normal
- Bone profile, Electrolytes, LFT, Magnesium: Normal
- Ammonia: 50 umol/L
- Lactate: 1mmol/L
- Blood gas: Normal
- Tandem Mass Spectrometry: Unremarkable
- Uric acid: 0.20 mmol/L (0.15 -0.35)
- Urine organic acids: unremarkable
- Lysosomal enzymes: unremarkable
- Serum pyridoxal phosphate: 206 nmol/L (35 -110)
- Plasma homocysteine: 7 umol/L (< 10)
- Urine sulfocystiene: Not detected
- Plasma amino acids: Unremarkable
- CSF Lactate: 1.6 mmol/L
- CSF Glucose: 3mmol/L (Blood glucose -5 mmol/L)
- CSF Amino acids -Slight decrease in Glycine (4,0 µmol/l Reference values 6.0-11.0) moderate increase in glutamine (606,0 µmol/l Reference values 333.9-575.5)
- CSF biogenic amines: Normal
- Serum Pipecolic acid: Normal
- CDG (Congenital disorder of glycosylation): Normal
- EEG: Abnormal for frequent generalized spike and wave discharges followed by brief period of suppression of background. Also independent epileptiform discharges arising from both temporal regions which become almost continuous at times. Also noticed to have asynchrony. The EEG is suggestive of early epileptic encephalopathy.
- MRI Brain: Cerebral atrophy with thin corpus callosum and delayed myelination
- MRI Brain: Generalized brain atrophy more marked in the supratentorial compartment with scanty white matter
- USG Abdomen: Normal

Last clinical review: She was still having daily brief seizures on multiple occasions. She had not attained any developmental milestones She is on nasogastric feeding with formula milk only. Examination showed a bedridden child with microcephaly, no vision and hearing, no facial asymmetry, generalized hypotonia with grade 3/5 power in both upper and lower limbs, DTR are just elicitable, and planters are -flexor bilaterally. Current medications: Calcium Folinate 5mg BID, Phenobarbitone 30 mg BID, which is 4.3 mg/kg/day Topiramate 25mg am and 50mg pm which is 5.4 mg/kg/day, Levetiracetam 250 mg BID, which is 36 mg/kg/day, Clonazepam 300mcg BID.

Individual 12: Individual 12 was born at term with unremarkable perinatal history. Growth parameters were normal. The parents were first-degree cousins. Two maternal uncles had global delay with intractable epilepsy and died at age of 1 and 4 years, respectively. At three months, the baby was noted to have episodic leg jerking which was confirmed to be epileptic seizures. With time, seizures became more frequent and daily, consisting of brief tonic seizures with uprolling of eyes. Several combinations of antiepileptic drugs were tried, but seizures remained intractable. The latest of which included phenobarbital, topiramate, and levetiracetam. Trial of pyridoxine was not helpful.

Comprehensive metabolic investigations were unrevealing. These included serum lactate, amino acids, renal and hepatic profiles, ammonia, transferrin isoelectric focusing, acyl carnitine profile and urine organic acids. EEG showed frequent generalized spikes during sleep associated with frequent independent sharp waves over frontal and central areas bilaterally. Trial of steroids – suspecting variant Landau Kluffner syndrome-was not helpful either. Brain MRI showed brain atrophy and developmental changes in the mesial temporal lobes. Long bone and chest X-rays showed osteopenia, leading to one event of femoral fracture. No abnormal storage was noted in skeletal bones or on femur MRI. Abdominal ultrasound showed borderline liver size but normal echogenicity. Thigh Muscle MRI showed possible moderate diffuse fatty changes involving both gluteal muscle groups and posterior thigh muscle compartment in both sides, with milder fatty changes in the anterior thigh compartment. Currently, at age 10, he is stroller bound, profoundly globally delayed in development. He is fed through nasogastric tube due to severe dysphagia. No organomegaly or major dysmorphic features are noted. His seizures are tonic, brief lasting seconds with up-rolling of eyes that happen daily, sometimes triggered by sound. They are more frequent upon awaking. He is not attentive to parents, both with sound or visual stimulation. Flash VEP showed delayed p100 wave and an abnormal electroretinogram. He is on multiple antiepileptic drugs including, toperamate, levetiracetam and phenobarbital as well as pyridoxine.

Individual 13: Individual 13 is the affected sister of individual 12. She was born at term with unremarkable perinatal course and normal birth growth parameters. The mother noticed seizures at the age of 5 months which were having semiology of infantile spasm, with flexion of the trunk and the upper limbs. Attacks were occurring in clusters. She was noted to be developmentally delayed as she was unable to support her neck when she was first evaluated at the age of 7 months. When examined, height, weight and head circumferences were between 10th and 50th percentiles. She was spastic with brisk reflexes. The rest of systemic examination was normal. MRI showed prominence of bilateral frontal horns with brain atrophy. EEG was abnormal showing paroxysmal epileptiform discharges but no classical hypsarrythmia. Brain auditory evoked potentials, electroretinography and visual evoked potentials of the left eye were normal while visual evoked potentials of the right eye showed reduced amplitude of p100. Comprehensive metabolic testing with serum, urine and CSF analysis were unrevealing. CSF/serum glucose ratio was normal excluding possibility of Glut-1 deficiency. WBC Electron microscopy for neuronal ceroid lipofuscinosis was negative. The patient was severely handicapped and seizures were difficult to control. She was treated with pyridoxine, levetiracetam and vigabatrin. At the age of 15 months, she died when she had a febrile illness with increased seizures. The cause of death was presumed aspiration with respiratory arrest at home.

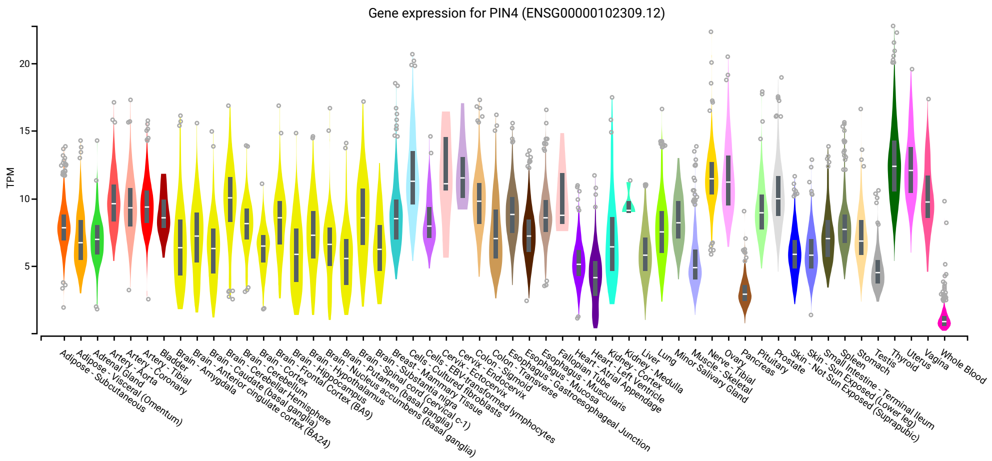
After finding UGP2 as the main candidate gene for both affected siblings, the parents of family 10 elected to pursue preimplantation genetic diagnosis and in-vitro-fertilization upon genetic counseling. Following controlled ovarian stimulation, fourteen oocytes were retrieved and nine were found to be suitable for biopsy on day 3. Karyomapping, haplotype chart and detailed haplotype analysis were reviewed and risk of contamination was excluded using AmpFlSTR® Identifiler® PCR Amplification Kit (following the manufacturer's instructions). Two embryos were selected for transfer, embryo #2 and #5. Genetic analysis indicated that embryo 2 is a carrier with the inheritance of the normal maternal allele whereas embryo 5 showed completely normal pattern. Both embryos were chromosomally normal (euploid) and resulted in the delivery of normal born twin (carrier male and normal female). Currently at 25 month both children are free from any disease symptoms.

Supplementary Note

The disease we here describe is caused by the loss of an isoform of an essential gene, due to an alteration affecting an isoform specific start codon. To investigate whether this same mechanism could apply to other essential genes that were previously not implicated in human genetic disease, we investigated the occurrence of homozygous or hemizygous ATG altering mutations using data mining of whole exome sequencing data from undiagnosed patients from our own data base, the Queen Square Genomic Center database and those from Centogene and GeneDx, focusing on the list of genes presented in Figure 7. This identified a number of currently genetically unexplained individuals with homozygous and hemizygous start codon altering variants, that we will report elsewhere in more detail.

We here briefly describe as an additional example of the mutational mechanism the occurrence of a hemizygous start codon altering variant in the peptidylproly cis/trans isomerase, NIMA-interacting-4 gene PIN4 (NM_006223.3:c.2T>A, p.Met1?). In the CentoMD data base, we identified 5 hemizygous patients, presenting with a shared phenotype of neurodevelopmental delay, microcephaly, seizures, inguinal hernia and a few other shared features, that we will describe elsewhere in full detail. Using routine clinical diagnostics, including whole exome and whole genome sequencing, no alternative disease explaining variant has been identified in these individuals. The variant is absent in gnomAD, and not found in our in house data bases. We did not identify any other LoF variant in this gene in our cohorts.

PIN4 encodes a member of the parvulin subfamily of the peptidyl-prolyl cis/trans isomerase family. It catalyzes the isomerization of peptidylprolyl bonds, and is proposed to play a role in cell cycle, chromatin remodeling, ribosome biogenesis and mitochondria function. Importantly, it has been shown to influence the formation of microtubules1. PIN4 is widely expressed amongst tissues, including different brain regions, according to data from the GTEX portal (**Figure**). Together, this makes PIN4 a strong candidate gene for a novel neurodevelopmental disorder.



1. Thiele, A. et al. Parvulin 17 promotes microtubule assembly by its peptidyl-prolyl cis/trans isomerase activity. J Mol Biol 411, 896-909 (2011).
2. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. Nat Genet 45, 580-5 (2013).

Supplementary Tables

https://link.springer.com/article/10.1007/s00401-019-02109-6

Supplementary Table 1: Extended clinical characteristics of 18 patients with homozygous UGP2 variants

Supplementary Table 2: RNA-seq data used in this study

Supplementary Table 3: Differentially expressed genes

Supplementary Table 4: Enrichment analysis

Supplementary Table 5: UGP2 variants in gnomAD

Supplementary Table 6: Genome-wide homology search results

Supplementary Table 7: gnomAD data of 247 disease candidate genes

Supplementary Table 8: Oligonucleotides used in this study

Supplementary Movies

https://link.springer.com/article/10.1007/s00401-019-02109-6

Supplemental Movie 1: Affected individual from family 11

Supplemental Movie 2: wild type zebrafish eye movements

Supplemental Movie 3: Ugp2a/b double mutant zebrafish eye movements

Supplementary Figures

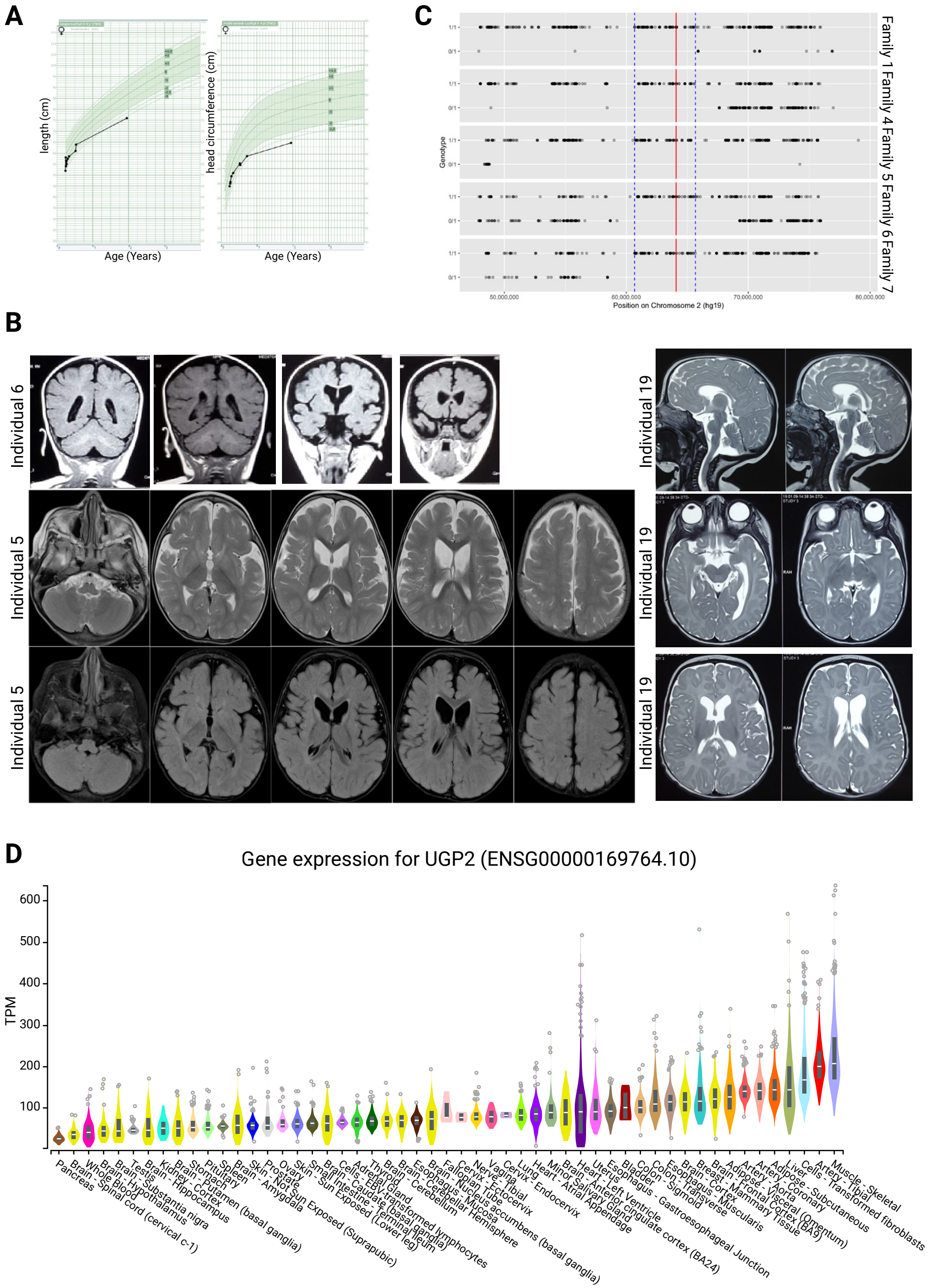


Figure S1 | A) Growth chart from individual 1 for length (left) and head circumference (right) in cm. Reference chart from the Dutch population are used (TNO) and regions between -2 and + 2 SD are shaded. **B)** MRI studies of individual 5 (at the age of 12 month) and individual 6, showing global brain atrophy. **C)** ROH comparison between affected individuals from family 1, 4, 5, 6 and 7, carrying the homozygous chr2:64083454A>G mutation. The red line indicates the UGP2 variant, and the blue lines demark the shared ROH region between the individuals (chr2:60679942-65667235). **D)** Violin plots showing distribution of gene expression (in TPM) amongst samples from the GTEx portal35 for tissues and cell lines. Samples are sorted with the highest median TPM on the right. Outliers are indicated by dots.

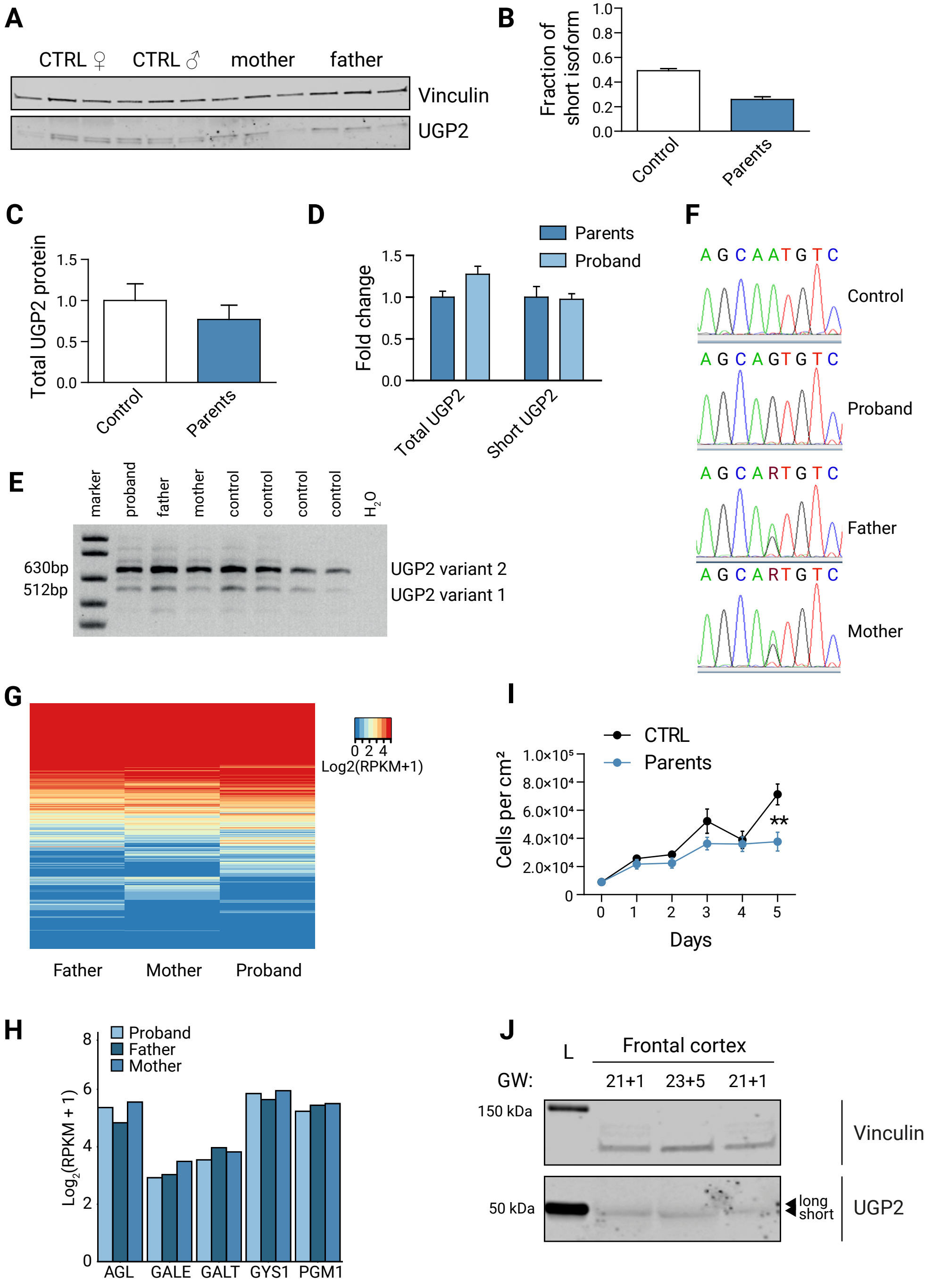


Figure S2 | A) Western blotting of cellular extracts derived from control fibroblasts or fibroblasts obtained from heterozygous parents of family 2, detecting the house keeping control vinculin or UGP2. Note the two separated isoforms of UGP2 that have a similar intensity in wild type cells. The shorter isoform shows reduced expression in fibroblasts from heterozygous parents. **B)** Quantification of the fraction of the short UGP2 protein isoform compared to total UGP2 expression in control, and heterozygous fibroblasts from family 2, as determined in three independent experiments. Error bars represent SEM. **C)** Western blotting quantification of total UGP2 protein levels, as determined by the relative expression to the housekeeping control vinculin. Bar graph showing the results from three independent experiments. Error bars represent SEM; no significant differences between control and parent samples, unpaired t-test, two-tailed. **D)** qRT-PCR analysis of total UGP2 or the short isoform in fibroblast from heterozygous parents or homozygous proband from family 1, normalized for the housekeeping control TBP. The mean fold change compared to heterozygous parents of two biological replicates and two technical replicates is shown; error bars represent SEM no significant differences between control and parent samples, unpaired t-test, two-tailed. **E)** Multiplex RT-PCR detecting relative expression of UGP2 isoform 1 and isoform 2 in peripheral blood from family 1 and unrelated wild type controls. **F)** Sanger sequencing of RT-PCR products from E), showing the expression of the homozygous and heterozygous chr2:64083454A>G UGP2 variant in the index proband, her parents and an unrelated control. **G)** Heat map showing genome-wide gene expression levels (in log2(RPKM+1)) in peripheral blood from heterozygous parents and homozygous proband from family 1. **H)** Gene expression levels (in log2(RPKM+1)) from RNA-seq in peripheral blood for a selected number of genes involved in metabolism. **I)** Cell proliferation experiment of fibroblast from heterozygous parents from family 2 and wild type controls, during a 5 days period. Error bars represent SEM, **= p<0.01, unpaired t-test, two-tailed. **J)** Western blotting detecting UGP2 in human frontal cortex from week 21 and 23 of gestation, showing the virtual absence of the long isoform expression in fetal brain. Vinculin is used as a housekeeping control.

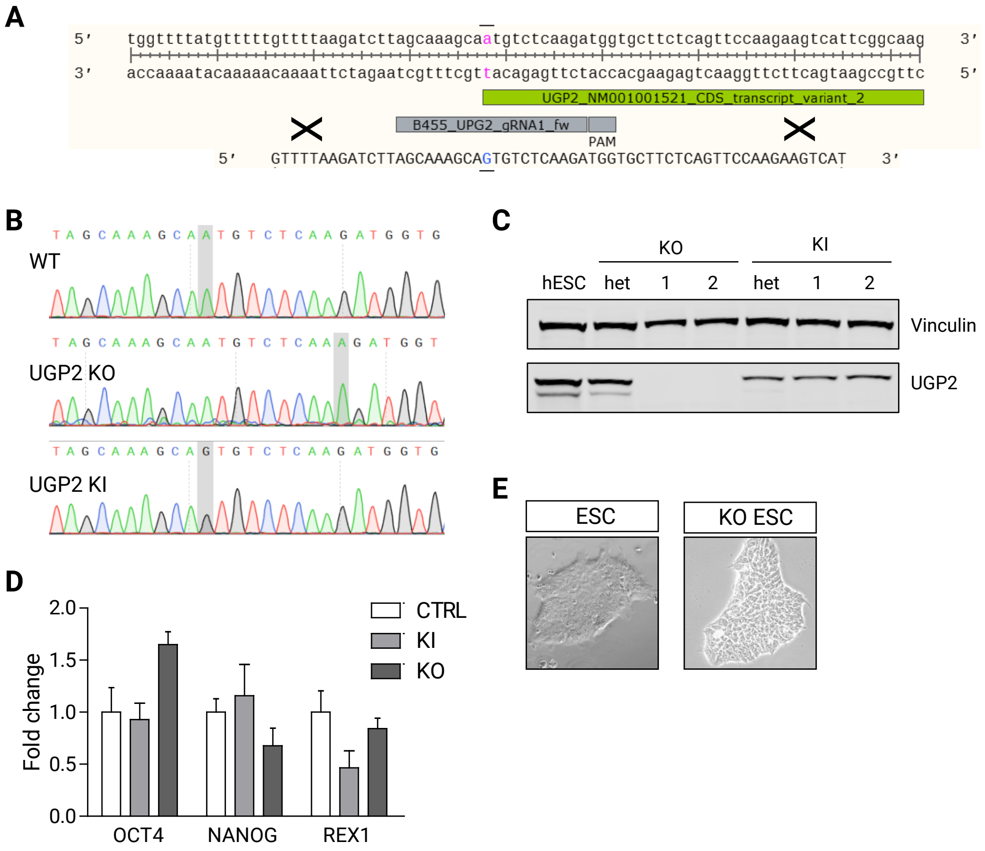


Figure S3 | Generation of mutant UGP2 H9 cell lines. A) Nucleotide sequence encompassing the ATG of UGP2 transcript isoform 2. Indicated are the coding sequence, the location of the gRNA, PAM sequence and ssODN used to introduce the C.1A>G, p.? mutation. **B)** Sanger sequencing traces of part of the UGP2 gene from wild type, UGP2 knock-out (KO) and UGP2 knock-in H9 ESCs (KI). The A at the start of the coding sequence of UGP2 isoform 2 (short isoform) is highlighted. The homozygous insertion of an additional A in knockout and the mutation into a G in knock-in cells are indicated. **C)** Western blotting detecting UGP2 and vinculin in wild type ESC, heterozygous and homozygous knockout and knock-in ESCs, as indicated. Note the complete loss of UGP2 in KO cells, and the loss of the short isoform in KI cells. **D)** RT-qPCR detecting the pluripotency factors *OCT4*, *NANOG* and *REX1* in H9 wild type, UGP2 knock-in (KI) and UGP2 knock-out (KO) ESCs, normalized for the house keeping control *TBP*. Mean fold change compared to wild type of two biological replicates and three technical replicates is shown; error bars represent SEM, *= p<0.05, unpaired t-test, two-tailed. **E)** Bright field image of a representative ESC colony from wild type parental and UGP2 KO ESCs.

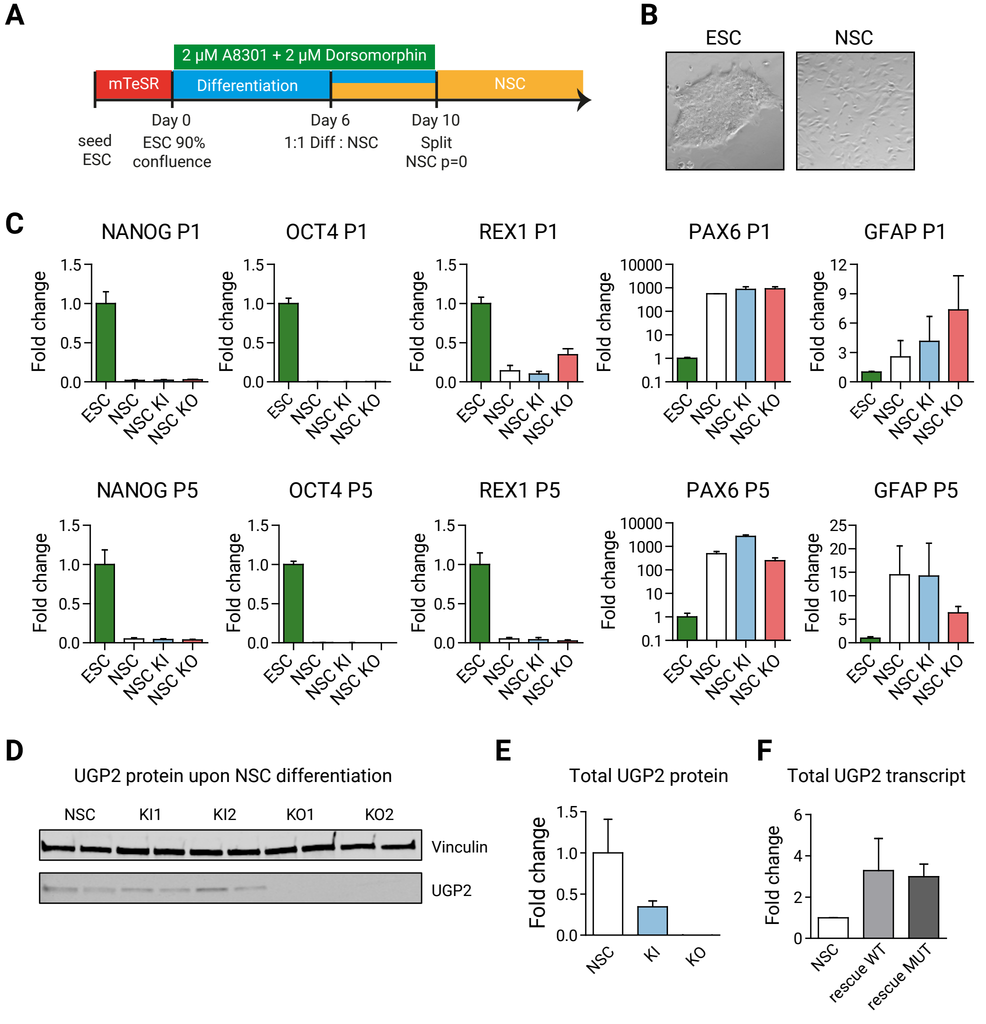
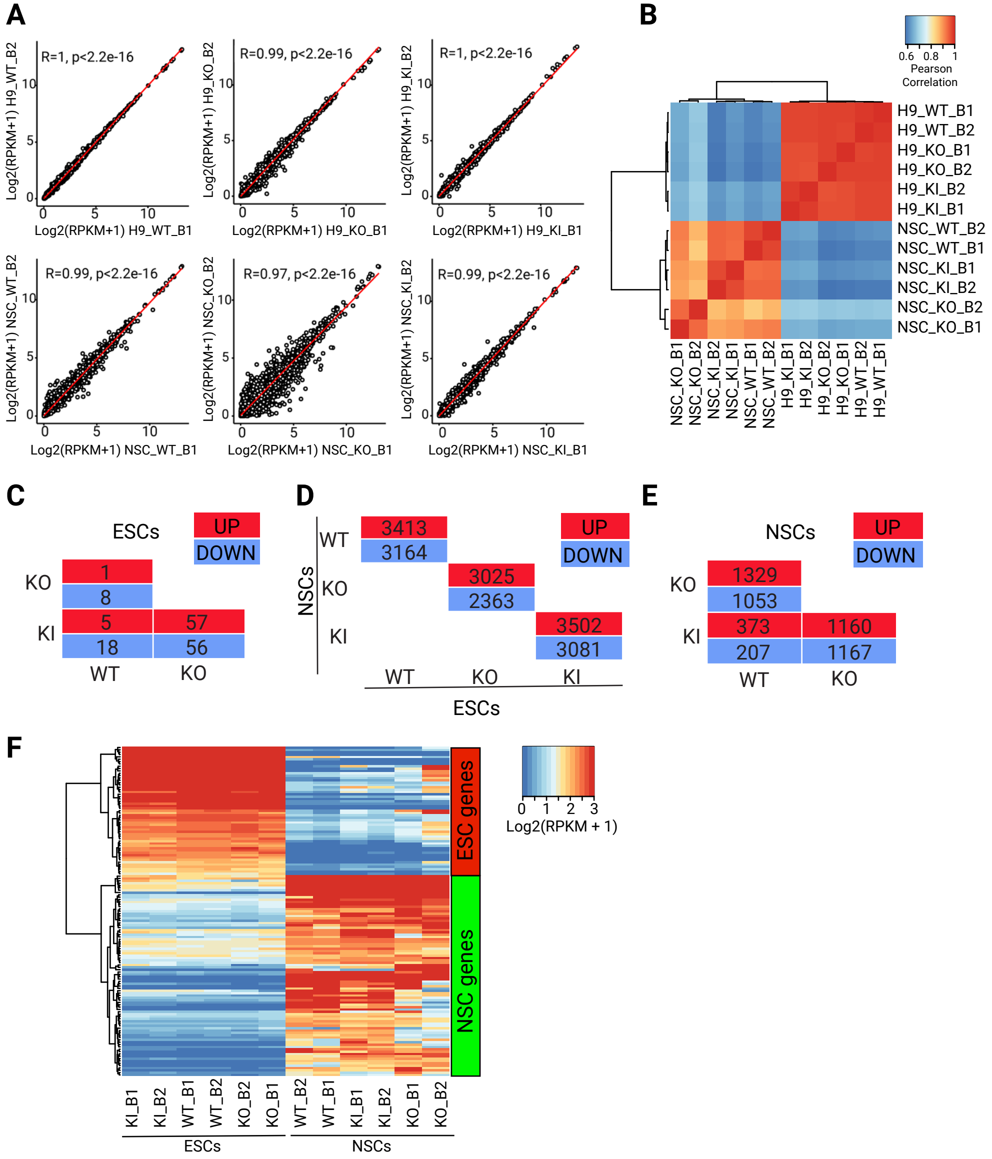


Figure S4 | NSC differentiation. A) Schematic drawing of the differentiation procedure, see online methods for details. **B)** Bright field image showing representative pictures from ESCs and differentiated NSCs. **C)** qRT-PCR analysis for pluripotency markers (*NANOG*, *OCT4* (*POU5F1*), *REX1*) and genes expressed in NSCs (*PAX6*, *GFAP*) in WT, UGP2 KO and KI differentiated NSCs at p1 and p5. Mean fold change compared to wild type of two biological replicates and two technical replicates is shown; error bars represent SEM. **D)** Western blotting showing UGP2 expression in WT, UGP2 KI and KO differentiated NSCs. Vinculin is used as a housekeeping control. **E)** Quantification of total UGP2 protein levels by Western blotting, as determined by the relative expression to the housekeeping control vinculin. Bar graph showing the results from two independent experiments; error bars represent SEM. **F)** qRT-PCR analysis of UGP2 in NSCs or KO NSCs rescued with either the long wild type or long mutant UGP2 isoform. Mean fold change compared to wild type is shown for two biological replicates and three technical replicates; error bars represent SEM.



C) ESCs

	WT	KO
KO	1 8	
KI	5 18	57 56

D) NSCs vs ESCs

	WT	KO	KI
WT	3413 3164		
KO		3025 2363	
KI			3502 3081

E) NSCs

	WT	KO
KO	1329 1053	
KI	373 207	1160 1167

Figure S5 | RNA-seq. A) Scatter plot showing the pair wise correlation between biological replicates. **B)** Heat map displaying Pearson correlation between biological replicates. **C)** Table summarizing up- (FDR<0.05 and LogFC>1) and down regulated (FDR<0.05 and LogFC<-1) genes in WT, KO and KI ESCs. **D)** Table summarizing up- (FDR<0.05 and LogFC>1) and down regulated (FDR<0.05 and LogFC<-1) genes in WT, KO and KI ESC upon differentiation in NSCs. **E)** Table summarizing up- (FDR<0.05 and LogFC>1) and down regulated (FDR<0.05 and LogFC<-1) genes in WT, KO and KI NSCs. **F)** Heat map visualizing gene expression (in log2(RPKM+1)) and clustering of WT, KO and KI ESCs and NSCs, for a panel of ESC and NSC specific genes (see methods)

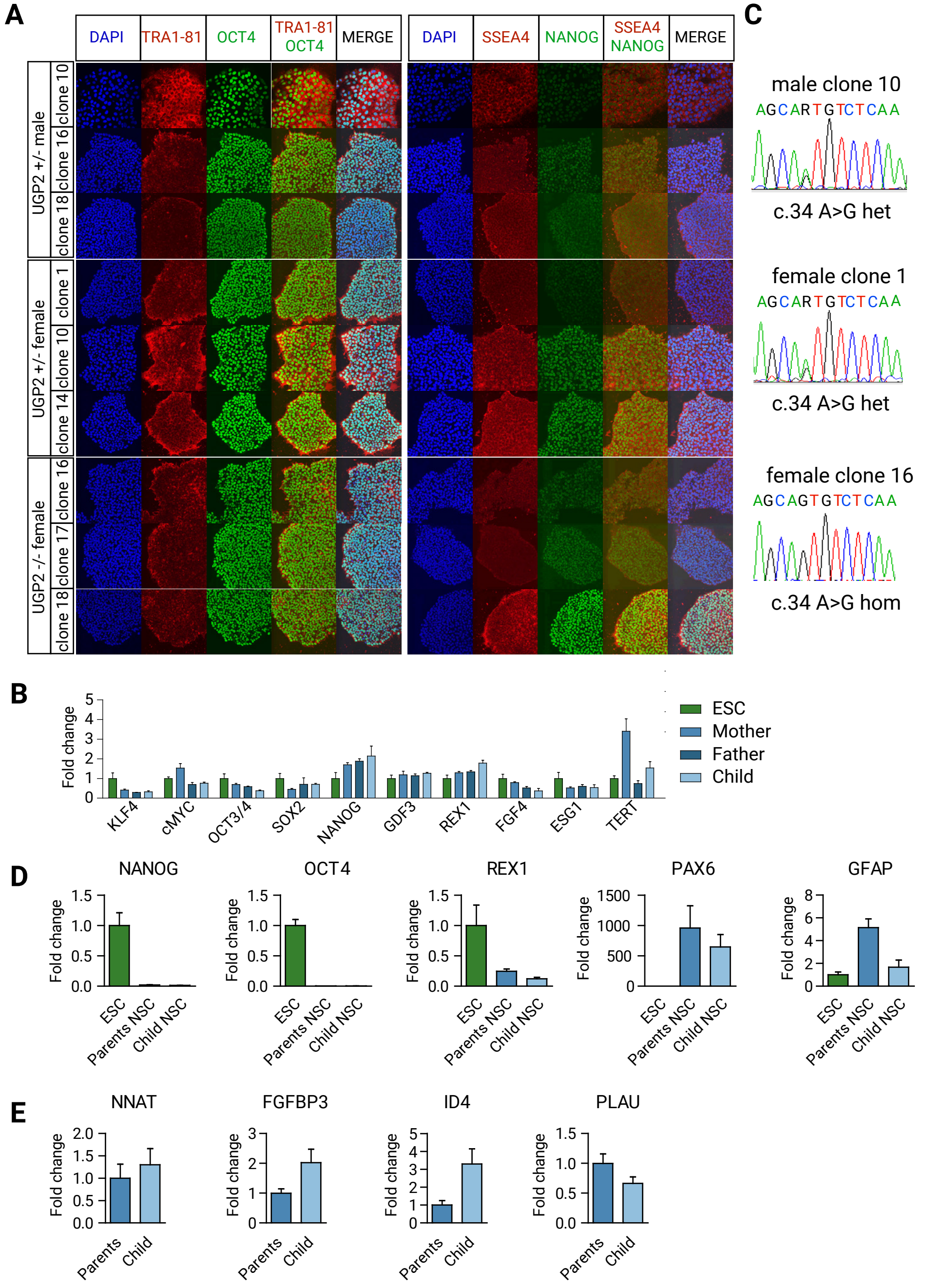


Figure S6 | UGP2 mutant iPSC. A) Immunofluorescence of iPSC clones used in this study derived from Family 1 (three clones per individual) showing iPSC colonies stained for the pluripotency markers TRA1-81 (red) and OCT4 (green) (left panel) or SSEA4 (red) and NANOG (green) (right panel). Nuclei are stained with DAPI (blue). **B)** qRT-PCR expression analysis for the indicated pluripotency associated genes in 4 wild type control human embryonic stem cell lines and the iPSCs derived from family 1. Mean fold change compared to human embryonic stem cells of three biological replicates (e.g. individual clones from A) and three technical replicates is shown; error bars represent SEM. No statistically significant differences were found, unpaired t-test, two-tailed. **C)** Sanger sequencing of representative iPSC clones confirming the presence of the chr2:64083454A>G UGP2 mutation in a heterozygous state in clones derived from parents and homozygous state in clones derived from the affected child. **D)** qRT-PCR PCR expression analysis upon differentiation for pluripotency (*NANOG*, *OCT4* (*POUF51*), *REX1*) and NSC markers (*PAX6*, *GFAP*), for H9 ESC control and heterozygous and homozygous iPSCs derived from family 1. Mean fold change compared to human embryonic stem cells of three biological replicates (e.g. individual clones from A)and two technical replicates is shown; normalized to *TBP*; error bars represent SEM. **E)** qRT-PCR expression analysis in iPSC-derived NSCs for genes that showed differential expression in RNA-seq experiments, e.g. *NNAT*, *FGFBP3*, *ID4* and *PLAU*. Mean fold change for cells obtained from the affected child compared to cells obtained from its parents (set to 1) of three biological replicates (e.g. individual clones from A) and two technical replicates is shown; normalized to *TBP*; error bars represent SEM.

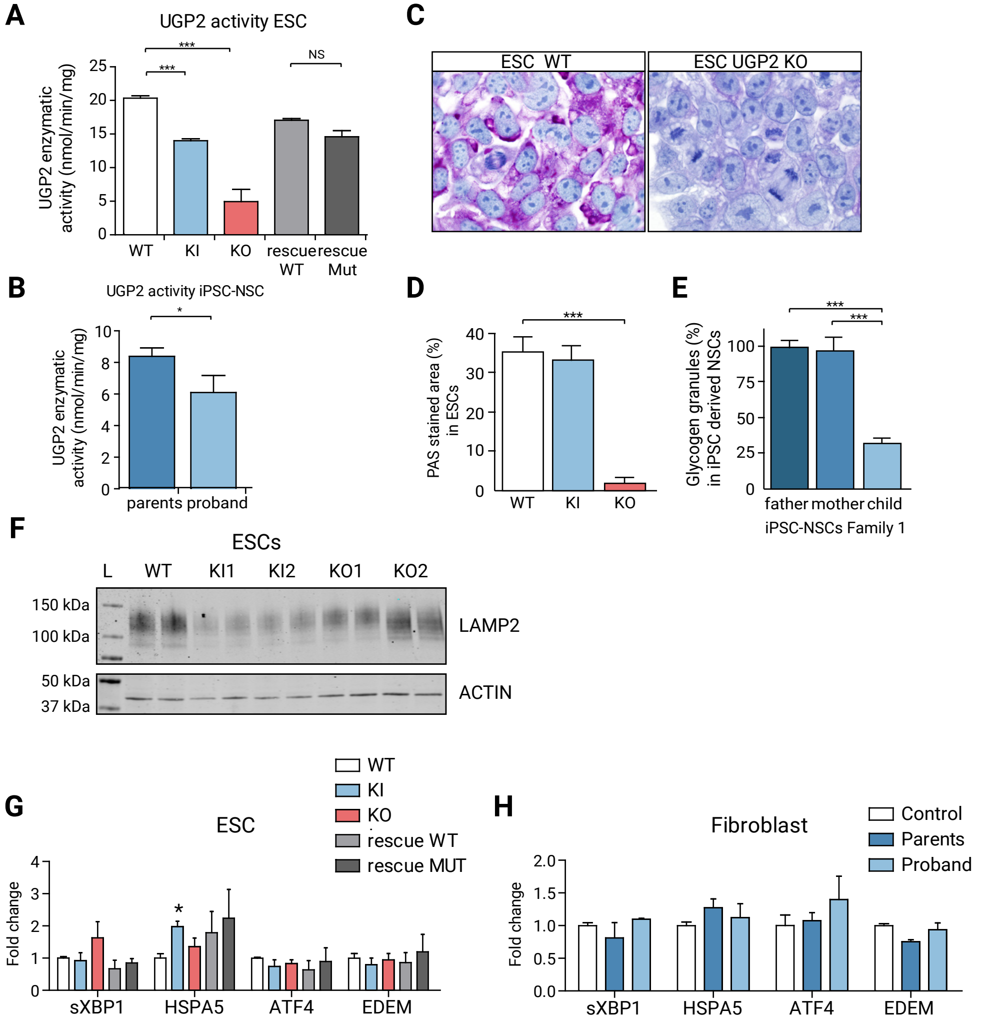


Figure S7 | **A)** UGP2 enzymatic activity in WT, UGP2 KI, KO and KO ESCs rescued with wildtype isoform 1 or mutant Met12Val isoform 1 of UGP2. Plotted is the mean from two replicate experiments, error bar is SEM. ***=p<0.001, unpaired t-test, two-tailed. **B)** UGP2 enzymatic activity in iPSC derived NSCs from family 1. Plotted is the mean from two replicate experiments, measuring each the results for the three clones for each individual, error bar is SEM. *=p<0.05; unpaired t-test, two-tailed. **C)** PAS staining in WT and UGP2 KO ESCs. Nuclei are counterstained with hematoxylin (blue). **D)** Quantification of the PAS stained area in WT, KI and KO ESCs. Shown is the average PAS positive area per genotype from two biological replicates, each stained in two experiments; error bars are SD. ***=p<0.001, unpaired t-test, two-tailed. **E)** Glycogen granules detected by PAS staining in iPSC-derived NSCs from family 1 after 48 hours culture under low-oxygen conditions. Number of granules for paternal cell line are set at 100%. Average of three biological and two technical replicates per genotype, with each n=80-100 cells counted. Error bars represent SD, ***=p<0.001, unpaired t-test, two-tailed. **F)** Western blotting detecting LAMP2 (upper panel) and the house keeping control actin (lower panel) in cellular extracts from ESCs, that are WT, UGP2 KI, or KO. Compare to Figure 5D. **G)** qRT-PCR expression analysis for UPR marker genes (spliced *XBP1*, *HSPA5*, *ATF4* and *EDEM*) in WT, UGP2 KI, KO and rescue ESCs. Shown is the mean fold change for the indicated genes compared to wild type, normalized for the housekeeping gene TBP. Results of two biological and three technical replicates are plotted from two experiments. Error bars represent SEM; *= p<0.05, unpaired t-test, two-tailed). **H)** qRT-PCR expression analysis for UPR marker genes (spliced *XBP1*, *HSPA5*, *ATF4* and *EDEM*) in in primary fibroblasts from family 1. Shown is the mean fold change for the indicated genes compared to wild type, normalized for the housekeeping gene TBP. Results of two experiments with each three technical replicates are plotted. Error bars represent SEM; *= p<0.05, unpaired t-test, two-tailed.

Part II

Active enhancer landscape in ESCs

Active enhancer landscape in NSCs

4

Active enhancer landscape in ESCs

- A catalog of functional enhancers in primed and naive hESCs
- Over 350,000 genome regions assessed with massively parallel reporter assay
- Identification of transcription factors and transposable elements linked to enhancers
- Detailed dissection of functional domains in super-enhancers

This chapter has been published in:

TS Barakat *et al.* (2018). Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. Cell Stem Cell 23, 276-288 e278.

Functional dissection of the enhancer repertoire in human embryonic stem cells

Tahsin Stefan Barakat*, Florian Halbritter*, Man Zhang+, André F. Rendeiro+, Elena Perenthaler, Christoph Bock, and Ian Chambers

* these authors contributed equally, + these authors contributed equally

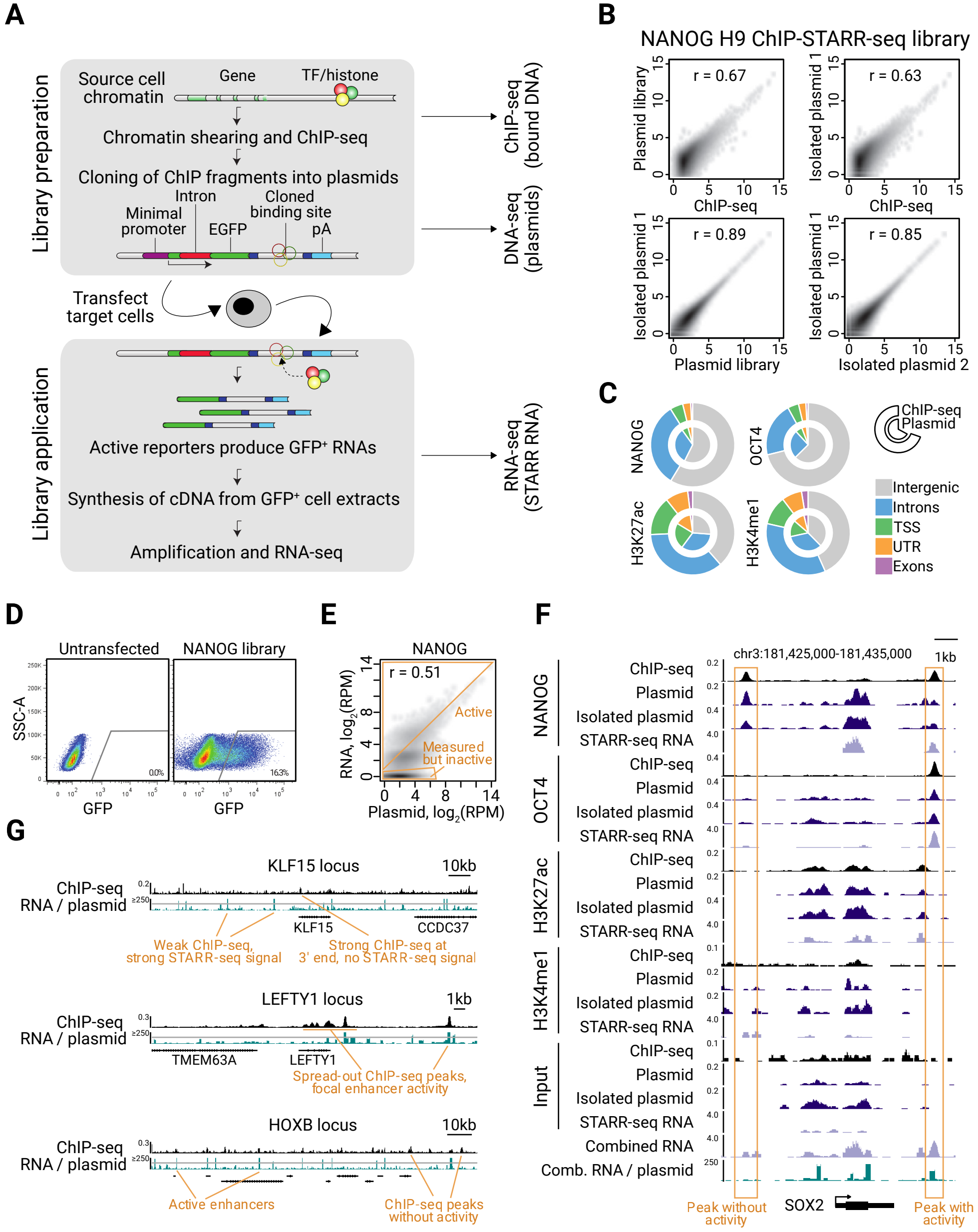
Enhancers are genetic elements that regulate spatiotemporal gene expression. Enhancer function requires transcription factor (TF) binding and correlates with histone modifications. However, the extent to which TF binding and histone modifications can functionally define active enhancers remains unclear. Here we combine chromatin immunoprecipitation with a massively parallel reporter assay to identify functional enhancers in human embryonic stem cells (ESCs) genome-wide in a quantitative unbiased manner. While active enhancers associate with TFs, only a minority of regions marked by NANOG, OCT4, H3K27ac and H3K4me1 function as enhancers, with activity changing markedly with culture conditions. Our analysis identifies an enhancer set associated with functions that extend to non-ESC-specific processes. Moreover, while transposable elements associate with putative enhancers only some exhibit activity. Similarly, within super-enhancers, large tracts are non-functional, with activity restricted to small sub-domains. This catalogue of validated enhancers provides a valuable resource for further functional dissection of the regulatory genome.

Introduction

Human embryonic stem cells (ESC) are a genetically tractable developmental model system with potential for stem-cell-based therapeutics. Understanding how ESC pluripotency is regulated by transcription factors (TFs) is central to achieving this promise. Gene expression is modulated by cis-regulatory elements such as enhancers[1] which can stimulate target gene expression in a position and orientation-independent manner, independent of their genomic context[2]. ESCs direct a specific gene expression program using a network of TFs including OCT4, SOX2 and NANOG. Compared to mouse ESCs, ESCs are more developmentally advanced with characteristics of post-implantation embryos. Recently, so-called naive ESCs with characteristics of pre-implantation embryos have been derived from established ESCs either by transient transgene expression[3-5] or by altering culture conditions[6-9]. Naive ESCs differ from primed ESCs in several ways including increased clonogenicity, different growth factor requirements, distinct energy metabolism, and altered morphology[10] but how naive and primed ESCs differ in enhancer usage is currently unclear.

The past decade of genomics research has focused on cataloguing cis-regulatory elements within the non-coding genome[11]. Technological advances have allowed genome-wide occupancy by TFs to be measured by chromatin immunoprecipitation (ChIP) followed by sequencing (ChIP-seq). Putative enhancer locations have been obtained by mapping histone modifications (e.g. H3K27ac, H3K4me1)[12,13] and by measuring chromatin accessibility[14]. However, not all predicted enhancers can be validated functionally. To assay enhancer activity, plasmid-based cell transfections can be used. Recent advances have enabled thousands of sequences to be tested simultaneously[15-19]. For instance, with Self-Transcribing Active Regulatory Region Sequencing (STARR-seq) compact, non-mammalian genomes can be screened quantitatively for enhancer activity by cloning randomly sheared DNA between a minimal-promoter-driven GFP open reading frame and a downstream polyA sequence. If an enhancer is active, this results in transcription of the enhancer sequence[15,20,21]. Similar approaches have recently been adapted to test chosen sequences with putative enhancer features[22-24], predicted TF binding sites[25], features of quantitative trait loci[26-28] or nucleosome-depleted sequences[29].

Application of STARR-seq to explore mammalian genomes is hindered by genome size which means enhancer sequences would be infrequently sampled. This issue can be alleviated by combining ChIP with STARR-seq[30]. Using a similar approach (that we refer to as ChIP-STARR-seq) we generate a resource of genome-wide activity



maps of functional enhancers in ESCs. This identifies highly active enhancers with major changes in activity patterns between primed and naive ESCs. Moreover, some transposable element (TE) families are enriched at highly active enhancers. Our data also identify the functional components within super-enhancers (SEs) and uncover a previously unidentified set of enhancers, including some associated with housekeeping functions. This resource encompasses an extensive collection of functional enhancer sequences in ESCs, providing a knowledge base for systematic analysis of the transcriptional circuitry underlying ESC maintenance and differentiation. Enhancer data are available from the STAR-methods and from a resource website (http://hesc-enhancers.computational-epigenetics.org).

Results

ChIP-STARR-seq: an effective strategy for genome-wide identification of functional enhancers

To generate a catalogue of genomic elements that regulate ESC biology we used a massively parallel reporter assay, called ChIP-STARR-seq. In ChIP-STARR-seq, DNA is co-immunoprecipitated and cloned en masse within the transcription unit of a STARR-seq plasmid, downstream of GFP driven by a minimal promoter and upstream of a polyA sequence (**Figure 1A**)[15]. The resulting libraries can be tested for enhancer activity by cell transfection. If a cloned sequence functions as an enhancer, the transfected GFP-positive cells can be purified by FACS. Since the assayed sequences lie upstream of the polyA signal, the transcribed mRNA will

Figure 1 | ChIP-STARR-seq in human embryonic stem cells. A) Outline of the ChIP-STARR-seq approach combining antibodies against TFs or histone modifications (colored balls) with the STARR-seq plasmid (Arnold et al., 2013). **B)** ChIP-STARR-seq for NANOG in H9. Scatterplots compare normalized read count (reads per million) per peak between datasets, obtained from ChIP-seq or DNA-seq of plasmid libraries pre- or post-transfection/recovery from ESCs (n=2); *r*, Pearson correlation. **C)** Genomic distribution of peaks called for ChIP-seq (outer chart) and corresponding plasmid libraries (inner chart); TSS, transcription start sites; UTR, untranslated region. **D)** FACS plots of single, DAPI-negative ESCs. Left, untransfected cells; right, cells transfected with a NANOG ChIP-STARR-seq plasmid library. **E)** Scatterplot (as in B) comparing the NANOG plasmid library and corresponding ChIP-STARR-seq RNA. The dense cluster of points in the lower left corresponds to library plasmids that did not produce RNAs. RPM, reads per million. **F)** Genome browser plot of *SOX2* showing tracks for ChIP-seq, DNA-seq of plasmid libraries pre- and post-transfection, and from RNA-seq of GFP⁺ cells transfected with the indicated libraries. Bottom: combination (maximum) of all STARR-seq RNA-seq tracks and ratio of normalized RNA-seq/ plasmid reads. **G)** Genome browser shots of *KLF15*, *LEFTY* and *HOXB* cluster, illustrating a broad variety of enhancers profiled in this functional enhancer catalogue.

contain the enhancer sequence. Therefore, both the identity and activity of captured regions can be determined quantitatively by sequencing mRNA (RNA-seq) from GFP-positive cells.

To investigate the functional potential of enhancers in ESCs, we first focused on primed H9 ESCs (**Figure S1A, B**) and performed ChIP for NANOG, OCT4, H3K4me1 and H3K27ac. ChIP-qPCR and ChIP-seq were similar to previous results (**Figure S1C, D**). While plasmid transfection can elicit an immune response in some cell types [31], the low expression of STING and CGAS in H1[31] and H9 (**Figure S1E**) suggests this does not apply to ESCs. ChIP-STARR-seq libraries were generated (**see STAR-Methods**). Sequencing precipitated DNA, plasmid libraries, and transcribed RNAs produced 2.7×10^9 reads in total. Each plasmid library consisted of 8.4-30.8×10^6 unique plasmids, with a mean insert size of 221 bp (**Table S1**). **Figure S2A** summarises the sequenced samples analysed in this study.

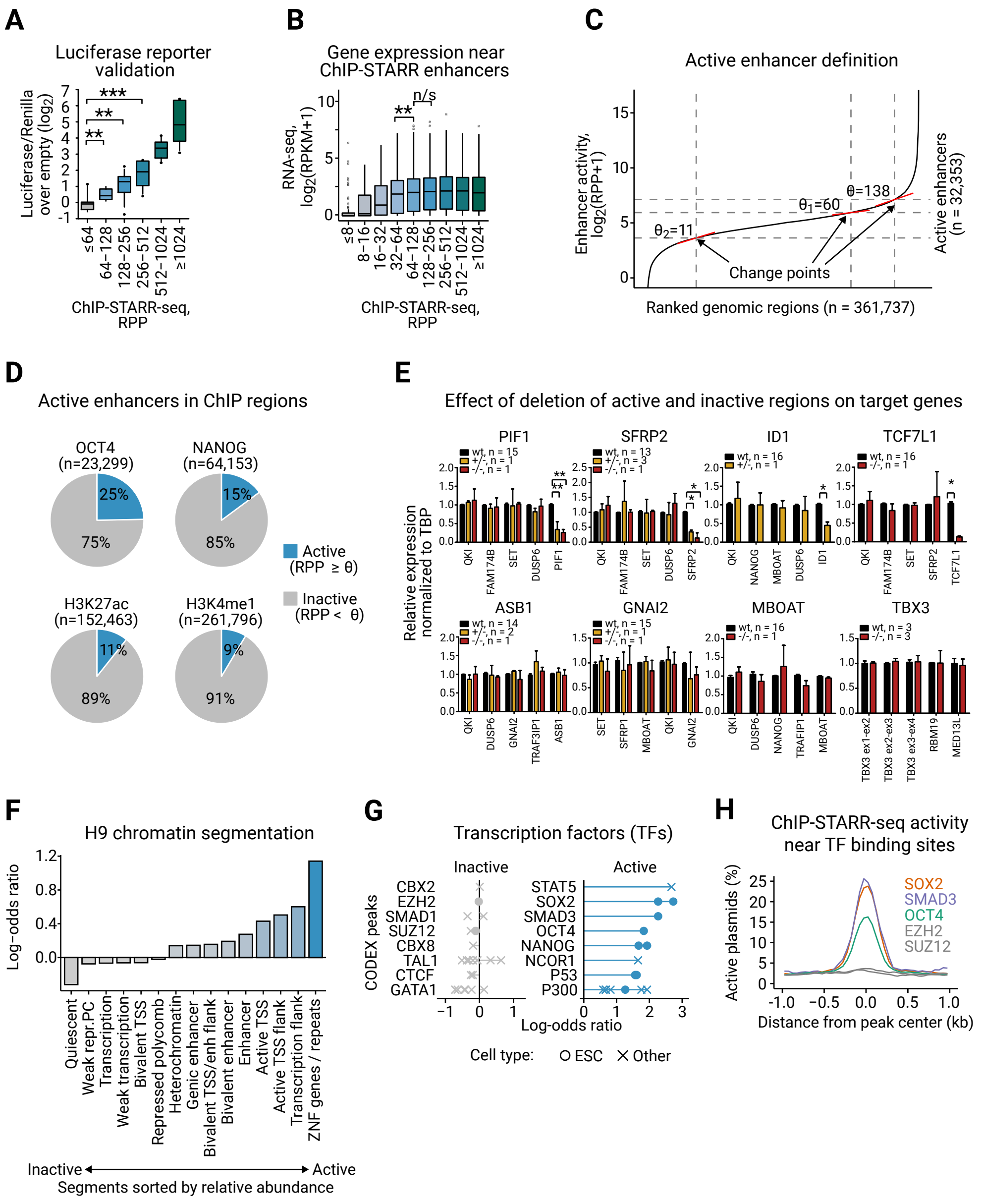
We first assessed whether the plasmid libraries achieved a good representation of the binding events captured by ChIP-seq (**File S1**). A good correlation between ChIP-seq coverage and the corresponding plasmid libraries was seen both pre- and post-transfection (**Figure 1B,C S2B,C**). Next, the ability of the plasmid libraries to drive GFP expression in primed ESCs was tested. Library transfections produced up to 20% GFP-positive cells compared to <1% GFP-positive cells obtained by transfection of the empty STARR-seq vector or ~50% cherry-positive in control transfections with a constitutively expressed mCherry plasmid (**Figure 1D** and data not shown). Therefore, a considerable proportion of cells contained plasmids with enhancer activity. 24h post-transfection, DNA was prepared from unsorted cells and RNA from FACS-purified GFP-positive cells was amplified for RNA-seq. DNA sequencing confirmed high consistency between the original plasmid libraries and plasmids re-isolated post-transfection (**Figure 1B, S2C**). Positive correlations were also observed between read coverage from STARR-RNA-seq and the respective plasmid libraries (**Figure 1E, S2D**) and between replicate STARR-RNA-seq datasets, with an increase for expressed plasmids sampled in replicates (mean correlation r =0.77 at read count ≥ 5). These results show that while abundant plasmids can produce more RNA, some plasmids produce RNA in excess of the plasmid count, indicating high enhancer activity. However, many plasmids transfected into cells did not produce RNA indicating that the ChIP-enriched DNA in these plasmids lacked enhancer activity.

Visual inspection of selected genomic regions illustrates the broad spectrum of enhancer activity measured by ChIP-STARR-seq (**Figure 1F,G**). For instance, ChIP-seq for NANOG indicates two strong binding sites up- and downstream of SOX2 (**Figure 1F**) but only the downstream binding site resulted in ChIP-STARR-seq RNA in excess of plasmid abundance.

Activity levels define classes of enhancers bound by distinct transcription factors

Using ChIP-STARR-seq, we assessed the functional capacity of 361,737 genomic regions (**Table S2**). Enhancer activity was defined as the ratio of RNA reads relative to plasmid reads after normalization (RPP, reads per plasmid and per million sequenced reads). Paired-end sequencing enabled unequivocal assignment of RNA reads to plasmids. The activity level of each region was recorded as the activity generated by the most active plasmid (from any library) within this region. The activities of sixty-eight genomic regions covering the full activity range were compared with luciferase-based assays, and included regions covered in ChIP-seq and evaluated as not active in the STARR-seq assay. DNAs from regions of <64 RPP had luciferase activities indistinguishable from empty vector. In contrast, regions with increasingly high ChIP-STARR-seq activity showed gradually higher luciferase activity (**Figure 2A**). Using different minimal promoters did not affect the activity calls of selected regions (**Figure S3A**). To assess the relationship of activity classifications to gene expression, each region was assigned to a putative target gene based on genomic distance. ChIP-STARR-seq regions with enhancer activity were associated with genes that showed significantly higher gene expression values than genes associated with regions lacking enhancer activity (**Figure 2B, S3B**). To simplify further analysis and ease interpretation, we defined thresholds for discriminating genuine enhancer activity from the activity of the minimal promoter in the STARR-seq by examining mathematical changepoints in the ranked curve of RPP values (**Figure 2C**). The greatest changepoint (θ ≥138) was taken as the threshold to define active enhancers. Based on these thresholds, ChIP-STARR-seq identified 32,353 active enhancers (**Figure 2C**, **File S1**).

Applying this threshold to regions bound by NANOG, OCT4, H3K4me1, H3K27ac or combinations of these factors, indicates that only a minority of ChIP-seq peaks showed enhancer activity (**Figure 2D, S3C**), with regions bound by OCT4 having the highest proportion of high activity enhancers. To determine whether activity predictions from the plasmid-based assay identified enhancers functional at the endogenous loci, ESCs with deletions of regions exhibiting or lacking STARR-seq activity were engineered using CRISPR-Cas9 (**Figure 2E, S3D**). Changes in gene



expression at each locus were observed only for the target gene and only when an active element was deleted. Removal of inactive regions was without effect.

The endogenous context of assessed regions was examined by comparing our data to public reference datasets starting with the H9 chromatin segmentation[32] (**Figure 2F**). Chromatin segments marked as enhancers, transcription start sites (TSSs), sites flanking transcription and repeat sequences were most overrepresented in active regions. The relative representation of TFs from 190 ChIP-seq datasets from CODEX were next assessed by LOLA enrichment analysis[33,34] (**Figure 2G**, **Table S3**). High activity enhancers were preferentially associated with pluripotency-related TFs (SOX2, SMAD3, OCT4, NANOG). Overlaps were also seen for regions bound in non-ESCs by STAT5 and NCOR1. In contrast, no TFs were enriched at inactive regions. Similar results were obtained by extending the analysis to 690 ChIP-seq datasets for TFs from ENCODE[11] (**Figure S3E**). Enhancer activity was strongest close to the binding peaks of enriched factors with activity lost quickly with increasing distance from the peak center (**Figure 2H, S3F**). These results suggest that binding of distinct TFs in close proximity may contribute to robust enhancer activity. How enhancer classes relate to chromatin state was further examined by LOLA analysis of ENCODE chromatin segmentations from H1 ESCs and various non-pluripotent cell types

Figure 2 | Activity levels define functional classes of enhancers. A) Luciferase activities of 68 genomic sequences in primed ESCs grouped by ChIP-STARR-seq activity. Boxes are interquartile range (IQR), line is median, whiskers are 10th to 90th percentile. *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$, Mann-Whitney test; n=2. **B)** Distribution of expression values (Takashima et al., 2014) of genes associated with enhancers grouped by activity level. Boxes are IQR, line is median, whiskers extend to 1.5xIQR, dots are outliers. **, $p<0.01$; ***, $p < 0.001$, unpaired t-test. **C)** Plot showing enhancer activity (enrichment of ChIP-STARR-seq RNA over plasmids; $\log_2$) ranked from lowest to highest across all measured enhancers (union of all peak calls). Enhancers were distinguished based on activity; dashed lines indicate thresholds (θ). **D)** Distribution of active (RPP ≥138) and inactive sequences (RPP <138) in peaks called for the indicated factors. **E)** qRT-PCR analysis of wild type (wt) and enhancer-deleted heterozygous (+/-) or homozygous (-/-) ESC clones. Indicated mRNAs are normalized to TBP (wt = 1) and average results for the indicated deletions are plotted relative to wild type, *n* = number of cell lines per genotype (see STAR methods for further details); *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$ (2-way ANOVA, Bonferroni post-test), error bars = SD. **F)** Relative enrichment of H9 chromatin segment overlaps (Kundaje et al., 2015) between regions with ChIP-STARR-seq activity and inactive regions (see C). TSS, transcription start site; enh, enhancer; PC, polycomb; ZNF, zinc-finger protein. **G)** Relative LOLA enrichment of TFs from CODEX (Sanchez-Castillo et al., 2015) in inactive regions and active enhancers. Odds ratios between observed frequencies of enhancers overlapping binding sites for the eight most enriched TFs in the respective groups relative to the percentage in the entire region set are shown, ranked by mean odds ratio. Each dot represents a TF ChIP-seq dataset. ChIP-seq datasets from non-ESCs are shown as crosses. **H)** Smooth line plots of the proportion of active plasmids (RPP ≥220) around the peak center for the indicated ChIP-seq binding sites.

(**Figures S3G,H**). This confirmed that active enhancers were enriched in segments annotated as H1 enhancers and promoters, while inactive regions occurred primarily in closed chromatin. Together, these results indicate that ChIP-STARR-seq can distinguish ChIP-seq peaks on the basis of enhancer activity and that enhancer activity reflects expression and regulatory function at the endogenous loci.

Sequence determinants of enhancer activity

To address what distinguishes active enhancers from inactive regions, a machine learning approach was used to train a classifier to distinguish both types of regions based on sequence features (conservation, GC content, dinucleotide frequencies) and TF binding motif occurrence (see **STAR-Methods**). Mediocre classifier performance was achieved (AUC= 0.72; **Figure 3A**). The most informative features for enhancer activity were sequence conservation, ESC-related TF binding motifs occurrence and various dinucleotide frequencies (**Figure 3B**), in line with recent observations from other MPRA data[35,36]. The top-3 enriched TFs were found in higher abundance at regions with increasing RPP (**Figure 3C**). Our analysis highlights sequence features influencing enhancer activity but indicates that computational analysis with the simple features assessed could not unequivocally predict activity.

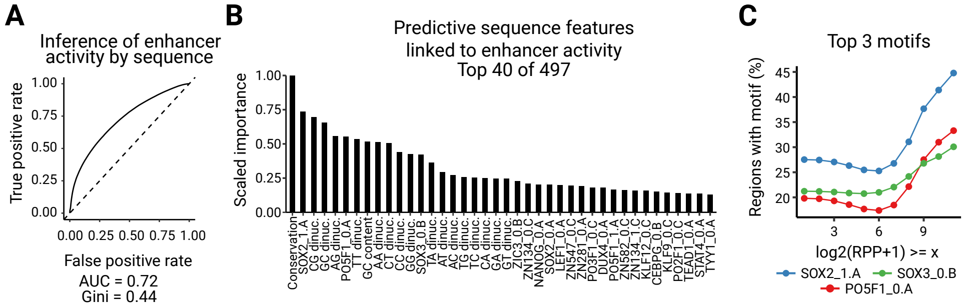
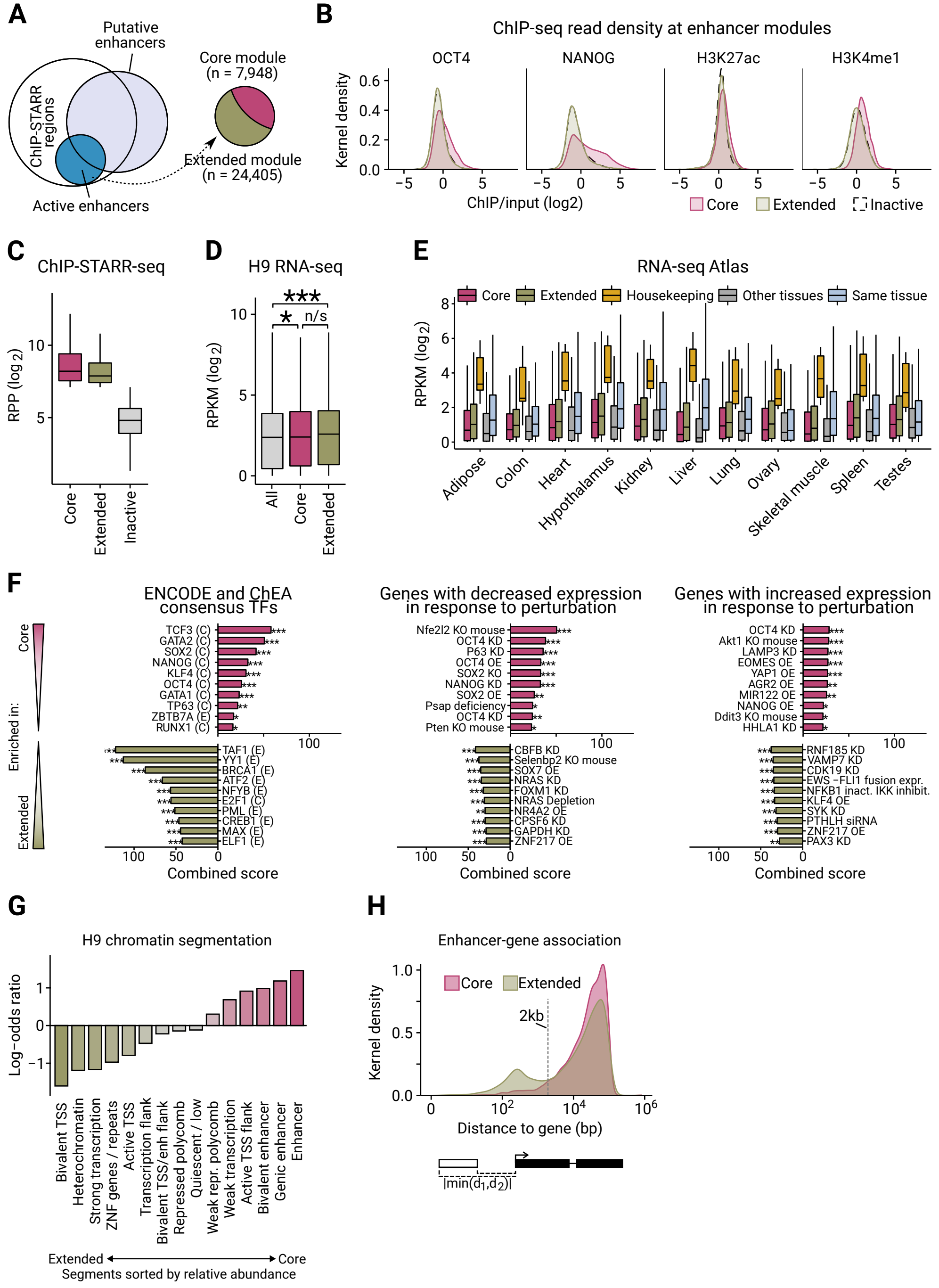


Figure 3 | Sequence determinants of enhancer activity. A) Receiver operating characteristic (ROC) curve of the random forest classifier performance; AUC, area under the curve. **B)** The top-40 sequence features used to distinguish active and inactive regions ordered by variable importance. HOCOMOCO motif IDs were shortened (Kulakovskiy et al., 2016). **C)** Line plots of the percentage of regions containing one of the top-3 motifs from HOCOMOCO as a function of enhancer activity. Each point is the fraction of regions with at least $\log_2$(RPP+1) also containing the respective motif.

Active ESC enhancers include an extended module containing enhancers associated with housekeeping functions

High-throughput sequencing studies have attempted to predict ESC enhancers on the basis of histone marks, TF binding or DNAseI hypersensitivity[13,37,38]. However, the overlap between enhancers predicted from these studies is limited (**Figure S4A**). Comparing the combination of three previously described enhancer maps with our dataset, 7,948 of the 32,353 active enhancers identified by ChIP-STARR-seq were among these predicted enhancers (n = 76,666; union of all datasets; **Table S2**). Several putative enhancers predicted by these previous studies that were inactive by ChIP-STARR-seq were tested in luciferase assays but none possessed enhancer activity in this assay (**Figure S4B**). Enrichment analysis using GREAT[39] showed that the active ChIP-STARR-seq enhancer subset overlapping with previously predicted enhancers had stronger enrichment for gene ontology (GO) terms related to ESC biology than terms identified from all predicted enhancers (**Table S3**). This “core enhancer module” (**Figure 4A**) includes enhancers in close proximity to ESC TFs (NANOG, OCT4) and signaling pathway genes (TGF-b, FGF, WNT signaling). The remaining 24,405 enhancers with high ChIP-STARR-seq activity, that were not predicted previously, had GO terms associated with more generic processes; e.g., regulation of transcription, chromosome organization, housekeeping processes and cytoskeleton organization. We therefore refer to these enhancers as the “extended enhancer module”.

A comparison of the ChIP-seq signal intensity for all peaks to peaks associated with either the core or extended module indicates that enhancers of the extended module generally had slightly lower association with H3K4me1, NANOG and OCT4 (**Figure 4B**). Reduced NANOG and OCT4 binding suggests that extended enhancers rely less on ESC-specific TFs, which is supported by a machine learning classifier attempt to discriminate enhancer modules based on sequence features (**Figure S4D,E**). This analysis demonstrated that core enhancers could be identified by CG dinucleotide frequency and GC content as well as the occurrence of OCT4 and NANOG binding motifs. Nonetheless, the extended module sequences are bona fide enhancers, as their activities are similar to core enhancers (**Figure 4C**). Similarly, the expression of genes associated with the core and extended enhancer modules was comparable, with both gene sets expressed significantly above average ($p < 0.05$; **Figure 4D, S4C**). Consistent with function in many cell types, expression of genes associated with the extended enhancer module was higher than core-module-associated genes in data from somatic tissues obtained from the RNA-seq Atlas[40] (**Figure 4E**) and GTEx[41]



(**Figure S4F**). To provide context, orthogonal "housekeeping" [42] and "tissue-specific" gene sets[43] were included in this analysis. Enrichment analysis using Enrichr[44] with data from ENCODE[11] or ChEA[45] showed that core module enhancers were enriched near genes bound by NANOG, TCF3, SOX2 and OCT4, whereas extended enhancer module enhancers showed preferential enrichment of broadly expressed factors, such as TAF1, YY1, BRCA1 and ATF2 (**Figure 4F**, **Table S3**). Core enhancers were often found in regions associated with enhancer-like chromatin in H9[32] (**Figure 4G**). In contrast, ~6% of extended module enhancers are annotated as heterochromatic or bivalent in H9 chromatin, suggesting that the activity of these enhancers may be suppressed by endogenous chromatin. The majority of enhancers from either the core or extended modules showed a similar distance distribution around TSSs, although a subset of extended module enhancers (n= 4,731) lie within 2 kb of TSSs (**Figure 4H**). GO terms associated with the TSS-proximal subset are enriched for terms related to metabolic processes and housekeeping functions, whereas terms associated with TSS-distal enhancers include cell fate and differentiation annotations (**Table S3**). This indicates that a subset of extended module enhancers may be linked to housekeeping genes. ChIP-STARR-seq therefore identified by function, previously unappreciated enhancer sequences characterized by lower enrichment of enhancer-associated histone modifications and pluripotency-related TFs but with comparable enhancer activity.

Figure 4 | Active enhancers include core and extended ESC-enhancer modules. A) The overlap between published putative enhancers (Hawkins et al., 2011; Rada-Iglesias et al., 2011; Xie et al., 2013) (light blue) and regions assessed by ChIP-STARR-seq (white) or called active (RPP ≥138; blue). We refer to ChIP-STARR-seq enhancers overlapping published putative enhancers as the core module and non-overlapping regions as the extended module. **B)** Kernel density plots of the distribution of enrichment values in ESCs for the indicated factor for peaks associated with the core or extended modules or for inactive regions. **C)** RPP values for all assessed genomic regions compared to enhancers from the core or extended modules. Boxes are IQR, line is median, whiskers extend to 1.5xIQR. **D)** RNA-Seq in H9 (Takashima et al., 2014) for all genes compared to genes associated with either core or extended enhancer modules. Boxes as in C. RPKM, reads per kilobase million. *** = p<0.001 (t-test). **E)** Gene expression in tissues from the RNA-seq Atlas (Krupp et al., 2012) for all genes linked to the core or extended modules. Housekeeping (Eisenberg and Levanon, 2013) and tissue-specific genes (Lachmann et al., 2017) are also shown. Tissue-specific genes are split into the one indicated (same; x-axis) or "other tissues". As no tissue-specific gene set was available for hypothalamus, whole-brain-specific genes were used. Boxes as in D. **F)** Enrichment analysis (Enrichr) testing genes associated with the core (top) and extended (bottom) modules. Top-10 results for TF binding sites from ENCODE and ChEA (left) and genes down-regulated (middle) or up-regulated (right) upon single-gene perturbations from GEO. **G)** Relative enrichment (log-odds ratio in ESCs compared to all) of H9 chromatin segments (Kundaje et al., 2015) in core and extended module enhancers. **H)** Kernel density plot of the distance to associated genes for core and extended module enhancers. Shortest distance from either enhancer region boundary was recorded.

Major changes in enhancer activity upon induction of naive pluripotency

To augment the catalogue of functional enhancers in ESCs and to gauge the dynamics of enhancer activity we applied ChIP-STARR-seq to a closely related cell type. Primed H9 ESCs were converted to naive ESCs (**Figure S5A-D**). Characterization of established cultures agreed with prior studies[7,46], as did ChIP-qPCR and ChIP-seq for NANOG, OCT4, H3K4me1 and H3K27ac (Figure S5E-I). ChIP-STARR-seq plasmid libraries generated from naive ESCs (**Figures 5A, S6**) were transfected into naive ESCs and for comparison, into primed ESCs. Transfections followed by RNA-seq readout yielded measurements of enhancer activity in naive ESCs comparable to those obtained previously in primed ESCs, albeit at slightly lower reproducibility (mean correlation r = 0.63 at read count ≥ 5). Enhancer activity was categorized using the threshold applied previously (**Table S2**, **File S1**). 359,880 regions covered by plasmids in naive ESCs (**Figure S7A**) were analysed, identifying 36,417 enhancers. Again, only a fraction of ChIP-seq peaks displayed activity with peaks marked by OCT4, H3K27ac and H3K4me1 showing the highest proportion of activity (**Figure S7B**). LOLA enrichment analysis of TFs from CODEX for the naive enhancer class (Figure 5B, Table S3), identified a similar TF profile as in primed ESCs (compare to Figure 2G). Sites bound by pluripotency-related TFs (e.g., SOX2 and NANOG) were also strongly represented at enhancers active in naive ESCs. Enrichment analysis of ENCODE ChIP-seq datasets (**Figure S7D**) and chromatin segmentations (**Figure S7E**)[32,47] confirmed overlap with ESC TF binding sites.

Having an extensive genome-wide enhancer maps for both pluripotent states allowed a global comparison of enhancer usage in both primed and naive ESCs (**Figure 5C**). Only 18% of enhancers active in primed ESCs maintained activity in naive ESCs (Active→Active), while 82% became inactive (Active→Inactive). Conversely, 9% of inactive regions in primed ESCs gained activity (Inactive→Active). Despite these extensive changes, the relative ranking of RPP values is stable, indicating that the highest and lowest activity score are comparable (**Figure S7F**). The changes in activity are not explicable by altered affinity of TF binding alone, as illustrated by discriminating peaks into strongly and weakly bound regions (**Figure S7G**) and applying the same analysis to ChIP-seq affinity values (**Figure S7H**). For instance, only 36% of regions that maintained strong enhancer activity in both states were also strongly bound in both states, while 15.3% of regions switched from strongly to weakly bound or vice versa. Enrichment analysis of enhancers maintaining or switching activity level (**Figure 5D, S7I, Table S3**) revealed that enhancers with high activity

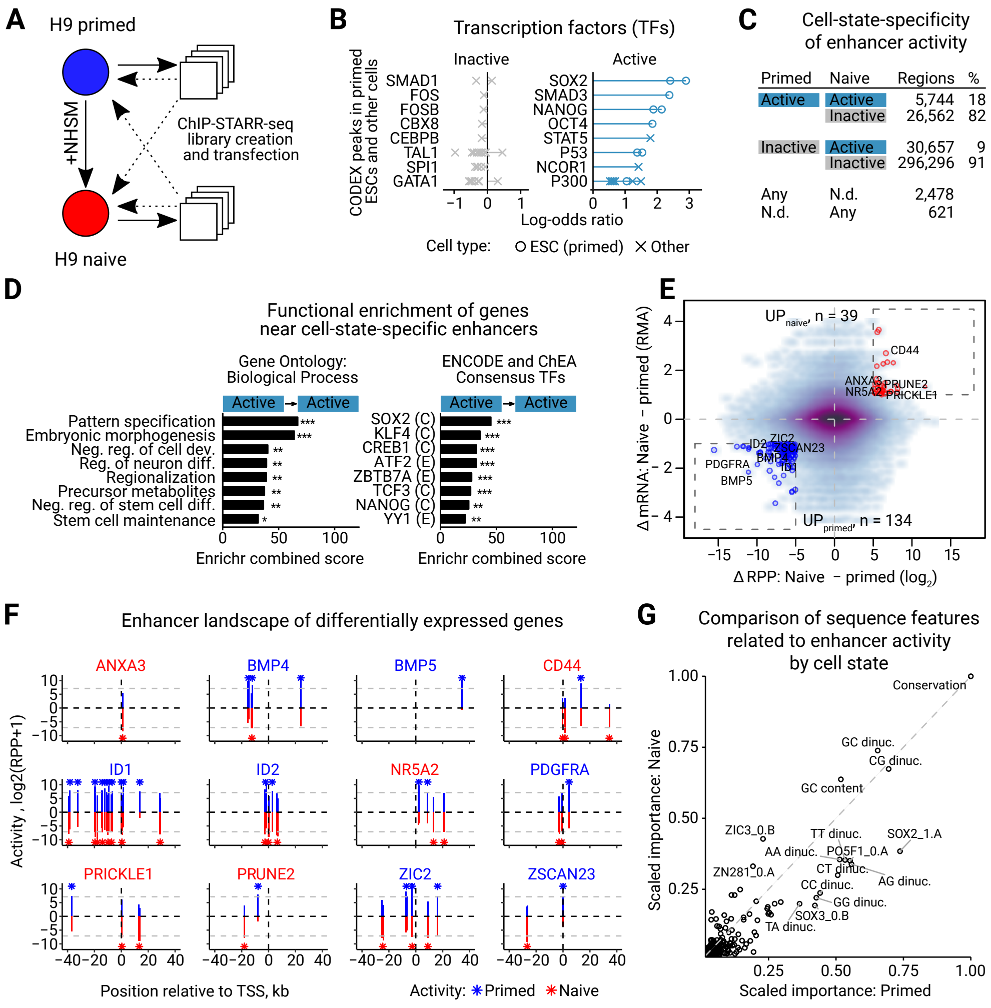


Figure 5 | Changes in enhancer activity upon induction of naive pluripotency. A) Overview of primed to naive conversion and ChIP-STARR-seq cross-over design. **B)** Relative enrichment of TFs from CODEX (Sanchez-Castillo et al., 2015) in inactive, active enhancers in naive hESCs. Plots as in Figure 2G. **C)** Table of relative changes in enhancer activity between primed and naïve ESCs. **D)** Enrichment analysis (Enrichr) to test genes near enhancers active in both primed and naive ESCs against GO assignments (left) or binding sites from ENCODE and ChEA ChIP-seq (right). **E)** Scatterplot contrasting average changes in enhancer activity with changes in associated gene expression. Genes with strong concordant changes in enhancer activity and gene expression are shown using the thresholds: |max(ΔRPP)| ≥ 5, |mean(ΔmRNA)| ≥ 1. **F)** Visualization of enhancer activity in ChIP-STARR-seq regions near selected genes (boxes in panel E; TSS+/-40kb) with differential expression in primed and naive ESCs. Bars indicate enhancer activity (RPP) in primed (blue) and naive (red) ESCs. Grey dashed bars indicate activity threshold for active enhancers. Active enhancers are highlighted with asterisks. Gene name color shows the state expressing the gene the highest. **G)** Scatterplot of scaled variable importance of sequence features used to discriminate active and inactive regions in primed and naive ESCs. In both cases, a random forest classifier was trained.

in both cell states (Active→Active) were related to suppression of differentiation processes and maintenance of stem cells, whereas genes near enhancers that lost activity (Active→Inactive) were annotated with generic expression-related terms. No significant GO terms were associated with enhancers that gained activity or regions that remained inactive, though this may be due to lack of annotation in naive ESCs. However, examining ChIP-seq data from ENCODE and ChEA indicated that enhancers that were active only in naive cells were enriched for transcriptional activators such as ATF2, TAF1 or BRCA1 that occur near target promoters. Comparative analysis of core and extended module enhancers (see **Figure 4**), showed that core enhancers were significantly ($p < 2.2\times10^{-16}$) more likely to be active in naive ESCs than either extended module enhancers or enhancers inactive in primed ESCs (**Figure S7J**).

To relate changes in enhancer activity to differences in the expression of regulated genes, the average difference in enhancer RPP levels between naive and primed ESCs was plotted against the expression of nearby genes (**Figure 5E**). We highlighted genes with at least one strong enhancer change. Detailed examination of the ChIP-STARR-seq regions in the proximity (<=40kb) of the TSS of these genes (**Figure 5F,** Supplemental Website) confirmed increased enhancer activities for several genes that were expressed higher in naive ESCs (e.g., CD44, ANXA3). In contrast, several genes were expressed more highly in primed ESCs and in each case enhancers with increased activity in primed ESCs could be identified that may drive preferential expression in primed ESCs (e.g., BMP4/5, ID1/2). Notably, some genes showed concordant changes in multiple adjacent enhancers that presumably jointly drive expression changes (PRICKLE1, BMP4), while other genes switched activity from one enhancer to another (PRUNE2, CD44, ZSCAN23). The catalogue of functional enhancers presented here will help to decipher the complexity of enhancer/target interactions directing gene expression.

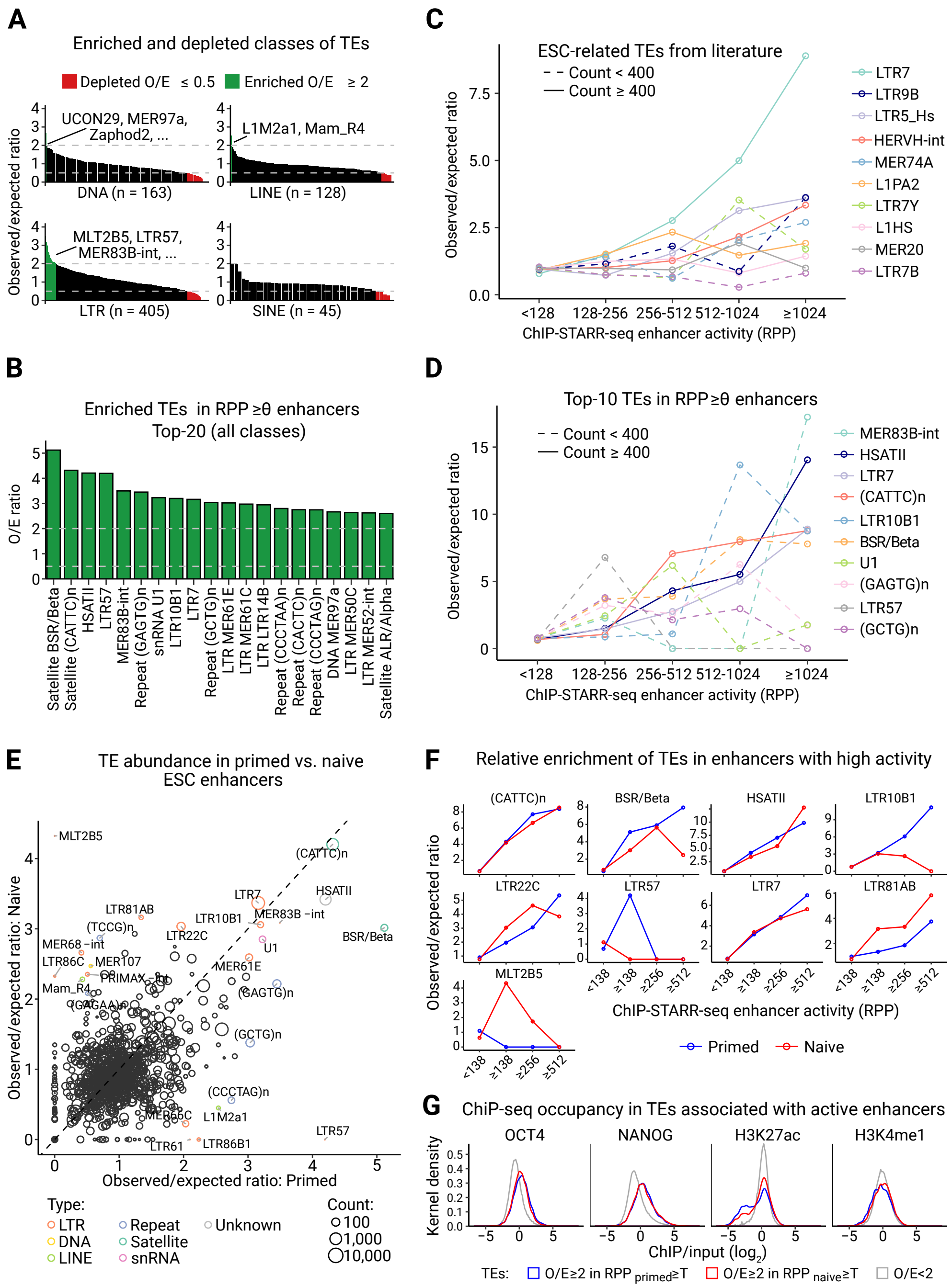
We next trained a classifier to discriminate active from inactive regions in naive ESCs and compared the results to those we obtained previously (**Figure 5G, S7K**; cp. **Figure 3**). We find a consistent contribution of evolutionary conservation and GC/CG dinucleotide frequencies to enhancer activity. Notably, the relative importance of TF binding motifs shifts slightly between naïve and primed: e.g., ZIC3 is linked to naïve ESCs[48] and SOX3 is linked to primed ESCs, in line with a recent report on primed pluripotent mouse cells[49].

The occurrence of various transposable elements is associated with enhancer activity

As chromatin associated with repetitive DNA was found in active enhancers (**Figure 2F**), we examined the link between repeats and enhancer activity more closely. Large portions of mammalian genomes are derived from TEs which are linked to TF binding sites[50-53], but whether this enrichment reflects enhancer activity has not been determined genome-wide. To assess ChIP-STARR-seq enhancers for the occurrence of TE sequences, we used the RepeatMasker annotation[54]. The number of TE-derived sequences in active and inactive regions was compared to the number detected in all genomic regions (**Figure 6**, **Table S4**). LTR-containing TEs, such as LTR57, were enriched in primed ESCs enhancers (**Figure 6A**). However, not all LTR-containing TEs were enriched at active enhancers. The most enriched elements were satellite repeats and LTR family members (**Figure 6B**). For TEs enriched for NANOG and OCT4 binding (e.g., LTR9B)[52] or TEs enriched at candidate human-specific regulatory loci (e.g., LTR7)[51], the observed enrichment increased further with increasing activity (**Figure 6C,D**). Indeed LTR7, LTR9B and HERVH-int show the strongest enrichment at the highest activity enhancers. In contrast, other TE families previously linked to human-specific TF binding sites[51], were either not (L1HS) or only weakly (L1PA2) enriched at active enhancers. Although many repeat families were found equally in primed and naive ESCs (e.g. LTR7, (CATTC)n), other families showed less or no enrichment in one of the two states (e.g., LTR81AB, LTR57; **Figure 6E, F**). In general, TEs that were overrepresented in active enhancers showed increased binding of NANOG and OCT4, but not H3K27ac or H3K4me1 (**Figure 6G**). These results indicate that certain families of TEs are overrepresented at active enhancers and that their enrichment correlates with enhancer activity in a cell-state-dependent manner. However, not all TEs of the same type are associated with active enhancers, nor do all TEs enriched in pluripotency TF binding sites occupy active enhancers.

ChIP-STARR-seq dissects super-enhancers into small functional units

Recently, large linear tracts of chromatin, referred to as SEs have been identified that function to regulate lineage-specific gene expression[55,56]. Compared to traditional enhancers, SEs have increased binding of Mediator, specific histone marks and lineage-specific TFs. Whether the full length of SEs is required for biological activity is a matter of debate[57-60]. We used our enhancer catalogue to dissect the regulatory potential of DNA underlying SE regions. SEs were first identified by H3K27ac



enrichment in primed (**Figure 7A**, **File S1**) and naive (**Figure S8A**) ESCs. Alignment of ChIP-STARR-seq data to these SEs showed that the H3K27ac intensity used to define SEs correlated to RPP levels (**Figures 7B, S8B,C**), supporting the notion that SE-likeness is an indicator of enhancer activity. SEs discovered here overlapped strongly between primed and naive ESCs (n= 824 SEs shared), containing many of the previously described H1 ESC SEs (**Figure S8D,E**)[55]. Detailed examination of the FGFR1 SE indicated strong RPP signals originating from small regions within the SE (**Figure 7C**). To exclude the possibility that this observation was due to limited coverage in our ChIP-STARR-seq libraries we included additional STARR-seq libraries made from BACs covering the FGFR1 SE and two other SEs providing robust coverage of the entire SEs plus flanking regions (**Figure 7C, S8F**). Luciferase assays confirmed spatially restricted enhancer activity of DNA in the neighborhood of the central active region of the FGFR1 SE. Strong activity was confined to a 596 bp region with other DNA elements from this SE devoid of enhancer activity (**Figure 7D**). Homozygous deletion of this region by CRISPR-Cas9 reduced expression of FGFR1 and WHSC1L1 significantly compared to wild type cells, without affecting expression of other flanking genes (**Figure S8G,H**). Homozygous deletion of two other parts of this SE did not affect gene expression of target and flanking genes. This indicates that the FGFR1 SE is composed of small units with enhancer activity. To test whether this finding is valid globally, the relative abundance of active plasmids (RPP ≥138) in SEs compared to "normal" enhancers (NEs) was examined. Most enhancers contained only a small percentage of active plasmids within their bounds (**Figures 7E,F**). Although this fraction was slightly higher in SEs than in NEs, it accounted for only a minority (2.8%) of the genome annotated as SEs. Therefore, only a small part of the large SEs has enhancer function (**Figure 7F, S8I**). Notably, regions within naive SEs or within SEs called in both primed and naive were more frequently active in both states (18.1% and 13.2%, respectively) than regions within

Figure 6 | Distinct transposable elements are associated with enhancers of differing activity in ESCs. A) Enrichment ratios for the occurrence of TE families (LTR, DNA, SINE, LINE) in high activity ChIP-STARR-seq enhancers (RPP ≥138). **B)** Top-25 most enriched TE families in active enhancers. **C)** Enrichment ratio versus activity level for distinct TE families. **D)** As C) but for the top-10 most enriched families of TEs in B. **E)** Comparison of the enrichment ratios in primed and naive ESCs. Each repeat element is shown by a dot with the size proportional to the number of overlaps with ChIP-STARR-seq regions. Elements with O/E≥ 3 in naive or primed, or with strong differences between both (O/E≥ 2 and $\Delta\log_2$(O/E)>=2) are labelled. **F)** Relative enrichment of selected TEs (from E) in primed (blue) and naive (red) ESCs as a function of enhancer activity level (RPP). **G)** Kernel density plots of coverage (ChIP-seq/input) in ESCs for the indicated factor for all TEs overrepresented (O/E>2) in active enhancers.

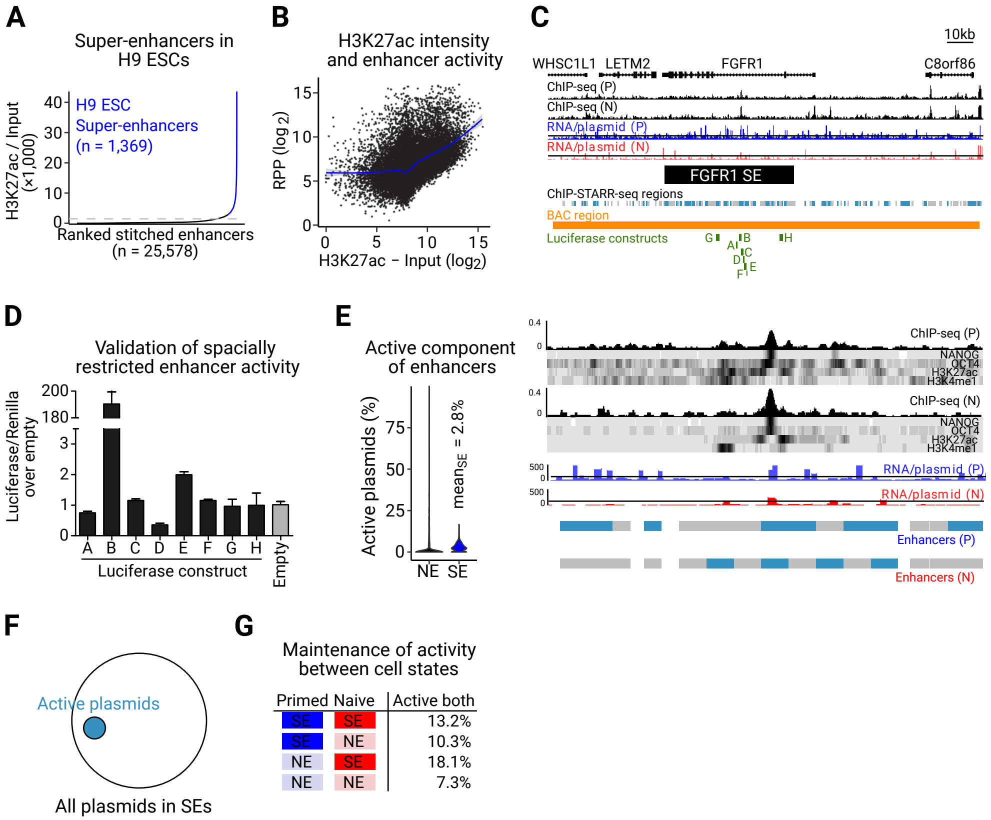


Figure 7 | ChIP-STARR-seq dissects super-enhancers into functional elements. A) SEs were called from H3K27ac ChIP-seq data using ROSE (Whyte et al., 2013). **B)** Scatterplot of SE intensity (H3K27ac enrichment over input) with ChIP-STARR-seq activity. *r*, Pearson correlation; blue line indicates a generalized additive model fit. **C)** SE overlapping *FGFR1*, with ChIP-seq tracks for the indicated factors in primed/naive ESCs. Top plot shows SE locus; bottom plot zooms into second intron. Shown are the positions of regions assessed by ChIP-STARR-seq (grey) and active enhancers (blue) from this study and coordinates of luciferase constructs matching selected enhancers (labeled A-H). Enhancer activities are concentrated at a single position within the SE. **D)** Luciferase assays of DNA sequences depicted in green in C); n=2, error bars represent SD. **E)** Violin plots of the proportion of active plasmids (RPP ≥138) for 1,369 SEs compared to normal enhancers (NE). **F)** Sketch of the active subspace (covered by plasmids with RPP ≥138) of the entire SE space (all plasmids occurring within SEs). **G)** Table of the percentage of ChIP-STARR-seq plasmids representing regions within SEs and NEs active in primed and naive ESCs (RPP ≥138). Groups of enhancers that were called SEs in both, in only one or in neither state are distinguished.

primed SEs or outside SEs (**Figure 7G**). Since only a subspace of SEs displayed enhancer activity, we investigated the relationship between active components and H3K27ac ChIP-seq peaks by repeating the SE calling without stitching disjoint peaks (ROSE stitching distance= 0). However, the fraction of active plasmids remained unaffected indicating that H3K27ac occupancy alone cannot identify active SE components (**Figure S8J**).

Discussion

In this study, we present a large-scale analysis of ESC enhancer activities. By using ChIP-STARR-seq we assessed the ability of sequences bound by OCT4, NANOG or marked by H3K4me1 and H3K27ac to function as enhancers. Our results show that only a subset of these sequences displayed enhancer activity. We find that TF binding is linked with enhancer activity, in line with recent reports[22,61,62], but that no individual TF, histone mark or combination thereof could unequivocally predict enhancer activity. Our study identified a previously unrecognized group of functional enhancers that are active in ESCs but are associated with generic cell processes. This extended enhancer module is characterized by reduced binding of pluripotency-associated TFs and histone marks. This reduced binding might have placed these regions below the detection threshold in previous ChIP-seq-based studies that lacked a functional read-out.

The use of an episomal-plasmid-based reporter system may be considered a limitation, as it does not fully recapitulate endogenous chromatin context[63]. It is also possible that in some cases cloned fragments might be too short to enable all the TF interactions that mediate enhancer function at the endogenous locus. However, the generally accepted definition of an enhancer focusses on the functional capacity of DNA to enhance transcription of a reporter gene in an orientation and position-independent manner[1]. Indeed, several lines of evidence argue for the broad usefulness of ChIP-STARR-seq as a high-throughput assay of enhancer function: 1) ChIP-STARR-seq confirmed the function of known enhancers; 2) genes near active enhancers tend to be more highly expressed; 3) active enhancers are marked by motifs of TF associated with ESCs; 4) active enhancers are enriched in genome annotations as enhancer chromatin; 5) deletion of active enhancers from endogenous loci decreases expression of linked target genes, whereas deletion of sequences devoid of enhancer activity in ChIP-STARR-seq did not affect gene expression.

Previous studies identified crucial roles for OCT4, NANOG and SMAD3, the latter of which are downstream mediators of TGF-β signaling in the maintenance of ESC pluripotency[64,65]. Enhancer activity is enriched near these binding peaks, suggesting that these TFs may act combinatorially to provide enhancer function. Other studies have shown that heterotypic clusters of different TF binding sites can increase enhancer activity[19] and that sequences marked by H3K122ac but lacking H3K27ac can act as transcriptional enhancers[66]. It will be of future interest to decipher the individual contributions of TFs to these active enhancers. Several classes of TEs were also enriched at active enhancers, as reported recently[61]. TEs are enriched in species-specific TF binding sites and have been hypothesized to shape the enhancer network in ESCs [51,52]. Our data indicate that only a limited number of TEs contribute to enhancer function and can do so in a cell-state-dependent manner.

Most enhancers studied to date lie within distal elements or intronic sequences. However, some sequences detected by ChIP-STARR-seq lie near TSSs (n= 3,283 active enhancers within 500 bp of a TSS). As tested enhancers are inserted downstream of the GFP ORF in STARR-seq (**Figure 1A**) GFP-positive transcripts cannot be made by initiating transcription in situ from an inserted TSS. Therefore, sequences near a TSS can exert enhancer activity, in line with recent reports[67,68]. Furthermore, a subset of extended module enhancers lies close (+/- 2 kb) to a TSS and display GO enrichments related to housekeeping genes and metabolic processes. This suggest that nearby enhancers may regulate some human housekeeping genes. It will be interesting to investigate links between enhancers and promoters that distinguish housekeeping genes from developmental genes, as identified in Drosophila[69-71].

Several groups have recently developed culture conditions supporting a more naive ESC state enabling contribution to interspecies chimaeras[5,7,8,72]. Here we have used one such culture condition to compare primed and naive ESCs and find that enhancer activity is altered substantially. Pluripotency in both states is established by differential use of regulatory elements that is partly reflected in gene expression changes. Further studies should clarify differences between states of pluripotency and how these relate to altered enhancer usage.

SEs are characterized by large domains marked by H3K27ac with increased binding of Mediator and other TFs. ChIP-STARR-seq analysis indicates that the majority of sequences within SEs lack enhancer activity. Rather, enhancer activity is limited to small domains within the SEs that frequently overlap with TF binding sites. This suggests that the observed chromatin signatures at SEs might be a consequence of enhancer activity from much smaller units. Recent reports suggest that SE constituents may function alternatively as either independent and additive enhancers[58,60], as constituents in a temporal and functional enhancer hierarchy[59], or as interdependent units[73] exhibiting synergy[74]. The large scale identification of such active constituents within SEs reported here will help to decipher the regulatory mechanisms contributing to SE formation and function.

The catalogue of functional enhancers presented here provides the means to refine models of the regulatory circuitry of ESCs and a framework for understanding transcriptional regulation in humans. Given the increasing appreciation of the importance of the regulatory genome in health and disease we expect that this resource and the more widespread use of MPRAs such as ChIP-STARR-seq will advance basic and translational research.

Experimental procedure

Cell lines

H9 female human embryonic stem cells were a gift of David Hay (Edinburgh). All cells were regularly karyotyped and checked for the presence of mycoplasm.

Cell Culture conditions

H9 human embryonic stem cells were cultured on Matrigel coated cell culture plates, using mTesR1 medium (Stem Cell Technology, 05850). Cells were routinely split (ratio 1:3-1:4) using 0.5mM EDTA (Invitrogen, 15575020). For transfection, single cells were obtained by Accutase treatment (Invitrogen, A1110501), in the presence of Rock inhibitor, Y-27632 (10uM, Cambridge bioscience, SM02-10). For conversion to the naive state, cells were split on irradiated MEFs on gelatin coated plates and media was changed to NHSM media, as described by Gafni et al. [7], containing knockout DMEM (Invitrogen), 20% knockout serum (Invitrogen), human insulin (Sigma, 12.5mg ml-1 final concentration), 20 ng ml-1 recombinant human LIF (Millipore), 8 ng ml-1 recombinant bFGF (Peprotech) and 1 ng ml-1 recombinant TGF-b1 (Peprotech), 1 mM glutamine (Invitrogen), 1% nonessential amino acids (Invitrogen), 0.1 mM beta-mercaptoethanol (Invitrogen), penicillin-streptomycin (Invitrogen) and small molecule inhibitors: PD0325901 (1mM, ERK1/2i, Axon Medchem); CHIR99021 (3mM, GSKbi, Axon Medchem); SP600125 (10mM, JNKi, Abcam ab120065) and SB203580 (10 mM, p38i,Abcam ab120638) Y-27632 (5mM, ROCKi) and protein kinase C inhibitor G06983 (5 mM, PKCi, Abcam, ab144414). Cells were 1:10 passaged using TrypLE^{TM} (Invitrogen, 12604021) in the presence of Rock inhibitor and maintained for more than 10 passages in NHSM media prior to analysis.

Experimental Design

All experiments were replicated. For the specific number of replicates done see either the figure legends or the specific section below. No aspect of the study was done blinded. Sample size was not predetermined and no outliers were excluded.

Chromatin immunoprecipitation

For chromatin immunoprecipitation, 2x10^7 H9 primed or naive ESC were harvested in 9 ml of medium and cross-linked by addition of 270 ml 37% Formaldehyde (Sigma, final concentration of 1%), for 10 min at room temperature under rotation. 1 ml of 1.25 M Glycine was added, cells were incubated on ice for 5 min and 3x washed with ice cold PBS. At this point, cross-linked cell pellets were snap-frozen and stored at -80°C, or immediately processed for sonication. Prior to sonication, cells were resuspended in 1ml TE-I-NP40 (10mM TRIS-HCl pH 8, 1mM EDTA, 0.5% NP40, 1mM PMSF, 1x Protease inhibitor complex (PIC, Complete tablets, 04693116001, Roche)) incubated on ice for 5 min and centrifuged for 5 min at 2500 rpm at 4°C in a refrigerated bench top centrifuge (Eppendorf). Supernatant was removed and nuclei were resuspended in 1 ml ice-cold lysis buffer (50mM TRIS-HCl pH 8, 10mM EDTA, 1% SDS, 1mM PMSF, 1x PIC) and transferred to a 15 ml Falcon tube for sonication, using a Diagenode Bioruptor Next Gen (40 cycles of 30” on, 30” off). After transfer to an Eppendorf tube and centrifugation for 10 min at 13200 rpm at 4°C, chromatin solution was aliquoted and used for immunoprecipitation or snap-frozen and stored at -80°C. A 20 µl sample was taken and served as a total input control. For immunoprecipitation, Protein Dynabeads G (10004D, Life Technologies) were washed with PBS and incubated for 6 hours with 5 mg of antibody, at 4°C on a rotating wheel. Antibodies used were: goat-anti-NANOG (AF1997, R&D Systems), rabbit-anti-OCT4 (AB19857, Abcam), rabbit-anti-H3K4me1 (AB8895, Abcam) and rabbit-anti-H3K27ac (AB4729, Abcam); as a control, respective IgG antibodies were used (rabbit-IgG: 10500C, Life Technology, goat-IgG: SC-2028, Santa Cruz Biotechnology). After washing with PBS, antibody-coupled beads were incubated with 200 ml chromatin solution, diluted to a final volume of 2 ml with dilution buffer (167mM NaCl, 16.7mM TRIS-HCl pH 8.1, 1.2mM EDTA, 0.01% SDS, 1.1% Triton-X100, 1mM PMSF, 1x PIC), overnight at 4°C on a rotating wheel. Washing of beads was performed by incubation with ice-cold 1 ml of washing buffer, for 5 min, at 4°C on a rotating wheel, followed by removal of supernatant using a magnetic stand, for each of the following: 2x with wash buffer 1 (10mM TRIS-HCl pH 7.6, 1mM EDTA, 0.1% SDS, 1% Triton-X100, 0.1% NaDeoxychloate), 2x with wash buffer 2 (10mM TRIS-HCl pH 7.6, 1mM EDTA, 0.1% SDS, 1% Triton-X100, 0.1% NaDeoxychloate, 150mM NaCl), 2x with wash buffer 3 (250mM LiCl, 0.5% NP40, 0.1% NaDeoxychloate), 1x with TE 1x with 0.2% TritonX-100 and 1x with TE 1x, after which beads were resuspended in 100ul TE1x. Immunoprecipitated chromatin and total input control were decross-linked, by addition of 3 ml of 10% SDS and 5 ml Proteinase K (20 mg/ml, Roche) and 10 ml RNAse A (50 mg/ml, Roche) to each tube and incubation overnight at 65°C on a shaking thermomixer block, 1400 rpm (Eppendorf). The next day, beads were briefly vortexed and supernatants were transferred to new tubes using the magnetic stand. 100ml of TE1x containing 500mM NaCl was added to the beads and briefly vortexed, after which the supernatant was added to the first fraction of collected supernatant. Following Phenol / chloroform extraction, DNA was precipitated using 1ml glycogen (20mg/ml), 1/10 vol NaOAc (3M) and 100% ice-cold Ethanol, at -20°C for 1 hour, followed by centrifugation at 13200 rpm for 1 hour at 4°C. After a final wash with 70% ethanol, the DNA pellet was dried and resuspended in 50ml H_2O. Concentration of ChIP DNA was determined by Qubit measurement following manufacturer’s instructions and sonication was assessed by gel-electrophoresis of total input DNA (target fragment size between 200 and 600 bp).

ChIP-qPCR

Concentration of ChIP and total input control DNA was assessed by Qubit measurement (LifeTech) according to manufacturer’s instructions and was diluted to 2 ng/ml. 2 ml of DNA was used per qPCR reaction, using a 2x Takyon qPCR master mix (No ROX SYBR, UF-NSMT-B0701, Takyon). qPCR reactions were run on a Roche Lightcycler 480 II (Roche), using the following cycle conditions: 95°C 3 min, (95°C 10 sec, 60°C 30 sec, 72°C 25 sec) x45, followed by a melting curve from 95° to 65°C. All data shown are averages of at least 2 biological replicates and 3 technical replicates. All primers used are shown in **Table S5.**

ChIP-seq, ChIP-STARR-seq plasmid library preparation

For ChIP-seq and ChIP-STARR-seq plasmid library generation, 10 ng of ChIP DNA was used as starting material. Using NEB Next ChIP-seq library preparation kit (E6200 or E6240, NEB), DNA was end-repaired, dA-tailed and adapter-ligated according to manufacturer’s instructions. After adapter ligation and purification using AMPure-XP beads (0.8x, Beckman Coulter) and elution into 30ml of 0.1xTE, 25 ml of the reaction product was used for ChIP-seq library preparation, by PCR amplification with Illumina index primers (7335 and 7500, NEB) using the NEB Next Q Hot start high fidelity master mix (M0543S, NEB) according to manufactures instructions (cycle conditions: 98°C 30 sec, (98°C 10 sec, 65°C 75 sec) x15, 65°C 5 min, 4°C hold). After an additional round of AMPureXP bead purification, DNA was eluted in 0.1xTE without further size selection. Quality and quantity of the prepared ChIP-seq libraries was assessed on an Agilent Tapestation. All sequencing occurred on an Illumina HiSeq 2500 platform, using 50 bp single-end sequencing.

The remaining 5 ml of purified adapter ligated DNA were used for ChIP-STARR-seq plasmid library generation. Therefore, DNA was diluted to a total volume of 10 ml in 0.1xTE and used as an input in 8 x 50 ml PCR reactions using Phusion Polymerase, High-fidelity buffer (M0530L, NEB) and primers 147 STARRseq libr FW (TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCT) and 148 STARRseq libr RV (GGCCGAATTCGTCGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT)[15], which prime on the adapter sequences and add a 5'and 3' 15 nucleotide homology sequence to the reaction products which are used for Gibson assembly. After PCR amplification (cycle conditions: 98° 2 min, (98°C 10 sec, 62°C 30 sec, 72°C 30 sec) x 15, 72°C 5 min, 4°C hold), PCR reactions were pooled, purified using AMPure XP beads (1.8x), eluted in 30 ml 0.1xTE and used for Gibson assembly. Therefore, 15 mg of the mammalian STARRseq plasmid (a kind gift of A.Stark)[15] were digested with AgeI-HF and SalI-HF (NEB) for 8h at 37°C, column purified (Nucleospin purification columns, 740609250, Machery-Nagel), eluted in 30 ml elution buffer and used as a vector in a Gibson reaction, using 2 ml of digested plasmid, 5 ml purified PCR product, 3 ml H20 and 10 ml of a home-made Gibson reaction (100mM Tris-HCl, 10mM MgCl2, 0.2 mM dNTP (each), 0.5U Phusion DNA polymerase (NEB), 0.16U 5' T5 exonuclease (Epicentre), 2 Gibson reactions per library. After incubation at 50°C for 1 hour, Gibson reaction were pooled and precipitated by addition of 1 ml Glycogen (20 mg/ml, Roche, 1090139300), 5 ml NaOAc (3M) and 125 ml ice-cold 100% ethanol, incubation at -20°C for 1 hour and centrifugation for 1 hour at 13200 rpm at 4°C, followed by a final wash in 70% ethanol. After air drying, DNA pellet was dissolved in 10 ml water and used for electroporation into electrocompetent MegaX DH10b E.coli bacteria (Invitrogen), according to manufacturer's instructions, using a Biorad pulser. A total of 5 electroporations per library were performed with each 2 ml of DNA. After recovery in 1 ml SOCS medium each, bacteria were grown for 1 hour at 37°C in a bacterial shaker in the absence of antibiotics. Then, bacteria were pooled together and 50 ml of a 1:100 and 1:10000 dilution was plated on Ampicillin containing Agar plates to enable estimation of the number of transformants after overnight growth at 37°C (Control electroporations with Mock-Gibson without addition of PCR product plated on Ampicillin, or digested STARRseq plasmid transformations on Ampicillin- and Ampicillin/Chloramphenicol-containing Agar plates were negative, confirming complete digestion of the STARR-seq plasmid and a functional Ccdb counter-selection in DH10βE.Coli). The remaining 5 ml of bacteria culture were incubated in a total volume of 2 liter of LB-media supplemented with Ampicillin and allowed to grow for 16 hours in a bacterial shaker at 37°C. Plasmid DNA was isolated using a Qiagen Maxiprep kit according to manufacturer's instructions and eluted in 500 ml 10mM Tris-HCl, pH 7.4. Concentration was determined by Nanodrop measurement. For BAC-STARR-seq of super enhancer regions, three BAC clones (RP11-357D8, RP11-100L8, RP11-713N22) were ordered at the BAC PAC resource center from CHORI. DNA was isolated according to standard procedures, mixed in equimolar quantities and subjected to sonication, after which 10 ng was used for end-repair, adapter ligation and cloning of plasmid libraries as described above for the ChIP-STARR-seq.

Transfection of plasmid libraries

Primed and naive H9 ESCs were transfected using either Nucleofection (Lonza, VPH-5022), or using Lipofectamine 3000 according to manufacturer's instructions. For each transfection, 6-10 million cells were used (approximately 2.5-4.2 x10^8 cells in total) and transfected with 8 mg of plasmid library DNA and 500 ng pmCherry-N1 plasmid (Clonetech) as transfection control. Cells were incubated in 10 cm dishes and 24h post-transfection, single cells were harvested and subjected to FACS. Non-transfected cells were used to set sorting gates, DAPI was used as a marker for dead cells. All percentages mentioned are relative to the fraction of DAPI-negative, single cells.

ChIP-STARR-seq RNA and DNA samples

A minimum of 400,000 GFP-positive, sorted cells were used to isolate total RNA using Trizol (Thermo Fisher) according to manufacturer's instructions. On average, 2 million GFP-positive cells were used per sample. The mRNA fraction was captured using Oligo (dT)25 beads (61002, Life Technologies) and DNAseI treated (18068-015, Life Technologies), followed by reverse transcription using 2 ml SuperscriptIII (18080-044, Life Technologies) using a GFP-mRNA specific primer (149 STARRseq rep RNA cDNA synth, CAAACTCATCAATGTATCTTATCATG) at 50°C for 90 minutes, in a total reaction volume of 21 ml. To repress residual plasmid DNA contamination, cDNA was PCR amplified using a combination of primers (152 STARR reporter specific primer 2 fw, GGGCCAGCTGTTGGGGTG*T*C*C*A*C and 153 STARR reporter specific primer 2 rv, CTTATCATGTCTGCTCGA*A*G*C, where * represent phosphorothioate bonds) spanning a synthetic intron in the STARR-seq plasmid, as previously described[15]. PCR was performed with Phusion polymerase and High-fidelity buffer, in 6 x 50 ml reactions (cycling conditions: 98°C 2 min, (98°C 10 sec, 62°C 30 sec, 72°C 70 sec) x15, 72°C 5 min, 4°C hold). PCR reactions were pooled, purified using AMPureXP beads (1.0x) and eluted in 18 ml 0.1xTE. Absence of significant plasmid contamination in the PCR amplified cDNA was assessed by qPCR using a primer-set amplifying an amplicon from the STARR-seq plasmid backbone (161 STARRseq detect plasmid backbone qPCR fw, CATCATCGGGAATCGTTCTT, and 162 STARRSeq detect plasmid backbone qPCR rv, TGAAGATCAACTGGGTGCAA), relative to a primer-set amplifying GFP (154 STARRseq GFP fw, ACGGCCACAAGTTCTCTGTC, and 155 STARRseq GFP rv, GCAGTTTGCCAGTAGTGCAG). PCR amplified cDNA was then used in a second round of PCR to add Illumina index primers (7335, 7500, NEB) using priming on the adapter sequences added during the plasmid library generation. PCR was performed in 1-4x 50 ml reactions using Phusion polymerase and High-fidelity buffer (NEB)(cycling conditions: 98°C 2 min, (98°C 10 sec, 65°C 30 sec, 72°C 30 sec) x13, 72°C 5 min, 4°C hold), after which PCR reactions were pooled, purified using AMPureXP beads (1.0x) and eluted in 15 ml 0.1xTE. Corresponding plasmid libraries were similarly amplified in a nested PCR, using primers detecting the STARR-seq plasmid (160 STARR reporter specific primer for plasmid DNA fw, GGGCCAGCTGTTGGGGTG, and 153 STARR reporter specific primer 2 rv, CTTATCATGTCTGCTCGA*A*G*C, where * represent phosphorothioate bonds) and Illumina index primers. In addition to sequencing libraries prepared from plasmid maxiprep DNA, we also sequenced plasmid libraries reisolated from transfected ESCs. For this, we transfected H9 ESCs as described above and harvested non-sorted cells 24h post-transfection, followed by plasmid reisolation using a Qiagen miniprep isolation kit and sequencing library preparation. Quantity and quality of generated sequencing libraries was assessed on an Agilent Tapestation. All sequencing occurred on an Illumina HiSeq 2500 platform, using 50 bp or 125 bp paired-end sequencing. Up to 22 RNA samples were pooled on a single lane. During data-processing all reads were trimmed to 50 bp length to improve consistency.

RT-qPCR

For RNA analysis of complete cultures, cells were lysed in Trizol (Thermo Fisher) and RNA was prepared according to manufacturer's instructions. 1 mg of RNA was treated with DNAseI (Invitrogen) to remove genomic DNA contamination and cDNA was obtained through reverse transcription using SuperScriptIII (Invitrogen) in the presence of RNAseOUT (Invitrogen). cDNA was diluted in DEPC-treated water to a final volume of 200 ml and 2 ml of cDNA was used per qPCR reaction, using a 2x Takyon qPCR master mix

(No ROX SYBR, UF-NSMT-B0701, Takyon). qPCR reactions were run on a Roche Lightcycler 480 II (Roche), using the following cycle conditions: 95°C 3 min, (95°C 10 sec, 60°C 30 sec, 72°C 25 sec) x45, followed by a melting curve from 95° to 65°C. All data shown are averages of at least 2 biological replicates and 3 technical replicates, normalized to TBP. All primers used are shown in **Table S5.**

Immunostaining

Cells were grown on culture dishes suitable for confocal microscopy (Ibidi, 81156) and fixed using 4% v/v Paraformaldehyde at room temperature for 10 min. After permeabilisation using 0.3% Triton/PBS and incubation with blocking solution (1% BSA, 3% Donkey serum, 0.1% triton in PBS), cells were incubated with primary antibody O/N at 4°C. After washing with PBS, cells were incubated with secondary antibody at RT for 1h, washed and counterstained with DAPI. Imaging occurred on a Leica SP8 STED-CW confocal microscope and images were processed using ImageJ software. Antibodies used are: goat-anti-NANOG (1: 200, AF1997, R&D Systems), rabbit-anti-OCT4 (1: 200, AB19857, Abcam). Secondary antibodies were Donkey-anti-goat conjugated to Alexa fluor488 (1:800, A11055, Invitrogen) and Donkey-anti-rabbit conjugated to Alexa fluor568 (1:1000, A10042, Invitrogen).

Western blotting

Whole cell protein extracts were isolated and Western blotting was performed using standard procedures using pre-cast 10% Bis-Tris Bolt gels (Invitrogen). Primary antibody used was goat-anti-NANOG (1: 500, 1mg/ml, AF1997, R&D Systems), secondary antibody conjugated to fluorophores was donkey-anti-goat-IRDey680 (1:500, 926-68074, Li-cor). Rabbit-anti-Laminin B (1:1000, AB16048, Abcam) served as a loading control and was detected by chemi-iluminescence. Imaging occurred on an Odyssey imager (Li-cor).

Luciferase assays

Enhancer sequences were PCR amplified from human genomic DNA using Phusion polymerase and cloned by Gibson assembly into a KpnI-NheI linearized Pgl3 promoter luciferase vector. For primer sequences, see **Table S5.** All constructs were sequence-verified by Sanger sequencing and co-transfected with a Renilla expressing plasmid using Lipofectamin 3000 into H9 ESCs. 48h post-transfection illuminescence was assessed using the Dual Glo luciferase kit (E2920, Promega) according to manufacturer's instructions, on a Promega Glumax Multidection system. All data shown are average from at least two biological replicates and two technical replicates, representing fold-change in luciferase activity compared to empty vector controls and normalized for Renilla transfection control.

Alternative promoter STARR-seq constructs

To replace the SCP1 minimal promoter from the original STARR-seq plasmid[15], the plasmid was linearized by restriction digestion using KpnI-ApaI (NEB) and used to ligate annealed oligonucleotides, coding for the adenovirus major late (AML) or CMV IE core promoter[75]. Test enhancer sequences were introduced by PCR amplification and Gibson assembly as done during library cloning. All constructs were verified by Sanger sequencing. Oligonucleotide sequences are given in **Table S5**. Constructs (1 mg of each plasmid) were transfected in H9 primed ESCs cultured in 6-well plates using Lipofectamine 3000 and fluorescents was assessed using flow cytometry. Shown are the results for two independent experiments (analyzing > 30.000 GFP positive cells each), comparing all identical tested enhancer sequences in constructs with the SCP1, AML or CMV minimal promoter transfected in parallel.

CRISPR/Cas9 genome editing

Oligonucleotides for gRNAs (**Table S5**) flanking the tested enhancers were annealed and cloned into a BbsI digested spCas9 plasmid, from which the gRNAs are separately expressed together with a eSpCas9(1.1)-t2a-mCherry or eSpCas9(1.1)-t2a-GFP (modified from Addgene plasmid #71814,[76]). All plasmids were sequence verified and 1 mg of each gRNA was used to transfect primed H9 ESCs in a 6-well plate using Lipofectamine 3000. 48h post-transfection, mCherry and GFP double positive cells were FACS sorted and cells were plated at low density in 10 cm dishes coated with Matrigel in conventional mTesR1 ESC medium. Emerging clones were expanded and genotyped by PCR using primers flanking the gRNA targets (**Table S5**). For the pos3_ID1 enhancer, a nested PCR using outer and inner primers was performed. All candidate clones were validated by Sanger sequencing of PCR products and correct clones were expanded.

ChIP-seq and ChIP-STARR-seq data processing

We trimmed possible adapter contaminants from reads using Skewer[77]. Trimmed reads were then aligned to the GRCh37/hg19 assembly of the human genome using Bowtie2[78] with the "--very-sensitive" parameter. Genome browser tracks were created from all aligned reads with the genomeCoverageBed command in BEDTools[79] and normalized such that each value represents the read count per base pair per million uniquely mapped reads. Finally, the UCSC Genome Browser's bedGraphToBigWig tool was used to produce a bigWig file.

ChIP-STARR-seq enhancer activity levels

For ChIP-seq and plasmid DNA-seq libraries, peak calling was performed with MACS2 version 2.1.0.20150420[80] with default parameters (narrow peak calling, fragment length detection from libraries, genome size $2.7x10^9$ bp, FDR < 0.05), using the respective input samples as background. Significant peaks (FDR < 0.05) were fixed to a width of 500 bp from the peak summit for transcription factors and 1000 bp for histone modifications. Peaks overlapping blacklisted features as defined by the ENCODE project[81] were removed. ChIP-seq peaks are given in **File S1**.

To define a non-redundant set of enhancers to compare in our analysis of ChIP-seq, plasmid DNA-seq and ChIP-STARR RNA-seq samples, we produced a set of regions by merging all peaks across cell types and experiment types (ChIP-seq and plasmid DNA-seq). This operation results in regions that can be very large. To preserve high genomic resolution for our analysis, large regions were split in half recursively until all regions were at most 1000 bp long. All further analysis were performed on these scaffold regions.

We initially quantified the intensity of ChIP-seq, plasmid DNA-seq and ChIP-STARR RNA-seq datasets in the enhancer peak regions by counting the number of aligned fragments (only properly paired, concordantly aligning and uniquely mapping fragments – i.e. both mate reads mapped to same chromosome with MAPQ >= 30 – were kept) overlapping each enhancer region. To get a more accurate and precise measure of plasmid reporter intensity for further analysis, we then made use of our paired-end sequencing data to unequivocally link RNA-seq reads to the plasmid that they came from. To do so, we matched RNA-seq reads to plasmid reads with the exact same start coordinate of the first read and the exact same end coordinate of the second read. Comparing the counts for both made it possible to define a measure of RNA-seq activity relative to the abundance of plasmids in the. To avoid distortion by differences in sequencing depth, we normalized the raw read counts for each plasmid library and all RNA-seq datasets derived from transfections of this library together using DESeq2[82]. The ratio of normalized RNA-seq and (plasmid) DNA-seq reads was used as a measure of enhancer activity (reads per plasmid, RPP). We then calculated the mean RPP of replicate measurements for the same plasmid position and used the maximum observed RPP value per region as an estimate of enhancer-peak-level activity. Since our individual replicate datasets were sparse, with the same plasmids infrequently measured in both replicates, but our overall coverage of enhancers was much better, we used RPP from all datasets generated in the same cell type (so specific to either primed or naive H9 ESCs) for this purpose. We could do so because the ChIP-STARR-seq plasmid libraries are independent from the antibody target used to pull down the enriched DNA fragments, thus the plasmids in all libraries jointly report the activity of the same genome. To objectively define a threshold for discriminating highly active and inactive genome regions, we looked at the curve of RPP ranks vs. RPP values (**Figure 2C**) and defined points of change in the mean and variation of the data using the changepoint package in R[83]. The highest value was used as a threshold for active enhancers (θ = 138). The coordinates of all genome regions assessed with activity calls are given in **File S1** and **Table S2**.

Motif enrichment analysis for ChIP-seq data

For de novo motif discovery (**Figure S5**), BED files of ChIP-seq data sets were generated with 500 bp sequences centered on the narrow ChIP-seq peak, and used for motif enrichment analysis using CentriMo (http://meme-suite.org/)[84], using default settings.

Assignment of enhancers to genes

We used GREAT, version 3.0.0[39] to assign regulatory elements identified in ChIP-STARR-seq to their putative target genes, using the following settings: basal plus extension, proximal 5kb upstream and 1kb downstream, plus distal up to 100kb. Publically available, processed RNA-seq data from primed human ESCs were downloaded[5,85,86] and their RPKM value distribution was plotted for the various ChIP-STARR-seq regions grouped by activity in RPP. For naive ESCs, we used publically available microarray data from the original study describing gene expression in naive cells cultured under NHSM conditions[7].

Comparison to previously published enhancers

The coordinates of putative enhancers were obtained from the supplementary data of Hawkins et al, Rada-Iglesias et al and Xie et al[13,37,38], and when necessary converted to the hg19 version of the human genome using the liftOver tool. Overlapping enhancers were merged into 76,666 putative enhancers and joint to our ChIP-STARR-seq enhancers using GenomicRanges[87] in R (see **Figure S4A, Table S2**). We refer to those enhancers that overlapped with previously published enhancers and showed a ChIP-STARR-seq activity of RPP>=138 as the core enhancer module (n=7,948). Conversely, we refer to active enhancers (RPP>= 138) that did not overlap with the previously published enhancers as the extended enhancer module (n = 24,405).

Functional enrichment analysis

To help understand the function and relevance of different groups of enhancers, we used three types of functional enrichment analysis (**Table S3**).

(a) We used LOLA[33] to determine the relative over-representation of ChIP-seq peaks related transcription factor binding and other elements of known regulatory function. To this end, we used the codex, encode_tfbs, and encode_segmentation databases contained in the LOLA Core database and tested for the enrichment of overlap in genome regions with a specific level of activity (high, low or inactive) over the background of all ChIP-STARR-seq peaks.

(b) We also used the Enrichr API (January 2018 version)[44] to test genes linked to enhancers of interest for significant enrichment in numerous functional categories. To comply with the web interface, we considered the 1000 genes closest to the tested peaks for enrichments. In all plots, we report the “combined score” calculated by Enrichr, which is a product of the significance estimate and the magnitude of enrichment (combined score c = log(p) * z, where p is the Fisher’s exact test p-value and z is the z-score deviation from the expected rank).

(c) We additionally used the GREAT web interface (version 3.0.0) (http://great.stanford.edu/public/html/)[39] for gene ontology analysis, using the following settings: basal plus extension, proximal 5kb upstream and 1kb downstream, plus distal up to 100kb, including curated regulatory domains, and whole genome (hg19) as background.

Machine learning

We used the random forest classifier implementation in the h2o R package (https://github.com/h2oai/h2o-3) to train models for predicting enhancer activity (“Active” vs. “Inactive”) in primed and naive ESCs and to discriminate enhancers from the Core and Extended module (“Core” vs. “Extended”). Three types of features based on the DNA underlying each ChIP-STARR-seq region were used as inputs: (a) sequence conservation. The maximum PhastCons score from overlaps with the UCSC Golden Path reference was used per region; (b) GC content calculated from alphabet frequency in R; (c) dinucleotide frequencies calculated with the bioconductor package Biostrings), taking the maximum on either forward or reverse strand; and (d) occurrence of known motifs from the HOCOMOCO database[88] (v11; limited to “excellent” [A] and “good” [B] quality motifs). The tool FIMO (v4.10.2)[89] was used (parameters: --no-qvalue --text --bgfile motif-fil) to scan DNA sequences for these motifs and regions with at least one hit ($p < 0.05$) were counted. Each classifier was trained on balanced classes from the complete set of ChIP-STARR-seq regions (excluding missing RPP values) or on all active enhancers (RPP>=138; for Core/Extended discrimination) using 10-fold cross-validation and evaluation 500 trees with 50 features sampled at each split and a maximum depth of 10 (parameters: mtries=50, nfolds=10, keep_cross_validation_predictions=T, balance_classes=T, ntrees=500, max_depth=10).

Enrichment analysis for transposable elements

The UCSC RepeatMask (hg19) was downloaded from the UCSC Table Browser, imported into Galaxy (https://usegalaxy.org/)[90] and joined to the ChIP-STARR-seq activity calls for primed or naive ESCs. The number of overlaps of each type of repeat ($n_{overlaps}$) with all ChIP-STARR-seq regions (n) was used to calculate the relative frequency (f_{all}= $n_{overlaps}$/n). Multiplication of the relative frequency with the number of regions (n_{test}, e.g. $n_{active,primed}$) in any tested groups yields the expected frequency (E). This number was compared with the actual observed frequency in the subgroups ($f_{test} = n_{overlap,test}/n_{test} = O$) to calculated the observed vs. expected ratio (O/E). We considered repeats with O/E<0.5 as depleted, or O/E>2 as enriched. For the subsequent data interpretation we only focused on transposable elements that were present multiple times ($n_{overlap}$>15) in all ChIP-STARR-seq regions.

Super-enhancer analysis

To call super-enhancers in primed and naive H9 ESCs, we used the ROSE software (v0.1)[56] to combine ("stitch") H3K27ac ChIP-seq peaks within 12.5 kb of each other and excluding 2.5 kb around known transcription start sites. An alternate analysis was also run with stitching distance d=0 for comparison. We then asked the software to quantify the ratio of the H3K27ac ChIP-seq signal in primed and naive ESCs over the total input control and to call super-enhancers. The coordinates of all stitched enhancers, as well as primed and naive super-enhancers are given in **File S1.**

Statistics for qPCR and luciferase assays

qPCR and luciferase assay figures were plotted and statistics were calculated using GraphPad Prism 5 software, p<0.05 was considered significant. Statistical tests used are indicated in the figure legends. For the qPCR analysis of CRISPR deleted enhancer clones in **Figure 2E**, we calculated expression as follows: in each graph (with the exception of TBX3), average results for the indicated enhancer deletion (heterozygous (+/-) or homozygous (-/-) as indicated) are plotted relative to wild type, n = number of cell lines per genotype. Wild type controls consisted of H9 parental, two untransfected H9 clones and all remaining clones that were wild type for the respective allele. Genes assessed were the presumed target gene and four randomly selected genes. For the TBX3 intronic deletion, three H9 wt and three -/- deletion clones were assessed for three amplicons detecting TBX3 mRNA and two flanking genes. All measurements occurred at two different passages, in two independent cultures measured in duplicate.

Data availability

High-throughput sequencing data generated in this study have been submitted to the Gene Expression Omnibus (GEO) under accession code GSE99631, and to the Sequence Read Archive (SRA) under accession codes SRP108517, SRP108518, SRP108519, and SRP108520. A BioProject for this study has also been registered (PRJNA389108).

Additional resources

Additional data, an interactive search tool for active enhancers in the proximity of genes and the genome browser track hub providing raw and processed ChIP-STARR-seq data for interactive visualization and processing with online tools such as Galaxy, are available from a supplemental website under the following URL: http://hesc-enhancers.computational-epigenetics.org

AUTHOR CONTRIBUTION

TSB and IC conceived the study. TSB performed the molecular biology and contributed to the bioinformatics analysis and cell culture. AR developed the primary data analysis pipeline. FH performed bioinformatics analysis, visualization and interpretation. MZ performed cell culture, immunofluorescence and generated CRISPR/Cas9 deleted clones, together with EP. CB supervised the bioinformatics analysis. TSB, FH and IC wrote the paper, with input from all authors.

ACKNOWLEDGEMENTS

We thank S. Pollard, D. O'Carroll, A. Soufi and S. Tomlinson for comments, E. Hall-Ponsele, F. Rossi and G. Schaaf for technical support and R. Pantier, J. Zhang and other colleagues for discussions. We thank A. Stark for the STARR-seq plasmid, F. Zhang for eSpCas9(1.1) and D. Hay for H9 ESCs. IC's lab is supported by the Medical Research Council (UK) and The Wellcome Trust. TSB was supported by fellowships from Niels Stensen, EMBO (EMBO-LTF) and Marie Sklodowska-Curie (H2020 MSCA-IF). FH was supported by a DFG fellowship (HA 7723/1-1). CB is supported by a New Frontiers Group award of the Austrian Academy of Sciences and an ERC Starting Grant (no. 679146).

DECLARATION OF INTEREST

The authors declare no competing interests.

References

1 Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299-308 (1981).

2 Spitz, F. & Furlong, E. E. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613-626 (2012).

3 Buecker, C. *et al.* A murine ESC-like state facilitates transgenesis and homologous recombination in human pluripotent stem cells. *Cell stem cell* **6**, 535-546, doi:10.1016/j.stem.2010.05.003 (2010).

4 Hanna, J. *et al.* Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 9222-9227, doi:10.1073/pnas.1004584107 (2010).

5 Takashima, Y. *et al.* Resetting transcription factor control circuitry toward ground-state pluripotency in human. *Cell* **158**, 1254-1269, doi:10.1016/j.cell.2014.08.029 (2014).

6 Chan, Y. S. *et al.* Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast. *Cell stem cell* **13**, 663-675, doi:10.1016/j.stem.2013.11.015 (2013).

7 Gafni, O. *et al.* Derivation of novel human ground state naive pluripotent stem cells. *Nature* **504**, 282-286, doi:10.1038/nature12745 (2013).

8 Theunissen, T. W. *et al.* Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. *Cell stem cell* **15**, 471-487, doi:10.1016/j.stem.2014.07.002 (2014).

9 Ware, C. B. *et al.* Derivation of naive human embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 4484-4489, doi:10.1073/pnas.1319738111 (2014).

10 Sperber, H., Mathieu, J., Wang, Y., Ferreccio, A. & Hesson, J. The metabolome regulates the epigenetic landscape during naive-to-primed human embryonic stem cell transition. **17**, 1523-1535, doi:10.1038/ncb3264 (2015).

11 ENCODE. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

12 Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics* **39**, 311-318, doi:10.1038/ng1966 (2007).

13 Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-283, doi:10.1038/nature09692 (2011).

14 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213-1218, doi:10.1038/nmeth.2688 (2013).

15 Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science (New York, N.Y.)* **339**, 1074-1077, doi:10.1126/science.1232542 (2013).

16 Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C. & Cohen, B. A. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 19498-19503, doi:10.1073/pnas.1210678109 (2012).

17 Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature biotechnology* **30**, 271-277, doi:10.1038/nbt.2137 (2012).

18 Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nature biotechnology* **30**, 265-270, doi:10.1038/nbt.2136 (2012).

19 Smith, R. P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature genetics* **45**, 1021-1028, doi:10.1038/ng.2713 (2013).

20 Arnold, C. D. *et al.* Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nature genetics* **46**, 685-692, doi:10.1038/ng.3009 (2014).

21 Shlyueva, D. *et al.* Hormone-responsive enhancer-activity maps reveal predictive motifs, indirect repression, and targeting of closed chromatin. *Molecular cell* **54**, 180-192, doi:10.1016/j.molcel.2014.02.026 (2014).

22 Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome research* **24**, 1595-1602, doi:10.1101/gr.173518.114 (2014).

23 Vanhille, L. *et al.* High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nature communications* **6**, 6905, doi:10.1038/ncomms7905 (2015).

24 Shen, S. Q. *et al.* Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome research* **26**, 238-255 (2016).

25 Verfaillie, A. *et al.* Multiplex enhancer-reporter assays uncover unsophisticated TP53 enhancer logic. *Genome research* **26**, 882-895, doi:10.1101/gr.204149.116 (2016).

26 Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519-1529, doi:10.1016/j.cell.2016.04.027 (2016).

27 Ulirsch, J. C. *et al.* Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **165**, 1530-1545, doi:10.1016/j.cell.2016.04.048 (2016).

28 Vockley, C. M. *et al.* Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome research* **25**, 1206-1214, doi:10.1101/gr.190090.115 (2015).

29 Murtha, M. *et al.* FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. *Nature methods* **11**, 559-565, doi:10.1038/nmeth.2885 (2014).

30 Vockley, C. M. *et al.* Direct GR Binding Sites Potentiate Clusters of TF Binding across the Human Genome. *Cell* **166**, 1269-1281.e1219, doi:10.1016/j.cell.2016.07.049 (2016).

31 Muerdter, F. *et al.* Resolving systematic errors in widely used enhancer activity assays in human cells. *Nature methods* **15**, 141-149 (2018).

32 Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

33 Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics (Oxford, England)* **32**, 587-589, doi:10.1093/bioinformatics/btv612 (2016).

34 Sanchez-Castillo, M., Ruau, D., Wilkinson, A. C. & Ng, F. S. CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. **43**, D1117-1123, doi:10.1093/nar/gku895 (2015).

35 Kreimer, A. *et al.* Predicting gene expression in massively parallel reporter assays: A comparative study. *Hum Mutat* **38**, 1240-1250 (2017).

36 Kreimer, A., Yan, Z., Ahituv, N. & Yosef, N. Meta-analysis of massive parallel reporter assay enables functional regulatory elements prediction. *bioRxiv* (2017).

37 Hawkins, R. D. *et al.* Dynamic chromatin states in human ES cells reveal potential regulatory sequences and genes involved in pluripotency. *Cell research* **21**, 1393-1409, doi:10.1038/cr.2011.146 (2011).

38 Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134-1148, doi:10.1016/j.cell.2013.04.022 (2013).

39 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**, 495-501, doi:10.1038/nbt.1630 (2010).

40 Krupp, M. *et al.* RNA-Seq Atlas--a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics (Oxford, England)* **28**, 1184-1185, doi:10.1093/bioinformatics/bts084 (2012).

41 GTEx. The Genotype-Tissue Expression (GTEx) project. *Nature genetics* **45**, 580-585, doi:10.1038/ng.2653 (2013).

42 Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet* **29**, 569-574 (2013).

43 Lachmann, A. *et al.* Massive Mining of Publicly Available RNA-seq Data from Human and Mouse. *bioRxiv* (2017).

44 Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics* **14**, 128, doi:10.1186/1471-2105-14-128 (2013).

45 Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics (Oxford, England)* **26**, 2438-2444, doi:10.1093/bioinformatics/btq466 (2010).

46 Barakat, T. S. *et al.* Stable X chromosome reactivation in female human induced pluripotent stem cells. *Stem Cell Reports* **4**, 199-208 (2015).

47 Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49, doi:10.1038/nature09906 (2011).

48 Warrier, S. *et al.* Direct comparison of distinct naive pluripotent states in human embryonic stem cells. *Nature communications* **8**, 15055 (2017).

49 Corsinotti, A. *et al.* Distinct SoxB1 networks are required for naive and primed pluripotency. *Elife* **6** (2017).

50 Bourque, G. *et al.* Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome research* **18**, 1752-1762, doi:10.1101/gr.080663.108 (2008).

51 Glinsky, G. V. Transposable Elements and DNA Methylation Create in Embryonic Stem Cells Human-Specific Regulatory Sequences Associated with Distal Enhancers and Noncoding RNAs. *Genome biology and evolution* **7**, 1432-1454, doi:10.1093/gbe/evv081 (2015).

52 Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature genetics* **42**, 631-634, doi:10.1038/ng.600 (2010).

53 Teng, L., Firpi, H. A. & Tan, K. Enhancers in embryonic stem cells are enriched for transposable elements and genetic variations associated with cancers. *Nucleic acids research* **39**, 7371-7379, doi:10.1093/nar/gkr476 (2011).

54 Kent, W. J. *et al.* The human genome browser at UCSC. *Genome research* **12**, 996-1006, doi:10.1101/gr.229102 (2002).

55 Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947, doi:10.1016/j.cell.2013.09.053 (2013).

56 Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319, doi:10.1016/j.cell.2013.03.035 (2013).

57 Dukler, N. & Gulko, B. Is a super-enhancer greater than the sum of its parts? **49**, 2-3, doi:10.1038/ng.3759 (2016).

58 Hay, D. & Hughes, J. R. Genetic dissection of the alpha-globin super-enhancer in vivo. **48**, 895-903, doi:10.1038/ng.3605 (2016).

59 Shin, H. Y. *et al.* Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nature genetics* **48**, 904-911, doi:10.1038/ng.3606 (2016).

60 Moorthy, S. D. *et al.* Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome research* **27**, 246-258, doi:10.1101/gr.210930.116 (2017).

61 Ernst, J. *et al.* Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. **34**, 1180-1190, doi:10.1038/nbt.3678 (2016).

62 Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome research* **23**, 800-811, doi:10.1101/gr.144899.112 (2013).

63 Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome research* **27**, 38-52 (2017).

64 James, D., Levine, A. J., Besser, D. & Hemmati-Brivanlou, A. TGFbeta/activin/nodal signaling is necessary for the maintenance of pluripotency in human embryonic stem cells. *Development (Cambridge, England)* **132**, 1273-1282, doi:10.1242/dev.01706 (2005).

65 Xu, R. H. *et al.* NANOG is a direct target of TGFbeta/activin-mediated SMAD signaling in human ESCs. *Cell stem cell* **3**, 196-206, doi:10.1016/j.stem.2008.07.001 (2008).

66 Pradeepa, M. M. *et al.* Histone H3 globular domain acetylation identifies a new class of enhancers. *Nature genetics* **48**, 681-686, doi:10.1038/ng.3550 (2016).

67 Engreitz, J. M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**, 452-455, doi:10.1038/nature20149 (2016).

68 Dao, L. T. M. *et al.* Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nature genetics* **49**, 1073-1081 (2017).

69 Zabidi, M. A. *et al.* Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556-559, doi:10.1038/nature13994 (2015).

70 Cubenas-Potts, C. *et al.* Different enhancer classes in Drosophila bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture. *Nucleic acids research*, doi:10.1093/nar/gkw1114 (2016).

71 Zabidi, M. A. & Stark, A. Regulatory Enhancer-Core-Promoter Communication via Transcription Factors and Cofactors. *Trends Genet* **32**, 801-814 (2016).

72 Wu, J. *et al.* Interspecies Chimerism with Mammalian Pluripotent Stem Cells. *Cell* **168**, 473-486. e415, doi:10.1016/j.cell.2016.12.036 (2017).

73 Hnisz, D. *et al.* Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Molecular cell* **58**, 362-370, doi:10.1016/j.molcel.2015.02.014 (2015).

74 Suzuki, H. I., Young, R. A. & Sharp, P. A. Super-Enhancer-Mediated RNA Processing Revealed by Integrative MicroRNA Network Analysis. *Cell* **168**, 1000-1014.e1015, doi:10.1016/j.cell.2017.02.015 (2017).

75 Juven-Gershon, T., Cheng, S. & Kadonaga, J. T. Rational design of a super core promoter that enhances gene expression. *Nature methods* **3**, 917-922 (2006).

76 Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science (New York, N.Y.)* **351**, 84-88, doi:10.1126/science.aad5227 (2016).

77 Jiang, H., Lei, R., Ding, S. W. & Zhu, S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC bioinformatics* **15**, 182, doi:10.1186/1471-2105-15-182 (2014).

78 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

79 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

80 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).

81 Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic acids research* **41**, 827-841, doi:10.1093/nar/gks1284 (2013).

82 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550 (2014).

83 Killick, R. E., IA. changepoint: An R Package for Changepoint Analysis. *Journal of Statistical Software* **58**, 1--19, doi:10.18637/jss.v058.i03 (2014).

84 Bailey, T. L. & Machanick, P. Inferring direct DNA binding from ChIP-seq. *Nucleic acids research* **40**, e128, doi:10.1093/nar/gks433 (2012).

85 Gifford, C. A. *et al.* Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell* **153**, 1149-1163, doi:10.1016/j.cell.2013.04.037 (2013).

86 Ji, X. *et al.* 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. *Cell stem cell* **18**, 262-275, doi:10.1016/j.stem.2015.11.007 (2016).

87 Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS computational biology* **9**, e1003118, doi:10.1371/journal.pcbi.1003118 (2013).

88 Kulakovskiy, I. V. *et al.* HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic acids research* **44**, D116-125, doi:10.1093/nar/gkv1249 (2016).

89 Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)* **27**, 1017-1018, doi:10.1093/bioinformatics/btr064 (2011).

90 Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic acids research* **44**, W3-w10, doi:10.1093/nar/gkw343 (2016).

Supplementary Material

Supplementary Tables

https://www.cell.com/cell-stem-cell/fulltext/S1934-5909(18)30296-0

Table S1: Data overview

Table S2: Enhancer activities and modules

Table S3: LOLA, Enrichr, and GREAT enrichments

Table S4: transposable elements observed/expected ratios

Table S5: Oligonucleotides used in this study

File S1: Genomic coordinates (BED files) of ChIP-seq peaks, ChIP-STARR-seq enhancer with activity level and of super-enhancers called in this study

Supplementary Figures

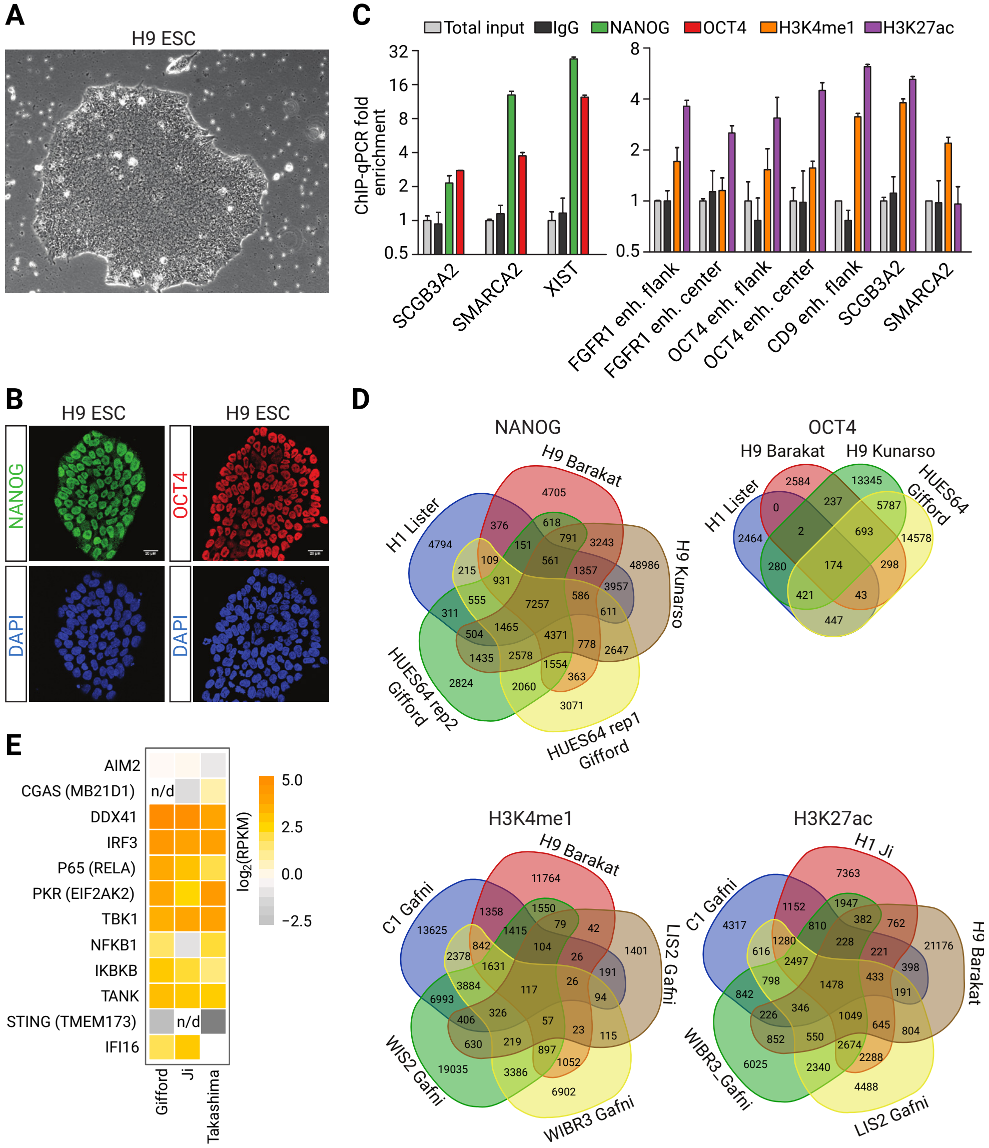
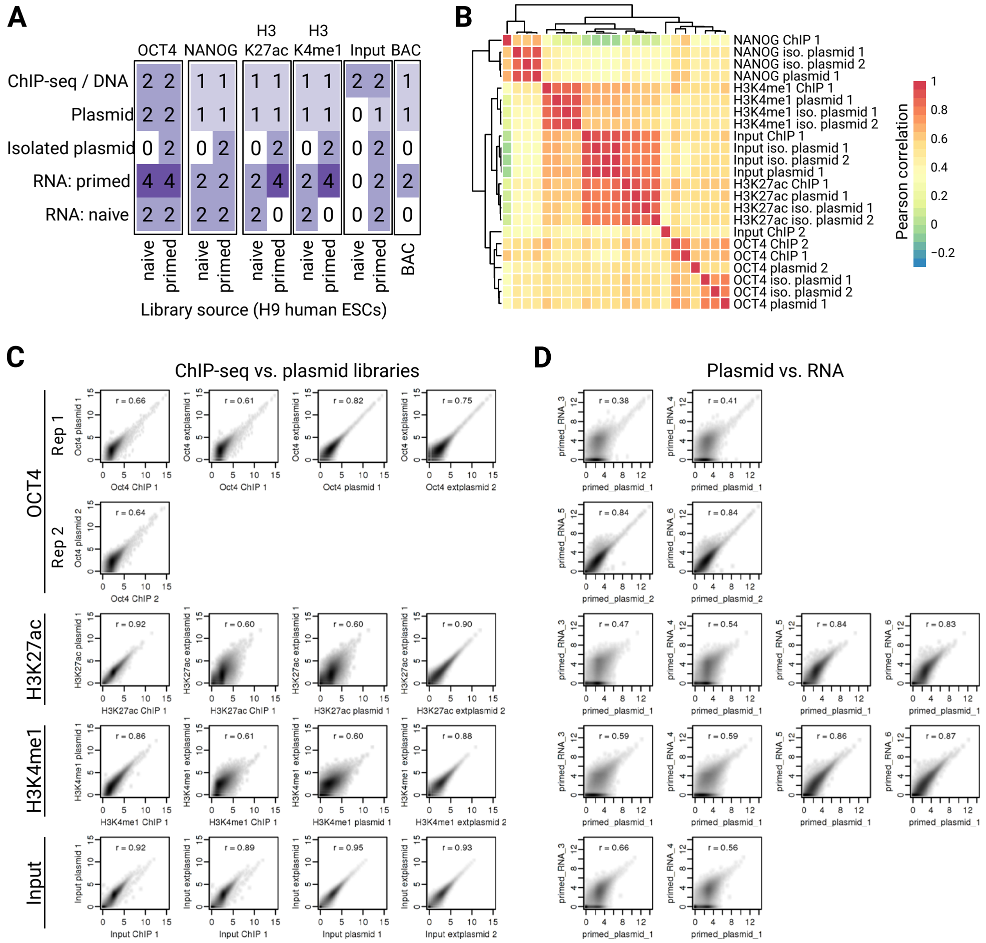


Figure S1 | ChIP-seq in primed H9 human embryonic stem cells. A) Brightfield microscopy of a representative colony of H9 ESCs cultured on Matrigel in standard ESC culture conditions. **B)** Immunofluorescence of primed H9 ESCs for NANOG (green) or OCT4 (red); DNA is stained with DAPI (blue). **C)** ChIP-qPCR in primed H9 ESC with anti-NANOG, anti-OCT4 or rabbit IgG at known OCT4 and NANOG binding sites in SCGB3A2, SMARCA (Kunarso et al., 2010) and XIST (left) or with anti-H3K4me1, anti-H3K27ac or rabbit IgG at binding sites near FGFR1 (at central and flanking locations), POU5F1 (at central and flanking locations), CD9, SCGB3A2, and SMARCA (Rada-Iglesias et al., 2011) (right). The mean fold-enrichment is shown relative to total input control DNA, normalized to a non-bound site in ACTB (for OCT4 and NANOG), or NCAPD2 (for H3K4me1 and H3K27ac). Error bars indicate standard deviations; n= 3. **D)** Venn diagrams of the overlap between ChIP-seq peaks in primed ESCs, indicating the cell line and study for NANOG, OCT4, H3K4me1 and H3K27ac. The numbers indicate overlapping peaks. **E)** Heatmap contrasting gene expression of immune response genes in human ESCs as determined by RNA-seq (Muerdter et al., 2018). Data from (Gifford et al., 2013; Ji et al., 2016; Takashima et al., 2014)



	OCT4		NANOG		H3 K27ac		H3 K4me1		Input		BAC
	naive	primed	naive	primed	naive	primed	naive	primed	naive	primed	BAC
ChIP-seq / DNA	2	2	1	1	1	1	1	1	2	2	1
Plasmid	2	2	1	1	1	1	1	1	0	1	1
Isolated plasmid	0	2	0	2	0	2	0	2	0	2	0
RNA: primed	4	4	2	2	2	4	2	4	0	2	2
RNA: naive	2	2	2	2	2	0	2	0	0	2	0

Library source (H9 human ESCs)

Figure S2 | Overview of generated datasets. A) Summary table of the high-throughput sequencing datasets generated in this study, indicating the number of ChIP-seq, plasmid (corresponding to ChIP-STARR-seq or BACSTARR-seq plasmid libraries prior to transfection) and isolated plasmid (plasmid libraries 24h after transfection) samples, as well as RNA-seq data from GFP-positive primed and/or naive ESCs (RNA: primed and RNA: naive, respectively). **B)** Pearson correlation matrix of samples from the indicated ChIP-seq, plasmid libraries and isolated plasmid libraries from primed ESCs. Rows and columns have been arranged by hierarchical clustering with complete linkage. **C-D)** Scatterplots contrasting normalized read counts (reads per million) per peak for different datasets. The Pearson correlation coefficient (r) for each comparison is indicated for each plot. **C)** Comparison between ChIP-seq, DNA-seq of ChIP-STARR-seq plasmid libraries and two replicates of DNA-seq for isolated plasmid libraries post transfection, for OCT4, NANOG, H3K37ac, H3K27ac and genomic DNA (input). **D)** Comparison between ChIP-STARR-seq plasmid libraries prior to transfection and the corresponding RNA-seq read counts generated from two replicates of GFP-positive cells after transfection. **E)** Pearson correlation coefficients (r) between replicate STARR-RNA-seq measurements from the same pool of ESCs transfected with the same ChIP-STARR-seq library, shown as a function of minimum read count in both replicates (0 = unfiltered, 1 = at least one read from the same plasmid measured in each replicate, etc).

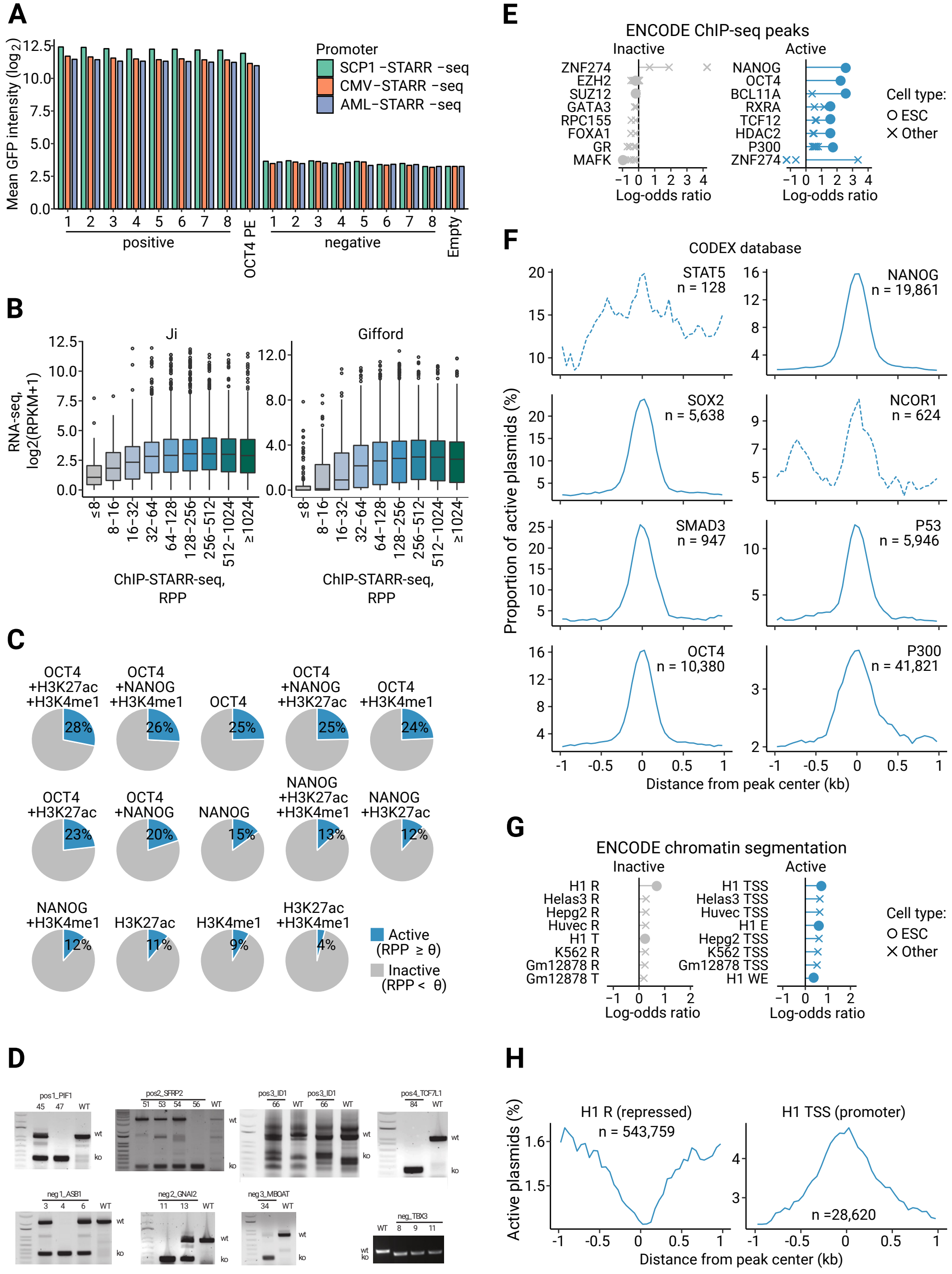


Figure S3 | ChIP-STARR-seq in primed H9 ESCs. A) Bar graph showing the mean GFP intensity as measured by flow cytometry for positive or negative ChIP-STARR-seq regions cloned into STARR-seq plasmids and used in transfection of H9 primed ESCs. Average results for two experiments are shown. Three different minimal promoters were assessed for each tested sequence. GFP mean intensity was indistinguishable from non-transfected cells for negative sequences in all constructs. B) Boxplots showing the distribution of RNA-seq RPKM values of genes associated with enhancers grouped by activity level. Boxplots represent the interquartile range (IQR), the line is the median, whiskers extend to 1.5xIQR and outliers are indicated as dots. The numbers on the x-axis indicate thresholds on the RPP activity level; RPKM, reads per kilobase million. RNA-seq datasets were from the following studies: H1 (Ji et al., 2016); HUES64 (Gifford et al., 2013). C) Distribution of active (RPP ≥138) and inactive sequences (RPP <138) in ChIP-STARRseq regions overlapping with ChIP-seq peaks of the indicated factor or combination of factors. **D)** PCR genotyping of H9 ESC wild type (WT) and targeted clones (numbers), that were transfected with Cas9 and gRNAs to delete DNA sequences with enhancer activity detected by ChIP-STARR-seq (pos, top) or inactive (neg, bottom) sequences. The putative interacting gene is indicated. Primers used for genotyping are located outside the gRNA targets; wt = wt allele, ko = knockout allele. **E)** Relative enrichment of DNA-binding proteins (DBPs) from the ENCODE database (2012) in inactive genome regions, as well as active enhancers (compare to Fig. 2G). Shown are the log2-odds ratios between observed percentages of enhancers overlapping binding sites of each given DBP in the respective groups over the percentage of overlaps in the entire enhancer dataset. Each dot represents one ChIP-seq dataset for the given DBP and the lines connect the most extreme dot with zero for visualization. For each category, the eight most enriched DBPs are shown ranked by their mean log-odds ratio. ChIP-seq datasets produced from ESCs are indicated as dots and those from other cell sources as crosses. Enrichments were calculated using LOLA (Sheffield and Bock, 2016). **F)** Smooth line plots showing the proportion of active plasmids (RPP ≥138) of all plasmids measured at the indicated distance from the peak center for ChIP-seq binding sites of factors from the CODEX database (Sanchez-Castillo et al., 2015) found preferentially associated at highly active enhancers (compare to Fig. 2H), averaged across all binding sites for the respective factor. The number of peaks (n) is indicated in each plot. **G)** LOLA enrichment plots as in panel E, but showing instead the relative over-presentation of ENCODE chromatin segments from different cell lines in enhancers with different activity levels. E, enhancer; PF, promoter-flanking region; R, repressed; T, transcribed; TSS, transcription start site; WE, weak enhancer. **H)** Line plots as in panel F, showing the proportion of active plasmids in a window around the center of repressed chromatin segments and enhancer chromatin segments from the ENCODE H1 chromatin segmentation (Hoffman et al., 2013).

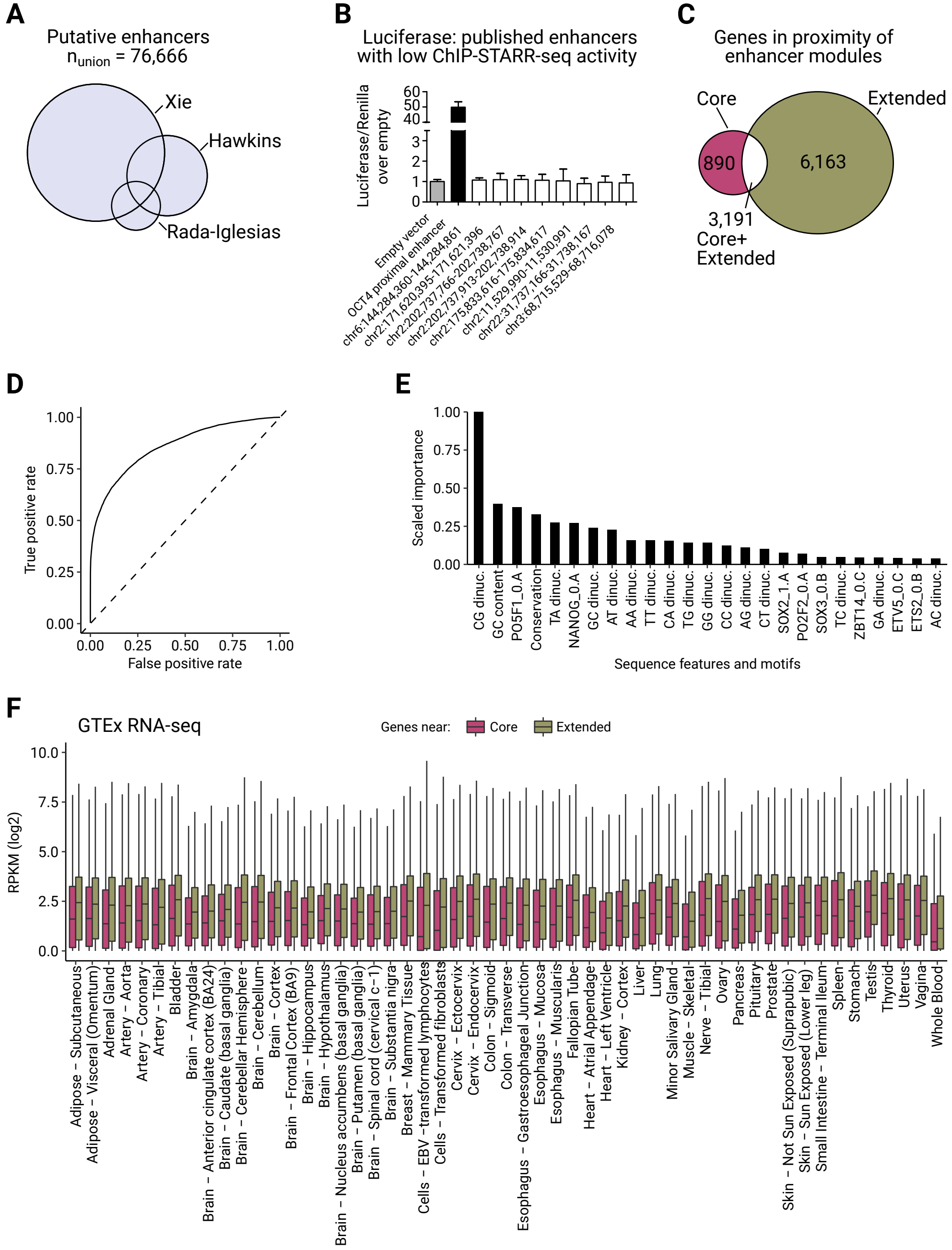


Figure S4 | Core and extended enhancer module. A) Illustration showing three source datasets (Hawkins et al., 2011; Rada-Iglesias et al., 2011; Xie et al., 2013) that contributed to the catalogue of putative ESC enhancers used in this study. We converted all enhancer coordinates to the same assembly (hg19) and then merged overlapping peaks resulting in a list of 76,666 putative enhancers. **B)** Luciferase assay in primed ESCs for eight putative enhancers that did not show activity in ChIP-STARR-seq. The OCT4 proximal enhancer (PE) is tested as a positive control. Luciferase activity is reported as the fold enrichment in luciferase counts over empty vector, normalised to the Renilla transfection control. Error bars indicate standard deviations, n=2. **C)** Illustration of the number of genes found in the proximity of core enhancer module (pink, left circle), or extended enhancer module (olive, right), or with enhancers of both types (white, overlap). Enhancer-gene assignments were performed with GREAT (McLean et al., 2010). **D)** Receiver operating characteristic (ROC) curve illustrating random forest classifier performance. **E)** Bar charts displaying the scaled variable importance of the top 25 sequence features used for the distinction of active enhancers from inactive regions. Motif IDs are shortened versions of the full ID from the HOCOMOCO database (v11) (Kulakovskiy et al., 2016). dinuc., dinucleotide frequency. **F)** Boxplots of gene expression (RNA-Seq; log2) in different tissues from the GTEx database (2013)for all genes linked to Core (pink) or Extended (olive) module enhancers or both (white). Boxplots represent the interquartile range (IQR), the line is the median, whiskers extend to 1.5xIQR and outliers are indicated as dots. RPKM, reads per kilobase million.

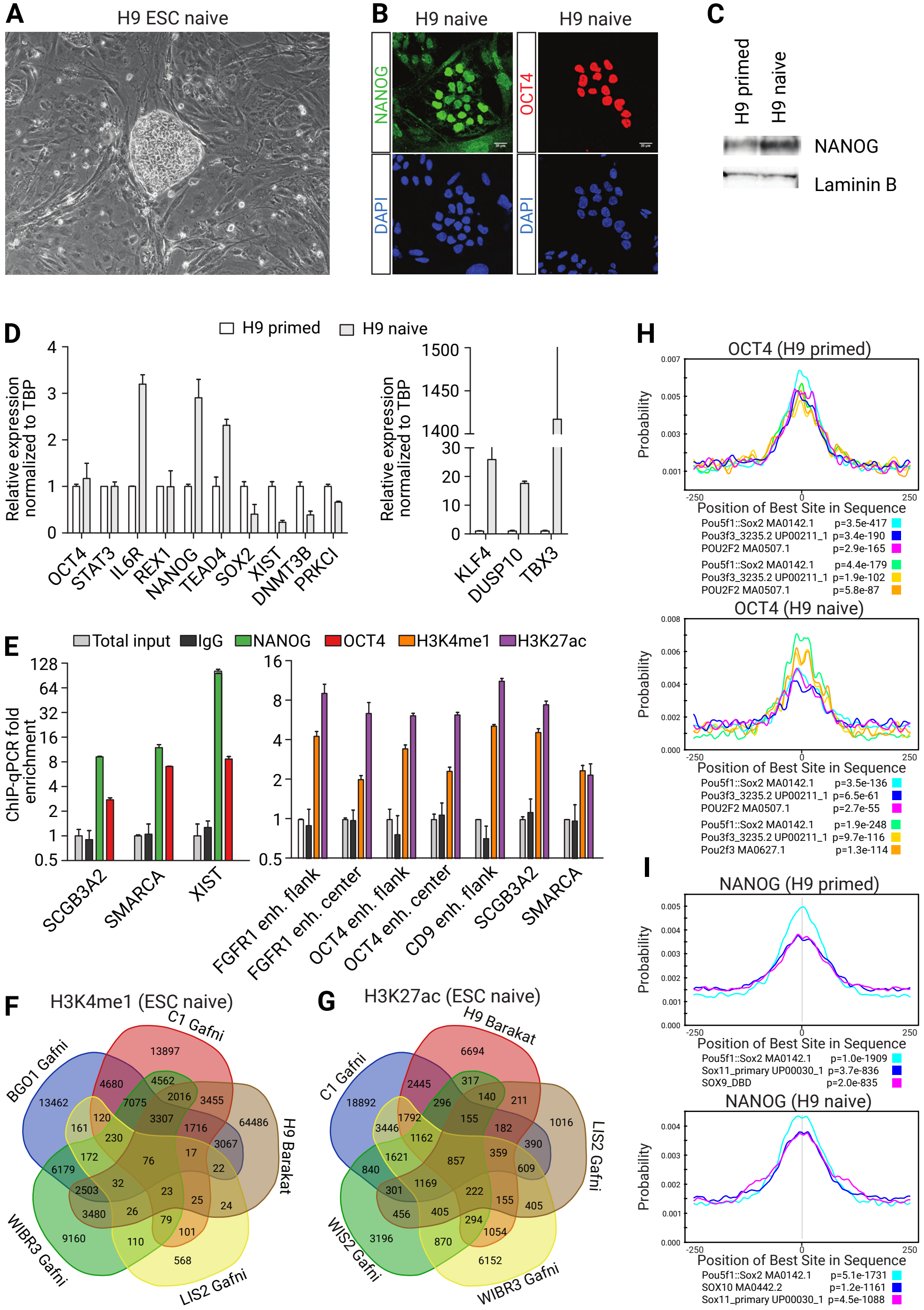


Figure S5 | Conversion of primed to naive H9 human embryonic stem cells. A) Brightfield microscopy of a representative colony of H9 ESCs cultured in naive culture conditions on feeders for 10 passages showing a more dome-shaped colony morphology (compare to Figure S1A). **B)** Immunofluorescence of naive H9 ESCs for NANOG (green) and OCT4 (red); DNA is stained with DAPI (blue). **C)** Immunoblot analysis of NANOG and LAMININ B in primed and naive ESCs. **D)** qRT-PCR of pluripotency-related genes in H9 ESCs cultured in primed or naive conditions (10 passages). Error bars indicate standard deviations; n=3. **E)** ChIP-qPCR in naive H9 ESC with anti-NANOG, anti-OCT4 or rabbit IgG at three OCT4 and NANOG binding sites (from primed ChIP-seq data) in SCGB3A2, SMARCA (Kunarso et al., 2010) and XIST (left) or with anti-H3K4me1, anti-H3K27ac or rabbit IgG at binding sites near FGFR1 (at central and flanking locations), POU5F1 (at central and flanking locations), CD9, SCGB3A2 and SMARCA (Rada-Iglesias et al., 2011) (right). The mean fold-enrichment is shown relative to total input control DNA, normalized to a non-bound site in ACTB (for NANOG and OCT4) or NCAPD2 (for H3K4me1 and H3K27ac). Error bars indicate standard deviations; n= 3. **F-G)** Venn diagrams of the overlap between ChIP-seq peaks in naive ESCs, indicating the cell line and study for F) H3K4me1 and G) H3K27ac. The numbers indicate overlapping peaks. **H)** Local motif enrichment analysis using CentriMo (Bailey and Machanick, 2012) for OCT4 ChIP-seq data generated in this study for primed (upper panel) and naive H9 ESCs (lower panel). Top-3 identified motifs and their p-values are indicated. **I)** as H, but now for NANOG.

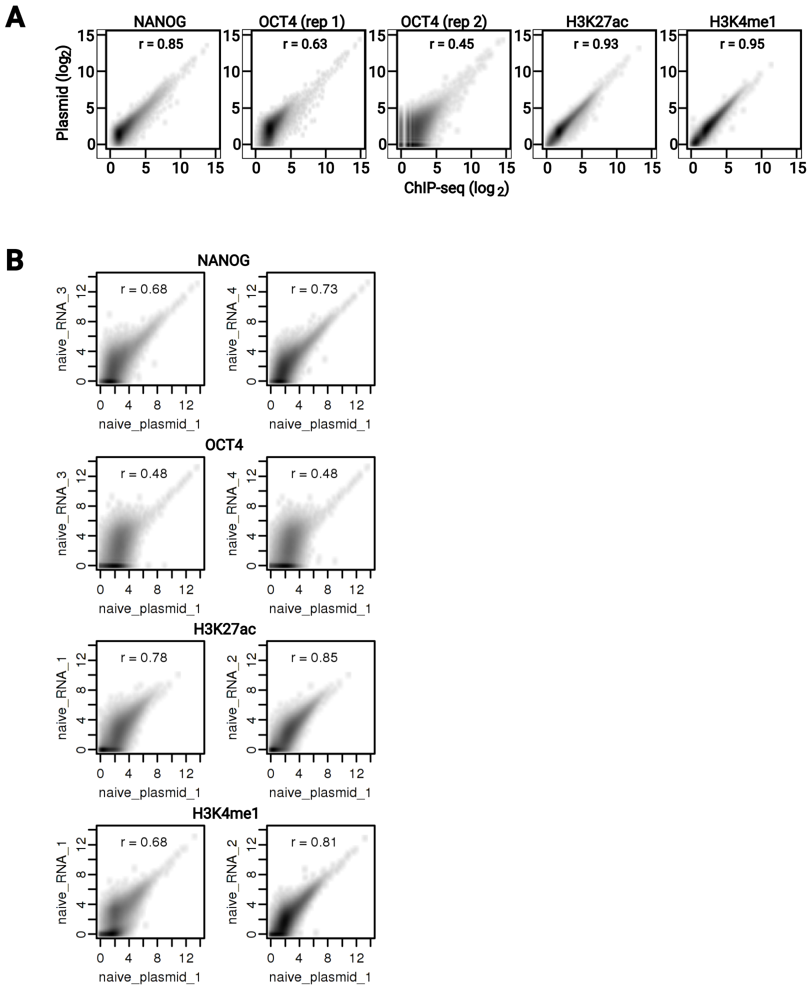
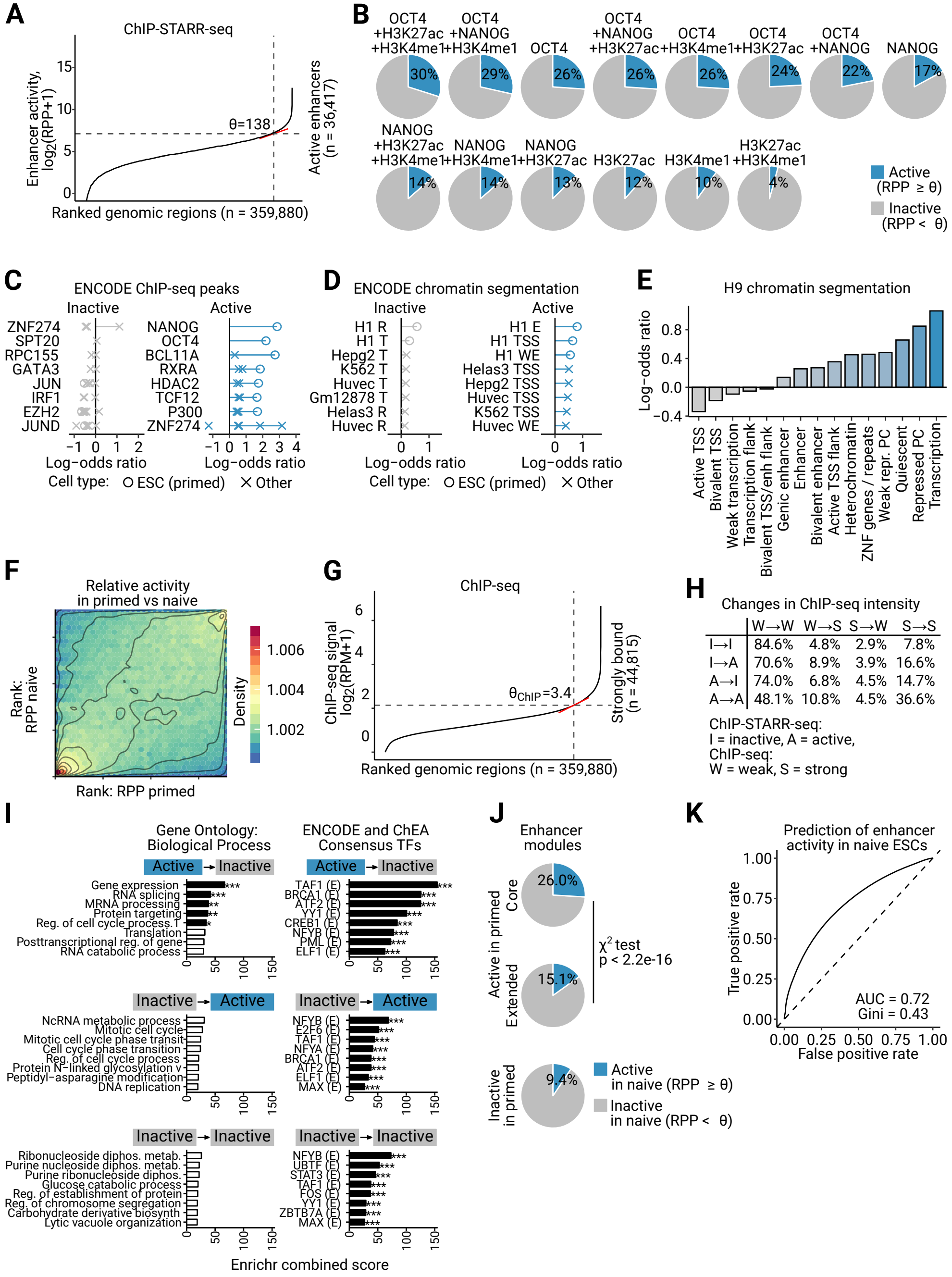


Figure S6 | ChIP-STARR-seq in naive ESCs. A-B) Scatterplots contrasting normalized read counts (reads per million) in naive H9 ESCs per peak for different datasets. The Pearson correlation coefficient (r) for each comparison is indicated for each plot. **A)** Comparison of ChIP-seq datasets and the corresponding ChIP-STARR-seq plasmid library. **B)** Comparison between ChIP-STARR-seq plasmid libraries prior to transfection and the corresponding RNA-seq read counts generated from two replicates of GFP-positive naive H9 ESCs after transfection. **C)** Pearson correlation coefficients (r) between replicate STARR-RNA-seq measurements from the same pool of either primed or naive ESCs transfected with a ChIP-STARR-seq library that was either generated in the same cell state (Primed→Primed and Naive→Naive) or in the respective other one (Primed→Naive and Naive→Primed), shown as a function of minimum read count in both replicates (0 = unfiltered, 1 = at least one read from the same plasmid measured in each replicate, etc.).



	W→W	W→S	S→W	S→S
I→I	84.6%	4.8%	2.9%	7.8%
I→A	70.6%	8.9%	3.9%	16.6%
A→I	74.0%	6.8%	4.5%	14.7%
A→A	48.1%	10.8%	4.5%	36.6%

ChIP-STARR-seq:
I = inactive, A = active,
ChIP-seq:
W = weak, S = strong

Figure S7 | Functional enhancers in naive ESCs. A) Plot showing enhancer activity (normalized ratio of ChIP-STARR RNA over plasmids; log2) in naive H9 ESCs ranked from lowest to highest across all measured enhancers (union of all peak calls). Active enhancers and inactive regions are discriminated by a threshold (θ) as indicated in the plot by a dashed line. RPP, reads per plasmid million. **B)** Distribution of active (RPP ≥138) and inactive sequences (RPP <138) in naive ESCs in ChIP-STARR-seq regions overlapping with ChIP-seq peaks of the indicated factor or combination of factors. **C)** Relative enrichment of DNA-binding proteins (DBPs) from the ENCODE database (2012) in inactive genome regions, as well as lowly active and highly active enhancers in naive H9 ESCs (compare to Fig. 5B). Shown are the log2-odds ratios between observed percentages of enhancers overlapping binding sites of each given DBP in the respective groups over the percentage of overlaps in the entire enhancer dataset. Each dot represents one ChIP-seq dataset for the given DBP and the lines connect the most extreme dot with zero for visualization. For each category, the eight most enriched DBPs are shown ranked by their mean log-odds ratio. ChIP-seq datasets produced from ESCs are indicated as dots and those from other cell sources as crosses. Enrichments were calculated using LOLA (Sheffield and Bock, 2016). **D)** LOLA enrichment plots as in panel C, but showing instead the relative over-presentation of ENCODE chromatin segments from different cell lines in enhancers with different activity levels. E, enhancer; PF, promoter-flanking region; R, repressed; T, transcribed; TSS, transcription start site; WE, weak enhancer. **E)** Barplots showing relative enrichment of H9 chromatin segment overlaps (Kundaje et al., 2015) in regions with ChIP-STARR-seq activity compared to inactive regions (see panel A). TSS, transcription start site; enh, enhancer; ZNF, zinc-finger protein. **F)** Density scatter plot comparing ranked activity (RPP) in primed and naive ESCs. Point density is represented by color and contours are shown to emphasize dense regions. The majority of points is located either in the lower left (inactive) or upper left (active in primed and naive) section of the plot. **G)** Plot equivalent to panel A, but showing ChIP-seq binding intensity instead of RPP. A change point analysis was performed to determine a threshold to distinguish strongly bound from weakly bound regions. RPM, reads per million. **H)** Table showing the percentage of ChIP-STARR-seq regions with a certain activity (e.g., Inactive in primed and naive ESCs) that remain weakly bound (W→W), gain binding (W→S), lose binding (S→W), or remain strongly bound (S→S) in the transition from primed to naive ESCs according to the threshold shown in panel G. W, weak ChIP-seq binding; S, strong ChIP-seq binding; I, inactive ChIP-STARR-seq region; A, active ChIP-STARR-seq enhancer. **I)** Functional enrichment analysis using Enrichr to test the relative over-representation of enhancers that lose ChIP-STARR-seq activity in the transition from primed to naive ESCs (Active→Inactive), that gain activity (Inactive→Active), or that remain inactive in both states (Inactive→Inactive). Enrichments were calculated for genes near these regions to test for GO assignments (left) or for ENCODE and ChEA ChIP-seq experiments (right). The x-axis reports the combined score calculated by Enrichr. **J)** Distribution of active (RPP ≥138) and inactive sequences (RPP <138) in naive ESCs in ChIP-STARR-seq regions of the Core module (top), Extended module (middle), or in regions that were inactive in primed ESCs. The p-value calculated by the χ2 test is indicated. **K)** Receiver operating characteristic (ROC) curve illustrating random forest classifier performance. The area under the curve (AUC) and Gini index are reported.

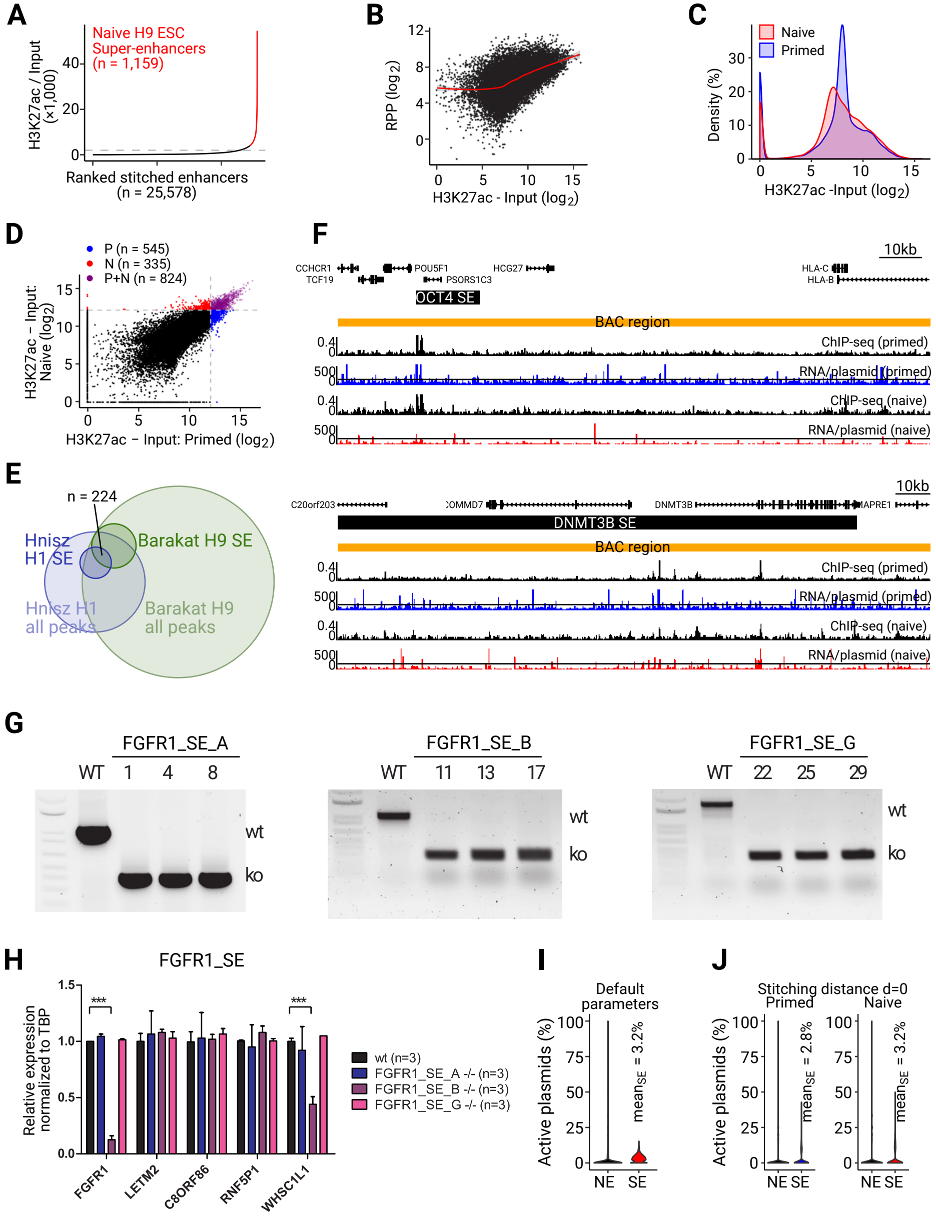


Figure S8 | Super-enhancers and ChIP-STARR-seq in primed and naive ESCs A) Super-enhancers in naive H9 ESCs (SEs) were called from H3K27ac ChIP-seq data on enhancers stitched within 12.5kb windows using the ROSE software (Whyte et al., 2013). **B)** Scatterplot contrasting SE intensity in naive H9 ESCs (H3K27ac signal divided by input) with ChIP-STARR-seq activity. The Pearson correlation coefficient (r) is indicated and the red line represents a generalised additive model fit to the data. **C)** Kernel density plots showing the distribution of SE intensity values (H3K27ac signal divided by input) in primed and naive H9 ESCs. **D)** Scatterplot contrasting SE intensity values (H3K27ac signal divided by input) in primed and naive H9 ESCs. Regions called as SEs in primed, naive, or both ESCs are indicated in blue, red, or purple, respectively. **E)** Comparison between super-enhancers and H3K27ac peaks from H1 ESCs (Hnisz et al., 2013) and this study. **F)** Genome browser plots showing two regions captured by our additional BAC-STARR-seq libraries. Shown are annotated genes in these regions, the BAC-covered region itself (in yellow), followed by tracks showing the combined ChIP-seq coverage and RNA over plasmid ratios in primed and naive ESCs. **G)** PCR genotyping of H9 wild type (WT) and targeted clones (numbers) that were transfected with Cas9 and gRNAs to delete three different constituents of the FGFR1 superenhancer (part A, B and G, see Figure 7). Primers used for genotyping are located outside the gRNA targets; wt = wt allele, ko = knockout allele. **H)** qRT-PCR analysis of wild type (wt) H9 ESCs or H9 ESCs with a homozygous deletion of FGFR1 super-enhancer part A, part B or part G for amplicons detecting FGFR1, LETM2, C8ORF86, RNF5P1 and WHSC1L1 (NSD3) mRNA. 3 clones per genotype were assessed, and expression was normalized for TBP and one wt set to 1. *** =p<0.001 (2-way ANOVA with Bonferoni post-test), error bars are SD. **I)** Violin plots showing the proportion of active plasmids (RPP ≥ 138) for 1,159 superenhancers (SE) compared to normal enhancers (NE) in naive ESCs. **J)** Violin plots as in panel I, for SEs and NEs in primed (left) and naive (right) ESCs with super-enhancers based on H3K27ac peaks with d=0kb stitching distance (called using ROSE).

YY1 Interactome

Dissecting the role of YY1 in gene regulation

Part III

Appendix

Summary

Samenvatting

Riassunto

Curriculum vitae

List of publications

PhD portfolio

Acknowledgements

Summary

Neurodevelopmental disorders are a group of complex and heterogeneous disorders affecting more than 3% of the children worldwide. Even when focussing on the genetic forms of NDDs, up to date about 50% of the patients still do not have a molecular diagnosis, which is essential both from a basic biology point of view, to better understand the molecular mechanisms at the basis of development and disease, but also from a clinical point of view, to provide better counselling and ultimately a therapeutic intervention tailored to each patient, moving the field towards precision medicine.

Thanks to the wider implementation of whole exome sequencing, more and more genes are identified to be causative of disease when their expression or function is altered. In the work I contributed to in these years, we could associate novel genes to disease (*BICRA* and *VPS41*, not included in this *Thesis*) and identify a new pathway, the nucleotide sugar metabolism, associated with developmental and epileptic encephalopathy. In **chapter 2**, I present such work, in which we report on 22 individuals presenting with intractable seizures, severe developmental delay and progressive microcephaly carrying a recurrent homozygous variant in the gene *UGP2* (chr2:64083454A > G). This variant leads to a tolerable missense variant in the longer *UGP2* isoform while leading to a loss of the start codon of the short isoform. The short isoform represents virtually all the UGP2 produced in brain, leading to a complete absence of the protein in this organ in patients. Having seen that the ratio of the two UGP2 isoforms differs in different cell types and tissues, we next wondered whether the regulation of their expression depends on different regulatory elements. In **chapter 3** we used two cell lines, embryonic stem cells (ESCs) and neural stem cells (NSCs), that express mainly the long and the short isoform, respectively, to try to identify drivers of differential isoform expression. We observed that the promoters of the two isoforms show differential activity in ESCs and NSCs. Furthermore, targeted chromatin capture analysis identified an enhancer region flanking *OTX1* that regulates *OTX1* expression and marginally affects UGP2 expression. Interestingly, silencing the *UGP2* promoter we observed a downregulation of both *UGP2* and *OTX1*, suggesting these two transcripts might indeed show mutual regulation in close 3D proximity.

Despite the investigation of the exome, many patients remain undiagnosed. These patients might carry alterations in the non-coding genome and, precisely, in enhancer regions that play a key role during embryonic development by regulating spatio-temporal gene expression. Our understanding of gene regulation has deepened

over the last decade, nevertheless we still lack the perfect mark to identify relevant and active enhancers, the prime targets where to search for variants in unsolved patients. Identifying active enhancers in a relevant cell type might help to solve the "missing heritability" and explain currently unexplained disorders. To achieve this, I contributed to the work of my fellow PhD student Soheil Yousefi, in which we performed an integrative computational analysis of virtually all currently available epigenome data sets related to human foetal brain to identify the differentially active enhancers that are likely important for early brain development (not included in this *Thesis*). However, it is still crucial to further develop functional high-throughput approaches for the functional validation of putative enhancer, so that more non-coding sequences and their variants can be functionally annotated, leading to a higher confidence in non-coding genome data resources and, ultimately, to clinical utility. In **chapter 4** we present such a method that combines chromatin immunoprecipitation with the massively parallel reporter assay STARR-seq, allowing the genome wide identification of functional enhancers. Here, we tested the method in human ESCs at two different states of pluripotency (naive and primed), while in **chapter 5** we applied this new method to identify the repertoire of functional enhancers in NSCs, which are of broader relevance for early brain development and NDDs. This work further helps to annotate functional non-coding sequences, including those that likely play a role in the regulation of NDD-relevant genes.

To further understand the mechanisms underlying regulation of gene expression, we focussed on the transcription factor YY1, alterations in which also lead to an NDD. It was recently described that in a variety of cell types YY1 is the protein that generally mediates enhancer-promoter interactions and thus, likely, gene activation. In **chapter 6** we investigated the YY1 protein interactome by affinity purification followed by mass-spectrometry to identify which protein partners and complexes might mediate the specificity of YY1 binding and cell-type specific gene regulation in ESCs and NSCs. In the final part of this thesis, we explored more deeply the functional mechanisms of gene expression regulation by enhancers and into what defines them as active. To better understand the role of YY1 in the flow of gene expression and enhancer activation, in **chapter 7** we generated human ESCs allowing rapid depletion of YY1 to investigate the consequences over time at various levels of the flow of information, including transcription, enhancer activity, histone acetylation (H3K27ac) and chromatin accessibility. Moreover, as YY1 seems to be particularly important during early neural induction and its haploinsufficiency leads to a neurodevelopmental disorder, we started the investigation of the same processes in NSCs.

The conclusive **chapter 8**, the general discussion, aims at merging all the findings of this thesis, broadening our current view on the genes and regulatory elements possibly involved in neurodevelopmental disorders and expanding our knowledge on the basic biology of active enhancers and gene regulation.

Samenvatting

Neuronale ontwikkelingsstoornissen zijn een groep complexe en heterogene aandoeningen die wereldwijd meer dan 3% van de kinderen treffen. Zelfs wanneer de aandacht wordt toegespitst op de genetische vormen van deze stoornissen, wordt bij ongeveer 50% van de patiënten nog steeds geen moleculaire diagnose gesteld. Het stellen van een diagnose is van essentieel belang, zowel vanuit fundamenteel biologisch oogpunt, om de moleculaire mechanismen die aan de basis liggen van de ontwikkeling en de ziekte beter te begrijpen, als vooral ook vanuit klinisch oogpunt, om een betere begeleiding te kunnen bieden en uiteindelijk een therapeutische interventie te kunnen toepassen die is toegesneden op elke patiënt, resulterende in precision medicine.

Dankzij de ruimere implementatie van “whole exome sequencing” wordt van steeds meer genen vastgesteld dat zij de oorzaak zijn van ziekte wanneer hun expressie of functie is veranderd. In het werk waaraan ik in deze jaren heb bijgedragen, konden we nieuwe genen associëren met ziekte (BICRA en VPS41, niet opgenomen in dit proefschrift) en een nieuwe pathway identificeren, het nucleotide suikermetabolisme, geassocieerd met ontwikkelings- en epileptische encephalopathie. In hoofdstuk 2 presenteer ik dit werk, waarin we verslag doen van 22 individuen met hardnekkige epileptische aanvallen, ernstige ontwikkelingsachterstand en progressieve microcefalie die een recidiverende homozygote variant in het gen UGP2 (chr2:64083454A > G) dragen. Deze variant leidt tot een tolereerbare missense variant in de langere UGP2 isovorm, terwijl deze leidt tot een verlies van het startcodon van de korte isovorm. De korte isovorm vertegenwoordigt vrijwel al het UGP2 dat in de hersenen wordt geproduceerd, hetgeen leidt tot een volledige afwezigheid van het eiwit in dit orgaan bij patiënten , hetgeen leidt tot de ziekte. Nu we gezien hebben dat de verhouding van de twee UGP2 isovormen verschilt in verschillende celtypen en weefsels, vroegen we ons vervolgens af of de regulatie van hun expressie afhangt van verschillende regulatoire elementen. In hoofdstuk 3 gebruikten we twee cellijnen, embryonale stamcellen (ESCs) en neurale stamcellen (NSCs), die respectievelijk voornamelijk de lange en de korte isovorm tot expressie brengen, om te proberen de regulatie van deze differentiële isovorm expressie te identificeren. We stelden vast dat de promotors van de twee isovormen een verschillende activiteit vertonen in ESCs en NSCs. Bovendien identificeerden gerichte chromatine capture analyse een enhancer regio die het gen OTX1 flankeert welke OTX1 expressie reguleert en UGP2 expressie marginaal beïnvloedt. Interessant is dat door het uitschakelen van de UGP2-promotor een downregulatie van zowel UGP2 als OTX1 werd waargenomen, wat suggereert dat deze twee transcripten inderdaad een wederzijdse regulatie dicht bij gelegen in de 3D ruimte van de celkern aangaan.

Ondanks het onderzoek van het exoom blijven vele patiënten ongediagnosticeerd. Deze patiënten zouden veranderingen kunnen dragen in het niet-coderende genoom en, precies, in enhancer regio’s die een sleutelrol spelen tijdens de embryonale ontwikkeling door het reguleren van spatio-temporele genexpressie. Ons begrip van genregulatie is het laatste decennium enorm uitgebreid, maar toch ontbreekt het ons nog aan de perfecte markering om relevante en actieve enhancers te identificeren. Terwijl die enhancers juist de voornaamste doelwitten zijn voor varianten in onopgeloste patiënten die zouden kunnen helpen om de “ontbrekende erfelijkheid” op te lossen en momenteel onverklaarbare aandoeningen te verklaren. Om dit te ondervangen heb ik bijgedragen aan het werk van mijn mede-promovendus Soheil Yousefi, waarin we een integratieve bioinformatische analyse hebben uitgevoerd van vrijwel alle momenteel beschikbare epigenoom-datasets met betrekking tot menselijke foetale hersenen om de differentieel actieve enhancers te identificeren die waarschijnlijk belangrijk zijn voor de vroege hersenontwikkeling (niet opgenomen in dit proefschrift). Naast dergelijk rekenkundig onderzoek, blijft het echter ook nog steeds van cruciaal belang om functionele high-throughput benaderingen voor de functionele validatie van mogelijke enhancers verder te ontwikkelen. Hierdoor kunnen meer niet-coderende sequenties en hun varianten functioneel worden geannoteerd, wat zal leiden tot betere databronnen voor het niet-coderende genoom welke klinisch kunnen worden toegepast. In hoofdstuk 4 presenteren we een dergelijke methode die chromatine immunoprecipitatie combineert met de massive parallel reporter assay STARR-seq, en die de genoombrede identificatie van functionele enhancers mogelijk maakt. Hier hebben we de methode getest in humane ESCs, gekweekt in twee verschillende stadia van pluripotentie (naïef en primed), terwijl we in hoofdstuk 5 deze nieuwe methode hebben toegepast om het repertoire van functionele enhancers in neurale stamcellen te identificeren, die van bredere relevantie zijn voor vroege hersenontwikkeling en neuronale ontwikkelingsstoornissen. Dit werk draagt verder bij tot de annotatie van functioneel relevante niet-coderende sequenties, met inbegrip van die sequenties die waarschijnlijk een rol spelen in de regulatie van genen betrokken bij neuronale ontwikkelingsstoornissen.

Om de mechanismen die ten grondslag liggen aan de regulatie van gen enhancers verder te begrijpen, hebben we ons gericht op de transcriptiefactor YY1, waarvan afwijkingen ook leiden tot een neuronale ontwikkelingsstoornis. Recent is beschreven dat in verschillende celtypen YY1 het eiwit is dat over het algemeen de enhancer-promoter interacties en daarmee, waarschijnlijk, gen activatie medieert. In hoofdstuk 6 hebben we het interactoom van het YY1 eiwit onderzocht door middel van affinity purification gevolgd door massaspectrometrie om te identificeren welke eiwitpartners en complexen mogelijk mediëren in de specificiteit van YY1 binding en celtype-specifieke genregulatie in ESCs en NSCs. In het laatste deel van dit proefschrift hebben we ons verder verdiept in de functionele mechanismen van genexpressieregulatie door enhancers en in wat hen definieert als actief. Om de rol van YY1 in de stroom van genexpressie en enhancer activatie beter te begrijpen, hebben we in hoofdstuk 7 humane ESCs gegenereerd die snelle depletie van YY1 mogelijk maken om de gevolgen in de tijd te onderzoeken op verschillende niveaus van die informatiestroom, waaronder transcriptie, enhanceractiviteit, histonacetylering (H3K27ac) en chromatine toegankelijkheid. Bovendien, aangezien YY1 bijzonder belangrijk lijkt te zijn tijdens de vroege neurale inductie en zijn haploinsufficiëntie leidt tot een neurologische ontwikkelingsstoornis, zijn we begonnen met het onderzoek van dezelfde processen in NSCs.

Het afsluitende hoofdstuk 8, de algemene discussie, heeft tot doel alle bevindingen van dit proefschrift samen te voegen, onze huidige kijk op de genen en regulatorische elementen die mogelijk betrokken zijn bij neurologische ontwikkelingsstoornissen te verbreden en onze kennis over de basisbiologie van actieve enhancers en genregulatie uit te breiden.

Riassunto

I disturbi del neurosviluppo sono un gruppo complesso ed eterogeneo di malattie che colpiscono più del 3% dei bambini nel mondo. Circa il 50% dei pazienti con un disturbo genetico non ha una diagnosi molecolare che è essenziale sia dal punto di vista biologico, per capire i meccanismi alla base della malattia, sia dal punto di vita clinico, per fornire una migliore consulenza genetica e per sviluppare terapie su misura per ogni paziente.

Grazie alla diffusione del sequenziamento completo dell'esoma (ovvero la parte del genoma che fornisce le istruzioni per la produzione di proteine), vengono identificati sempre più geni i cui difetti di espressione o funzione portano allo sviluppo di una malattia. Durante il mio lavoro di dottorato ho contribuito all'identificazione di nuovi geni collegati a disturbi del neurosviluppo, tra cui BICRA e VPS41, non inclusi in questa *Tesi*, e UGP2, descritto nel **capitolo 2**. Il gene *UGP2* viene trascritto in due isoforme di RNA messaggero, una lunga e una corta. La mutazione identificata nei 22 pazienti inclusi nel nostro studio causa una perdita del codone di inizio nell'isoforma corta che porta alla totale perdita della proteina UGP2 corta. L'isoforma corta rappresenta tutta la UGP2 che viene prodotta nel cervello, dove l'isoforma lunga non è espressa, e la sua totale assenza nei pazienti causa un'encefalopatia epilettica caratterizzata da crisi epilettiche resistenti a farmaci, severo ritardo dello sviluppo e microcefalia progressiva. In questo studio abbiamo osservato che diversi organi espirimono le due isoforme di UGP2 in diverse percentuali, per questo nel **capitolo 3** abbiamo studiato i meccanismi che portano alla differente espressione di isoforma lunga e corta. Abbiamo usato due linee cellulari, cellule staminali embrionali (ESC) e neurali (NSC), che esprimono rispettivamente una maggioranza di isoforma lunga o corta, e abbiamo osservato che i promotori delle due isoforme hanno un diverso livello di attività nelle due linee cellulari. Al momento stiamo cercando di determinare le cause di questa differenza in attività. Abbiamo intoltre identificato un enhancer che fiancheggia e regola l'espressione del gene *OTX1* che regola marginalmente anche l'espressione di *UGP2*. Sorprendentemente, silenziando i promotori di *UGP2* osserviamo una riduzione nell'espressione di *OTX1*, suggerendo ulteriormente che questi due geni si potrebbero regolare a vicenda.

Come scritto in precedenza, nonostante il sequenziamento completo dell'esoma, molti pazienti rimangono senza una diagnosi molecolare. La nostra ipotesi è che questi pazienti possano avere delle mutazioni nella parte del genoma che non esprime proteine in sé, ma ne regola l'espressione. Tra queste regioni, troviamo gli enhancer che, regolando il preciso momento e luogo in cui i geni sono espressi, hanno un ruolo cruciale durante lo sviluppo embrionale. La nostra conoscenza dell'espressione genica è aumentata drasticamente nell'ultimo decennio, ma nonostante ciò, manca ancora un marcatore che identifichi con sicurezza enhancer attivi in determinati tipi cellulari o tessuti. Nel **capitolo 4** presentiamo un metodo che, combianando due tecniche ampiamente utilizzate nel campo, permette di identificare contemporaneamente e in larga scala enhancer attivi. Abbiamo applicato questo metodo in ESC e nel **capitolo 5** in NSC, che rappresentano un modello cellulare rilevante per lo studio dello sviluppo embrionale del cervello umano.

Nell'ultima parte di questa *Tesi*, abbiamo studiato più nel dettaglio i meccanismi molecolari per cui gli enhancer regolano l'espressione genica. YY1 è una proteina che media tale funzione e quando il suo livello viene ridotto causa un disturbo del neurosviluppo. Nel **capitolo 6** abbiamo identificato i partner di YY1 in ESC e in NSC per capire come YY1 può regolare l'espressione genica specifica dei due tipi cellulari. Nel **capitolo 7** abbiamo generato una linea cellulare in cui è possibile indurre la rapida degradazione di YY1 per studiarne le conseguenze nel tempo a livello di espressione genica e attività di enhancer, sia in ESC che in NSC.

CURRICULUM VITAE

26/10/1992 Trento - Italy elena.perenthaler@gmail.com

WORK EXPERIENCE

PhD STUDENT
Aug 2017 – Dec 2021 (4 years 5 months)
Barakat laboratory - Clinical Genetics department
Erasmus University Medical Center - Rotterdam (NL)

ERASMUS + RESEARCH INTERN
JAN 2017 - JUL 2017 (6 months)
Gillingwater laboratory - Centre for integrative physiology & Euan MacDonald centre for motor neurone disease research
University of Edinburgh (UK)

RESEARCH ASSISTANT
Feb 2015 – Aug 2015 (6 months)
Viero laboratory of translational architectomics
Institute of biophysics - National research Council (IBF-CNR; Italy)

EDUCATION

MSc IN CELLULAR AND MOLECULAR BIOTECHNOLOGY - NEUROBIOLOGY
UNIVERSITY OF TRENTO, ITALY; OCT 2014 - OCT 2016
Thesis "In vivo translatome analyses reveal translational defects in early symptomatic Spinal Muscular Atrophy" supervised by Prof. A. Quattrone

BSc IN BIOMOLECULAR SCIENCES AND TECHNOLOGY
UNIVERSITY OF TRENTO, ITALY; SEP 2011 - SEP 2014
Thesis "A possible new role for the RNA binding protein CELF3 in translation regulation and neurites outgrowth" supervised by Prof. A. Quattrone

HIGH SCHOOL - LICEO SCIENTIFICO BILINGUE
LICEO BERTRAND RUSSELL, CLES, ITALY; SEP 2006- JUN 2011

LIST OF PUBLICATIONS

1. **Perenthaler E**, Brouwer RWW, Yousefi S, Nikoncuk A, Deng R, Lanko K, van IJcken WFJ, and Barakat TS. Investigating the chromatin architecture of the *UGP2* locus by targeted chromatin conformation capture. *In preparation*
2. **Perenthaler E***, Yousefi S*, Deng R*, Nikoncuk A, van IJcken WFJ, Mulugeta E and Barakat TS. Identification of the active enhancer landscape in Neural Stem Cells by ChIP-STARR-seq. *In preparation*
3. **Perenthaler E**, Yousefi S, Deng R, Nikoncuk A, Lanko K, van IJcken WFJ, Mulugeta E, Demmers J and Barakat TS. YY1 interacts with cell-type specific complexes in embryonic and neural stem cells. *In preparation*
4. **Perenthaler E***, Deng R*, Yousefi S*, Nikoncuk A, Kan TW, Pelicano de Almeida M, van Ijken WFJ, Mulugeta E and Barakat TS. Dissecting the role of YY1 in determining gene expression and enhancer activity. *In preparation*
5. Yousefi S, Deng R*, Lanko K*, Medico Salsench E*, Nikoncuk A*, van der Linde HC, **Perenthaler E**, van Ham TJ, Mulugeta E#, Barakat TS#. (2021) Comprehensive multi-omics integration identifies differentially active enhancers during human brain development with clinical relevance. Genome Med Oct 19;13(1):162.
6. Sanderson LE*, Lanko K*, Alsagob M*, Almass R*, Al-Ahmadi N*, Najafi M, Al-Muhaizea MA+, Alzaidan H+, AlDhalaan H+, **Perenthaler E**, ..., Schmidts M#, Barakat TS#, van Ham TJ#, Kaya N#. (2021) Bi-allelic variants in HOPS complex subunit VPS41 cause cerebellar ataxia and abnormal membrane trafficking. Brain 144(3):769-780
7. Barish S*, Barakat TS*, Michel BC*, Mashtalir N*, Phillips JB, Valencia AM, Ugur B, Wegner J, Scott TM, Bostwick B, Undiagnosed Diseases Network, Murdock DR, Dai H, **Perenthaler E**, ..., Bellen HJ. (2020) BICRA, a SWI/SNF Complex Member, Is Associated with BAF-Disorder Related Phenotypes in Humans and Model Organisms. Am J Hum Genet 07, 096-2
8. Verheul TCJ*, van Hijfte L*, **Perenthaler E***, Barakat TS. (2020) The Why of YY1: Mechanisms of Transcriptional Regulation by Yin Yang. Front Cell Dev Biol 8: 59264
9. Lauria F*, Bernabò P*, Tebaldi T*, Groen EJN*, **Perenthaler E**, Maniscalco F, Rossi A, Donzel D, Clamer M, Marchioretto M, Omersa N, Orri J, Dalla Serra M, Anderluh G, Quattrone A, Inga A, Gillingwater TH# and Viero G#. (2020) SMN-primed ribosomes modulate the translation of transcripts related to spinal muscular atrophy. Nat Cell Biol 22: 239-25
10. **Perenthaler E**, Nikoncuk A*, Yousefi S*, Berdowski WM*, Alsagob M*, ..., Barakat TS. (2020) Loss of UGP2 in brain leads to a severe epileptic encephalopathy, emphasizing that bi-allelic isoform-specific start-loss mutations of essential genes can cause genetic diseases. Acta Neuropathol 39: 45-442
11. **Perenthaler E***, Yousefi S*, Niggl E*, Barakat TS. (2019) Beyond the Exome: The Non-coding Genome and Enhancers in Neurodevelopmental Disorders and Malformations of Cortical Development. Front Cell Neurosci 3, 352
12. Clamer M, Tebaldi T, Lauria F, Bernabò P, Gómez-Biagi RF, Marchioretto M, Kandala DT, Minati L, **Perenthaler E**, Gubert D, Pasquardini L, Guella G, Groen EJN, Gillingwater TH, Quattrone A, and Viero G. (2018) Active Ribosome Profiling with RiboLace. Cell Rep 25: 097-08 e095
13. Groen EJN, **Perenthaler E**, Courtney NL, Jordan CY, Shorrock HK, van der Hoorn D, Huang Y, Murray LM, Viero G, Gillingwater TH. (2018) Temporal and tissue-specific variability of SMN protein levels in mouse models of spinal muscular atrophy. Hum Mol Genet 27: 285-2862.
14. Barakat TS*, Halbritter F*, Zhang M+, Rendeiro AF+, **Perenthaler E**, Bock C, Chambers I. (2018) Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. Cell Stem Cell 23: 276-288 e278
15. Bernabò P*, Tebaldi T*, Groen EJN*, Lane FM*, **Perenthaler E**, Mattedi F, Newbery HJ, Zhou H, Zuccotti P, Potrich V, Shorrock HK, Muntoni F, Quattrone A, Gillingwater TH, Viero G. (2017) In Vivo Translatome Profiling in Spinal Muscular Atrophy Reveals a Role for SMN Protein in Ribosome Biology. Cell Rep 2: 953-965

* these authors contributed equally
+ these authors contributed equally
# these authors jointly supervised this work

PhD PORTFOLIO

Courses

2018 - Optogenetics
2018 - Epigenetic regulation in heath and disease
2018 - Scientific integrity
2018 - Safely working in the laboratory
2018 - Stem cells, organoids and regenerative medicine
2020 - Multiomics data analysis using R
2021 - Microscopic image analysis: from theory to practice

Workshops and conferences

2018 - MGC workshop (Texel - NL; short presentation)
2018 - UCSC Gene Browsing workshop
2019 - MGC workshop (Maastricht - NL; poster)
2021 - MGC workshop (online; presentation)
2021 - Boost your research career with a personal grant

Meetings and conferences

2017-2021 - Weekly clinical genetics department research meetings
2018-2021 - Weekly cell biology department research meetings
2017-2021 - Clinical genetics department lectures
2017 - 27th MGC symposium (Rotterdam - NL)
2017 - Dutch neurodevelopmental disorders day (Rotterdam - NL)
2018 - Gene transcription in health and disease symposium (Rotterdam - NL)
2019 - Brain malformations: a roadmap for future research conference (Rehovot - IL; poster)
2019 - 29th MGC symposium (Rotterdam - NL)
2019 - ACE SBM and SCORE day (Rotterdam - NL)
2019 - Mini symposium CRISPR-Cas9 (Rotterdam - NL)
2020 - ESHG (online; presentation)
2021 - Neuro-MIG conference (online; presentation)
2021 - Sophia research day (online; presentation)
2021 - Brain prize meeting (online)
2021 - ESHG (online; presentation)

Teaching

Supervision of a Bachelor student
Supervision of a Master student (2x)
Supervision of an intern (2x)

Acknowledgements

> *"Pride can lead those who achieve their dreams to delude themselves into thinking that success was just the result of their actions"*
>
> *"Chi realizza il sogno può illudersi con superbia che il successo sia soltanto il risultato del proprio agire"*
>
> *(Samantha Cristoforetti - Diary of an apprentice astronaut)*

…but that's absolutely not the case! I would have never reached the stage of having a thesis without the support and the help of so many people and I'll try my best to mention them all here.

Firstly, I'd like to thank all the **patients and their families** for contributing to our research and for giving us the permission to use their material and data. I hope our work has contributed to a better understanding of genetic diseases and has taken us a step closer to the development of treatments.

My deepest gratitude goes to my copromotor. Dear **Stefan**, thanks for giving me the opportunity and the honour of being the first PhD student in your lab. Thanks for your support over the past years and for helping me grow as a scientist. I hope I did not let you down!

I would like to thank also my late promotor Prof.dr. Robert Hofstra and my promotor Prof.dr. Ype Elgersma. Dear **Robert**, thank you for making me reach the end of every single meeting happier of my achievements than I was at the start of it. You are greatly missed. Dear **Ype**, thank you for supporting me during the last and hardest period of my PhD journey and thesis writing.

I would also like to extend my gratitude to all the members of my **thesis committee**, Prof.dr. Danny Huylebroeck, Prof.dr. Niels Geijsen, Dr. Grazia Mancini, Prof.dr Elfride De Baere, Prof.dr. Sarah Vergult, Dr. Tjakko van Ham, and Dr. Annelies de Klein, thank you for reading this book and for being willing to be on the panel.

Anita and Mari, a very special thanks for being my **paranymphs**. I love knowing you will be by my side.

In the past five years the Barakat lab expanded a lot, going from one single bench for 3 people to a full "U". But one person has been there all the time. **Anita**, words will never be enough to describe how happy I am to have met you. The first time I saw you I though "she does not talk a lot and neither do I, how are we going to deal with this?". Well, it took only 5 minutes for this feeling to change. You are the best lab partner and friend I could ever wish for. We understand each other despite using half of the words a proper sentence requires (or even a single sound)! I definitely wouldn't be anywhere without you and your encouragement, all the things "I" did in these years, "we" did. So….cheers to our achievement, FREND!!!

The interplay between wet-lab and dry-lab has been fundamental for this work. And due to my complete inability to write a single line of code, I need to thank my angels. Dear **Soheil**, dear **Ruizhi** without you this thesis would have been a collection of meaningless, useless data. Thank you from the bottom of my heart for your help, your support and for always answering the most stupid questions and requests I had! Also, thanks **Eskeww** and **Rutger** for your analyses and for your support!

Dear **Kristina**, dear comrade, dear UPIR. You joined us relatively recently but you have been my guiding light in times of hard experimental failures! Thank you for discussing with me a million solutions. Most failed but…eventually things came together! And thank you for feeding us with amazing cakes! Give a big -unappreciated- cuddle to Bulochka 😊

To all the other members of the group, **Eva M**, **Yuwei**, **Sarah**, **Leslie**, **Ellis**. Thank you for all your input and for the amazing time we had! And I cannot forget about all the students, Mariana, Thomas, Daria and Maria. Dear **Mariana**, unfortunately you experienced only the failures of the degron project, but fortunately enough you also missed 2 more years of failures! With only with 3 years delay, results also came. So, thank you sooo much for your resilience with cloning! Dear **Thomas**, thank you for constantly bringing joy to the lab, we had a great time when you were around! Dear **Daria**, you stayed with us only 3 months, but I got to know you as the sweetest, kindest girl. You grew so much as a scientist in such a little time and you kick-started the *UGP2* project, thank you!! 😊

I want to thank also all the people with whom I was lucky enough to work before my doctoral journey. If I decided to pursue a PhD it's definitely also thanks to you! Dear **Gabriella**, my first role-model of a researcher. Thank you for always inspiring me and for transmitting me your passion. Dear **Paola**, this thesis is dedicated to you. Thank you for constantly supporting me every time I had a single doubt in the past NINE years. Dear **Tom** and **Ewout**, thank you for hosting me and for guiding me through 6 months in beautiful Edinburgh (I strongly suggest anyone that did not visit the city to do so, you will absolutely not regret it!). You taught me so much! And without this experience I would have never come across this PhD position in Rotterdam so… double thanks!

I want to thank also all the friends I met along the way. Dear **Shami**, thank you for being the most amazing "U"-neighbour. Thank you for always making me laugh, for always being up for a coffee or a trip to AH, for showing me your beautiful brain stainings, for organizing social activities, etc. Simply thanks for everything! Dear **Eva V**, thanks for your friendship, for sharing a flat with me, for all the dinners you cooked, for enjoying what I cooked, for all the movies we watched together, for making COVID-quarantine so enjoyable and for all the fun we had! But also, thank you for being so organized. I would have never gotten anything done during quarantine, if it weren't for your extreme organization 😊 **Fabio**, **Ale**, **EliGelli**, **Rodrigo**, **Pablo**, **Douglas**, **Stjin**, **Claudia**, **Eva N**, **Isa** thank you for all the happy memories with a beer, or two, and a couple of G&Ts. You were so much fun! And a much-needed distraction from the work-stress!

Tjakko, **Herma**, **Woutje** and all the other members of the van Ham lab, thank you for your amazing zebrafish work! **Tsung Wai**, thank you so much for all the hours spent sorting and your patience. It took ages, but we made it! **Wilfred** and the **Biomics Core Facility**, **Jeroen** and the **Proteomics Core Facility**, thanks for your crucial help!

Adriana, **Annelies**, **Atze**, **Gerben**, **Grazia**, **Maria**, **Pim**, **Rob**, **Vincenzo**, **Wim**, **Almira**, **Ana**, **Bianca**, **Chantal**, **Christina**, **Claudine**, **Daphne**, **Erik**, **Erwin**, **Esmay**, **Federico**, **Fenne**, **Guido**, **Jonathan**, **Jordy**, **Katherine**, **Kirke**, **Kyra**, **Laura K**, **Laura V**, **Lies-Anne**, **Martyna**, **Mike**, **Monica**, **Naomi**, **Natasha**, **Niko**, **Nina**, **Nynke**, **Quishi**, **Rachel**, **Rob V**, **Roy**, **Saif**, **Tom**, **Valerie**, **Wim Q**, **Wojtek**, **Yu-Ying**, the other past and present colleagues in the **Clinical genetics department**, **Eveline**, **John**, **Joost**, **Marjoleine**, **Mehrnaz**, the **iPS Core facility**, all the other colleagues on the 9th floor, and all the colleagues of the **Cell Biology department**, thank you for your support and for your input!

Dear **Jeanne**, you are 100% my luckiest "pick" in the Netherlands! Deciding to share a flat with a stranger was scary, but I had the most amazing time! Starting from the first day I set foot in this country, you always amazed me in all the ways you tried (and succeeded) to help me: you made sure I knew the neighbourhood and the way to work, you helped me finding a bike and filling in all the taxes paperwork, you supported me in every single sad or happy moment and you even made sure I would get my last flat!! You are simply amazing and I wish you all the best in your future! I hope I will get to show you round Trento like you showed me round Rotterdam 😊

Marti, I'll miss so much cooking for a "pozzo senza fondo" like you that appreciates everything! Thank you for always trying to make me want to socialize and thank you for never giving up! Every time you'll come back to Italy, my door (and the one of my fridge) will always be open for you 🖤

Amicicci biotech, **Bus**, **Dennis**, **Gabri**, **Giacomino**, **Mari**, **Mic**, **Simic**, grazie per esserci sempre nonostante la distanza. Dennis, Mari, Gabri grazie per essere stati la mia via di fuga preferita dall'Olanda e grazie per aver bivaccato nel mio soggiorno. You are happiness <3

Grazie a tutti gli amici e all'amica di sempre, **Laura**. Mi sei mancata un sacco in questi anni, ma ci sei sempre stata e sempre ci sarai. E io per te. Adesso recupereremo il tempo perso 😌

Cinquanta. I biglietti aerei per vederti ancora salvati sul mio telefono. E chissà quanti ne mancano! **Mattia**, grazie di tutto, per sopportarmi e supportarmi sempre e comunque. Senza di te sarei persa. E grazie a **Zero** 🐱

A tutta la mia famiglia, i miei genitori, i miei fratelli, mia sorella, i miei nipoti, i miei cognati: GRAZIE!

Elena