EXPLORING MISSING HERITABILITY IN NEURODEVELOPMENTAL DISORDERS

LEARNING FROM REGULATORY ELEMENTS

Soheil Yousefi

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: 978-94-6419-706-8
Author: Soheil Yousefi
Cover design and layout: Soheil Yousefi and Tooba Abbassi-Daloii
Printed by: Gildeprint
Printing of this thesis has been kindly supported by the department of Clinical Genetics.

EXPLORING MISSING HERITABILITY IN NEURODEVELOPMENTAL DISORDERS

LEARNING FROM REGULATORY ELEMENTS

ONDERZOEK NAAR ONTBREKENDE ERFELIJKHEID BIJ NEUROLOGISCHE ONTWIKKELINGSSTOORNISSEN

LEREN VAN 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

11th April 2023 at.13:00 hrs

by

Soheil Yousefi
born in Gorgan, Iran

Erasmus University Rotterdam

Doctoral Committee:

Promotor:	Prof.dr. Y. Elgersma
Other members:	Prof.dr. J. Gribnau Dr. J.E.M.M. de Klein Prof.dr. P.A.C. 't Hoen
Copromotors:	Dr. T.S. Barakat Dr. ir.E. Mulugeta

Contents

Chapter 1

Introduction

Parts of this Introduction have been published in:

Elena Perenthaler[1*], Soheil Yousefi[1*], Eva Niggl[1*] and Tahsin Stefan Barakat[1]. Beyond the Exome: The Non-coding Genome and Enhancers in Neurodevelopmental Disorders and Malformations of Cortical Development. *Frontiers in Cellular Neuroscience. 2019; 13:352.*

[1]Department of Clinical Genetics, Erasmus MC University Medical Center, Rotterdam, The Netherlands.

[*] Contributed equally

The brain and causes of neurodevelopmental disorders

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, behavior and social interactions. How such a complex organ is formed during development has fascinated biologists for centuries, and we are now living in a technology driven era where knowledge gained through various disciplines such as medicine, biotechnology, computational biology and neuroscience enables us for the first time to get a glimpse on how these intricate processes are genetically regulated. Understanding the developmental biology of the human brain is not only of utmost importance for satisfying our own natural curiosity of what makes us human, but also promises improvements and therapeutic options for various disorders that affect the human brain, including neurodevelopmental and neurodegenerative diseases, which, together, are a burden on society and have a negative effect on the quality of lives of individuals[1,2].

One of the traditionally best studied parts of the human brain is the cerebral cortex, which is responsible for cognition and sensorimotor activity. The development of the cerebral cortex is a complex and dynamic process organized in three major steps: (I) neural stem cell proliferation, (II) neuronal migration towards the cortical plate, and (III) post-migratory organization (for further review see[3-5]). Alterations in any of these complex developmental stages can be responsible for the development of neurodevelopmental disorders (NDDs), which are a heterogeneous group of disorders, affecting more than 3% of children worldwide[6,7]. Disorders belonging to this group present with various clinical features, that include amongst others neurodevelopmental delay, intellectual disability, autism, epilepsy and malformations of cortical development[8,9].

Although the list of causes of NDDs is long, and covers a wide spectrum ranging from various environmental exposures including infections, to injuries to the central nervous system such as perinatal asphyxia and traumata, many NDDs have a genetically encoded cause. These can vary from chromosomal aneuploidies, microdeletion and duplication syndromes, polygenic and oligogenic causes, as well as monogenetic diseases following all possible modes of Mendelian inheritance[7,10]. Establishing a genetic diagnosis is crucial, as this allows counseling about the disease and its prognosis, offers reproductive choices to parents and family members, and increasingly more often leads to changes in clinical disease management, enabling tailored care and personalized medicine[7]. Potential disease-causing genetic alterations can be

detected using SNP-arrays that allow the detection of chromosomal imbalances, by targeted analysis of genes that are high in the differential diagnosis, and by whole exome sequencing (WES), an agnostic method of sequencing all protein-coding exons in a genome, that enables the detection of disease-causing variants in virtually all protein-coding genes. Although the implementation of WES in the diagnostic process improved the diagnostic yield of Mendelian disorders to ~25-30%[11] and has greatly accelerated disease gene discovery[10,12-17], still many cases of NDDs remain genetically unexplained, which is a major problem in the field of human genetics. This holds true even for cases where multiple affected individuals are found in the same family, or other environmental causes have been excluded, strongly hinting at a genetic cause.

A particular relevant group of NDDs where this applies to are developmental and epileptic encephalopathies (DEEs), which are a large group of individually rare, severe genetic disorders, presenting with intractable seizures, severe to profound developmental impairment (with an IQ usually below 40) and a wide range of comorbidities, including psychiatric, sleep, gastrointestinal and gait disorders[18]. Patients are typically empirically treated with multiple anti-seizure drugs, which are associated with substantial toxicity[19]. Studies of DEEs limited to onset under age 18 months found a combined incidence of 1:2,000 births[20,21]. Currently, according to Online Mendelian Inheritance in Man (OMIM) (https://omim.org/), more than 300 genes are known to cause DEEs, many of which have been identified in the last decade using WES[12-15,22]. Despite each of these separate disease entities being extremely rare[23], DEEs as a group are the cause of significant morbidity affecting quality of life[24-26] and mortality in childhood, with approximately 20% dying by 20 years. If newborn onset, 53% of infants die by 2 years of age[27,28]. Furthermore, DEEs pose a significant economic burden on communities with most patients being dependent on daily care requiring lifelong support, thereby accounting for example for a major portion of the estimated $12.5 billion epilepsy cost to Australia[29-31], with a similar impact per capital in EU countries[32-34]. Progress has been made with personalized medicine and tailored therapies, but to fully exploit these promises, accurate and early diagnosis of all patients is needed. It thus remains crucial to determine the genetic cause of DEEs, as this is the requisite first step towards development of tailored treatments that specifically target the disease cause and not just address the symptoms[35-40].

Given the genetic heterogeneity of DEEs, state-of-the-art genetic diagnostic tools that are applied include gene-panel based and whole exome sequencing (WES) and, less frequently, whole genome sequencing (WGS). Previous works[27,41-45] have shown

that the diagnostic yield in DEE is at best between 30-55% even when using WGS, leaving currently more than half of affected individuals without a genetic diagnosis, and thereby excluding them from personalized treatments that require the genetic cause to be known[37,39,40]. A lack of a genetic diagnosis also makes patients ineligible for precision medicine trials, excludes their parents from genetically informed reproductive choices, and prevents accurate estimation of epilepsy causes. It is thus of utmost importance to increase the diagnostic yield amongst these rare disease patient groups, to increase options for treatments and tailored care, but also to better understand the natural histories of defined genetic DEE entities and thereby improve prognosis prediction and counseling.

Currently, even though WGS is increasingly being used in DEE diagnostics, analysis of clinical WGS remains exome-focused, as protein-coding exons remain by far the most knowledge-dense areas of our genome despite only comprising ~2% of all our genetic information. Increasing evidence supports that genomic alterations outside coding genes, located within the 98% of non-coding genome can cause genetic disease[46-50] and thus likely explain at least part of the missing heritability (e.g. the lack of finding a disease-causing genetic variant despite the high suspicion of a genetic disorder). This includes 1) deep intronic variants affecting mRNA splicing; 2) single nucleotide variants (SNVs), insertions and deletions (indels), copy number (CNV) and structural variants (SVs) disturbing the regulatory landscape of protein-coding genes; 3) alterations affecting the expression and function of long non-coding RNAs (lncRNAs) that can either be directly implicated in disease or indirectly affect regulation of disease implicated genes[51,52]; and 4) epigenetic alterations such as aberrant DNA-methylation leading to gene expression perturbation. However, none of these non-coding mechanisms of genetic disease are currently routinely assessed in DEE diagnostics, or in the diagnostics of other NDDs, and this in fact might explain at least part of the missing heritability that is observed in the clinical genetics field.

The hypothesis that disease-causing variants might be located in non-coding regions of the genome, in particular those involved in regulation of protein-coding genes, 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[53], therefore potentially in regions involved in transcriptional regulation. Second, the last decade has witnessed enormous progress in our understanding of mechanisms involved in gene regulation that find their origin in the non-coding genome, and it has become clear that aberrant gene regulation can cause a variety of genetic disorders[47,54,55]. Key

elements in the non-coding genome such as promoters, insulators, and enhancers, the latter also referred to as non-coding regulatory elements (NCREs), ensure that genes are turned on or off at the right moment in time. When this tight spatio-temporal regulation is disturbed, gene expression can be affected, resulting in a genetic disorder. Although only very few large-scale genetic studies have investigated the role of the non-coding genome in genetic disorders[56-58] it is clear from the number of excellent studies that have recently been published[59-69], 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 this phenotypic variability could likely be influenced by genetic variations outside of coding genes influencing gene expression[55,70-73]. Therefore, it thus remains crucial to gain more detailed information on the functional relevance of the non-coding genome and its variants from a basic science point of view, which promises translational progress leading to improved diagnostics for NDDs and new avenues leading to future therapy development. In the following paragraphs, I will discuss the role of the non-coding genome in gene regulation (with a particular focus on enhancers), provide examples of non-coding alterations causing genetic diseases, and review recently developed technologies and computational approaches that facilitate current and future investigations of the non-coding genome.

The non-coding genome and non-coding regulatory elements

According to the central dogma of molecular biology, there are three main processes taking place in the cell: replication of the genetic information, transcription of DNA into RNA, and translation of the RNA molecule into the final functional product, the protein[74]. As one of the main surprises from the Human Genome Project, it is now well established that more than 98% of the human genome does not encode proteins[75]. These non-protein-coding regions were initially considered as 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[76,77]. However, several structural elements of non-coding DNA have now been described that regulate gene expression, by determining the 3D genomic organization critical for correct gene regulation. 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 in order to acquire a broad range of phenotypes, while maintaining the same genotype. This is achieved by a tight spatiotemporal 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, unlike coding DNA, needs to be seen from a non-linear, 3D perspective where regulatory elements need to interact with target genes on long distances.

Chromatin organization

Genomic organization comprises efficient DNA packaging in the limited space of the nucleus while allowing for DNA replication and gene expression. First, nucleosomes are formed, in which 147 base pairs of DNA are wrapped around 8 histone proteins linked to each other by DNA stretches of various lengths. This beads-on-a-string organization forms the basis of a 10-nanometer chromatin fiber that is typical of open chromatin, also known as euchromatin. This differs from tightly packaged heterochromatin, where multiple histones wrap into a 30-nanometer fiber consisting of nucleosome arrays in their most compact form. As a result of this, chromatin is organized into active and inactive compartments that are either open or condensed and which vary in size between 1 to 10 megabases (Mb). Inactive compartments are often found in association with the nuclear lamina, whereas active compartments are more likely to be found in other regions of the nucleus[78]. Regulatory elements such as enhancers and promoters and actively transcribed genes are located in open-chromatin regions, so that they are accessible for the transcriptional machinery. Various post-translational epigenetic modifications of histones put in place by chromatin modifying enzymes can alter the accessibility of chromatin and can thereby influence how chromatin is packaged and whether it is more or less likely to be active. For example, histone acetylation results in increased chromatin accessibility and makes chromatin more available for the binding of regulatory proteins, such as transcription factors (TFs). Many studies focused on a wide variety of histone modifications[79,80], and have led to a draft of a histone code, where various histone modifications are indicative of the functional role that the chromatin has at those places that are modified. 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). Insulators are responsible for organizing chromatin at a sub-compartment level. They are often bound by the TF CTCF (also known as 11-zinc finger protein or CCCTC-binding factor)[81] and establish the boundaries of so-called topologically associating domains (TADs). TADs are usually <1 Mb in size and delineate those regions of our chromosomes in which sequences interact preferentially with each other rather than with

elements in other regions of the genome. The prevailing model is that these TADs are formed by the dimerization of two CTCF molecules binding the boundaries of a TAD and are stabilized by the interaction with the ring-shaped cohesin complex through a process called loop extrusion[82-84]. Inside TADs, smaller DNA loops are formed to allow enhancer-promoter interactions and thus regulation of transcription[82,85]. 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 1**)[86,87].

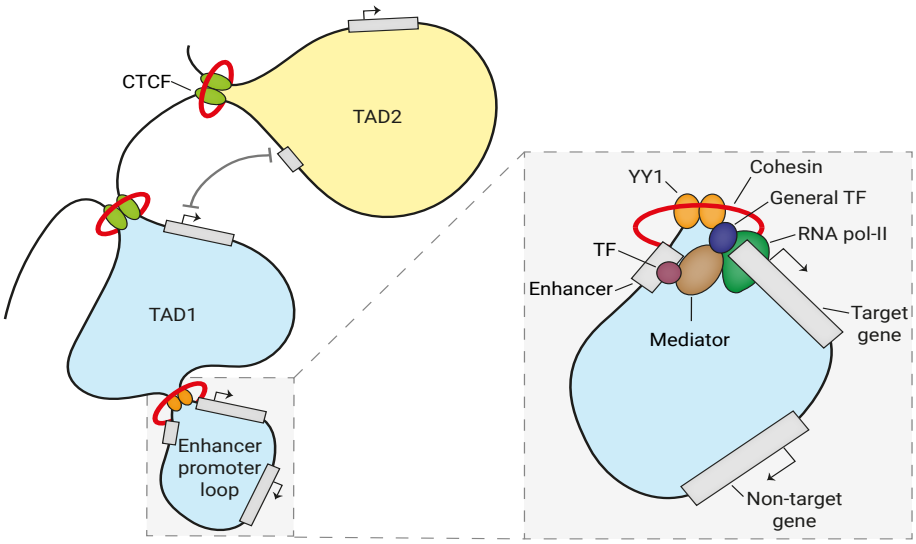


Figure 1: **Regulatory enhancer-promoter interactions are restricted within the TAD region**. The genome (here represented as a black line) is tightly packaged and organized in topologically associating domains (TADs) established by the binding of CTCF to insulator elements, followed by dimerization 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. 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. Mediator regulates the phosphorylation of the RNA Pol II in order to release it from the promoter and start transcription.

Enhancers and their role in gene regulation

Correct spatiotemporal gene expression is ensured by the activity of promoters and enhancers, two crucial classes of *cis*-regulatory elements. Promoters are located around the transcriptional start site (TSS) of genes and are essential to initiate transcription. Enhancers are positive regulators of transcription[88], whose location rela-

tive to the TSS of the gene they control varies from adjacent to the promoter, to many kilobases (kb) upstream or downstream of it, and can even be located in introns, also of other genes. Moreover, 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[60]. In addition, one enhancer can regulate several genes, and at the same time each gene can be regulated by multiple enhancers. This creates redundancy in the system that results in phenotypic robustness, and probably gives advantages during evolution[89]. Therefore, the positions, identities, and arrangements of enhancers ultimately determine the time and place that each gene is transcribed. On a mechanistic level, enhancers directly influence the recruitment of the transcriptional machinery to the TSS of genes[90,91]. Crucial for this long-range control of gene expression by enhancers is the formation of enhancer-promoter loops which preferentially occur within the neighborhood of a TAD, by DNA bending. The general TFs and the RNA polymerase II bind to the promoter sequence, whereas the enhancer sequences are bound by TFs, which orchestrate the rate of transcription initiation.

Transcription factors

TFs are proteins that regulate gene transcription. TFs have binding domains that allow them to bind to specific DNA sequences. Enhancers include TF binding sites (TFBSs) that typically consist of DNA motifs found at multiple sites in the genome, but that are not necessarily all equally likely to be bound by the recognizing TF[92]. To provide higher than background activity, homotypic or heterotypic dimerization of transcription regulators increases their DNA binding affinity and specificity (Funnell and Crossley, 2012). TF binding itself can also be influenced by DNA methylation, which is established by DNA methyltransferases (DNMTs)[93]. Moreover, if TF binding prevents the DNA from re-wrapping around the nucleosome, it increases the likelihood that a second transcription regulator binds the DNA, increasing the cooperative effect to the extent of displacing the histone core of the nucleosome[94,95]. Multiple TFs have been found to bind cooperatively in TF binding site "hotspots"[96], later called stretch enhancers[97] or super-enhancers (SEs)[98]. The latter are described as long regions with an increased density of enhancer elements characterized by a strong enrichment of H3K27ac, and of TFs and Mediator binding[98,99]. On the one hand, several studies suggest that SEs represent a novel class of NCREs that maintain, define, and control mammalian cell identity and whose transcriptional regulatory output is larger than that of the individual enhancer constituents[100-102]. On the other hand, an increasing number of 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[103,104]. Therefore, the debate on whether SE is a new class of NCREs or whether they simply reflect a clustering of normal NCREs within proximity remains to be settled.

What is clear from the above, is that our knowledge of complex gene regulatory mechanisms has increased dramatically over the last decade and has provided insights into many sophisticated processes that need to occur correctly for development to proceed normally. Aberrations in many of the steps described above can result in genetic disorders. For example, in recent years a large number of disorders have been described that are caused by mutations in chromatin modifying enzymes or proteins involved in 3D chromatin regulation[105-108]. Given the complexity of gene regulation and the many contributing factors acting at different stages of this process, it seems likely that many more will be discovered in the near future.

Enhancer in the context of genetic disease

As discussed, an increasing number of studies suggests that a high fraction of causative mutations in neurodevelopmental disorders such as intellectual disability and autism, belong to pathways of transcriptional regulation and chromatin remodeling[105,109,110]. Besides mutations in *trans-acting* factors such as TFs or chromatin modifiers, also mutations of NCREs *in cis* have been proven to be causative of disease in an increasing number of cases. Since 1983, when an enhanceropathy causing the mis-regulation of the β-globin gene in patients with β-thalassemia was reported [111], many disease-causing enhancer alterations have been reported, examples include phenotypes, as diverse as ranging from oncology to limb malformations[49,60,112-120]. This wide range of NCRE alterations can vary from point mutations affecting the binding of crucial TFs, deletions or duplications of NCRE sequences, shuffling of the genomic location of NCREs affecting their function (e.g. enhancer adoption), or alterations in the global chromatin landscape disrupting borders of TADs, just to mention a few. In the following, I will discuss some of these disorders, mainly those affecting the brain, caused by alterations to NCREs (**Table 1**).

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*	[121]
Aniridia	Point mutation	*PAX6*	[122]
Polymicrogiria in the Sylvian fissure	Deletion	*GPR56*	[123]
Parkinson's disease	SNP	*SNCA*	[124]
Schizophrenia	Tandem duplications	*VIPR2*	[125]
Adult-onset demyelinating leukodystrophy	Deletion of TAD boundary and deletions	*LMNB1*	[126,127]
Intellectual disability	CNV	*ARX*	[128]

A classic example is a pre-axial polydactyly caused by alterations of the zone of polarizing activity regulatory sequence (ZRS), a long-distance enhancer that regulates Sonic hedgehog (*SHH*) expression in the embryonic limb[59,129]. Next to point mutations, also copy number variations (CNVs) such as duplications[130], and insertions[131] in this region have all been shown to cause polydactyly phenotypes, illustrating the wide range of alterations that can affect enhancer function and thereby result in a phenotype. 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) was identified, located 460 kb upstream of the *SHH* gene in a patient with an identical phenotype[121]. 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, mutations in *SIX3* can lead to holoprosencephaly[132]. A disease-causing enhancer mutation is also found in the congenital eye malformation aniridia, that is often caused by haploinsufficiency of the TF PAX6, that also plays crucial roles in neural stem cells. A point mutation in the *PAX6* eye-enhancer was found to disrupt PAX6 binding, thereby affecting *PAX6* expression[122]. 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[123]. *GPR5* mutation 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.

Besides influencing disorders presenting early in life, diseases emerging later in life, such as neurodegenerative disorders and schizophrenia, are increasingly linked to NCREs 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[124] and in another study, an Alzheimer's disease-risk variant overlapped with the microglia-specific BIN1 enhancer[133]. 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[125]. Also, CNVs overlapping with NCREs in other schizophrenia related genes might be implicated in the disease pathogenesis, influencing the disease vulnerability[134].

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[135]. Besides changing gene dosage, CNVs can also change the dosage and position of NCREs, as well as the higher-order chromatin organization of a locus[47,136]. 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[127]. 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[126]. Other rare inherited structural variants in *cis*-regulatory elements might influence the risk for children of developing autism spectrum disorders (ASDs), depending on the parental origin of the structural variant[137]. Another study on autism using WGS on more than 2000 individuals found that probands carry more gene-disruptive CNVs and SNVs resulting in severe missense mutations and mapping to predicted fetal brain promoters and embryonic stem cell enhancers[138]. In addition, CNVs covering the regulatory elements of the *ARX* gene might cause an intellectual disability phenotype[128], and rare non-coding CNVs near previously known epilepsy genes were enriched in a cohort of 198 individuals affected with epilepsy compared to controls[139]. Similar findings are reported for multiple system atrophy[140] and non-coding variants

might influence the expression of *GLUT1* causing epilepsy[141].

Two large-scale analyses focused on NCREs 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 from the Deciphering Developmental Disorders (DDD) study and their parents[58]. The three classes of regulatory elements that they assessed consisted of 4,307 highly evolutionarily conserved non-coding elements[142], 595 experimentally validated enhancers[143], and 1,237 putative heart enhancers[144], 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, whereas in experimentally validated enhancers, heart enhancers, and intronic controls *de novo* variants were not enriched. When focusing only on conserved non-coding elements that had evidence of activity in the 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 fetal 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 on neurodevelopmental disorders. Doan and colleagues performed a similar targeted sequencing approach assessing so-called human accelerated regions (HARs)[56]. HARs are conserved regions with elevated divergence in humans and this 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 2,100 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. They then analyzed 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 (AF <0.5%) bi-allelic HAR alleles (43% excess compared to unaffected, p=0.008), and this enrichment further increased when only taking HARs into consideration that were likely active as regulatory elements in brain. Using massively parallel reporter assays (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 HARs is promising to solve currently genetically unexplained disease cases.

Together, these examples support the increasing relevance of understanding NCREs and their location in the non-coding genome from a disease point of view. In the next sections, I will discuss technologies that are used to annotate and identify these non-coding regulatory elements.

Genome-wide identification of putative enhancers

As introduced, transcriptional enhancers were first described as DNA sequences that are able to enhance gene expression on an episomal plasmid (e.g. a non-integrating, extra chromosomal circular DNA), irrespective of their location and orientation relative to the TSS[88,145]; 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. The most widely applied experimental techniques for genome-wide identification of putative enhancers at the endogenous genomic locus today do not rely directly on this functional property, but rather on features that distinguish enhancers from non-regulatory regions at the chromatin level. Indeed, enhancers are bound by TFs and transcription coactivators and are located in open chromatin regions that are 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 gene regulation[146]. This led to the development of high-throughput functional screenings, overall known as MPRAs that quantify the enhancer activity of millions of sequences. In the next paragraphs, I will discuss the most widely used techniques to identify putative regulatory regions (**Figure 2, Table 2**), and in the following section, I will focus on high-throughput functional screens.

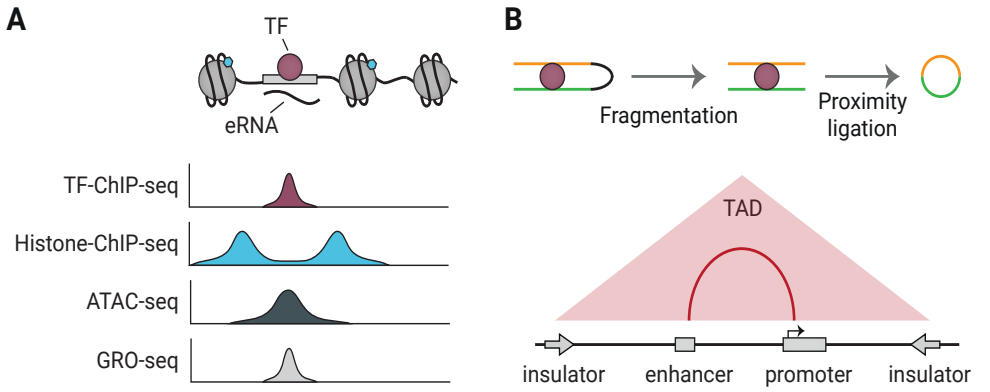


Figure 2: **Overview of the main techniques currently used to identify putative enhancer sequences and their interacting genes**. (**A**) Schematic drawing on a TF-bound enhancer, located in nucleosome depleted DNA from which eRNA is transcribed. Below are representative genome browser tracks shown, 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.

Table 2: Methods for the identification of non-coding regulatory elements (NCRE).

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 or 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 located 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 two 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.

Chromatin Immunoprecipitation

Chromatin immunoprecipitation (ChIP) was first introduced more than 30 years ago to study protein-DNA interactions[147] 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 NGS technologies allowed genome-wide mapping of these protein-DNA binding sites (ChIP-seq)[148,149]. 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[150] and H3K27ac[151], and transcriptional coactivators, such as the histone acetyltransferase p300/CBP[152] and Mediator[98]. However, neither the binding of a TF nor the presence of histone modifications provides 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[153], and not all H3K27ac marked DNA sequences show enhancer activity when functionality tested[154]. Several studies have used ChIP-seq for histone modifications to predict enhancers during human brain development[155,156] and in the adult brain[157-160], and some have made direct comparisons to brains from other primates, providing important insights in the evolution of humans[155,157].

Identification of open chromatin regions

As abovementioned, *cis*-regulatory sequences like enhancers are enriched in chromatin regions depleted from nucleosomes[161], as nucleosomes would impede TF binding[162]. 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[163]. 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[164]. 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 cross-linking reversal steps and is therefore less prone to loss of material[165]. However, other regulatory elements such as insulators or promoters are also located in accessible chromatin[161]. Therefore, ATAC-seq should be used in combination with other techniques that are more selective for enhancers. Moreover, these methods qualitatively identify putative enhancers and do not allow the quantification of their activity; indeed, also inactive enhancers can be in open-chromatin regions[146,166]. 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 the first glimpse of open chromatin dynamics during fetal brain development[167].

eRNA

Transcription of enhancer sequences was first reported in the early nineties in the Locus Control Region (LCR) of the β-globin gene cluster[168-170], 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[171], followed by sequencing of nascent RNA (GRO-seq, CAGE) in different cell types[172-175]. Enhancer RNAs are generally bidirectionally transcribed and not polyadenylated[171] but reports of unidirectional transcription and polyadenylation of eRNAs exist[176]. Enhancer transcription was shown to correlate with the presence of other enhancer marks such as histone tail post-translational modifications and p300/CBP and RNApolII binding[172-174], but whether their expression is a cause, or a consequence of gene transcription is still debated[177]. If eRNA transcription has a direct functional role and is not just noise due to the recruitment of RNApolII, 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[178,179]. However, even if eRNA presence correlates with enhancer activity at some loci, it seems that it is neither required nor sufficient

in all instances[146]. For example, a recent study assessing eRNAs in the 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[180]. The FANTOM project has found a similar small number of eRNAs in the brain, although the majority of those are not overlapping with those from Yao and colleagues[175]. 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.

Long-distance chromatin interactions

All methods described until now identify putative enhancers but understanding which genes they regulate remains a challenge. Indeed, despite often regulating nearby genes, enhancers 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, as abovementioned, enhancers and promoters need to be brought in close proximity in order for transcription to take place. In the past ~20 years several techniques have been developed to address this question (reviewed in[181,182]). 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[183]. 3C was developed to study whether two known regions are interacting with each other and is thus described as a "one vs one" method[181]. Further advances in 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[184,185], while 5C, "many vs many", investigates all contacts that are happening in a specific locus[186]. Finally, high-throughput contact identification became possible with Hi-C[187]. Hi-C allows the identification of genome-wide interactions thanks to the introduction of biotin-labeled 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, TADs

can be defined. To investigate all the genome-wide interactions involving a specific protein of interest, HiChIP was developed, by introducing a chromatin immunoprecipitation step[188]. This method has the advantage that it requires less input material and less sequencing reads. A conceptually similar method is called PLAC-seq[189].

Despite their capacity to identify enhancer-promoter interactions and thereby pieces of chromatin with putative regulatory roles, chromatin conformation techniques have the disadvantage of not directly measuring functional 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. Finally, the spatial resolution at which interactions can be determined is heavily influenced by the sequencing depth of Hi-C experiments. Hence, there remains a need for more functional tests to validate the regulatory activity of the identified interactions. A recent study has generated Hi-C maps from cortical plate (CP) and germinal zone (GZ) of the human fetal brain and from gestational weeks 17-18 of human brain development, a critical time period for cortex development[190], permitting the large-scale annotation of previously uncharacterized regulatory interactions relevant to the evolution of human cognition and disease. 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.

High-throughput functional identification of enhancers

As previously highlighted, most of the commonly used techniques to identify regulatory elements are merely predictive, and do not directly measure enhancer activity. Although there is no doubt that techniques such as ChIP-seq, open chromatin mapping and expression analysis have been of tremendous use to globally characterize the gene regulatory landscape of the non-coding genome, it is still clear that there is a need for improved techniques. In many instances, the identified putative enhancer sequences fail to perform as enhancers in functional validation experiments, giving rise to false positive enhancer predictions[191]; see[192] 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, for example, the assignment of functional roles of nucleotide variants found in the human population. Finally, many of the currently used techniques take into consideration previ-

ously identified knowledge on associations between epigenetic marks and putative enhancers. This potentially excludes other regions of the genome to be functionally assessed as they lack these associations but might nevertheless be functionally relevant[193]. Direct high-throughput functional tests of enhancer activity, such as massively-parallel reporter assays and CRISPR-Cas9 based screens have the potential to address these shortcomings (**Figure 3**), as I will explain in this section.

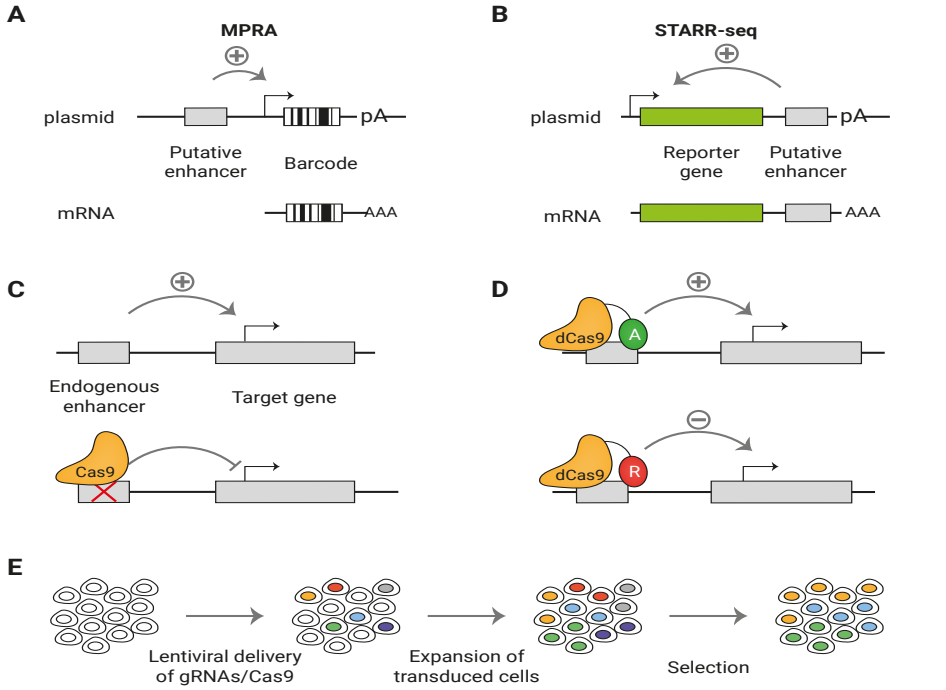


Figure 3: **Methods for functional identification of enhancers**. (**A**) Massively parallel reporter assays (MPRA) to test enhancer activity in an episomal setup. The putative enhancer sequence is cloned upstream of a minimal promoter that drives the expression of a reporter gene and a unique barcode. (**B**) With STARR-seq the putative enhancer sequence is cloned downstream of the reporter gene and upstream of the polyA signal. When the enhancer sequence is active, it 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.

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[88]. 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[98,166,194-199]. 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[166]. STARR-seq takes advantage of the fact that enhancers act in a position-independent fashion. 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[200] and autonomous promoter activity[201].

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 ChIP[154], ATAC[202] or other techniques[203,204]. In our laboratory's application, we have combined ChIP with STARR-seq to generate genome-wide enhancer activity maps in various types of human embryonic stem cells[154].

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[205]. 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 assessed by CRISPR-based deletions[154], or when a set of enhancers is assessed on both plasmids and integrated at multiple genomic locations[206]. 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[207].

One way to overcome these possible limitations of transgenic reporter assays, is to use the recently developed CRISPR-Cas9 system to manipulate NCREs at the endogenous chromatin context. Cas9 is an RNA-guided DNA endonuclease that is able to induce double-strand breaks that, in the absence of a donor template for homology directed repair, are repaired by the error-prone non-homologous end-joining (NHEJ). The enhancer sequence can thus be deleted, by targeting Cas9 with guide-RNAs (gRNAs) flanking the enhancer sequence or be mutated by the introduction of indels via NHEJ, allowing to test the effect of the enhancer ablation on gene expression in the endogenous chromatin environment. Whereas this approach can be used to study a selected enhancer of interest, as our laboratory did studying enhancers involved in pluripotency of human embryonic stem cells[154], it can also be used in high-throughput screenings with large libraries of gRNAs that are introduced in cells expressing Cas9. 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 disadvantage for the cells, and therefore gRNAs targeting relevant functional NCREs will be depleted over time. By comparing sequencing reads after and prior to the selection, it is possible to determine which gRNAs are lost over time, and as the targets of the gRNAs are known, the relevant NCRE 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 tilling screen, functional domain within these enhancer sequences were precisely mapped[208]. 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[209]. Other studies investigated putative enhancers involved in oncogene induced senescence[210], regulation of the *HPRT* gene involved in Lesch-Nyhan syndrome[211]

and regulation of the *POU5F1* gene in embryonic stem cells[212,213] , amongst others[214-217]. Besides genome engineering, CRISPR-Cas9 can also be applied to edit the epigenome, and also this can be coupled to high-throughput screening. Indeed, by fusing a catalytically dead Cas9 (dCas9), that lacks endonuclease activity, to various functional domains it is possible to alter the status of a NCRE forcing its activation or inactivation, referred to as CRISPRa and CRISPRi, respectively. Functional additions to dCas9 leading to NCRE activation include transcription activating domains such as multiple repeats of the herpes simplex VP16 activation domain (VP64)[218,219], the nuclear factor-κB (NF-κB) trans-activating subunit activation domain (p65) and human heat-shock factor 1 (HSF1)[220], the ten-eleven translocation methylcytosine dioxygenase 1 (TET1)[221], and the p300 acetyltransferase[222]. Opposingly, transcription repressive domains that can be used to silence NCREs include Krüppel-associated box (KRAB) domain[223,224], four concatenated mSin3 domains (SID4X)[225], cytosine-5-methyltransferase 3A (DNMT3A)[226], Histone deacetylase 3 (HDAC3)[227], and the lysine-specific histone demethylase 1A (KDM1A), called dCas9-LSD1[228]. Several of these dCas9 fusion have been used to activate or repress NCREs, and a number of studies have used them in high-throughput screening approaches, most of which focused on NCRE repression[229-232] but some included also NCRE activation[233,234]. It seems only a matter of time till more similar studies editing NCREs 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.

Computational enhancer prediction

As it has become clear from the discussion above, currently used enhancer prediction techniques heavily depend on computational data analysis, most often involving the analysis of next-generation sequencing data. Besides the direct use of computational analysis for biological data processing, more and more efforts are being undertaken to use computational power to predict functional NCREs *in silico*. We broadly summarized these methods applied for genome-wide enhancer prediction in three topics, focusing on 1) comparative genomics and evolutionary conservation, 2) clustering of motifs and epigenome features and machine learning approaches, and 3) techniques that deal with the processing of data obtained in functional genomic

screens. Here, we mainly focus on the advantages and disadvantages of some of these methods and highlight several resources that can be used to obtain information on genomic enhancer locations. We refer those readers who are interested in a more detailed discussion on the various options for machine learning and other prediction tools to a number of excellent recent reviews[235-238].

Comparative genomics and evolution in enhancer prediction

Functional sequences are expected to be more conserved compared to DNA stretches that are not expected to have any role, as changing of nucleotide composition is expected to alter function. This characteristic is exploited by comparative genomics approaches that aim to identify enhancers by looking at the most conserved sequences across different species. This was one of the first computational approaches to identify NCREs[239,240]. Nevertheless, different studies showed how some NCREs are strongly conserved, while others are rapidly changing also in closely related-species, rendering the solely use of comparative genomics techniques insufficient. For example, Arnold and colleagues[241] showed by STARR-seq of different Drosophila species how, in the majority of the cases, enhancer function is conserved across species, and the highly conserved enhancers are thought to play an important role during key processes such as embryonic development, and especially in the developing nervous system[239]. However, several other studies suggest that a portion of enhancers undergo rapid evolution, and that this might be a crucial driver of human evolution[242-245]. A subset of active enhancers in human embryonic stem cells is even enriched in human specific transposable elements, and those functional regions would be missed if one were to use only conservation as a key feature for enhancer selection[154]. Therefore, although sometimes useful, evolutionary conservation alone for the discovery of NCRE is not recommended as a sole criterion, as it would miss all the newly evolved enhancers. Another extreme example of this are so-called ultraconserved elements, stretches of DNA sequences that are more than 200 bp long and that are 100% identical in multiple species, such as human, rat and mouse[246]. Whereas some of these sequences were shown to play a role as enhancers[247,248], others can be removed from the genome without an obvious phenotype[249], and it is speculated that some of these sequences might contribute to genome stability[250]. Enhancers can also be identified by the presence of specific TFBS, as TF binding is a key characteristic of these regulatory sequences. Indeed, combining conservation with TFBS site discovery can further increase the predictive power of comparative genomic approaches. However, even this does not guarantee enhancer identification, as during evolution novel TFBS can appear which execute similar functions as the ones in the ancestry sequence[251].

Enhancer prediction algorithms

Several types of enhancer predicting algorithms have been developed for integrating multiple types of data, such as TF motifs, ChIP-seq, DNase-seq, ATAC-seq, and P300 binding data sets for enhancer prediction by using clustering and machine learning approaches, which include supervised and non-supervised algorithms. Supervised machine learning algorithms rely on high-confidence positive and negative training sets (e.g. known- and non-enhancers) to build models that can maximize the differentiation between enhancer and non-enhancer sets. Examples of supervised algorithms that can identify enhancers include CSI-ANN[252], ChromaGenSVM[253], RFECS[254], EnhancerFinder[255], DEEP[237], DELTA[256], PEDLA[257], REPTILE[258], eHMM[259], PREPRINT[260], CoRE-ATAC[261], PEREGRINE[262], EnhancerPred[263], GenoSTAN[264], CRUP[265], DBN[266] and ReFeaFi[267]. Unlike supervised methods, unsupervised methods do not require any training data and can identify hidden and unknown patterns directly from data. Unsupervised algorithms such as Segway[268] and ChromHMM[269] integrate multiple types of epigenome data to define chromatin segmentation that can be used to assign functional roles for various parts of chromatin. Also other machine learning models based on convolutional neural networks have been developed which identify and classify enhancers based on their strength such as iEnhancer-2L[270], iEnhancer-EL[271], iEnhancer-ECNN[272], iEnhancer-EBLSTM[273], iEnhancer-GAN[274] and iEnhancer-RD[275].

One of the main problems of all these prediction programs is that we still lack a detailed understanding of the underlying regulatory code in the non-coding genome. Despite all the advances made over the last decade, we are yet to pinpoint a feature that can identify enhancers (and their activity) in all cell types. As most programs rely on previously generated training sets or on static features such as DNA sequence motifs which by their own do not necessarily predict enhancers in each instance, it is more than logical that despite the large number of efforts that are undertaken, enhancer prediction programs are far from perfect. For example, although chromatin segmentation is very intuitive and access to these segments can be easily obtained from the UCSC genome browser, it is rather worrying that a recent study testing more than 2000 sequences classified as enhancers using these methods did not detect regulatory activity in 74% of the sequences tested[191]. Also, the overlap between individual predictions from various programs is rather poor[237]. Quite intuitively, programs that take into account multiple features for enhancer prediction tend to perform better[255,258,276]. Therefore, it is tempting to speculate that future large-scale meta-analyses of all currently available enhancer data might enable the further

fine-tuning of enhancer prediction tools in the near future.

Another area where further progress needs to be achieved is the prediction of enhancer-promoter (E-P) interactions that can be used to assign NCREs to their target genes. Recent advances in high-throughput experimental technologies such as HiC[187], ChIA-PET[277] and promoter capture Hi-C[278], have been developed to assess E-P interactions. However, the genomic resolutions to study E-P interactions are often low and performing these approaches are technically challenging in some tissues and cell types[279]. An alternative to these experimental approaches are computational methods that predict E-P interactions based on either epigenomic data-based methods or DNA sequence-based characteristics[279,280]. To the former group belong algorithms such as ELMER 2 and InTAD. ELMER 2 computes the correlation between the enhancer and target genes by combining both DNA methylation and gene expression data derived from the same dataset[281], but is limited by the fact that correlations are restricted to the closest neighboring gene, which does not necessarily present the real biological relevant target gene[282]. InTAD is a tool to detect genes located upstream and downstream of the enhancer in the same TAD boundary and it can support different types of data as input. The TAD information comes from available Hi-C datasets[283], but the currently available ones have a low resolution and, until now, have included only a limited number of cell types. Other epigenomic data-based machine learning methods are RIPPLE[284], TargetFinder[285], EpiTensor[286], JEME[287] and FOCS[288], which were trained using combinations of chromatin accessibility data, histone modifications, and gene expression data to predict E-P interactions. DNA sequence-based methods such as PEP[289], EPIVAN[280], SEPT[279], SPEID[290] and EP2vec[291], identify E-P interactions by training different statistical models based on DNA sequence features of given enhancers and promoters in one cell type, and allow to predict possible E-P interactions in other cell types. Although these methods can provide insights into putative E-P interactions, there are still a number of limitations. These algorithms can work well when both training and test data come from the same experiment, but perform worse when this is not the case. In addition, many of the required input data for these tools are only available for a limited number of cell lines or tissues, limiting their utility. One approach would be to integrate all available data from different tissues or cell lines to improve the above limitations, but the redundancy of data among cell types, how to integrate experimental data from different biological experiments and DNA sequence features into a single prediction model, or how to optimize feature distribution across various cell types, are still major challenges. Recently, using such a holistic approach, an Activity-by-Contact

(ABC) model was established that allows to identify both enhancers and their target genes based on chromatin accessibility (ATAC-seq), histone modifications (H3K-27ac), and chromatin conformation (HiC)[292]. This method first defines enhancers based on ATAC-seq and H3K27ac ChIP-seq data, and subsequently predicts E-P interactions based on HiC data. But given that the presently available HiC data have low resolution and have so far only been generated for a limited number of cell types, also such a combined model remains with its limitations. Furthermore, despite that ATAC-seq and H3K27ac ChIP-seq data can predict putative enhancers, defining those that show enhancer activity is still an additional challenge. Therefore, to fully identify enhancer-promoter interactions, it will be important to come up with novel experimental procedures that will enable us to directly test the biological relevance of enhancer-promoter predictions.

Finally, several programs have been developed that aim to predict the functional relevance and possible pathogenicity of variants in NCREs. These include, amongst others, RegulomeDB[293], HaploReg[294], CADD[295], GWAVA[296], GenoCanyon[297], Genomiser[298], and INFERNO[299]. In addition, several databases have been generated, including HEDD[300], DiseaseEnhancer[301] and EnDisease[302] which have collected NCREs that are related to diseases based on the current literature. It will be crucial to further expand and curate these collections of disease-relevant enhancers in the future, and combine them with improved ways of variant interpretation, to fully exploit the relevance of the non-coding genome in disease.

Enhancer databases

As the available information on the non-coding genome is increasing rapidly over the last decade, more and more resources are available online that can help to localize NCREs and to interpret their functional roles. In the next part, I summarize a selection of databases and resources that are currently available and can be used to find NCREs of relevance for brain development (**Table 3**).

Table 3: Enhancer databases

Database	Source
VISTA	http://genome.lbl.gov/vista/index.shtml
EnhancerAtlas 2.0	http://www.enhanceratlas.org/
FANTOM5	http://slidebase.binf.ku.dk/human_enhancers/presets
PsychENCODE	http://development.psychencode.org/
dbSUPER	http://asntech.org/dbsuper/

One of the first resources of experimentally tested NCREs was the Vista enhancer database[143]. Based on comparative genomics, a large selection of putative NCREs from mouse and human was selected and tested in transgenic mouse embryo assay, to determine their *in vivo* enhancer activity, as determined by LacZ expression. The database provides detailed information on the genomic localization of the tested sequences, likely associated genes, and images of transgenic mouse embryos identifying the localization of enhancer driven LacZ expression. Based on the 14/07/2022 update, this database contains 550 tested enhancers active in human forebrain, hindbrain and midbrain. In addition, the VISTA tool portal can be used as a comparative tool and users can submit their own sequences to conduct comparison against multiple species[143], thereby possibly identifying conserved functional NCREs.

EnhancerAtlas 2.0 is a database that has collected putative NCREs based on publicly available data obtained from ChIP-seq for different histone modification, TFs, EP300, and POLII, CAGE and eRNA expression, interaction studies by ChIA-PET (a method that combines 3C with chromatin immunoprecipitation) and chromatin accessibility as determined by FAIRE and DNase-seq. Each putative enhancer is supported by at least three independent high-throughput data sets although the database does not contain any direct functional validations. It contains more than 4,506,217 putative enhancers from 8,573 datasets of 179 human tissue/cells, which through an interactive website can be easily accessed. The 49,925 human fetal brain and 17,103 cerebellum enhancers were predicted using DHS, CAGE, and H3K4me1 and H3K27ac deposition[303].

The FANTOM5 database is the latest version of the FANTOM project, that aims to generate an atlas of mammalian regulatory elements, transcriptomes and long-non-coding RNAs. NCREs are predicted from sequencing data from Cap Analysis of Gene Expression (CAGE) along with RNA-Seq data from multiple tissues and cell types from different developmental time points[175]. In total, the database contains more than 43,000 putative enhancers, of which 639 are expressed in the brain and 376 were found in neuronal stem cells.

Recently, the PsychENCODE consortium has released data from a large multi-center effort trying to map NCREs during brain development[156,159,160]. Using analysis of transcriptome, methylation status, histone modifications and even single cell/nucleus-level (transcriptome) genomic data, NCREs were discovered across multiple brain regions over the entire span of human neurodevelopment and from adult brains, and an integrative data analysis was performed. These data, generated from age- and

often donor-matched samples, represent the most comprehensive multi-platform functional genomic analysis of the developing human brain performed so far. The analysis resulted in 79,056 enhancers identified from adult brains enriched for H3K-27ac and depleted for H3K4me3[160]. In addition, 96,375 enhancers were shown to interact with protein-coding genes during fetal brain development and during *in vitro* differentiation of brain organoids[156]. Of the latter, 46,735 enhancers were active only in the fetal cortex.

Several super-enhancer databases have been generated such as dbSUPER[304], SEA[305], and SEdb[306] providing annotation, genomic coordinates and length of super-enhancers, and their possible associated genes. Among those, dbSUPER is one of the most popular databases with 82,234 super-enhancers from multiple human and mouse cell types. In this database, there are 6,002 and 1,114 super-enhancers detected by H3K27ac enrichment from seven human and three mouse brain tissues and cell types, respectively[304].

Finally, GeneHancer is a database of human enhancers and their inferred target genes[307]. Integrating enhancer predictions from ENCODE, Ensembl, FANTOM and VISTA yielded more than 280 thousand candidate regions that were assigned to their target genes based on co-expression correlation, expression of quantitative trait loci and capture Hi-C.

Although all of these databases can easily be accessed and are user-friendly, it is important to realize when using them that it is still difficult to judge which of the sources provides the user with those sequences that are indeed most likely to be of functional biological relevance. To illustrate this, it is interesting to compare the overlap between predicted enhancers from the various resources. When we compare putative brain enhancers from VISTA, EnhancerAtlas 2.0, FANTOM5, PsychENCODE and dbSUPER, the overlap between the various enhancer predictions is rather limited, even when considering a single nucleotide as the required overlap (**Figure 4A**). The same holds true when assessing the overlap between ChIP-seq peaks for H3K27ac from key adult brain related data sets (**Figure 4B**). Intuitively, one would expect that those NCREs that are found in multiple data sets are more likely to have a true biological role, although this might be an oversimplification, as the brain is a very heterogeneous organ with many different cell types that might differ in NCRE landscape, and technical limitations might still hinder us from detecting all relevant NCREs in each cell type. Ideally, future studies should generate genome-wide functional activity maps of NCREs for all cell types during brain development.

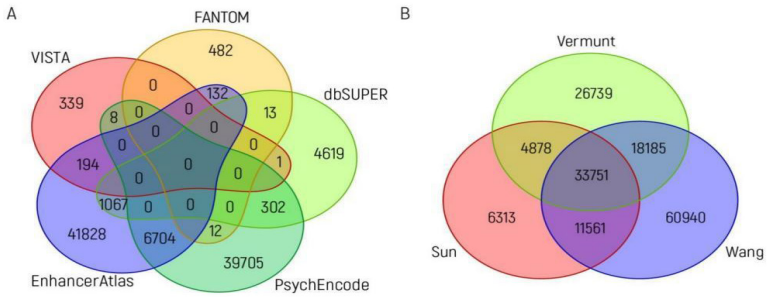


Figure 4: Overlap between brain enhancer databases. (A) Venn diagram showing the overlap between brain enhancers (in genome build hg19) from different databases as collected on 2019: VISTA (n=542)[143], dbSUPER (n=6,002)[304], FANTOM5 (n=639)[175], PsychENCODE (n=46,731; the 46,735 enhancers mentioned in the text are in genome build hg38 and the difference of 4 loci is due to liftover to hg19)[156] and EnhancerAtlas 2.0 (n=49,925)[303]. (**B**) Venn diagram showing the overlap between ChIP-seq peaks for H3K27ac from adult brain, as identified in three studies: Sun (n=56,503)[158], Vermunt (n=83,553)[157] and Wang (n=12,4437)[160]. The intersection between different sources (in the same order as above) was performed using bedops and bedtools. In both graphs, the minimum overlap of a single nucleotide is required.

A challenge that is probably easier to address on paper than in practice in the near future.

The future ahead

As I have discussed in this introduction, our understanding of gene regulation has deepened over the last decade. NCREs have been identified as crucial modifiers of gene expression, and more and more examples of their involvement in human genetic disorders, when mutant, are being reported. In routine clinical practice, genetic analysis has mainly focused on the ~2 percent of the genome that directly encodes for proteins. Most of the non-coding part has been instead neglected, and only recently we could witness a shift of attention towards these sequences. It seems intuitive that, if human evolution resulted in a large and subsequently maintained expansion of the non-coding genome, this should have a functional role, and alterations of these sequences should influence their function and lead to genetic disorders. Given the fact that NDDs are often genetically unexplained despite the routine use of WES, it would be surprising if, in the near future, no genetic alterations of non-coding sequences will be identified in those currently unexplained patients. In order to achieve this, it is crucial to develop novel diagnostic approaches focusing on non-coding regions. Will WGS be useful to find disease causes in those unexplained NDD pa-

tients in a clinical setting? Theoretically yes as it will enable the identification of all detectable variants genome-wide, but our current understanding of genomic variation outside exons severely hampers its routine implementation. As a matter of fact, most studies that have used WGS in a clinical setting, have limited their analysis to those nucleotides covering exons, deep intronic variants not covered in WES and copy number and structural variants[308-311]. Therefore, it remains crucial to gain more detailed information on the functional relevance of NCREs and their variants from a basic science point of view. Although the characterization of epigenomic marks such as histone modifications has shown to be useful to identify functional NCREs, it is clear from the discussion above that there are still some pitfalls, as we still lack the perfect mark to identify relevant and active NCREs. One particular concern is that many studies assume that investigating a single histone modification, such as H3K27ac, gives sufficient evidence to call a region a functional NCRE, but this is certainly an oversimplification. As I have argued above, predicted NCREs should remain classified as putative NCREs till they are functionally validated, or at least predicted in multiple studies ideally using different techniques to obtain a higher level of confidence in their function. In current studies, functional validations of putative NCREs are often performed only for a selected number of regions of interest and results of these few validations are extrapolated to the complete data set generated. Even though this is understandable from a pragmatic experimental point of view, it might be one of the reasons for the broad level of variation between predicted enhancers from different sources. Hence, it is crucial to further develop high-throughput approaches for functional validation studies so that more sequences and their variants can be directly functionally tested, leading to a higher confidence in the data resources. The future application of direct functional assays, such as MPRAs and CRISPR-Cas9 based screens, is expected to further add on to our current understanding, even though also these methods are far from perfect yet. Besides these emerging experimental techniques, it is also crucial to develop novel computational tools that outperform currently available programs for NCRE prediction and disease annotation. Similarly, it is also important to further improve the linking between NCREs and their target genes, going beyond the current resolution of chromatin conformation capture or correlation between activity of putative enhancers and expression of possibly linked genes. Until we will have all these ideal tools widely available, in our opinion the best practice to study the role of the non-coding genome in genetic disorders such as NDDs is to study genetic variation outside exomes in well-defined, exome-negative patients and preferably combine this with a direct readout of gene expression in a disease-relevant tissue. For the functional annotation of the non-cod-

ing variants found in patients, it is essential to use as many sources of information as possible, enabling the highest level of confidence in defining a certain region a regulatory sequence. And last but certainly not least, a detailed clinical phenotyping of patients prior to any genetic investigation remains crucial as it allows the comparison of patients with similar non-coding variants and shared phenotypes. Even in an era where it is cheap to sequence a whole genome, reverse phenotyping of patients remains essential to learn more about the consequences of the genetic variants and to further mature our understanding of the non-coding genome beyond the borders of the exome.

Aim of this thesis

As mentioned, about 50% of individuals affected by neurodevelopmental disorders still do not have a molecular diagnosis. In addition, around 98% of the human genome is non-coding and contains regulatory elements such as enhancers but is currently not assessed in clinical routine diagnostics. Genetic variants in these enhancers might cause disease and explain part of the missing heritability observed in clinical genetics. Investigating these enhancers and their non-coding variants might therefore help to improve molecular diagnosis of currently genetically unexplained patients. However to achieve this, first, we need to obtain an improved functional annotation of non-coding genomic regions. Studies in this thesis focus on two main subjects aiming to reduce missing heritability in neurodevelopmental disorders, by 1) identifying new disease genes and deciphering their regulation by non-coding sequences, and 2) by investigating putative functional enhancers in fetal brain and neural stem cells, which might provide new targets to explain causes of currently unexplained neurodevelopmental disorders.

Part 1

In **chapter 2A**, we focus on the identification of a novel cause of developmental and epileptic encephalopathy, due to a homozygous variant in the UDP-glucose pyrophosphorylase (*UGP2*) gene. This variant causes a start-loss of the shorter UGP2 isoform, which is the only isoform expressed in brain, and therefore causes the brain-specific absence of this essential protein in the brain of patients with epileptic encephalopathy. Using bioinformatics approaches, we identify additional genes, of which the isoform-specific loss of an essential protein is predicted to cause human disease. In **chapter 2B**, we focus on the molecular mechanisms underlying the gene regulation of the long and short UGP2 isoforms by the non-coding genome in dif-

ferent cell types, using targeted chromatin conformation capture (T2C) technologies and multi-omics.

Part 2

In **chapter 3**, we present a computational method to define putative functional enhancers during the developmental stages of the human fetal brain by integrating all previously published enhancer and epigenome data, identifying ~39 thousands enhancers that show dynamic epigenome rearrangement during development of which many are linked to human disease genes. **Chapter 4** provides a genome-wide identification of active functional enhancers in neural stem cells by combining chromatin immunoprecipitation with the massively parallel reporter assay STARR-seq (ChIP-STARR-seq). In **chapter 5**, we present a graphical interface application to visualize in a user-friendly manner all enhancer-related information obtained from chapters 3 and 4.

Finally, **Chapter 6** provides a general discussion of our findings in the context of recent literature.

References

1 Genereaux, D., van Karnebeek, C. D. & Birch, P. H. Costs of caring for children with an intellectual developmental disorder. Disabil Health J 8, 646-651, doi:10.1016/j.dhjo.2015.03.011 (2015).
2 Jonsson, U. et al. Annual Research Review: Quality of life and childhood mental and behavioural disorders - a critical review of the research. J Child Psychol Psychiatry 58, 439-469, doi:10.1111/jcpp.12645 (2017).
3 Agirman, G., Broix, L. & Nguyen, L. Cerebral cortex development: an outside-in perspective. FEBS Lett 591, 3978-3992, doi:10.1002/1873-3468.12924 (2017).
4 Van Essen, D. C., Donahue, C. J. & Glasser, M. F. Development and Evolution of Cerebral and Cerebellar Cortex. Brain Behav Evol 91, 158-169, doi:10.1159/000489943 (2018).
5 Seto, Y. & Eiraku, M. Human brain development and its in vitro recapitulation. Neurosci Res 138, 33-42, doi:10.1016/j.neures.2018.09.011 (2019).
6 Lin, C. H., Lin, W. D., Chou, I. C., Lee, I. C. & Hong, S. Y. Heterogeneous neurodevelopmental disorders in children with Kawasaki disease: what is new today? BMC Pediatr 19, 406, doi:10.1186/s12887-019-1786-y (2019).
7 Parenti, I., Rabaneda, L. G., Schoen, H. & Novarino, G. Neurodevelopmental Disorders: From Genetics to Functional Pathways. Trends Neurosci 43, 608-621, doi:10.1016/j.tins.2020.05.004 (2020).
8 Happ, H. C. & Carvill, G. L. A 2020 View on the Genetics of Developmental and Epileptic Encephalopathies. Epilepsy Curr 20, 90-96, doi:10.1177/1535759720906118 (2020).
9 Gallop, K., Lloyd, A. J., Olt, J. & Marshall, J. Impact of developmental and epileptic encephalopathies on caregivers: A literature review. Epilepsy Behav 124, 108324, doi:10.1016/j.yebeh.2021.108324 (2021).
10 Cardoso, A. R. et al. Essential genetic findings in neurodevelopmental disorders. Hum Genomics 13, 31, doi:10.1186/s40246-019-0216-4 (2019).
11 Yang, Y. et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med 369, 1502-1511, doi:10.1056/NEJMoa1306555 (2013).
12 Veeramah, K. R. et al. Exome sequencing reveals new causal mutations in children with epileptic encephalopathies. Epilepsia 54, 1270-1281, doi:10.1111/epi.12201 (2013).
13 Rochtus, A. et al. Genetic diagnoses in epilepsy: The impact of dynamic exome analysis in a pediatric cohort. Epilepsia 61, 249-258, doi:10.1111/epi.16427 (2020).
14 Epi, K. C. De Novo Mutations in SLC1A2 and CACNA1A Are Important Causes of Epileptic Encephalopathies. Am J Hum Genet 99, 287-298, doi:10.1016/j.ajhg.2016.06.003 (2016).
15 Rydzanicz, M. et al. A recurrent de novo variant supports KCNC2 involvement in the pathogenesis of developmental and epileptic encephalopathy. Am J Med Genet A 185, 3384-3389, doi:10.1002/ajmg.a.62455 (2021).
16 Specchio, N. & Curatolo, P. Developmental and epileptic encephalopathies: what we do and do not know. Brain 144, 32-43, doi:10.1093/brain/awaa371 (2021).
17 Manivannan, S. N. et al. De novo FZR1 loss-of-function variants cause developmental and epileptic encephalopathies. Brain 145, 1684-1697, doi:10.1093/brain/awab409 (2022).
18 Scheffer, I. E. et al. ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology. Epilepsia 58, 512-521, doi:10.1111/epi.13709 (2017).
19 Perucca, P. & Gilliam, F. G. Adverse effects of antiepileptic drugs. Lancet Neurol 11, 792-802, doi:10.1016/S1474-4422(12)70153-9 (2012).
20 Bhakdi, S., Valeva, A., Walev, I., Zitzer, A. & Palmer, M. Pore-forming bacterial cytolysins. Symp Ser Soc Appl Microbiol 27, 15S-25S, doi:10.1046/j.1365-2672.1998.0840s115s.x (1998).
21 Symonds, J. D. et al. Incidence and phenotypes of childhood-onset genetic epilepsies: a prospective population-based national cohort. Brain 142, 2303-2318, doi:10.1093/brain/awz195 (2019).
22 Minardi, R. et al. Whole-exome sequencing in adult patients with developmental and epileptic encephalopathy: It is never too late. Clin Genet 98, 477-485, doi:10.1111/cge.13823 (2020).
23 McTague, A., Howell, K. B., Cross, J. H., Kurian, M. A. & Scheffer, I. E. The genetic landscape of the epileptic encephalopathies of infancy and childhood. Lancet Neurol 15, 304-316, doi:10.1016/S1474-4422(15)00250-1 (2016).
24 Deuschl, G. et al. The burden of neurological diseases in Europe: an analysis for the Global Burden

of Disease Study 2017. Lancet Public Health 5, e551-e567, doi:10.1016/S2468-2667(20)30190-0 (2020).
25 Nickels, K. C. & Wirrell, E. C. Cognitive and Social Outcomes of Epileptic Encephalopathies. Semin Pediatr Neurol 24, 264-275, doi:10.1016/j.spen.2017.10.001 (2017).
26 Jakobsen, A. V., Moller, R. S., Nikanorova, M. & Elklit, A. The impact of severe pediatric epilepsy on experienced stress and psychopathology in parents. Epilepsy Behav 113, 107538, doi:10.1016/j.yebeh.2020.107538 (2020).
27 Howell, K. B. et al. The severe epilepsy syndromes of infancy: A population-based study. Epilepsia 62, 358-370, doi:10.1111/epi.16810 (2021).
28 Harini, C. et al. Mortality in infantile spasms: A hospital-based study. Epilepsia 61, 702-713, doi:10.1111/epi.16468 (2020).
29 Howell, K. B. et al. A population-based cost-effectiveness study of early genetic testing in severe epilepsies of infancy. Epilepsia 59, 1177-1187, doi:10.1111/epi.14087 (2018).
30 Doran, C. M. et al. How much does intellectual disability really cost? First estimates for Australia. J Intellect Dev Disabil 37, 42-49, doi:10.3109/13668250.2011.648609 (2012).
31 Collaborators, G. B. D. N. Global, regional, and national burden of neurological disorders, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol 18, 459-480, doi:10.1016/S1474-4422(18)30499-X (2019).
32 Jennum, P., Pickering, L., Christensen, J., Ibsen, R. & Kjellberg, J. Welfare cost of childhood- and adolescent-onset epilepsy: A controlled national study. Epilepsy Behav 61, 72-77, doi:10.1016/j.yebeh.2016.04.044 (2016).
33 Jedrzejczak, J. et al. Economic and social cost of epilepsy in Poland: 5-year analysis. Eur J Health Econ 22, 485-497, doi:10.1007/s10198-021-01269-1 (2021).
34 Riechmann, J. et al. Costs of epilepsy and cost-driving factors in children, adolescents, and their caregivers in Germany. Epilepsia 56, 1388-1397, doi:10.1111/epi.13089 (2015).
35 Striano, P. & Minassian, B. A. From Genetic Testing to Precision Medicine in Epilepsy. Neurotherapeutics 17, 609-615, doi:10.1007/s13311-020-00835-4 (2020).
36 Nabbout, R. & Kuchenbuch, M. Impact of predictive, preventive and precision medicine strategies in epilepsy. Nat Rev Neurol 16, 674-688, doi:10.1038/s41582-020-0409-4 (2020).
37 Bayat, A., Bayat, M., Rubboli, G. & Moller, R. S. Epilepsy Syndromes in the First Year of Life and Usefulness of Genetic Testing for Precision Therapy. Genes (Basel) 12, doi:10.3390/genes12071051 (2021).
38 Guerrini, R., Balestrini, S., Wirrell, E. C. & Walker, M. C. Monogenic Epilepsies: Disease Mechanisms, Clinical Phenotypes, and Targeted Therapies. Neurology 97, 817-831, doi:10.1212/WNL.0000000000012744 (2021).
39 Marini, C. & Giardino, M. Novel treatments in epilepsy guided by genetic diagnosis. Br J Clin Pharmacol 88, 2539-2551, doi:10.1111/bcp.15139 (2022).
40 Byrne, S., Enright, N. & Delanty, N. Precision therapy in the genetic epilepsies of childhood. Dev Med Child Neurol 63, 1276-1282, doi:10.1111/dmcn.14929 (2021).
41 Palmer, E. E., Howell, K. & Scheffer, I. E. Natural History Studies and Clinical Trial Readiness for Genetic Developmental and Epileptic Encephalopathies. Neurotherapeutics 18, 1432-1444, doi:10.1007/s13311-021-01133-3 (2021).
42 Palmer, E. E. et al. Diagnostic Yield of Whole Genome Sequencing After Nondiagnostic Exome Sequencing or Gene Panel in Developmental and Epileptic Encephalopathies. Neurology 96, e1770-e1782, doi:10.1212/WNL.0000000000011655 (2021).
43 Sheidley, B. R. et al. Genetic testing for the epilepsies: A systematic review. Epilepsia 63, 375-387, doi:10.1111/epi.17141 (2022).
44 Chen, W. et al. Next generation sequencing in children with unexplained epilepsy: A retrospective cohort study. Brain Dev 43, 1004-1012, doi:10.1016/j.braindev.2021.05.014 (2021).
45 Kim, S. Y. et al. Genetic diagnosis of infantile-onset epilepsy in the clinic: Application of whole-exome sequencing following epilepsy gene panel testing. Clin Genet 99, 418-424, doi:10.1111/cge.13903 (2021).
46 Carullo, N. V. N. & Day, J. J. Genomic Enhancers in Brain Health and Disease. Genes (Basel) 10, doi:10.3390/genes10010043 (2019).
47 Spielmann, M., Lupianez, D. G. & Mundlos, S. Structural variation in the 3D genome. Nat Rev

Genet 19, 453-467, doi:10.1038/s41576-018-0007-0 (2018).
48 Turner, T. N. & Eichler, E. E. The Role of De Novo Noncoding Regulatory Mutations in Neurodevelopmental Disorders. Trends Neurosci 42, 115-127, doi:10.1016/j.tins.2018.11.002 (2019).
49 Smith, E. & Shilatifard, A. Enhancer biology and enhanceropathies. Nat Struct Mol Biol 21, 210-219, doi:10.1038/nsmb.2784 (2014).
50 D'Haene, E. & Vergult, S. Interpreting the impact of noncoding structural variation in neurodevelopmental disorders. Genet Med 23, 34-46, doi:10.1038/s41436-020-00974-1 (2021).
51 Gil, N. & Ulitsky, I. Regulation of gene expression by cis-acting long non-coding RNAs. Nat Rev Genet 21, 102-117, doi:10.1038/s41576-019-0184-5 (2020).
52 Hezroni, H., Perry, R. B. T. & Ulitsky, I. Long Noncoding RNAs in Development and Regeneration of the Neural Lineage. Cold Spring Harb Symp Quant Biol 84, 165-177, doi:10.1101/sqb.2019.84.039347 (2019).
53 Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190-1195, doi:10.1126/science.1222794 (2012).
54 Zeng, Y. et al. Aberrant gene expression in humans. PLoS Genet 11, e1004942, doi:10.1371/journal.pgen.1004942 (2015).
55 Smith, M. & Flodman, P. L. Expanded Insights Into Mechanisms of Gene Expression and Disease Related Disruptions. Front Mol Biosci 5, 101, doi:10.3389/fmolb.2018.00101 (2018).
56 Doan, R. N. et al. Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. Cell 167, 341-354 e312, doi:10.1016/j.cell.2016.08.071 (2016).
57 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, doi:10.1007/s00439-018-1925-9 (2018).
58 Short, P. J. et al. De novo mutations in regulatory elements in neurodevelopmental disorders. Nature 555, 611-616, doi:10.1038/nature25983 (2018).
59 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, doi:10.1073/pnas.112212199 (2002).
60 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, doi:10.1093/hmg/ddg180 (2003).
61 Benko, S. et al. Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. Nat Genet 41, 359-364, doi:10.1038/ng.329 (2009).
62 Smemo, S. et al. Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. Hum Mol Genet 21, 3255-3263, doi:10.1093/hmg/dds165 (2012).
63 Weedon, M. N. et al. Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. Nat Genet 46, 61-64, doi:10.1038/ng.2826 (2014).
64 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, doi:10.1016/j.ajhg.2017.03.012 (2017).
65 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, doi:10.1093/hmg/ddx251 (2017).
66 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, doi:10.1097/MCD.0000000000000216 (2018).
67 Gloss, B. S. & Dinger, M. E. Realizing the significance of noncoding functionality in clinical genomics. Exp Mol Med 50, 1-8, doi:10.1038/s12276-018-0087-0 (2018).
68 Mehrjouy, M. M. et al. Regulatory variants of FOXG1 in the context of its topological domain organisation. Eur J Hum Genet 26, 186-196, doi:10.1038/s41431-017-0011-4 (2018).
69 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, doi:10.1038/gim.2018.18 (2018).
70 Castel, S. E. et al. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. Nat Genet 50, 1327-1334, doi:10.1038/s41588-018-0192-y (2018).
71 Niemi, M. E. K. et al. Common genetic variants contribute to risk of rare severe neurodevelopmen-

tal disorders. Nature 562, 268-271, doi:10.1038/s41586-018-0566-4 (2018).
72 Claringbould, A. & Zaugg, J. B. Enhancers in disease: molecular basis and emerging treatment strategies. Trends Mol Med 27, 1060-1073, doi:10.1016/j.molmed.2021.07.012 (2021).
73 Zeng, X. et al. The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: a systematic review. J Evid Based Med 8, 2-10, doi:10.1111/jebm.12141 (2015).
74 Crick, F. H. On protein synthesis. Symp Soc Exp Biol 12, 138-163 (1958).
75 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74, doi:10.1038/nature11247 (2012).
76 Ohno, S. So much "junk" DNA in our genome. Brookhaven Symp Biol 23, 366-370 (1972).
77 Kuska, B. Should scientists scrap the notion of junk DNA? J Natl Cancer Inst 90, 1032-1033, doi:10.1093/jnci/90.14.1032 (1998).
78 van Steensel, B. & Belmont, A. S. Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. Cell 169, 780-791, doi:10.1016/j.cell.2017.04.022 (2017).
79 Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. Cell Res 21, 381-395, doi:10.1038/cr.2011.22 (2011).
80 Hyun, K., Jeon, J., Park, K. & Kim, J. Writing, erasing and reading histone lysine methylations. Exp Mol Med 49, e324, doi:10.1038/emm.2017.11 (2017).
81 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, doi:10.1016/s0092-8674(00)81967-4 (1999).
82 Dowen, J. M. et al. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell 159, 374-387, doi:10.1016/j.cell.2014.09.030 (2014).
83 Ong, C. T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. Nat Rev Genet 15, 234-246, doi:10.1038/nrg3663 (2014).
84 Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. Nat Rev Genet 19, 789-800, doi:10.1038/s41576-018-0060-8 (2018).
85 Hnisz, D., Day, D. S. & Young, R. A. Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. Cell 167, 1188-1200, doi:10.1016/j.cell.2016.10.024 (2016).
86 Beagan, J. A. et al. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. Genome Res 27, 1139-1152, doi:10.1101/gr.215160.116 (2017).
87 Weintraub, A. S. et al. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. Cell 171, 1573-1588 e1528, doi:10.1016/j.cell.2017.11.008 (2017).
88 Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell 27, 299-308, doi:10.1016/0092-8674(81)90413-x (1981).
89 Osterwalder, M. et al. Enhancer redundancy provides phenotypic robustness in mammalian development. Nature 554, 239-243, doi:10.1038/nature25461 (2018).
90 Stadhouders, R. et al. Transcription regulation by distal enhancers: who's in the loop? Transcription 3, 181-186, doi:10.4161/trns.20720 (2012).
91 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, doi:10.1038/nrg3484 (2013).
92 Lambert, S. A. et al. The Human Transcription Factors. Cell 172, 650-665, doi:10.1016/j.cell.2018.01.029 (2018).
93 Yin, Y. et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science 356, doi:10.1126/science.aaj2239 (2017).
94 Iwafuchi-Doi, M. et al. The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. Mol Cell 62, 79-91, doi:10.1016/j.molcel.2016.03.001 (2016).
95 Iwafuchi-Doi, M. The mechanistic basis for chromatin regulation by pioneer transcription factors. Wiley Interdiscip Rev Syst Biol Med 11, e1427, doi:10.1002/wsbm.1427 (2019).
96 Siersbaek, R. et al. Extensive chromatin remodelling and establishment of transcription factor 'hotspots' during early adipogenesis. EMBO J 30, 1459-1472, doi:10.1038/emboj.2011.65 (2011).
97 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, doi:10.1073/pnas.1317023110 (2013).

98 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).
99 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).
100 Young, R. A. Control of the embryonic stem cell state. Cell 144, 940-954, doi:10.1016/j.cell.2011.01.032 (2011).
101 Whyte, W. A. et al. Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. Nature 482, 221-225, doi:10.1038/nature10805 (2012).
102 Loven, J. et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell 153, 320-334, doi:10.1016/j.cell.2013.03.036 (2013).
103 Hay, D. et al. Genetic dissection of the alpha-globin super-enhancer in vivo. Nat Genet 48, 895-903, doi:10.1038/ng.3605 (2016).
104 Shin, H. Y. et al. Hierarchy within the mammary STAT5-driven Wap super-enhancer. Nat Genet 48, 904-911, doi:10.1038/ng.3606 (2016).
105 Gregor, A. et al. De novo mutations in the genome organizer CTCF cause intellectual disability. Am J Hum Genet 93, 124-131, doi:10.1016/j.ajhg.2013.05.007 (2013).
106 Ball, A. R., Jr., Chen, Y. Y. & Yokomori, K. Mechanisms of cohesin-mediated gene regulation and lessons learned from cohesinopathies. Biochim Biophys Acta 1839, 191-202, doi:10.1016/j.bbagrm.2013.11.002 (2014).
107 Mirabella, A. C., Foster, B. M. & Bartke, T. Chromatin deregulation in disease. Chromosoma 125, 75-93, doi:10.1007/s00412-015-0530-0 (2016).
108 Holsten, T. et al. Germline variants in SMARCB1 and other members of the BAF chromatin-remodeling complex across human disease entities: a meta-analysis. Eur J Hum Genet 26, 1083-1093, doi:10.1038/s41431-018-0143-1 (2018).
109 De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature 515, 209-215, doi:10.1038/nature13772 (2014).
110 Gabriele, M. et al. YY1 Haploinsufficiency Causes an Intellectual Disability Syndrome Featuring Transcriptional and Chromatin Dysfunction. Am J Hum Genet 100, 907-925, doi:10.1016/j.ajhg.2017.05.006 (2017).
111 Kioussis, D., Vanin, E., deLange, T., Flavell, R. A. & Grosveld, F. G. Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. Nature 306, 662-666, doi:10.1038/306662a0 (1983).
112 Baars, M. J. D. et al. Dysregulated RASGRP1 expression through RUNX1 mediated transcription promotes autoimmunity. Eur J Immunol 51, 471-482, doi:10.1002/eji.201948451 (2021).
113 Shen, T. et al. An enhancer variant at 16q22.1 predisposes to hepatocellular carcinoma via regulating PRMT7 expression. Nat Commun 13, 1232, doi:10.1038/s41467-022-28861-0 (2022).
114 Reyes-Palomares, A. et al. Remodeling of active endothelial enhancers is associated with aberrant gene-regulatory networks in pulmonary arterial hypertension. Nat Commun 11, 1673, doi:10.1038/s41467-020-15463-x (2020).
115 Lin, X. et al. Genome-wide analysis of aberrant methylation of enhancer DNA in human osteoarthritis. BMC Med Genomics 13, 1, doi:10.1186/s12920-019-0646-9 (2020).
116 Ooi, W. F. et al. Integrated paired-end enhancer profiling and whole-genome sequencing reveals recurrent CCNE1 and IGF2 enhancer hijacking in primary gastric adenocarcinoma. Gut 69, 1039-1052, doi:10.1136/gutjnl-2018-317612 (2020).
117 Kragesteen, B. K., Brancati, F., Digilio, M. C., Mundlos, S. & Spielmann, M. H2AFY promoter deletion causes PITX1 endoactivation and Liebenberg syndrome. J Med Genet 56, 246-251, doi:10.1136/jmedgenet-2018-105793 (2019).
118 Al-Qattan, M. M., Al-Thunayan, A., Alabdulkareem, I. & Al Balwi, M. Liebenberg syndrome is caused by a deletion upstream to the PITX1 gene resulting in transformation of the upper limbs to reflect lower limb characteristics. Gene 524, 65-71, doi:10.1016/j.gene.2013.03.120 (2013).
119 Spielmann, M. et al. Homeotic arm-to-leg transformation associated with genomic rearrangements at the PITX1 locus. Am J Hum Genet 91, 629-635, doi:10.1016/j.ajhg.2012.08.014 (2012).
120 Taub, R. et al. Translocation of the c-myc gene into the immunoglobulin heavy chain locus in human Burkitt lymphoma and murine plasmacytoma cells. Proc Natl Acad Sci U S A 79, 7837-7841, doi:10.1073/pnas.79.24.7837 (1982).

121 Jeong, Y. et al. Regulation of a remote Shh forebrain enhancer by the Six3 homeoprotein. Nat Genet 40, 1348-1353, doi:10.1038/ng.230 (2008).
122 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, doi:10.1016/j.ajhg.2013.10.028 (2013).
123 Bae, B. I. et al. Evolutionarily dynamic alternative splicing of GPR56 regulates regional cerebral cortical patterning. Science 343, 764-768, doi:10.1126/science.1244392 (2014).
124 Soldner, F. et al. Parkinson-associated risk variant in distal enhancer of alpha-synuclein modulates target gene expression. Nature 533, 95-99, doi:10.1038/nature17939 (2016).
125 Vacic, V. et al. Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. Nature 471, 499-503, doi:10.1038/nature09884 (2011).
126 Nmezi, B. et al. Genomic deletions upstream of lamin B1 lead to atypical autosomal dominant leukodystrophy. Neurol Genet 5, e305, doi:10.1212/NXG.0000000000000305 (2019).
127 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, doi:10.1093/hmg/ddv065 (2015).
128 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, doi:10.1007/s00439-015-1594-x (2015).
129 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, doi:10.1002/ajmg.a.31563 (2007).
130 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, doi:10.1136/jmg.2007.055699 (2008).
131 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, doi:10.1002/humu.22097 (2012).
132 Wallis, D. E. et al. Mutations in the homeodomain of the human SIX3 gene cause holoprosencephaly. Nat Genet 22, 196-198, doi:10.1038/9718 (1999).
133 Nott, A. et al. Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. Science 366, 1134-1139, doi:10.1126/science.aay0793 (2019).
134 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, doi:10.1080/15622975.2017.1395072 (2019).
135 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, doi:10.1038/s41431-019-0335-3 (2019).
136 Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. Nat Rev Genet 7, 85-97, doi:10.1038/nrg1767 (2006).
137 Brandler, W. M. et al. Paternally inherited cis-regulatory structural variants are associated with autism. Science 360, 327-331, doi:10.1126/science.aan2261 (2018).
138 Turner, T. N. et al. Genomic Patterns of De Novo Mutation in Simplex Autism. Cell 171, 710-722 e712, doi:10.1016/j.cell.2017.08.047 (2017).
139 Monlong, J. et al. Global characterization of copy number variants in epilepsy patients from whole genome sequencing. PLoS Genet 14, e1007285, doi:10.1371/journal.pgen.1007285 (2018).
140 Hama, Y. et al. Genomic copy number variation analysis in multiple system atrophy. Mol Brain 10, 54, doi:10.1186/s13041-017-0335-6 (2017).
141 Liu, Y. C. et al. Evaluation of non-coding variation in GLUT1 deficiency. Dev Med Child Neurol 58, 1295-1302, doi:10.1111/dmcn.13163 (2016).
142 Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15, 1034-1050, doi:10.1101/gr.3715005 (2005).
143 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, doi:10.1093/nar/gkl822 (2007).

144 May, D. et al. Large-scale discovery of enhancers from human heart tissue. Nat Genet 44, 89-93, doi:10.1038/ng.1006 (2011).
145 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, doi:10.1093/nar/9.22.6047 (1981).
146 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, doi:10.1101/gad.310367.117 (2018).
147 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, doi:10.1016/s0092-8674(88)90469-2 (1988).
148 Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. Science 316, 1497-1502, doi:10.1126/science.1141319 (2007).
149 Robertson, G. et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4, 651-657, doi:10.1038/nmeth1068 (2007).
150 Heintzman, N. D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39, 311-318, doi:10.1038/ng1966 (2007).
151 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, doi:10.1073/pnas.1016071107 (2010).
152 Visel, A. et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457, 854-858, doi:10.1038/nature07730 (2009).
153 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, doi:10.1101/gr.129817.111 (2012).
154 Barakat, T. S. et al. Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. Cell Stem Cell 23, 276-288 e278, doi:10.1016/j.stem.2018.06.014 (2018).
155 Reilly, S. K. et al. Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. Science 347, 1155-1159, doi:10.1126/science.1260943 (2015).
156 Amiri, A. et al. Transcriptome and epigenome landscape of human cortical development modeled in organoids. Science 362, doi:10.1126/science.aat6720 (2018).
157 Vermunt, M. W. et al. Large-scale identification of coregulated enhancer networks in the adult human brain. Cell Rep 9, 767-779, doi:10.1016/j.celrep.2014.09.023 (2014).
158 Sun, W. et al. Histone Acetylome-wide Association Study of Autism Spectrum Disorder. Cell 167, 1385-1397 e1311, doi:10.1016/j.cell.2016.10.031 (2016).
159 Li, M. et al. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. Science 362, doi:10.1126/science.aat7615 (2018).
160 Wang, D. et al. Comprehensive functional genomic resource and integrative model for the human brain. Science 362, doi:10.1126/science.aat8464 (2018).
161 Boyle, A. P. et al. High-resolution mapping and characterization of open chromatin across the genome. Cell 132, 311-322, doi:10.1016/j.cell.2007.12.014 (2008).
162 Lidor Nili, E. et al. p53 binds preferentially to genomic regions with high DNA-encoded nucleosome occupancy. Genome Res 20, 1361-1368, doi:10.1101/gr.103945.109 (2010).
163 John, S. et al. Genome-scale mapping of DNase I hypersensitivity. Curr Protoc Mol Biol Chapter 27, Unit 21 27, doi:10.1002/0471142727.mb2127s103 (2013).
164 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, doi:10.1016/j.ymeth.2009.03.003 (2009).
165 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, doi:10.1038/nmeth.2688 (2013).
166 Arnold, C. D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 339, 1074-1077, doi:10.1126/science.1232542 (2013).
167 de la Torre-Ubieta, L. et al. The Dynamic Landscape of Open Chromatin during Human Cortical

Neurogenesis. Cell 172, 289-304 e218, doi:10.1016/j.cell.2017.12.014 (2018).
168 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, doi:10.1002/j.1460-2075.1990.tb08100.x (1990).
169 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, doi:10.1073/pnas.89.23.11219 (1992).
170 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, doi:10.1101/gad.11.19.2494 (1997).
171 Kim, T. K. et al. Widespread transcription at neuronal activity-regulated enhancers. Nature 465, 182-187, doi:10.1038/nature09033 (2010).
172 Wang, D. et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature 474, 390-394, doi:10.1038/nature10006 (2011).
173 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, doi:10.1101/gr.152306.112 (2013).
174 Kaikkonen, M. U. et al. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. Mol Cell 51, 310-325, doi:10.1016/j.molcel.2013.07.010 (2013).
175 Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455-461, doi:10.1038/nature12787 (2014).
176 Koch, F. et al. Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. Nat Struct Mol Biol 18, 956-963, doi:10.1038/nsmb.2085 (2011).
177 Lam, M. T., Li, W., Rosenfeld, M. G. & Glass, C. K. Enhancer RNAs and regulated transcriptional programs. Trends Biochem Sci 39, 170-182, doi:10.1016/j.tibs.2014.02.007 (2014).
178 Barakat, T. S. & Gribnau, J. X chromosome inactivation and embryonic stem cells. Adv Exp Med Biol 695, 132-154, doi:10.1007/978-1-4419-7037-4_10 (2010).
179 Natoli, G. & Andrau, J. C. Noncoding transcription at enhancers: general principles and functional models. Annu Rev Genet 46, 1-19, doi:10.1146/annurev-genet-110711-155459 (2012).
180 Yao, P. et al. Coexpression networks identify brain region-specific enhancer RNAs in the human brain. Nat Neurosci 18, 1168-1174, doi:10.1038/nn.4063 (2015).
181 de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. Genes Dev 26, 11-24, doi:10.1101/gad.179804.111 (2012).
182 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, doi:10.1038/nmeth.4146 (2017).
183 Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. Science 295, 1306-1311, doi:10.1126/science.1067799 (2002).
184 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, doi:10.1038/ng1896 (2006).
185 Zhao, Z. et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet 38, 1341-1347, doi:10.1038/ng1891 (2006).
186 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, doi:10.1101/gr.5571506 (2006).
187 Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289-293, doi:10.1126/science.1181369 (2009).
188 Mumbach, M. R. et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods 13, 919-922, doi:10.1038/nmeth.3999 (2016).
189 Fang, R. et al. Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. Cell Res 26, 1345-1348, doi:10.1038/cr.2016.137 (2016).
190 Won, H. et al. Chromosome conformation elucidates regulatory relationships in developing human brain. Nature 538, 523-527, doi:10.1038/nature19847 (2016).
191 Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing

of ENCODE segmentation predictions. Genome Res 24, 1595-1602, doi:10.1101/gr.173518.114 (2014).
192 Halfon, M. S. Studying Transcriptional Enhancers: The Founder Fallacy, Validation Creep, and Other Biases. Trends Genet 35, 93-103, doi:10.1016/j.tig.2018.11.004 (2019).
193 Pradeepa, M. M. et al. Histone H3 globular domain acetylation identifies a new class of enhancers. Nat Genet 48, 681-686, doi:10.1038/ng.3550 (2016).
194 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, doi:10.1038/nbt.2137 (2012).
195 Patwardhan, R. P. et al. Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol 30, 265-270, doi:10.1038/nbt.2136 (2012).
196 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, doi:10.1101/gr.144899.112 (2013).
197 Dickel, D. E. et al. Function-based identification of mammalian enhancers using site-specific integration. Nat Methods 11, 566-571, doi:10.1038/nmeth.2886 (2014).
198 Ernst, J. et al. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nat Biotechnol 34, 1180-1190, doi:10.1038/nbt.3678 (2016).
199 Murtha, M. et al. FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. Nat Methods 11, 559-565, doi:10.1038/nmeth.2885 (2014).
200 Arnold, C. D. et al. Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. Nat Biotechnol 35, 136-144, doi:10.1038/nbt.3739 (2017).
201 van Arensbergen, J. et al. Genome-wide mapping of autonomous promoter activity in human cells. Nat Biotechnol 35, 145-153, doi:10.1038/nbt.3754 (2017).
202 Wang, X. et al. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. Nat Commun 9, 5380, doi:10.1038/s41467-018-07746-1 (2018).
203 Vanhille, L. et al. High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. Nat Commun 6, 6905, doi:10.1038/ncomms7905 (2015).
204 Shen, S. Q. et al. Massively parallel cis-regulatory analysis in the mammalian central nervous system. Genome Res 26, 238-255, doi:10.1101/gr.193789.115 (2016).
205 Inoue, F. et al. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. Genome Res 27, 38-52, doi:10.1101/gr.212092.116 (2017).
206 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, doi:10.1038/nbt.4285 (2018).
207 Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet 15, 272-286, doi:10.1038/nrg3682 (2014).
208 Korkmaz, G. et al. Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. Nat Biotechnol 34, 192-198, doi:10.1038/nbt.3450 (2016).
209 Sanjana, N. E. et al. High-resolution interrogation of functional elements in the noncoding genome. Science 353, 1545-1549, doi:10.1126/science.aaf7613 (2016).
210 Han, R. et al. Functional CRISPR screen identifies AP1-associated enhancer regulating FOXF1 to modulate oncogene-induced senescence. Genome Biol 19, 118, doi:10.1186/s13059-018-1494-1 (2018).
211 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, doi:10.1016/j.ajhg.2017.06.010 (2017).
212 Diao, Y. et al. A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. Genome Res 26, 397-405, doi:10.1101/gr.197152.115 (2016).
213 Diao, Y. et al. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. Nat Methods 14, 629-635, doi:10.1038/nmeth.4264 (2017).
214 Canver, M. C. et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature 527, 192-197, doi:10.1038/nature15521 (2015).
215 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, doi:10.1038/ng.3793 (2017).
216 Rajagopal, N. et al. High-throughput mapping of regulatory DNA. Nat Biotechnol 34, 167-174, doi:10.1038/nbt.3468 (2016).
217 Sen, D. R. et al. The epigenetic landscape of T cell exhaustion. Science 354, 1165-1169, doi:10.1126/science.aae0491 (2016).
218 Maeder, M. L. et al. CRISPR RNA-guided activation of endogenous human genes. Nat Methods 10, 977-979, doi:10.1038/nmeth.2598 (2013).
219 Mali, P. et al. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nat Biotechnol 31, 833-838, doi:10.1038/nbt.2675 (2013).
220 Konermann, S. et al. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature 517, 583-588, doi:10.1038/nature14136 (2015).
221 Liu, X. S. et al. Editing DNA Methylation in the Mammalian Genome. Cell 167, 233-247 e217, doi:10.1016/j.cell.2016.08.056 (2016).
222 Hilton, I. B. et al. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat Biotechnol 33, 510-517, doi:10.1038/nbt.3199 (2015).
223 Gilbert, L. A. et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell 154, 442-451, doi:10.1016/j.cell.2013.06.044 (2013).
224 Thakore, P. I. et al. Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. Nat Methods 12, 1143-1149, doi:10.1038/nmeth.3630 (2015).
225 Konermann, S. et al. Optical control of mammalian endogenous transcription and epigenetic states. Nature 500, 472-476, doi:10.1038/nature12466 (2013).
226 Vojta, A. et al. Repurposing the CRISPR-Cas9 system for targeted DNA methylation. Nucleic Acids Res 44, 5615-5628, doi:10.1093/nar/gkw159 (2016).
227 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, doi:10.1038/ncomms15315 (2017).
228 Kearns, N. A. et al. Functional annotation of native enhancers with a Cas9-histone demethylase fusion. Nat Methods 12, 401-403, doi:10.1038/nmeth.3325 (2015).
229 Fulco, C. P. et al. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. Science 354, 769-773, doi:10.1126/science.aag2445 (2016).
230 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, doi:10.1016/j.cels.2017.08.011 (2017).
231 Gasperini, M. et al. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell 176, 377-390 e319, doi:10.1016/j.cell.2018.11.029 (2019).
232 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, doi:10.1016/j.molcel.2017.03.007 (2017).
233 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, doi:10.1038/nbt.3853 (2017).
234 Simeonov, D. R. et al. Discovery of stimulation-responsive immune enhancers with CRISPR activation. Nature 549, 111-115, doi:10.1038/nature23875 (2017).
235 Wang, C., Zhang, M. Q. & Zhang, Z. Computational identification of active enhancers in model organisms. Genomics Proteomics Bioinformatics 11, 142-150, doi:10.1016/j.gpb.2013.04.002 (2013).
236 Suryamohan, K. & Halfon, M. S. Identifying transcriptional cis-regulatory modules in animal genomes. Wiley Interdiscip Rev Dev Biol 4, 59-84, doi:10.1002/wdev.168 (2015).
237 Kleftogiannis, D., Kalnis, P. & Bajic, V. B. Progress and challenges in bioinformatics approaches for enhancer identification. Brief Bioinform 17, 967-979, doi:10.1093/bib/bbv101 (2016).
238 Lim, L. W. K., Chung, H. H., Chong, Y. L. & Lee, N. K. A survey of recently emerged genome-wide computational enhancer predictor tools. Comput Biol Chem 74, 132-141, doi:10.1016/j.compbiolchem.2018.03.019 (2018).
239 Pennacchio, L. A. et al. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444, 499-502, doi:10.1038/nature05295 (2006).

240 Li, L., Zhu, Q., He, X., Sinha, S. & Halfon, M. S. Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses. Genome Biol 8, R101, doi:10.1186/gb-2007-8-6-r101 (2007).
241 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. Nat Genet 46, 685-692, doi:10.1038/ng.3009 (2014).
242 Degner, J. F. et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482, 390-394, doi:10.1038/nature10808 (2012).
243 Shibata, Y. et al. Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. PLoS Genet 8, e1002789, doi:10.1371/journal.pgen.1002789 (2012).
244 Villar, D. et al. Enhancer evolution across 20 mammalian species. Cell 160, 554-566, doi:10.1016/j.cell.2015.01.006 (2015).
245 Glinsky, G. & Barakat, T. S. The evolution of Great Apes has shaped the functional enhancers' landscape in human embryonic stem cells. Stem Cell Res 37, 101456, doi:10.1016/j.scr.2019.101456 (2019).
246 Bejerano, G. et al. Ultraconserved elements in the human genome. Science 304, 1321-1325, doi:10.1126/science.1098119 (2004).
247 Visel, A. et al. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat Genet 40, 158-160, doi:10.1038/ng.2007.55 (2008).
248 Dickel, D. E. et al. Ultraconserved Enhancers Are Required for Normal Development. Cell 172, 491-499 e415, doi:10.1016/j.cell.2017.12.017 (2018).
249 Ahituv, N. et al. Deletion of ultraconserved elements yields viable mice. PLoS Biol 5, e234, doi:10.1371/journal.pbio.0050234 (2007).
250 McCole, R. B., Erceg, J., Saylor, W. & Wu, C. T. Ultraconserved Elements Occupy Specific Arenas of Three-Dimensional Mammalian Genome Organization. Cell Rep 24, 479-488, doi:10.1016/j.celrep.2018.06.031 (2018).
251 Ludwig, M. Z., Bergman, C., Patel, N. H. & Kreitman, M. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403, 564-567, doi:10.1038/35000615 (2000).
252 Firpi, H. A., Ucar, D. & Tan, K. Discover regulatory DNA elements using chromatin signatures and artificial neural network. Bioinformatics 26, 1579-1586, doi:10.1093/bioinformatics/btq248 (2010).
253 Fernandez, M. & Miranda-Saavedra, D. Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. Nucleic Acids Res 40, e77, doi:10.1093/nar/gks149 (2012).
254 Rajagopal, N. et al. RFECS: a random-forest based algorithm for enhancer identification from chromatin state. PLoS Comput Biol 9, e1002968, doi:10.1371/journal.pcbi.1002968 (2013).
255 Erwin, G. D. et al. Integrating diverse datasets improves developmental enhancer prediction. PLoS Comput Biol 10, e1003677, doi:10.1371/journal.pcbi.1003677 (2014).
256 Lu, Y., Qu, W., Shan, G. & Zhang, C. DELTA: A Distal Enhancer Locating Tool Based on AdaBoost Algorithm and Shape Features of Chromatin Modifications. PLoS One 10, e0130622, doi:10.1371/journal.pone.0130622 (2015).
257 Liu, F., Li, H., Ren, C., Bo, X. & Shu, W. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. Sci Rep 6, 28517, doi:10.1038/srep28517 (2016).
258 He, Y. et al. Improved regulatory element prediction based on tissue-specific local epigenomic signatures. Proc Natl Acad Sci U S A 114, E1633-E1640, doi:10.1073/pnas.1618353114 (2017).
259 Zehnder, T., Benner, P. & Vingron, M. Predicting enhancers in mammalian genomes using supervised hidden Markov models. BMC Bioinformatics 20, 157, doi:10.1186/s12859-019-2708-6 (2019).
260 Osmala, M. & Lahdesmaki, H. Enhancer prediction in the human genome by probabilistic modelling of the chromatin feature patterns. BMC Bioinformatics 21, 317, doi:10.1186/s12859-020-03621-3 (2020).
261 Thibodeau, A. et al. CoRE-ATAC: A deep learning model for the functional classification of regulatory elements from single cell and bulk ATAC-seq data. PLoS Comput Biol 17, e1009670, doi:10.1371/journal.pcbi.1009670 (2021).

262 Mills, C. et al. PEREGRINE: A genome-wide prediction of enhancer to gene relationships supported by experimental evidence. PLoS One 15, e0243791, doi:10.1371/journal.pone.0243791 (2020).
263 Jia, C. & He, W. EnhancerPred: a predictor for discovering enhancers based on the combination and selection of multiple features. Sci Rep 6, 38741, doi:10.1038/srep38741 (2016).
264 Zacher, B. et al. Accurate Promoter and Enhancer Identification in 127 ENCODE and Roadmap Epigenomics Cell Types and Tissues by GenoSTAN. PLoS One 12, e0169249, doi:10.1371/journal.pone.0169249 (2017).
265 Ramisch, A. et al. CRUP: a comprehensive framework to predict condition-specific regulatory units. Genome Biol 20, 227, doi:10.1186/s13059-019-1860-7 (2019).
266 Bu, H., Gan, Y., Wang, Y., Zhou, S. & Guan, J. A new method for enhancer prediction based on deep belief network. BMC Bioinformatics 18, 418, doi:10.1186/s12859-017-1828-0 (2017).
267 Umarov, R. et al. ReFeaFi: Genome-wide prediction of regulatory elements driving transcription initiation. PLoS Comput Biol 17, e1009376, doi:10.1371/journal.pcbi.1009376 (2021).
268 Hoffman, M. M. et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods 9, 473-476, doi:10.1038/nmeth.1937 (2012).
269 Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods 9, 215-216, doi:10.1038/nmeth.1906 (2012).
270 Liu, B., Fang, L., Long, R., Lan, X. & Chou, K. C. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics 32, 362-369, doi:10.1093/bioinformatics/btv604 (2016).
271 Liu, B., Li, K., Huang, D. S. & Chou, K. C. iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach. Bioinformatics 34, 3835-3842, doi:10.1093/bioinformatics/bty458 (2018).
272 Nguyen, Q. H. et al. iEnhancer-ECNN: identifying enhancers and their strength using ensembles of convolutional neural networks. BMC Genomics 20, 951, doi:10.1186/s12864-019-6336-3 (2019).
273 Niu, K. et al. iEnhancer-EBLSTM: Identifying Enhancers and Strengths by Ensembles of Bidirectional Long Short-Term Memory. Front Genet 12, 665498, doi:10.3389/fgene.2021.665498 (2021).
274 Yang, R., Wu, F., Zhang, C. & Zhang, L. iEnhancer-GAN: A Deep Learning Framework in Combination with Word Embedding and Sequence Generative Adversarial Net to Identify Enhancers and Their Strength. Int J Mol Sci 22, doi:10.3390/ijms22073589 (2021).
275 Yang, H., Wang, S. & Xia, X. iEnhancer-RD: Identification of enhancers and their strength using RKPK features and deep neural networks. Anal Biochem 630, 114318, doi:10.1016/j.ab.2021.114318 (2021).
276 Dogan, N. et al. Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. Epigenetics Chromatin 8, 16, doi:10.1186/s13072-015-0009-5 (2015).
277 Li, G. et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148, 84-98, doi:10.1016/j.cell.2011.12.014 (2012).
278 Javierre, B. M. et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell 167, 1369-1384 e1319, doi:10.1016/j.cell.2016.09.037 (2016).
279 Jing, F., Zhang, S. W. & Zhang, S. Prediction of enhancer-promoter interactions using the cross-cell type information and domain adversarial neural network. BMC Bioinformatics 21, 507, doi:10.1186/s12859-020-03844-4 (2020).
280 Hong, Z., Zeng, X., Wei, L. & Liu, X. Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. Bioinformatics 36, 1037-1043, doi:10.1093/bioinformatics/btz694 (2020).
281 Silva, T. C. et al. ELMER v.2: an R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles. Bioinformatics 35, 1974-1977, doi:10.1093/bioinformatics/bty902 (2019).
282 Kvon, E. Z. et al. Genome-scale functional characterization of Drosophila developmental enhancers in vivo. Nature 512, 91-95, doi:10.1038/nature13395 (2014).
283 Okonechnikov, K., Erkek, S., Korbel, J. O., Pfister, S. M. & Chavez, L. InTAD: chromosome

conformation guided analysis of enhancer target genes. BMC Bioinformatics 20, 60, doi:10.1186/s12859-019-2655-2 (2019).
284 Roy, S. et al. A predictive modeling approach for cell line-specific long-range regulatory interactions. Nucleic Acids Res 43, 8694-8712, doi:10.1093/nar/gkv865 (2015).
285 Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat Genet 48, 488-496, doi:10.1038/ng.3539 (2016).
286 Zhu, Y. et al. Constructing 3D interaction maps from 1D epigenomes. Nat Commun 7, 10812, doi:10.1038/ncomms10812 (2016).
287 Cao, Q. et al. Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. Nat Genet 49, 1428-1436, doi:10.1038/ng.3950 (2017).
288 Hait, T. A., Amar, D., Shamir, R. & Elkon, R. FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map. Genome Biol 19, 56, doi:10.1186/s13059-018-1432-2 (2018).
289 Yang, Y., Zhang, R., Singh, S. & Ma, J. Exploiting sequence-based features for predicting enhancer-promoter interactions. Bioinformatics 33, i252-i260, doi:10.1093/bioinformatics/btx257 (2017).
290 Singh, S., Yang, Y., Poczos, B. & Ma, J. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. Quant Biol 7, 122-137, doi:10.1007/s40484-019-0154-0 (2019).
291 Zeng, W., Wu, M. & Jiang, R. Prediction of enhancer-promoter interactions via natural language processing. BMC Genomics 19, 84, doi:10.1186/s12864-018-4459-6 (2018).
292 Fulco, C. P. et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. Nat Genet 51, 1664-1669, doi:10.1038/s41588-019-0538-0 (2019).
293 Boyle, A. P. et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 22, 1790-1797, doi:10.1101/gr.137323.112 (2012).
294 Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 40, D930-934, doi:10.1093/nar/gkr917 (2012).
295 Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46, 310-315, doi:10.1038/ng.2892 (2014).
296 Ritchie, G. R., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. Nat Methods 11, 294-296, doi:10.1038/nmeth.2832 (2014).
297 Lu, Q. et al. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Sci Rep 5, 10576, doi:10.1038/srep10576 (2015).
298 Smedley, D. et al. A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. Am J Hum Genet 99, 595-606, doi:10.1016/j.ajhg.2016.07.005 (2016).
299 Amlie-Wolf, A. et al. INFERNO: inferring the molecular mechanisms of noncoding genetic variants. Nucleic Acids Res 46, 8740-8753, doi:10.1093/nar/gky686 (2018).
300 Wang, Z. et al. HEDD: Human Enhancer Disease Database. Nucleic Acids Res 46, D113-D120, doi:10.1093/nar/gkx988 (2018).
301 Zhang, G. et al. DiseaseEnhancer: a resource of human disease-associated enhancer catalog. Nucleic Acids Res 46, D78-D84, doi:10.1093/nar/gkx920 (2018).
302 Zeng, W., Min, X. & Jiang, R. EnDisease: a manually curated database for enhancer-disease associations. Database (Oxford) 2019, doi:10.1093/database/baz020 (2019).
303 Gao, T. et al. EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. Bioinformatics 32, 3543-3551, doi:10.1093/bioinformatics/btw495 (2016).
304 Khan, A. & Zhang, X. dbSUPER: a database of super-enhancers in mouse and human genome. Nucleic Acids Res 44, D164-171, doi:10.1093/nar/gkv1002 (2016).
305 Wei, Y. et al. SEA: a super-enhancer archive. Nucleic Acids Res 44, D172-179, doi:10.1093/nar/gkv1243 (2016).
306 Jiang, Y. et al. SEdb: a comprehensive human super-enhancer database. Nucleic Acids Res 47, D235-D243, doi:10.1093/nar/gky1025 (2019).
307 Fishilevich, S. et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford) 2017, doi:10.1093/database/bax028 (2017).

308 Stavropoulos, D. J. et al. Whole Genome Sequencing Expands Diagnostic Utility and Improves Clinical Management in Pediatric Medicine. NPJ Genom Med 1, doi:10.1038/npjgenmed.2015.12 (2016).
309 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, doi:10.1038/gim.2017.119 (2018).
310 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, doi:10.1126/scitranslmed.aat6177 (2019).
311 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, doi:10.1038/s41525-018-0076-1 (2019).

Chapter 2A

Loss of UGP2 in brain leads to a severe DEE

Elena Perenthaler[1], Anita Nikoncuk[1*], Soheil Yousefi[1*], Woutje M. Berdowski[1*], Maysoon Alsagob[2*], et al. Loss of UGP2 in brain leads to a severe epileptic encephalopathy, emphasizing that biallelic isoformspecific startloss mutations of essential genes can cause genetic diseases. *Acta Neuropathologica. 2020; 139, 415-442*.

[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.

[*]Contributed equally.

Extended author information available on the last page of this chapter.

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 3,000-7,000 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.

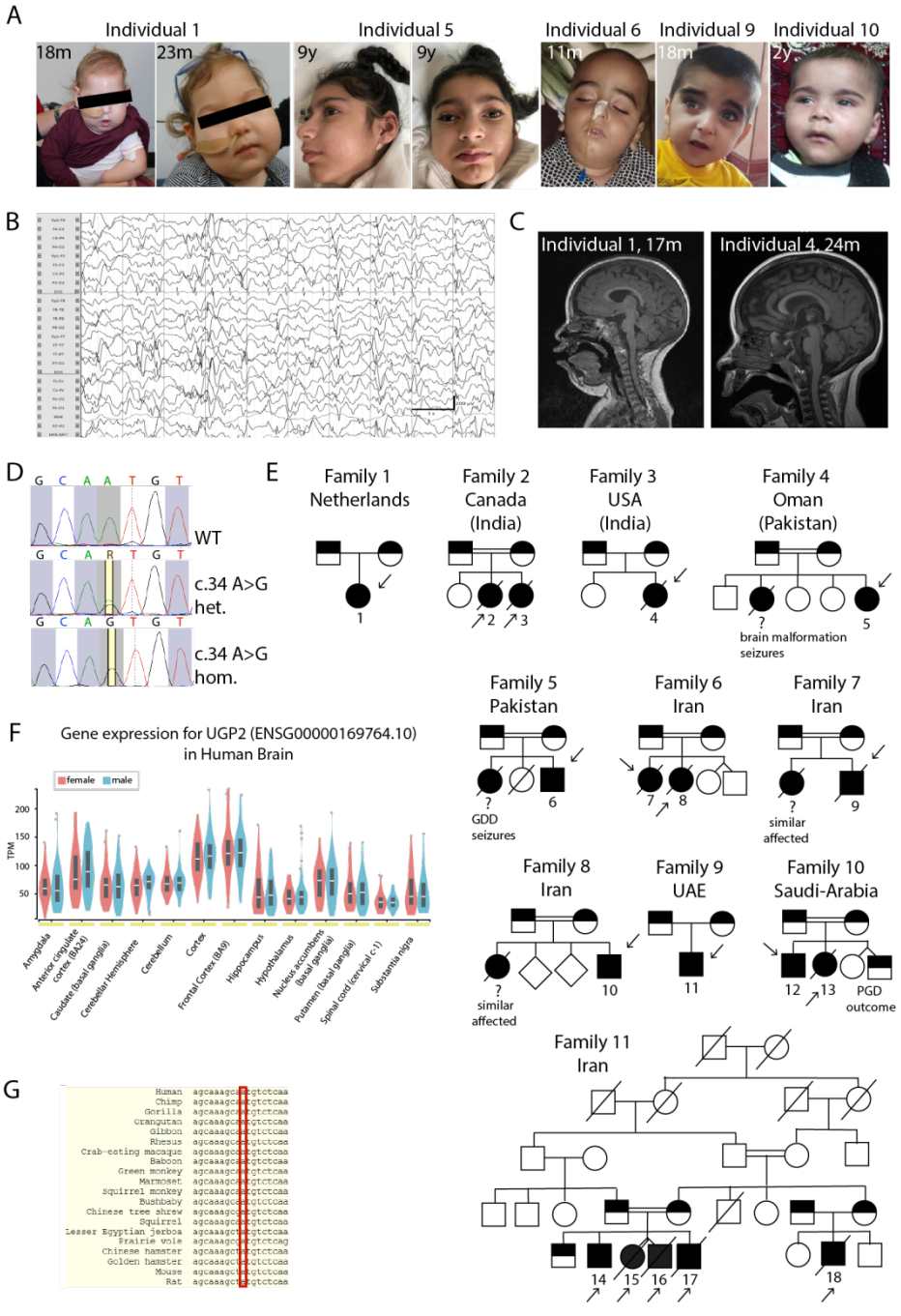
Results

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 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, Supplementary Figure 1B**). 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 (**Supplementary Figure 1A**). 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 pro-

nounced 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, Supplementary Figure 1B, 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] (**Supplementary Figure 1D**). 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


A
Individual 1
Individual 5
Individual 6
Individual 9
Individual 10
18m
23m
9y
9y
11m
18m
2y
B
C
Individual 1, 17m
Individual 4, 24m
D
WT
c.34 A>G het.
c.34 A>G hom.
E
Family 1 Netherlands
Family 2 Canada (India)
Family 3 USA (India)
Family 4 Oman (Pakistan)
brain malformation seizures
Family 5 Pakistan
GDD seizures
Family 6 Iran
Family 7 Iran
similar affected
Family 8 Iran
similar affected
Family 9 UAE
Family 10 Saudi-Arabia
PGD outcome
Family 11 Iran
F
Gene expression for UGP2 (ENSG00000169764.10) in Human Brain
female
male
TPM
Amygdala
Anterior cingulate cortex (BA24)
Caudate (basal ganglia)
Cerebellar Hemisphere
Cerebellum
Cortex
Frontal Cortex (BA9)
Hippocampus
Hypothalamus
Nucleus accumbens (basal ganglia)
Putamen (basal ganglia)
Spinal cord (cervical c-1)
Substantia nigra
G
Human
Chimp
Gorilla
Orangutan
Gibbon
Rhesus
Crab-eating macaque
Baboon
Green monkey
Marmoset
Squirrel monkey
Bushbaby
Chinese tree shrew
Squirrel
Lesser Egyptian jerboa
Prairie vole
Chinese hamster
Golden hamster
Mouse
Rat

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 (**Supplementary Figure 1C**). 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].

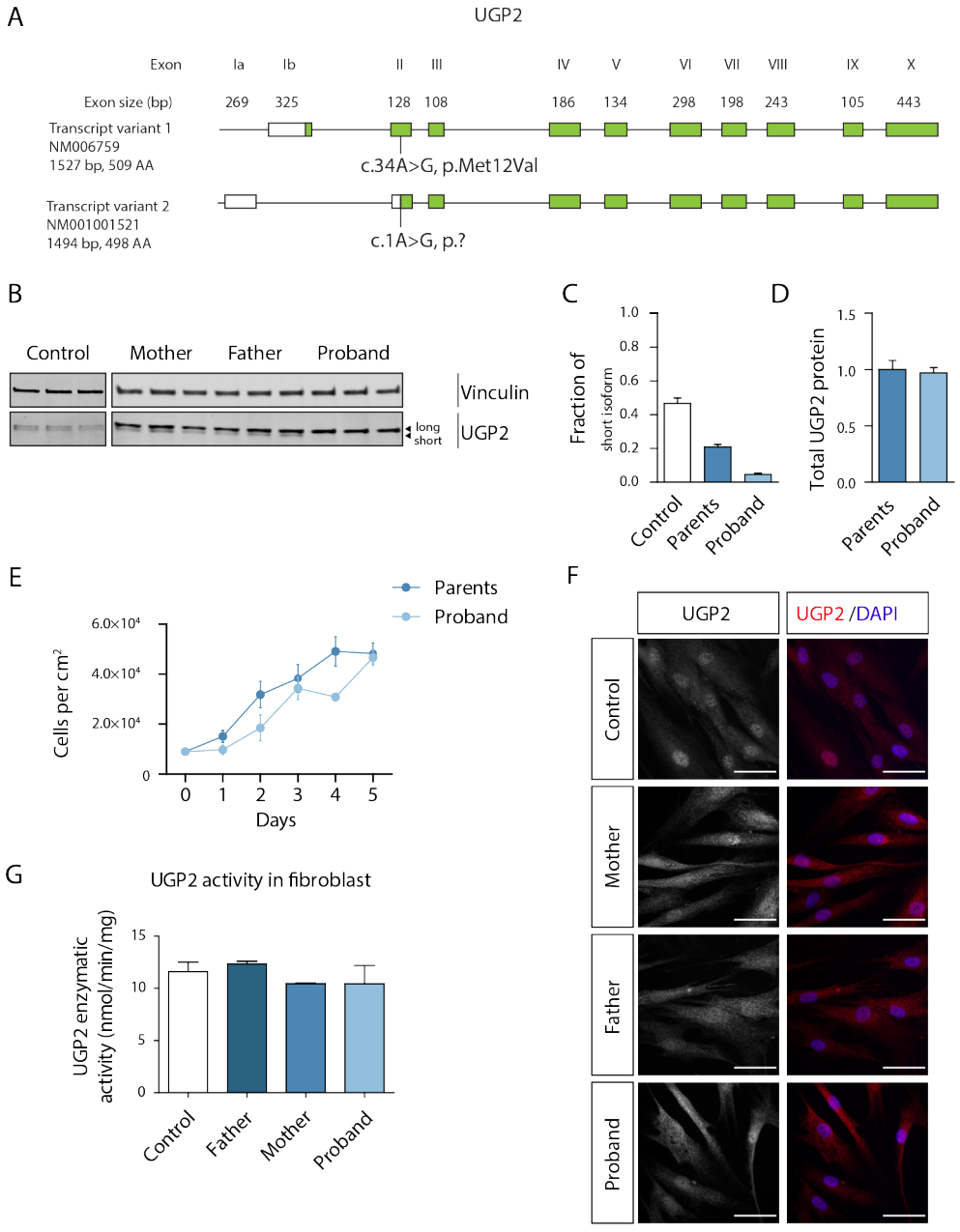
Short UGP2 isoform is predominantly expressed in brain and absent in patients with ATG mutations

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**).

Figure 1. UGP2 homozygous variants in 20 individuals with severe epileptic encephalopathy. A) Facial pictures of individual 1 (at 18 and 23 months), individual 5 (at 9 years), individual 6 (at 11 months), individual 9 (at 18 months), individual 10 (at 2 years) and individual 19 (at 13 months). 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 months 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 (at 17 months) and individual 4 (at 24 months) 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 heterozygous and homozygous states 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 not-tested affected siblings presenting with similar phenotypes are indicated with a question mark. Consanguineous parents are indicated with a double connection line. Males are squares, females are circles; unknown sex is 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 portal 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.

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 (**Supplementary figure 2A, B**), and was absent in cells from the affected individual 1 (**Figure 2 B, 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, Supplementary Figure 2C**). 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 (**Supplementary Figure 2D, E, F**). RNA-seq on peripheral blood samples of family 1 did not identify altered splicing events of *UGP2* and the 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 (**Supplementary Figure 2G, H**). Both homozygous and heterozygous fibroblasts had a similar proliferation rate compared to wild type fibroblasts (**Figure 2E, Supplementary Figure 2I**), 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.

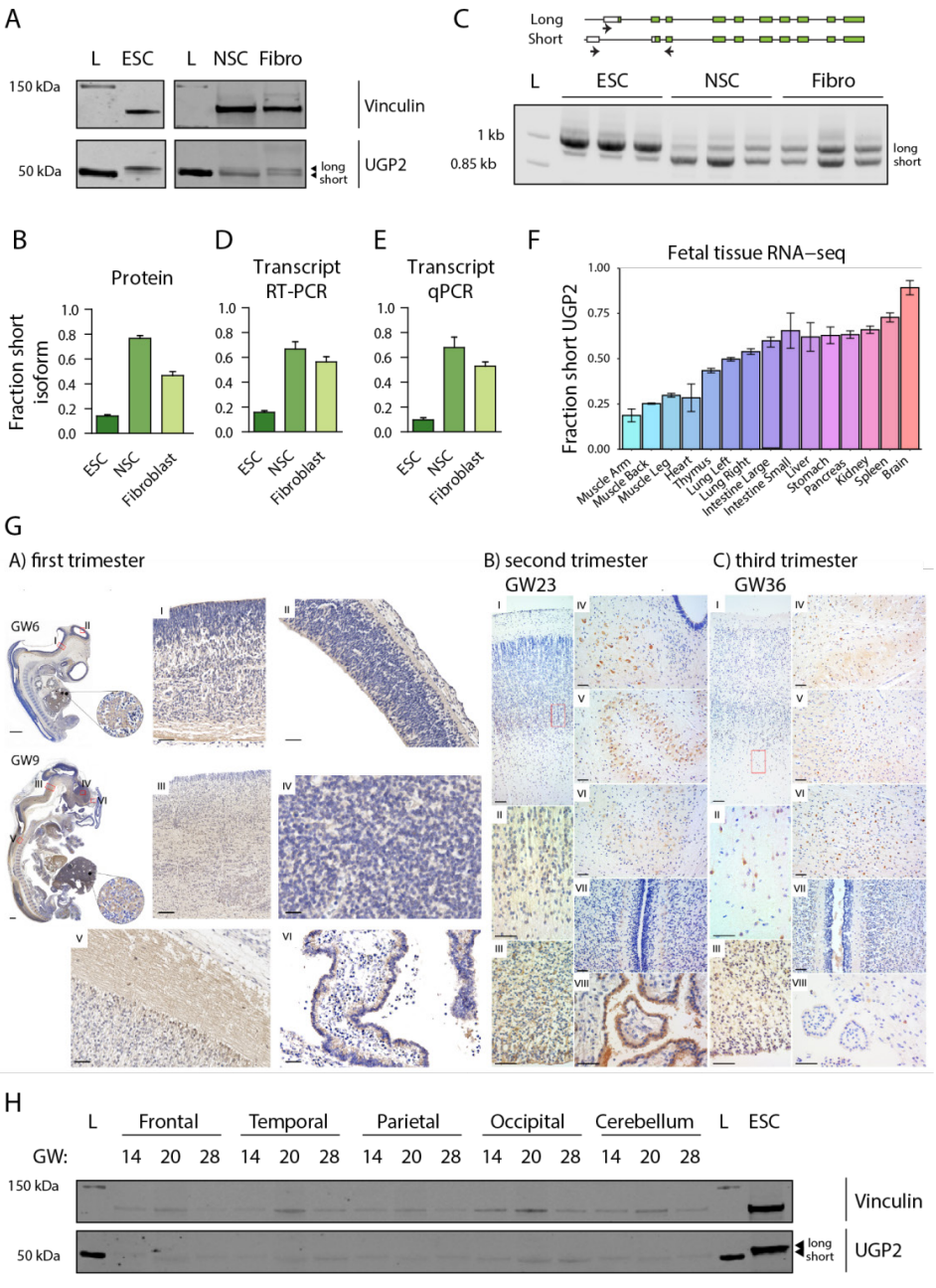
As the A>G variant results in a functional long UGP2 isoform but abolishes the translation of the shorter UGP2 isoform, we next investigated whether the ratio between short and long isoform differs amongst tissues. If so, the homozygous A>G variant would lead to depletion of UGP2 in tissues where mainly 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 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


A
UGP2
Exon
Ia
Ib
II
III
IV
V
VI
VII
VIII
IX
X
Exon size (bp)
269
325
128
108
186
134
298
198
243
105
443
Transcript variant 1
NM006759
1527 bp, 509 AA
c.34A>G, p.Met12Val
Transcript variant 2
NM001001521
1494 bp, 498 AA
c.1A>G, p.?
B
Control
Mother
Father
Proband
Vinculin
long
short
UGP2
C
Fraction of short isoform
1.0
0.8
0.6
0.4
0.2
0.0
Control
Parents
Proband
D
Total UGP2 protein
1.5
1.0
0.5
0.0
Parents
Proband
E
Parents
Proband
Cells per cm²
6.0×10⁴
4.0×10⁴
2.0×10⁴
0
0
1
2
3
4
5
Days
F
UGP2
UGP2 /DAPI
Control
Mother
Father
Proband
G
UGP2 activity in fibroblast
UGP2 enzymatic activity (nmol/min/mg)
15
10
5
0
Control
Father
Mother
Proband

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**).

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 and fibroblasts obtained from family 1, detecting the housekeeping control vinculin and 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 blot quantification of the fraction of 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 blot 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 fibroblasts from heterozygous parents and homozygous proband from family 1, during a 5-day period, determined in three independent experiments. Error bars represent SEM. **F)** Immunocytochemistry on cultured control and UGP2 heterozygous and homozygous mutant fibroblasts derived from family 1, detecting UGP2 (red). Nuclei are stained with DAPI. Scale bar 50 μm. **G)** Enzymatic activity of UGP2 in control and UGP2 heterozygous and homozygous mutant fibroblasts 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


A
L ESC L NSC Fibro
150 kDa
50 kDa
Vinculin
UGP2
long
short
C
Long
Short
L ESC NSC Fibro
1 kb
0.85 kb
long
short
B
Protein
Fraction short isoform
ESC NSC Fibroblast
D
Transcript RT-PCR
E
Transcript qPCR
F
Fetal tissue RNA–seq
Fraction short UGP2
Muscle Arm
Muscle Back
Muscle Leg
Heart
Thymus
Lung Left
Lung Right
Intestine Large
Intestine Small
Liver
Stomach
Pancreas
Kidney
Spleen
Brain
G
A) first trimester
B) second trimester
GW23
C) third trimester
GW36
GW6
GW9
H
L Frontal Temporal Parietal Occipital Cerebellum L ESC
GW: 14 20 28 14 20 28 14 20 28 14 20 28 14 20 28
150 kDa
50 kDa
Vinculin
UGP2
long
short

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 (**Supplementary Figure 2J**). 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) (**Supplementary Figure 3A, 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 (**Supplementary Figure 3C**). All derived ESCs had a normal morphology and remained pluripotent as assessed by marker expression (**Supplementary Figure 3D, 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

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 blot quantification of the fraction of 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 and 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 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* expressions come 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 weeks 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.

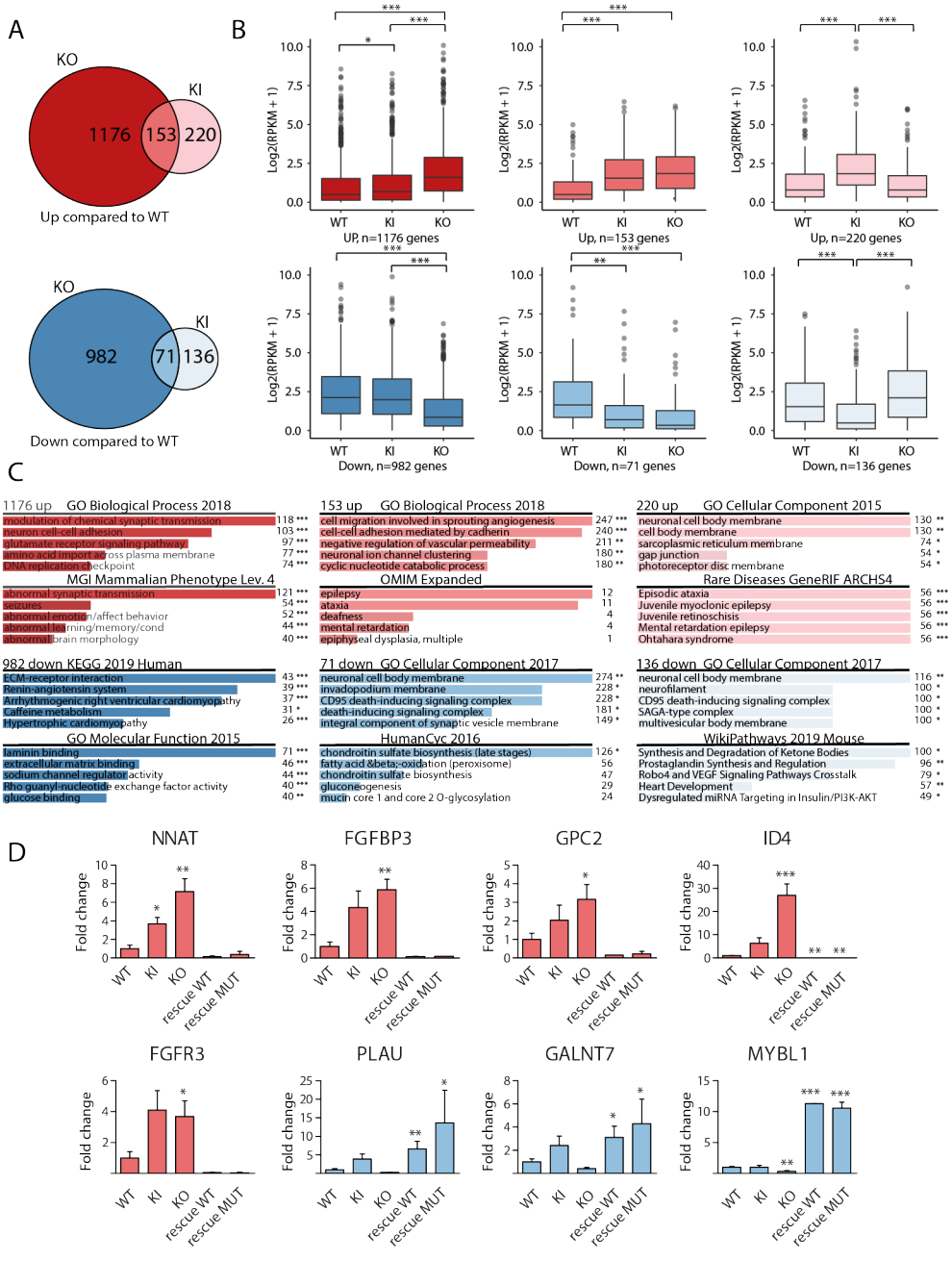
an essential gene in ESCs[52,53]. We differentiated wild type, KI and KO ESCs into NSCs, using dual SMAD inhibition (**Supplementary Figure 4 A-C**). Wild type cells could readily differentiate into NSCs, having a normal morphology and marker expression, whereas differentiation of KI and KO cells was 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 (**Supplementary Figure 4D, 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, Supplementary Figure 5, 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 (**Supplementary Figure 5C, 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 (**Supplementary Figure 5F**), 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, Supplementary Figure 5D, 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 (**Supplementary Figure 4F**). 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 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 (**Supplementary Figure 6**). 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.

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

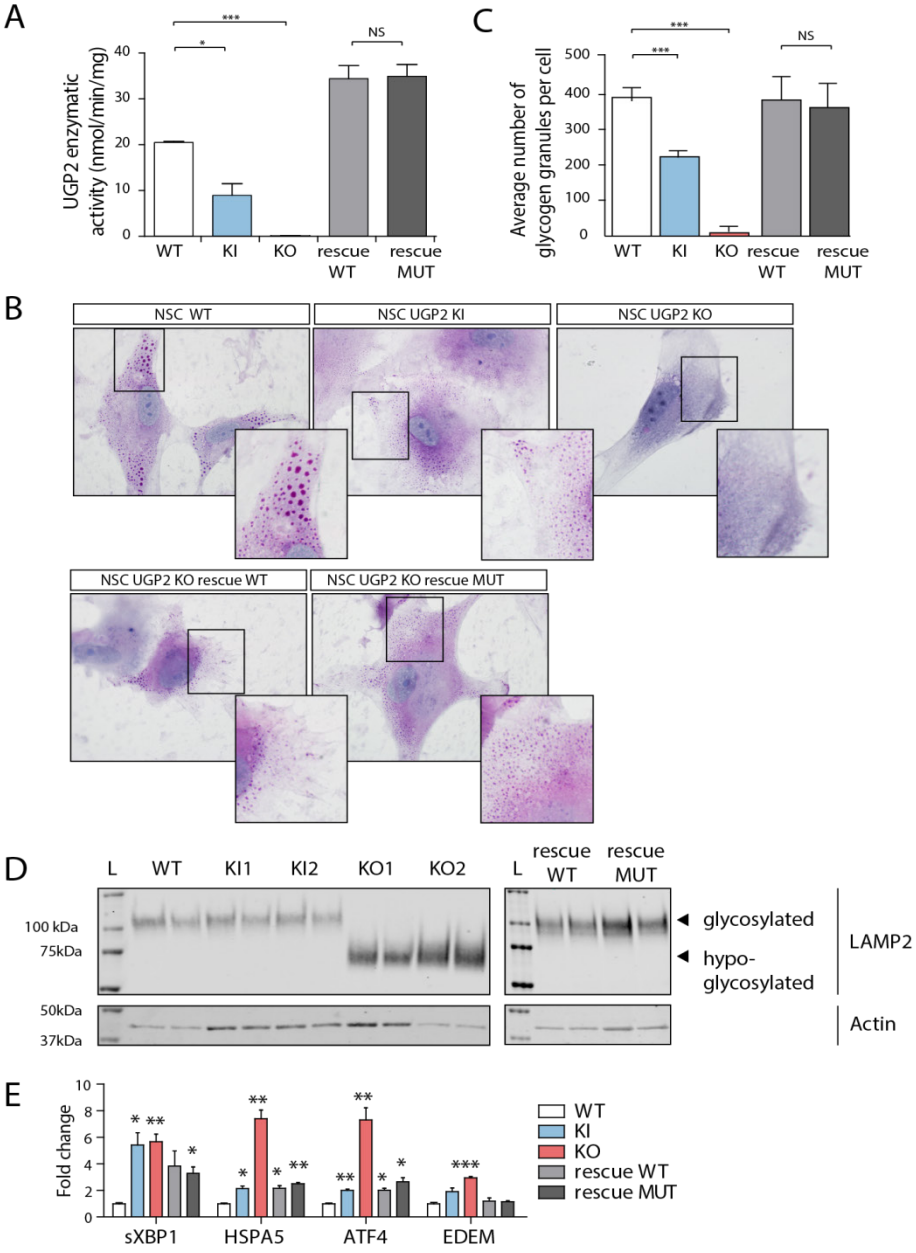
To investigate how reduced UGP2 expression levels in KO and KI cells would impact on NSC metabolism, we investigated the capacity to produce UDP-glucose in the presence of exogenously supplied glucose-1-phospate and UTP. KO NSCs showed a severely reduced ability to produce UDP-glucose (**Figure 5A)**. 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**),


A
KO
KI
1176
153
220
Up compared to WT
KO
KI
982
71
136
Down compared to WT
B
Log2(RPKM + 1)
WT
KI
KO
Up, n=1176 genes
Up, n=153 genes
Up, n=220 genes
Down, n=982 genes
Down, n=71 genes
Down, n=136 genes
C
1176 up GO Biological Process 2018
modulation of chemical synaptic transmission 118 ***
neuron cell-cell adhesion 103 ***
glutamate receptor signaling pathway 97 ***
amino acid import across plasma membrane 77 ***
DNA replication checkpoint 74 ***
153 up GO Biological Process 2018
cell migration involved in sprouting angiogenesis 247 ***
cell-cell adhesion mediated by cadherin 240 ***
negative regulation of vascular permeability 211 **
neuronal ion channel clustering 180 **
cyclic nucleotide catabolic process 180 **
220 up GO Cellular Component 2015
neuronal cell body membrane 130 **
cell body membrane 130 **
sarcoplasmic reticulum membrane 74 *
gap junction 54 *
photoreceptor disc membrane 54 *
MGI Mammalian Phenotype Lev. 4
abnormal synaptic transmission 121 ***
seizures 54 ***
abnormal emotion/affect behavior 52 ***
abnormal learning/memory/cond 44 ***
abnormal brain morphology 40 ***
OMIM Expanded
epilepsy 12
ataxia 11
deafness 4
mental retardation 4
epiphyseal dysplasia, multiple 1
Rare Diseases GeneRIF ARCHS4
Episodic ataxia 56 ***
Juvenile myoclonic epilepsy 56 ***
Juvenile retinoschisis 56 ***
Mental retardation epilepsy 56 ***
Ohtahara syndrome 56 ***
982 down KEGG 2019 Human
ECM-receptor interaction 43 ***
Renin-angiotensin system 39 ***
Arrhythmogenic right ventricular cardiomyopathy 37 ***
Caffeine metabolism 31 *
Hypertrophic cardiomyopathy 26 ***
71 down GO Cellular Component 2017
neuronal cell body membrane 274 **
invadopodium membrane 228 *
CD95 death-inducing signaling complex 228 *
death-inducing signaling complex 181 *
integral component of synaptic vesicle membrane 149 *
136 down GO Cellular Component 2017
neuronal cell body membrane 116 **
neurofilament 100 *
CD95 death-inducing signaling complex 100 *
SAGA-type complex 100 *
multivesicular body membrane 100 *
GO Molecular Function 2015
laminin binding 71 ***
extracellular matrix binding 46 ***
sodium channel regulator activity 44 ***
Rho guanyl-nucleotide exchange factor activity 40 ***
glucose binding 40 **
HumanCyc 2016
chondroitin sulfate biosynthesis (late stages) 126 *
fatty acid β-oxidation (peroxisome) 56
chondroitin sulfate biosynthesis 47
gluconeogenesis 29
mucin core 1 and core 2 O-glycosylation 24
WikiPathways 2019 Mouse
Synthesis and Degradation of Ketone Bodies 100 *
Prostaglandin Synthesis and Regulation 96 **
Robo4 and VEGF Signaling Pathways Crosstalk 79 *
Heart Development 57 **
Dysregulated miRNA Targeting in Insulin/PI3K-AKT 49 *
D
NNAT
FGFBP3
GPC2
ID4
FGFR3
PLAU
GALNT7
MYBL1
Fold change
WT
KI
KO
rescue WT
rescue MUT

but a more strongly reduced activity in NSCs compared to wild type (**Figure 5A**), correlating with total UGP2 expression levels (**Supplementary Figure 4D, E**). Surprisingly, contrary to KO NSCs, KO ESC showed some residual capacity to produce UDP-glucose despite the complete absence of UGP2 (**Supplementary Figure 7A**). 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 (**Supplementary Figure 7B**). 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 (**Supplementary Figure 7C, 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 (**Supplementary Figure 7D)**. 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 (**Supplementary Figure 7E**).

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

Figure 4. RNA-seq of UGP2 mutant H9-derived neural stem cells. 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 1.5 × the IQR (*$p < 0.05$; **$p < 0.01$, ***$p < 0.001$, unpaired *t* test, two tailed). **C)** Enrichment analysis using Enrichr [52] of up- or downregulated genes in NSCs from **(A)** for selected gene ontology sets, showing the five 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 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, from **(A)**. Error bars represent SEM; (*$p < 0.05$; **$p < 0.01$, ***$p < 0.001$, unpaired *t* test, one-tailed).

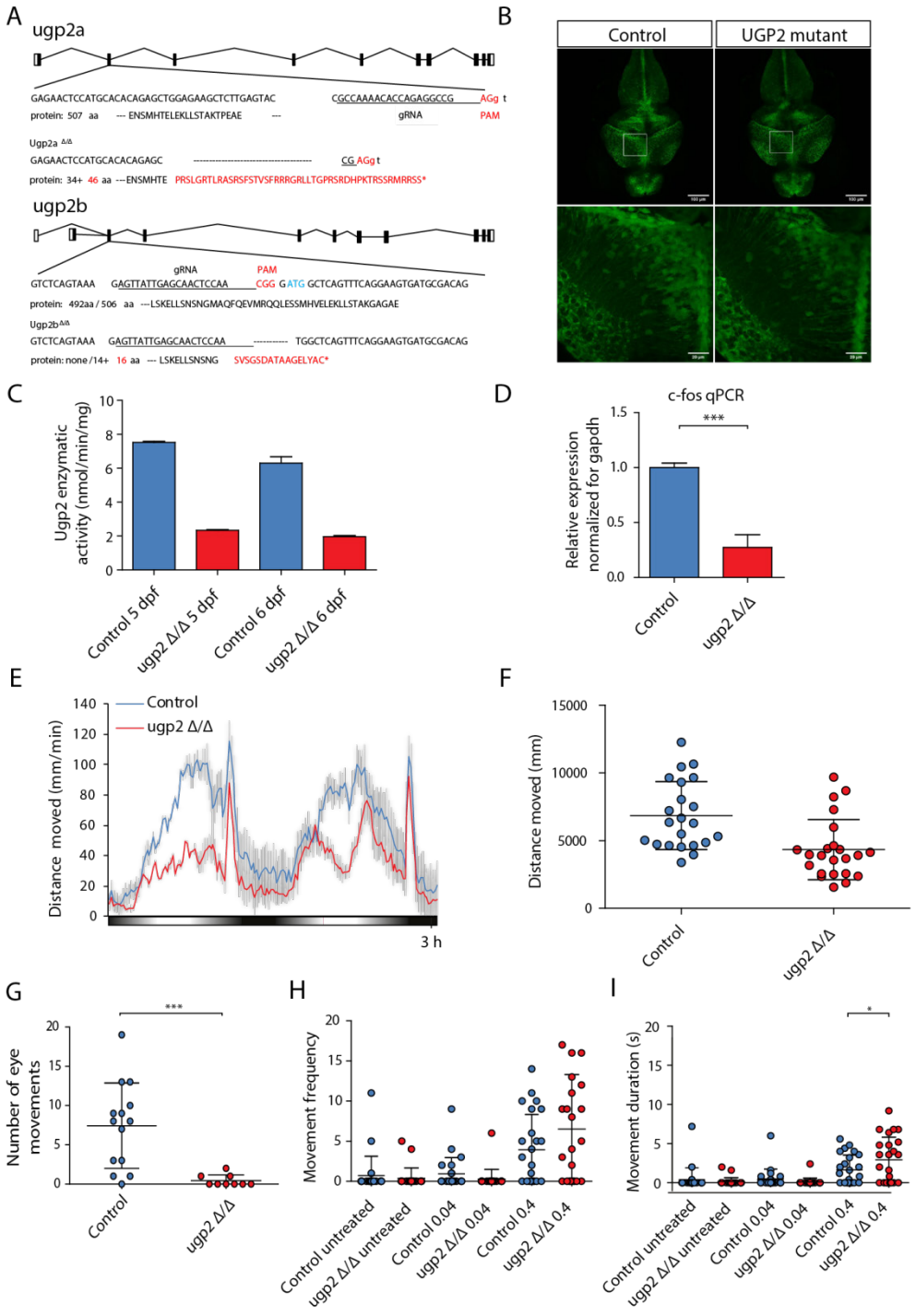

A
UGP2 enzymatic activity (nmol/min/mg)
WT
KI
KO
rescue WT
rescue MUT
C
Average number of glycogen granules per cell
B
NSC WT
NSC UGP2 KI
NSC UGP2 KO
NSC UGP2 KO rescue WT
NSC UGP2 KO rescue MUT
D
L
WT
KI1
KI2
KO1
KO2
100 kDa
75kDa
50kDa
37kDa
glycosylated
hypo-glycosylated
LAMP2
Actin
E
Fold change
sXBP1
HSPA5
ATF4
EDEM

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 (**Supplementary Figure 7F**). 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 (**Supplementary Figure 7G**), 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 (**Supplementary Figure 7H**). 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.

Figure 5. Metabolic changes upon UGP2 loss. A) UGP2 enzymatic activity in WT, UGP2 KI, KO and KO NSCs rescued with WT or MUT (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 rescued 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 WT or MUT (Met12Val) isoform 1 of UGP2. **C)** Quantification of the number of glycogen granules per cell in WT, UGP2 KI, KO and rescued NSCs, after 48 h 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 H9-derived NSCs that are WT, UGP2 KI, KO and KO cells rescued with WT or MUT (Met12Val) isoform 1 of UGP2. Glycosylated LAMP2 runs at ~110 kDa, whereas hypo-glycosylated LAMP2 is detected around 75 kDa. The absence of 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 (see Supplementary Fig. 5d, online resource). **E)** qRT-PCR expression analysis for UPR marker genes (spliced *XBP1*, *HSPA5*, *ATF4* and *EDEM*) in WT, KI, KO and rescued 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.

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.

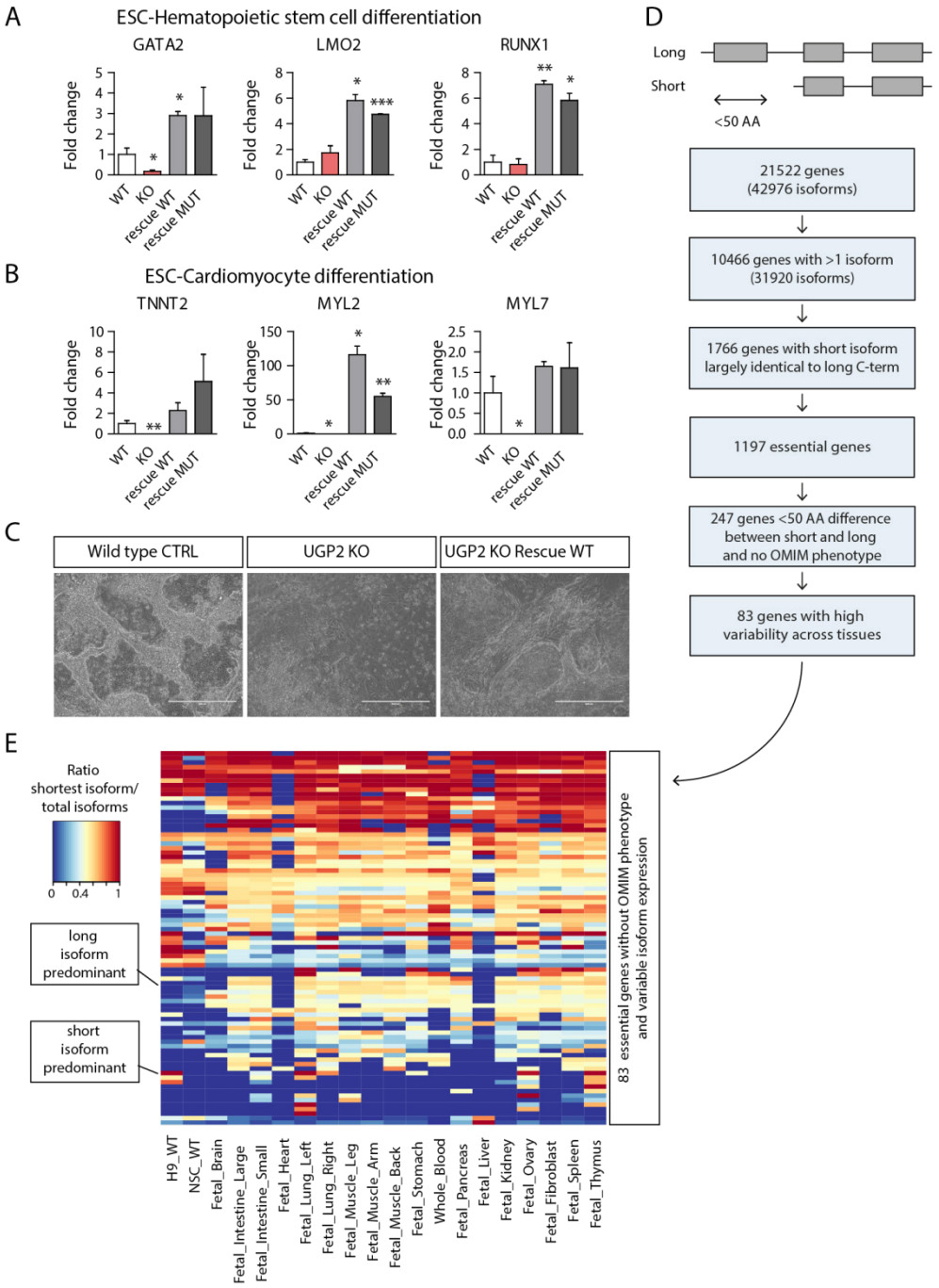

A
ugp2a
GAGAACTCCATGCACACAGAGCTGGAGAAGCTCTTGAGTAC
CGCCAAAACACCAGAGGCCG
AGg t
protein: 507 aa
--- ENSMHTELEKLLSTAKTPEAE
---
gRNA
PAM
Ugp2a Δ/Δ
GAGAACTCCATGCACACAGAGC
CG AGg t
protein: 34+ 46 aa ---ENSMHTE PRSLGRTLRASRSFSTVSFRRRGRLLTGPRSRDHPKTRSSRMRRSS*
ugp2b
gRNA
PAM
GTCTCAGTAAA GAGTTATTGAGCAACTCCAA CGG GATG GCTCAGTTTCAGGAAGTGATGCGACAG
protein: 492aa / 506 aa ---LSKELLSNSNGMAQFQEVMRQQLESSMHVELEKLLSTAKGAGAE
Ugp2b Δ/Δ
GTCTCAGTAAA GAGTTATTGAGCAACTCCAA ---------- TGGCTCAGTTTCAGGAAGTGATGCGACAG
protein: none /14+ 16 aa --- LSKELLSNSNG SVSGSDATAAGELYAC*
B
Control
UGP2 mutant
C
Ugp2 enzymatic activity (nmol/min/mg)
Control 5 dpf
ugp2 Δ/Δ 5 dpf
Control 6 dpf
ugp2 Δ/Δ 6 dpf
D
c-fos qPCR
***
Relative expression normalized for gapdh
Control
ugp2 Δ/Δ
E
Control
ugp2 Δ/Δ
Distance moved (mm/min)
3 h
F
Distance moved (mm)
Control
ugp2 Δ/Δ
G
***
Number of eye movements
Control
ugp2 Δ/Δ
H
Movement frequency
Control untreated
ugp2 Δ/Δ untreated
Control 0.04
ugp2 Δ/Δ 0.04
Control 0.4
ugp2 Δ/Δ 0.4
I
*
Movement duration (s)
Control untreated
ugp2 Δ/Δ untreated
Control 0.04
ugp2 Δ/Δ 0.04
Control 0.4
ugp2 Δ/Δ 0.4

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 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].

Figure 6. Zebrafish disease modeling. A) Schematic drawing of the *ugp2a* and *ugp2b* loci in zebrafish and the generated mutations. **B)** Confocal images (maximum projection of confocal Z-stacks) of the brain of wild type (left) and *ugp2a*Δ/Δ; *ugp2b*Δ/Δ mutant zebrafish larvae (right), both in an *slc1a2b*-citrine reporter background, at 4 days post-fertilization (dpf). The 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 h 12 min), $n=24$ larvae per genotype. Gray 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-min 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-min observation. Each dot represents a single larva; results of two experiments are shown, within total 24 larvae per condition. **I)** Quantification of the movement duration 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-min observation. Each dot represents a single larva; results of two experiments are shown, with in total 24 larvae per condition. *$p<0.05$, two-way ANOVA with Bonferoni pos*t* test.

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 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


A
ESC-Hematopoietic stem cell differentiation
GATA2
LMO2
RUNX1
Fold change
WT
KO
rescue WT
rescue MUT
B
ESC-Cardiomyocyte differentiation
TNNT2
MYL2
MYL7
C
Wild type CTRL
UGP2 KO
UGP2 KO Rescue WT
D
Long
Short
<50 AA
21522 genes
(42976 isoforms)
10466 genes with >1 isoform
(31920 isoforms)
1766 genes with short isoform
largely identical to long C-term
1197 essential genes
247 genes <50 AA difference
between short and long
and no OMIM phenotype
83 genes with high
variability across tissues
E
Ratio
shortest isoform/
total isoforms
0
0.4
1
long
isoform
predominant
short
isoform
predominant
83 essential genes without OMIM phenotype
and variable isoform expression
H9_WT
NSC_WT
Fetal_Brain
Fetal_Intestine_Large
Fetal_Intestine_Small
Fetal_Heart
Fetal_Lung_Left
Fetal_Lung_Right
Fetal_Muscle_Leg
Fetal_Muscle_Arm
Fetal_Muscle_Back
Fetal_Stomach
Whole_Blood
Fetal_Pancreas
Fetal_Liver
Fetal_Kidney
Fetal_Ovary
Fetal_Fibroblast
Fetal_Spleen
Fetal_Thymus

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.

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 structure 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 isoforms differ by less than 50 amino acids at the N-terminal.

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 150bp 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 100bp 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 extract-

ed 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 phosphorothiorate 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 differen-

tiation 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 knockout 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, 42,976 transcripts from 21,522 genes (Human genes GRCh38.p12) were extracted using BioMart of Ensembl (biomaRt R package). 11,056 out of 21,522 genes had only 1 transcript and the remaining 31,920 transcripts from 10,466 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). 1,766 genes meet these criteria. We then filtered these genes for published essential genes[6], leaving us with 1,197 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 1,197 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 5 Mb 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 $10^{\wedge 3}$ to $10^{\wedge 8}$ 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 $CaCl_2 \cdot 2H_2O$, 9.78 g $MgCl_2 \cdot 6\ H_2O$). 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 publically 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. We would like to thank Reviewer 1 for proposing the name “Barakat-Perenthalersyndrome of developmental epileptic encephalopathy” for this new disorder. DP was supported by an Erasmus+Traineeship Programme. MAS was supported by the King Saud University (RSP-2019/38). AGES was supported by the Yale Center for Mendelian Genomics (NIH Grant M#UM1HG006504-05). HH is supported by the Rosetree Trust, Ataxia UK, MSA Trust, Brain Research UK, Muscular Dystrophy UK, Muscular Dystrophy Association, Higher Education Commission of Pakistan, The MRC (MR/S01165X/1, MR/S005021/1, G0601943), Wellcome Trust (WT093205MA, WT104033AIA, Synaptopathies Strategic Award, 165908) and National Institute for Health Research University College London Hospitals Biomedical Research Centre. Families 5–8 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, 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 Scientifc Research (ZonMW Veni, Grant 91617021), a NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation.

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 McTague. *et al*.The genetic landscape of the epileptic encephalopathies of infancy and childhood. Lancet Neurol 15:304–316 (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 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 antiepileptic 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.

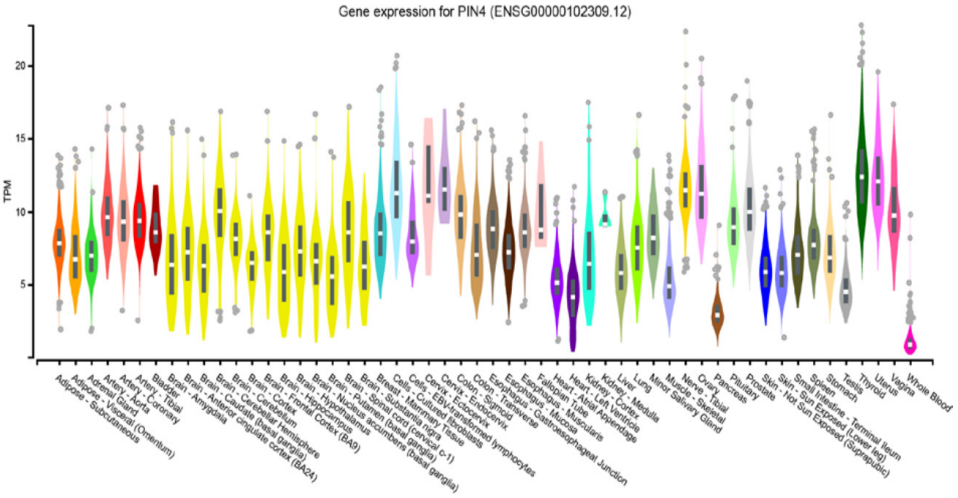
Individual 20: 2 year old male. Regression from 3 month of age with neurodevelopmental delay. Focal onsetseizures, generalized seizures, epileptic encephalopathy. Abnormal EEG Hypsarrhythmia. Brain MRI showing brain atrophy. Consanguineous parents. 2 affected sisters deceased at 8 days and 1 year. WES identified 2 homozygous pathogenic variants in TTLL5 and ARMC4, consistent with a genetic

diagnosis of autosomal recessive cone-rod dystrophy type 19 and autosomal recessive primary ciliary dyskinesia type 23. None of these however explain the neurological phenotype. Upon re-analysis, the recurrent homozygous variant in *UGP2* (UGP2 NM_001001521.1:c.1A>G NM_001001521.1:p.Met1?) was identified. Both parents were heterozygous carriers.

Individual 22: Female. Epileptic encephalopathy, regression, NDD.

Supplementary Note

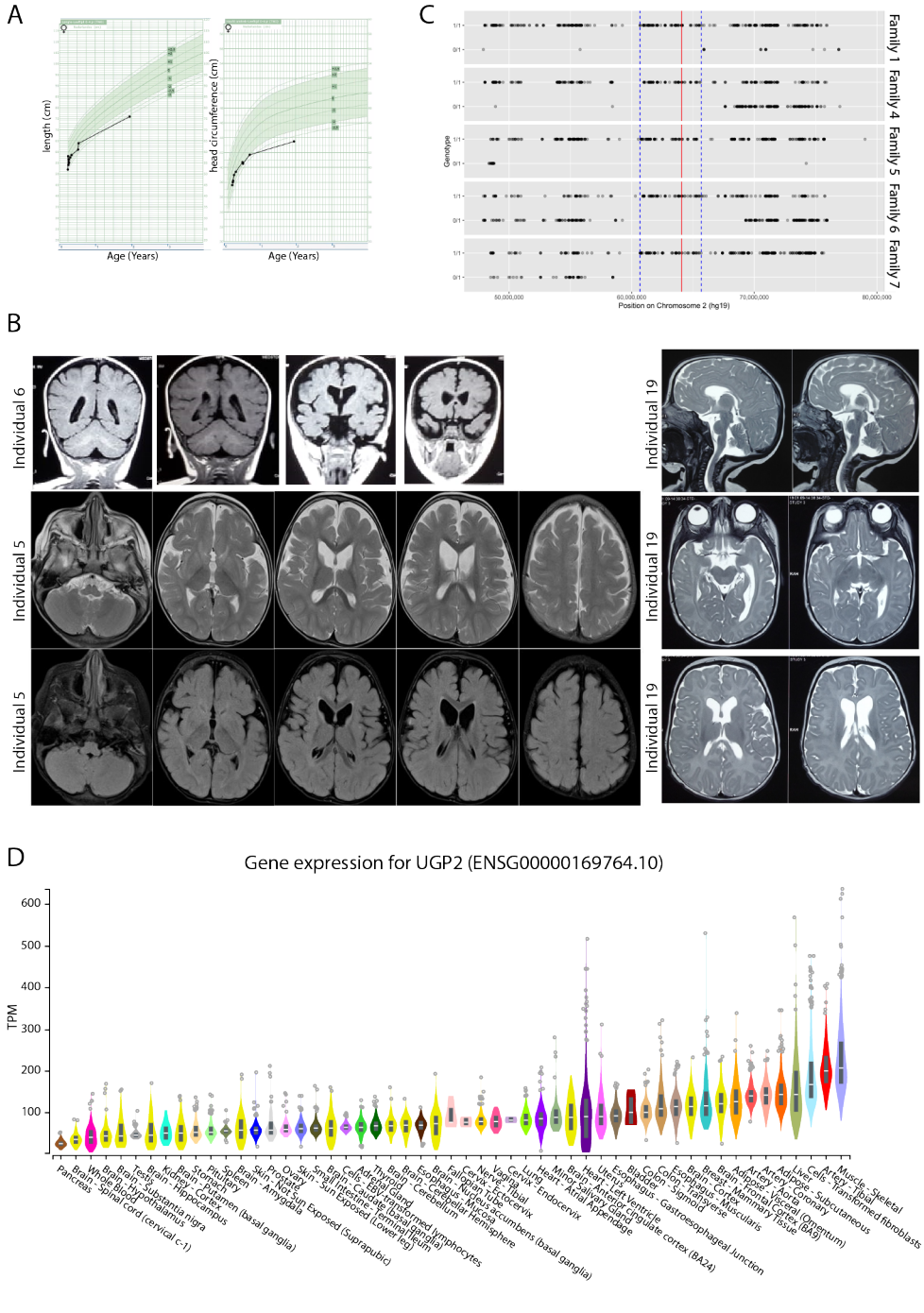
The disease we here describe is caused by the loss of an isoform of an essential gene, due to an alteration affecting an isoform specific start codon. To investigate whether this same mechanism could apply to other essential genes that were previously not implicated in human genetic disease, we investigated 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 microtubules[2]. *PIN4* is widely expressed amongst tissues, including different brain regions, according to data from the GTEX portal (**Figure**)[1]. Together, this makes *PIN4* a strong candidate gene for a novel neurodevelopmental disorder.



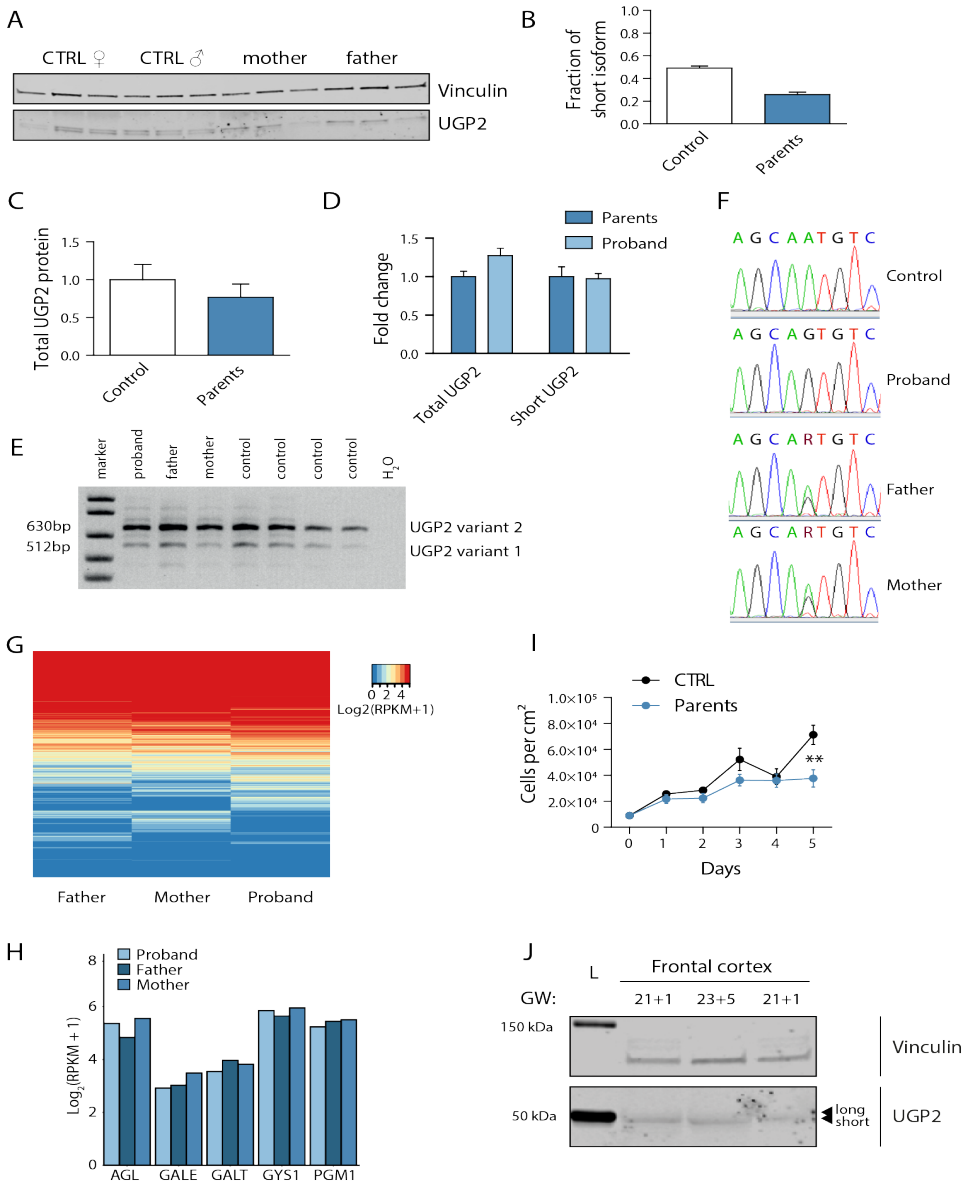
[1] Consortium GT (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45: 580- 585

[2] Thiele A, Krentzlin K, Erdmann F, Rauh D, Hause G, Zerweck J, Kilka S, Posel S, Fischer G, Schutkowski Met al (2011) Parvulin 17 promotes microtubule assembly by its peptidyl-prolyl cis/trans isomerase activity. J Mol Biol 411: 896-90

Supplementary Figures

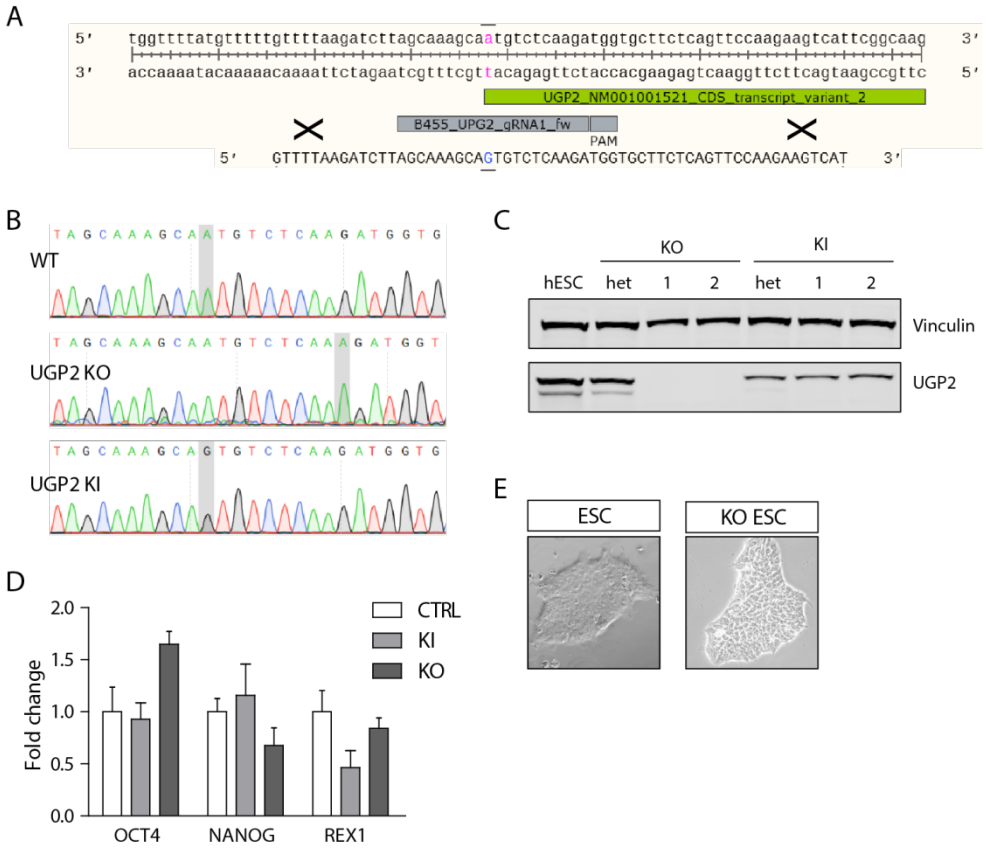


Supplementary Figure 1. A) Growth chart from individual 1 for length (left) and head circumference (right) in cm. Reference chart from the Dutch population are used (TNO) and regions between -2 and + 2 SD are shaded. **B)** MRI studies of individual 6, individual 5 (at the age of 12 months), and individual 19 (at the age of 4 months), 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 portal [1] for tissues and cell lines. Samples are sorted with the highest median TPM on the right. Outliers are indicated by dots.

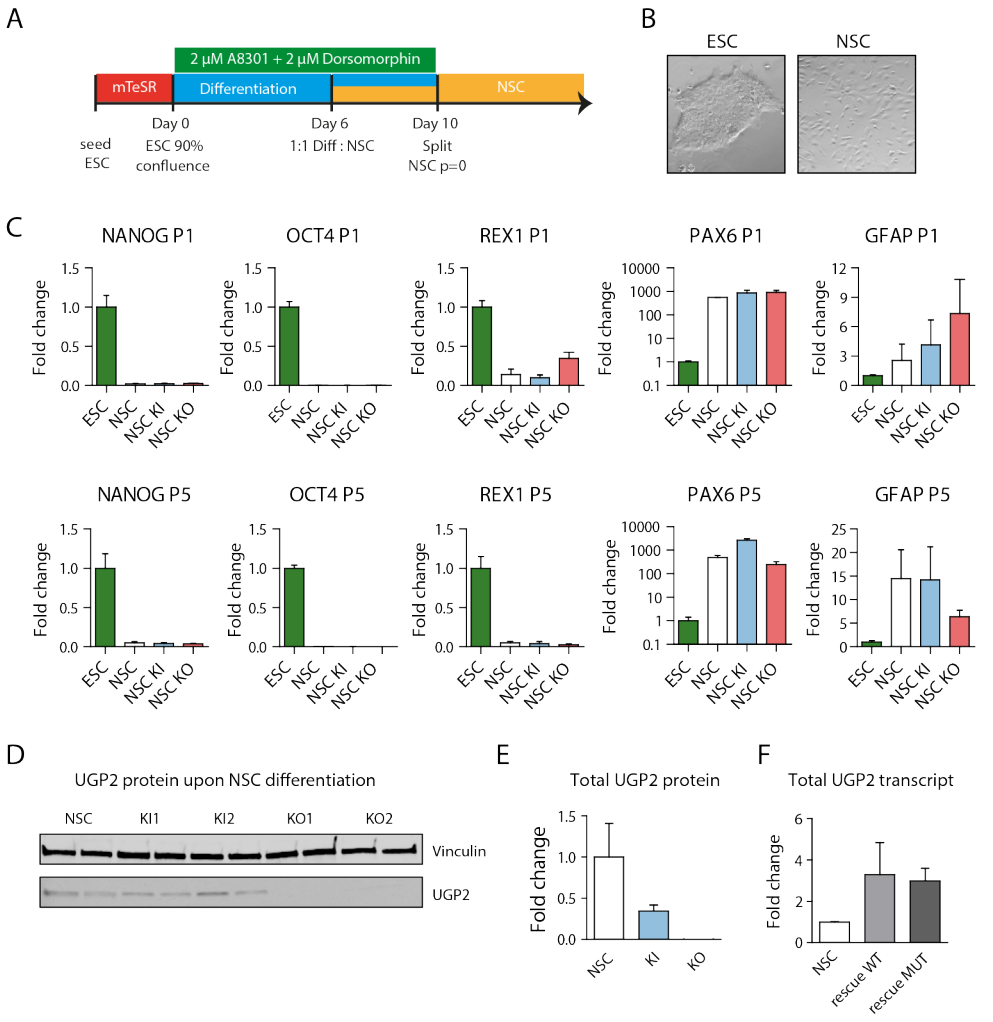


Supplementary Figure 2. **A)** Western blotting of cellular extracts derived from control fibroblasts or fibroblasts obtained from heterozygous parents of family 2, detecting the house keeping control vinculin or UGP2. Note the two separated isoforms of *UGP2* that have a similar intensity in wild type 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 blot 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,

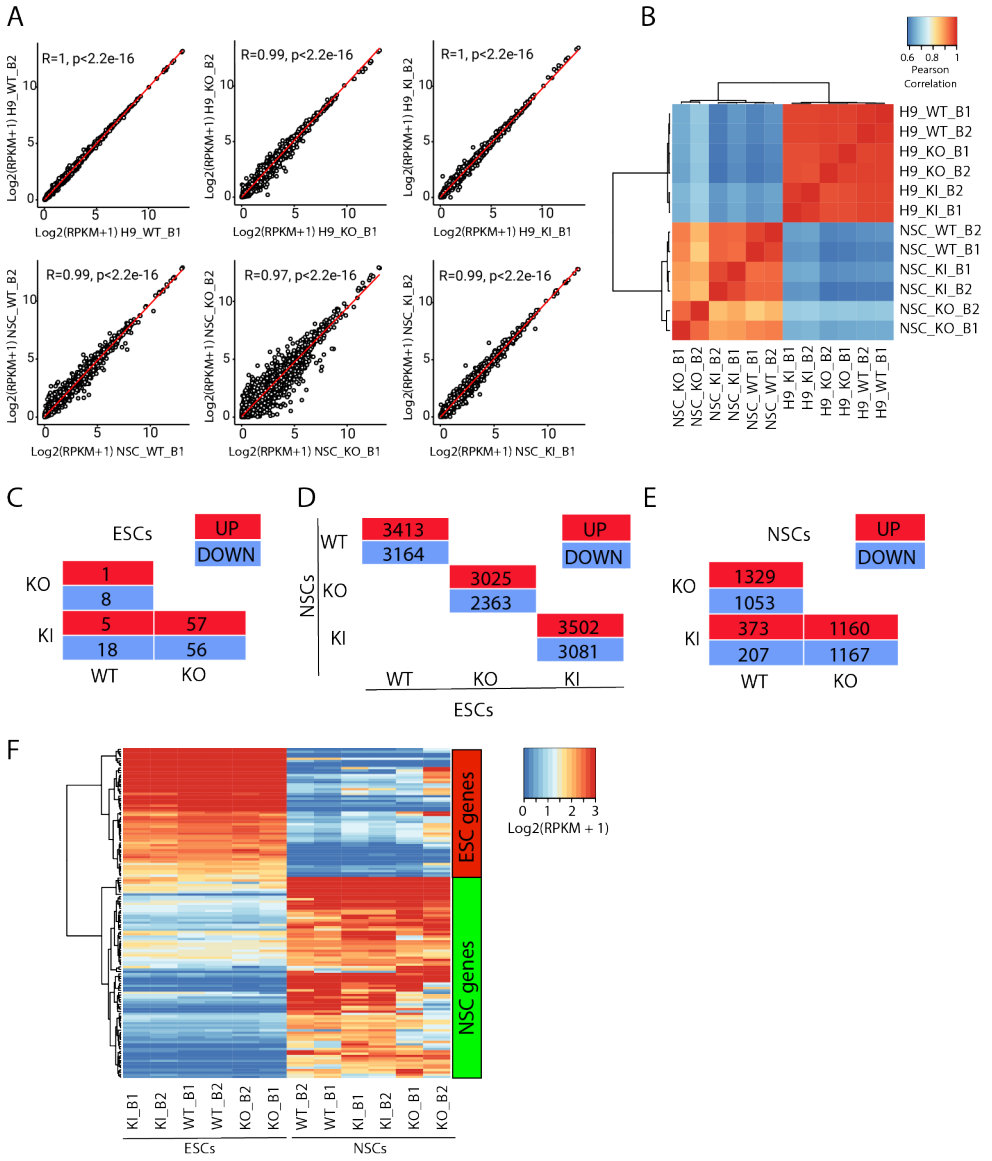
unpaired t-test, two-tailed. **D)** qRT-PCR analysis of total *UGP2* or the short isoform in fibroblast from heterozygous parents or homozygous proband from family 1, normalized for the housekeeping control *TBP*. The mean fold change compared to heterozygous parents of two biological replicates and 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.



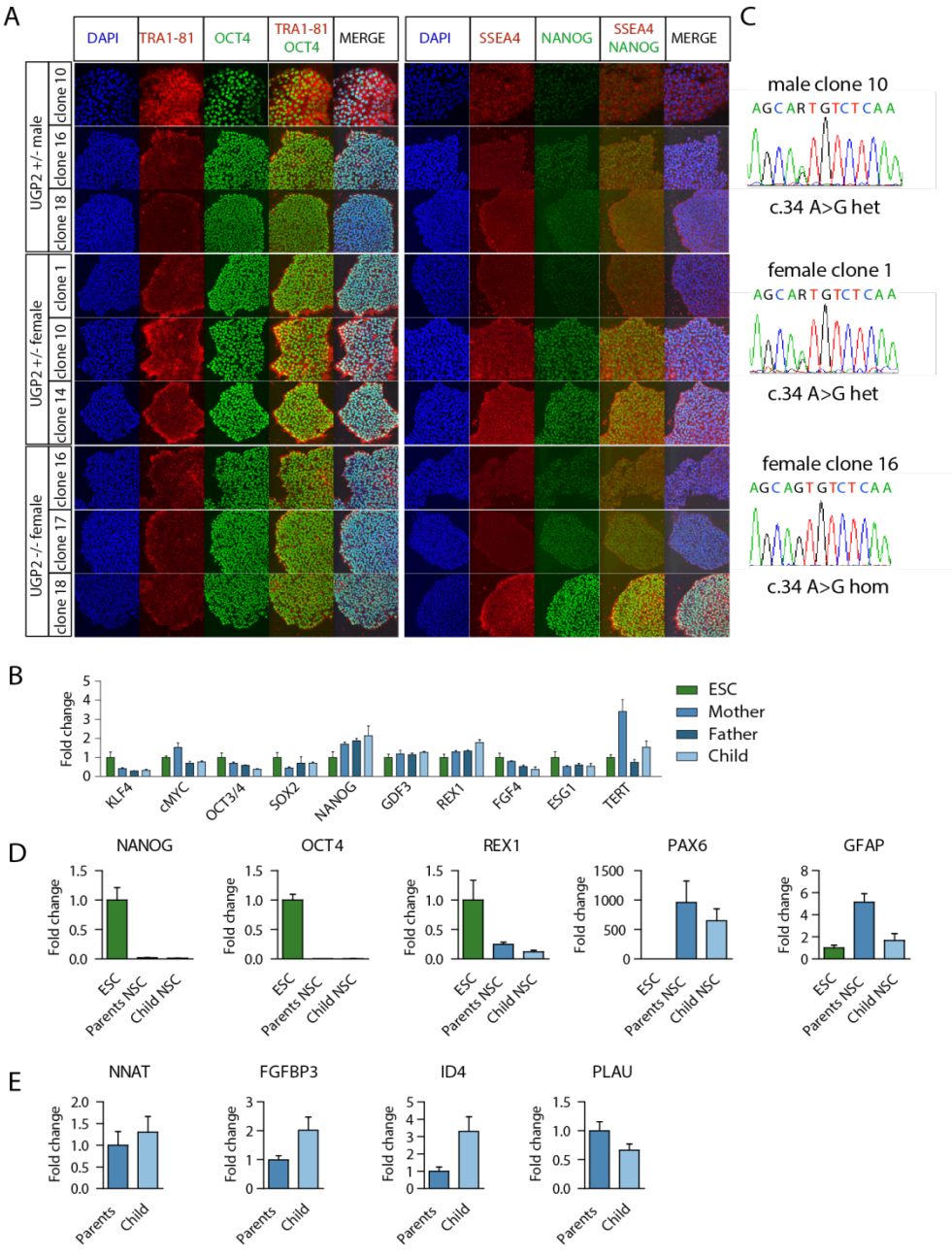
Supplementary Figure 3.Generation of mutant *UGP2* H9 cell lines. A) Nucleotide sequence encompassing the ATG of *UGP2* transcript isoform 2. Indicated are the coding sequence, the location of the gRNA, PAM sequence and ssODN used to introduce the C.1A>G, p.? mutation. **B)** Sanger sequencing traces 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 blot 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.



Supplementary Figure 4. 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 blot, 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.

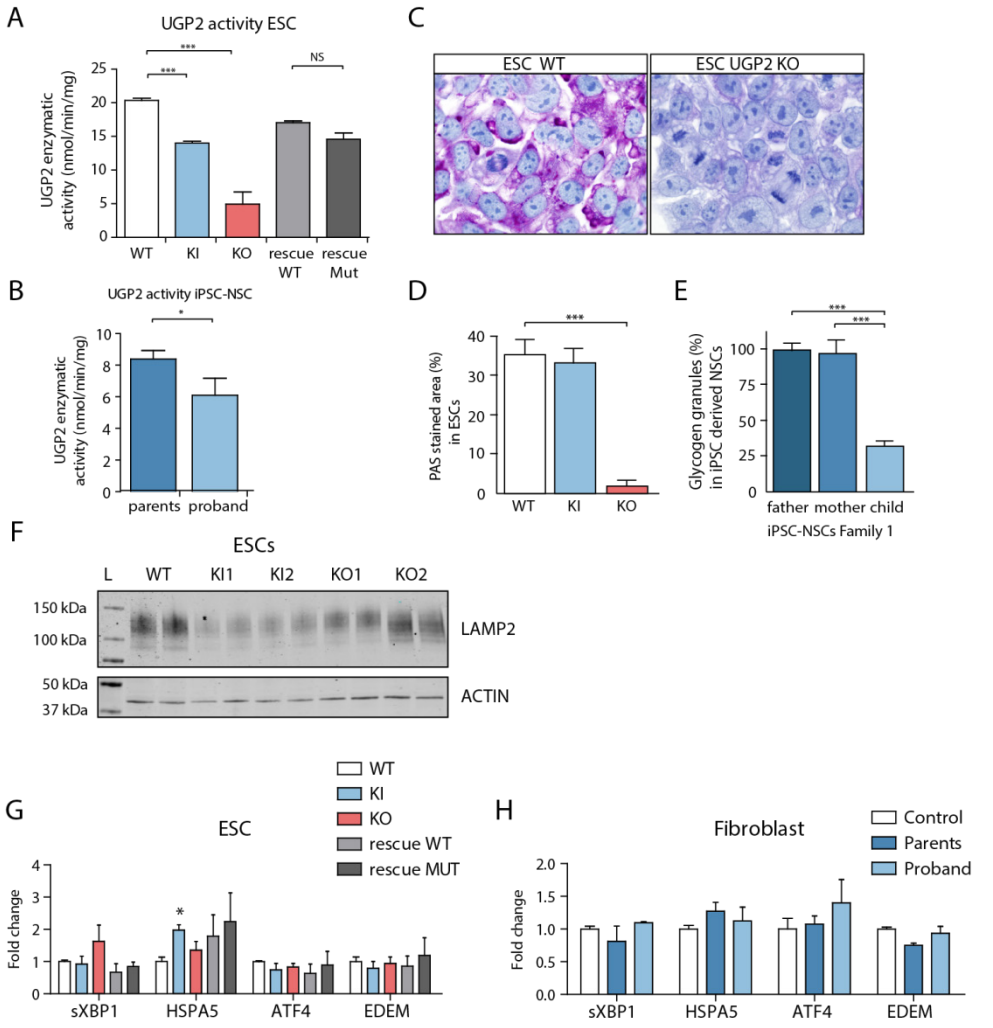


Supplementary Figure 5. RNA-seq. A) Scatter plot showing the pair wise correlation between biological replicates. **B)** Heat map displaying Pearson correlation between biological replicates. **C)** Table summarizing up- (FDR1) and down regulated (FDR<- 1) genes in WT, KO and KI ESCs. **D)** Table summarizing up- (FDR1) and down regulated (FDR<- 1) genes in WT, KO and KI ESC upon differentiation in NSCs. **E)** Table summarizing up- (FDR1) and down regulated (FDR<- 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)



Supplementary Figure 6. UGP2 mutant induced pluripotent stem cells. 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 anal-

ysis 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.



Supplementary Figure 7. A) UGP2 enzymatic activity in WT, UGP2 KI, KO and KO ESCs rescued with wild type 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.

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 *gnom*AD

Supplementary Table 6: Genome-wide homology search results

Supplementary Table 7: *gnom*AD 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

Supplementary Movie 1: Affected individual from family 11

Supplementary Movie 2: Wild type zebrafish eye movements

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

Affiliations

Elena Perenthaler[1], Anita Nikoncuk[1], Soheil Yousef[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 IJcken7, Walter G. de Valk[1], Evita Medicivan 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 Azri[15], ErikJan Kamsteeg[16], Majid Mojarrad[17,18,19] , Atieh Eslahi[17,20], Zaynab Khazaei[19], Fateme Massinaei Darmiyan[21], Mohammad Doosti[22], Ehsan Ghayoor Karimiani[23,24], Jana Vandrovcova[25], Faisal Zafar[26], Nuzhat Rana[26], Krishna K. Kandaswamy[27], Jozef Hertecant[28], Peter Bauer[27], Mohammed A. AlMuhaizea[29], Mustafa A. Salih[30], Mazhor Aldosary[2], Rawan Almass[2], Laila AlQuait[2], Wafa Qubbaj[31], Serdar Coskun[31], Khaled O. Alahmadi[32], Muddathir H. A. Hamad[30], Salem Alwadaee[31], Khalid Awartani[33], Anas M. Dababo[31], Futwan Almohanna[34], Dilek Colak[35], Mohammadreza Dehghani[36,37], Mohammad Yahya Vahidi Mehrjardi[38], Murat Gunel[39], A. Gulhan ErcanSencicek[39,40], Gouri Rao Passi[41], Huma Arshad Cheema[42], Stephanie Efthymiou[25], Henry Houlden[25], Aida M. BertoliAvella[27], Alice S. Brooks[1], Kyle Retterer[12], Reza Maroofan[25], 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, Novi Sad, Serbia. [4]iPS Cell Core Facility, Erasmus MC University Medical Center, Rotterdam, The Netherlands. [5]Department of (Neuro)Pathology, Amsterdam Neuroscience, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands. [6]Stichting Epilepsie Instellingen Nederland (SEIN), Zwolle, 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, ON L8S 4J9, Canada. [10]Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA 94035, USA. [11]Division of Medical Genetics, Department of Pediatrics, 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 Human Genetics, Radboud University Medical Centre, 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 Medical Genetics, Next Generation Genetic Polyclinic, Mashhad, Iran. [23]Molecular and Clinical Sciences Institute, St. George's University of London, Cranmer Terrace, London SW17 0RE, UK. [24]Innovative Medical Research Center, Mashhad Branch, Islamic Azad University, Mashhad, Iran. [25]Department of Neuromuscular Disorders, UCL Queen Square Institute of Neurology, London WC1N 3BG, UK. [26]Department of Paediatric Neurology, Children's Hospital and Institute of Child Health, Multan 60000, Pakistan. [27]CENTOGENE AG, Rostock, Germany. [28]Department of Pediatrics, Tawam Hospital, and College of Medicine and Health Sciences, UAE University, Al-Ain, UAE. [29]Department of Neurosciences, King Faisal Specialist Hospital and Research Centre, Riyadh 11211, Kingdom of Saudi Arabia. [30]Neurology Division, Department of Pediatrics, College of Medicine, King Saud University, Riyadh 11461, Kingdom of Saudi Arabia. [31]Department of Pathology and Laboratory Medicine, King Faisal Specialist Hospital and Research Centre, Riyadh 11211, Kingdom of Saudi Arabia. [32]Radiology Department, King Faisal Specialist Hospital and Research Centre, Riyadh 11211, Kingdom of Saudi Arabia. [33]Obstetrics/Gynecology Department, King Fais-

al Specialist Hospital and Research Centre, Riyadh 11211, Kingdom of Saudi Arabia. [34]Department of Cell Biology, King Faisal Specialist Hospital and Research Centre, Riyadh 11211, Kingdom of Saudi Arabia. [35]Department of Biostatistics, Epidemiology and Scientifc Computing, King Faisal Specialist Hospital and Research Centre, Riyadh 11211, Kingdom of Saudi Arabia. [36]Medical Genetics Research Center, Shahid Sadoughi University of Medical Sciences, Yazd, Iran. [37]Yazd Reproductive Sciences Institute, Shahid Sadoughi University of Medical Sciences, Yazd, Iran. [38]Diabetes Research Center, Shahid Sadoughi University of Medical Sciences, Yazd, Iran. [39]Department of Neurosurgery, Program On Neurogenetics, Yale School of Medicine, Yale University, New Haven, CT, USA. [40]Masonic Medical Research Institute, Utica, NY, USA. [41]Department of Pediatrics, Pediatric Neurology Clinic, Choithram Hospital and Research Centre, Indore, Madhya Pradesh, India. [42]Pediatric Gastroenterology Department, Children's Hospital and Institute of Child Health, Lahore, Pakistan.

Chapter 3

Meta-analysis of putative enhancers in fetal brain

Soheil Yousefi[1], Ruizhi Deng[1*], Kristina Lanko[1*], Eva Medico Salsench[1*], Anita Nikoncuk[1*], Herma C. van der Linde[1], Elena Perenthaler[1], Tjakko van Ham[1], Eskeatnaf Mulugeta[2#] and Tahsin Stefan Barakat[1#.] Comprehensive multi-omics integration identifies differentially active enhancers during human brain development with clinical relevance. *Genome Medicine. 2021; 13, 162*.

[1]Department of Clinical Genetics, Erasmus MC University Medical Center, Rotterdam, The Netherlands

[2]Department of Cell Biology, Erasmus MC University Medical Center, Rotterdam, The Netherlands

*contributed equally

#corresponding authors

Non-coding regulatory elements (NCREs), such as enhancers, play a crucial role in gene regulation and genetic aberrations in NCREs can lead to human disease, including brain disorders. The human brain is a complex organ that is susceptible to numerous disorders; many of these are caused by genetic changes, but a multitude remain currently unexplained. Understanding NCREs acting during brain development has the potential to shed light on previously unrecognised genetic causes of human brain disease. Despite immense community-wide efforts to understand the role of the non-coding genome and NCREs, annotating functional NCREs remains challenging.

Here we performed an integrative computational analysis of virtually all currently available epigenome data sets related to human fetal brain.

Our in-depth analysis unravels 39,709 differentially active enhancers (DAEs) that show dynamic epigenomic rearrangement during early stages of human brain development, indicating likely biological function. Many of these DAEs are linked to clinically relevant genes, and functional validation of selected DAEs in cell models and zebrafish confirms their role in gene regulation. Compared to enhancers without dynamic epigenomic rearrangement, DAEs are subjected to higher sequence constraints in humans, have distinct sequence characteristics and are bound by a distinct transcription factor landscape. DAEs are enriched for GWAS loci for brain related traits and for genetic variation found in individuals with neurodevelopmental disorders, including autism.

This compendium of high-confidence enhancers will assist in deciphering the mechanism behind developmental genetics of human brain and will be relevant to uncover missing heritability in human genetic brain disorders.

Introduction

Non-coding regulatory elements (NCREs), such as enhancers, play a pivotal role in gene regulation[1,2]. Enhancers ensure correct spatio-temporal gene expression, and it is increasingly recognized that genetic aberrations disturbing enhancer function can lead to human disease, including brain disorders[3-6]. Such non-coding genetic variants are expected to explain a considerable fraction of so-called missing heritability (e.g. the absence of a genetic diagnosis despite a high genetic clinical suspicion). These developments are pushing genetic diagnostic investigations to shift from whole-exome sequencing to whole-genome sequencing, and the number of potentially pathogenic non-coding variants found in patients is expected to rise[4]. It is therefore of urgent clinical interest to understand where functionally relevant non-coding sequences are located in the human genome, as this will help to interpret the effects on health and disease.

Despite tremendous progress over the last decades, our understanding of the underlying mechanisms of enhancer biology remains limited due to challenges in annotating functional enhancers genome-wide. Large scale community driven efforts [7-11] and an uncountable plethora of individual studies have produced a vast amount of epigenome data sets, such as profiles of histone modifications, chromatin accessibility and chromatin interactions for different human tissues and cell types, that can be used to predict putative enhancers at large scale. More recently, new technologies such as massively parallel reporter assays and CRISPR-Cas9 based screens have entered the stage[12-14], providing novel means to directly test the functionality of non-coding regions. In addition, computational prediction algorithms[15, 16], trained on epigenome and experimental data, are improving the capability to predict functional sequences and the effects of variants in these regions.

One of the inherent problems with this increasing amount of data is the difficulty in keeping track of individual data sets and the ability to integrate data from various sources. Usually, individual studies focus on a limited number of cell types or tissues and compare their findings to a small number of previously published data sets. Although this is a logical step, it does not leverage the potential to fine-tune enhancer predictions which integrating all available enhancer data could have. This is illustrated by our previous findings that the overlap between individual enhancer predictions from several studies tends to be quite poor[4]. This is likely caused by heterogeneity of starting biological samples, limitations of current technologies, and differences in data analysis. Although the first two are difficult to change, analyzing these data in a

similar way could avoid some of the noise and difference generated by data analysis.

Here we undertook such an integrative effort, focusing on human brain development (**Figure 1A, Supplementary Figure 1**). We retrieved virtually all previously published putative enhancers for brain (from PubMed and enhancer databases, n = ~1.6 million putative enhancers) (**Supplementary Table 1**)[9, 11, 17-27], and performed an integrative analysis of relevant available epigenome data sets (n=494)[9, 10, 27-32] (**Supplementary Table 2**), after re-analysing the data. Using this approach, we identify around 200 thousand putative critical regions (pCRs) in reported brain enhancers, of which around 40 thousand show dynamic epigenomic rearrangement during fetal brain development, indicating switching on and off of regulatory elements during development. We thus refer to these regions as differentially active enhancers (DAEs). Compared to their non-variable counterparts (nDAEs), DAEs have a higher level of sequence constraint, regulate genes that are expressed during fetal brain development and are associated with brain developmental processes. DAEs are enriched for binding sites of brain relevant transcription factors, brain related GWAS loci and are regulating disease relevant *Online Mendelian Inheritance in Man* (OMIM) genes. We validate a selected number of DAEs using *in vitro* reporter assays and CRISPRi in cell lines, and reporter assays during zebrafish development. Together, this provides an easily accessible and comprehensive resource of NCREs that are likely functional during human brain development.

Results

Integrative data analysis identifies differentially active regions during fetal brain development

We started our analysis by collecting relevant fetal brain epigenome data sets and previously published putative enhancers (**Supplementary Table 1, Supplementary Table 2**). Epigenome data sets included ChIP-seq for various histone modifications, DNase- and ATAC-seq data from various developmental time points and anatomical regions of human fetal brain, generated by several independent studies, including Roadmap, PsychENCODE and other publications[9, 10, 27-32, 86]. All primary data were re-analyzed using identical computational pipelines, and in total we processed 494 data sets. Scrutinizing through previously published literature on enhancers in brain and neuronal cell types, we collected 1,595,292 putative brain enhancers (**Supplementary Table 1**). These included enhancers retrieved from various enhancer databases, such as VISTA, FANTOM and EnhancerAtlas, enhancer predictions from the PsychENCODE consortium, human accelerated regions, ultra-conserved regions

and others[9, 11, 17-27]. We first analyzed the overlap between the different putative enhancers, and found only a small overlap between enhancer predictions from different studies (**Supplementary Table 1**). We reasoned that if different enhancer prediction methods used in the individual studies identified the same enhancers that only differ by the exact location or length, by merging the overlaps between different studies we could identify functional relevant parts of enhancers. We thus proceeded to determine putative critical regions (pCRs), by determining the unifying overlaps between all putative enhancers (**Figure 1A, step 1**). In this analysis, we kept those putative enhancers that were only found in a single study, merged the overlaps between multiple studies and eliminated those regions that were located within 2 kb upstream and 1 kb downstream of a transcriptional start site (TSS) or which had < 10 reads in epigenome data (see Methods). This resulted in 202,163 pCRs, with a total length of 93 Mb, an average size of 460 bps and most pCRs located between 5 and 50 kb away from the closest gene TSS (**Supplementary Figure 2A, B**).

We assumed that enhancers that have functional relevant roles during brain development would show dynamic changes in the levels of histone modifications and chromatin accessibility correlating with their function. To investigate this, we next intersected all pCRs with all epigenome data sets from different time points of fetal brain development and calculated the read count for each pCR region. After TMM-normalization, we performed differential accessibility analysis (for ATAC-seq and DNase data) and generated differential histone modification profiles (for H3K27ac, H3K27me3, H3K4me1, H3K4me2, H3K4me3) using edgeR[39]. This resulted in 39,709 pCRs that showed a high variability for these features across developmental time points (**Figure 1B, Supplementary Figure 2C, Supplementary Table 3, see Methods**) which we refer to as differentially active enhancers (DAEs). In contrast, the remaining 162,454 pCRs showed a more constant epigenome pattern and we thus refer to them as not-differentially active enhancers, nDAEs (**Figure 1B, Supplementary Figure 2C, Supplementary Table 3**).

Gene ontology analysis using GREAT[36] showed that DAEs were significantly enriched for terms related to brain development, including processes such as forebrain neuron fate commitment, dorsal/ventral axon guidance and spinal cord development (**Figure 1B, Supplementary Table 4**). nDAEs appeared to be enriched for more general terms, including various chromatin modifications and receptor mediated endocytosis (**Figure 1B, Supplementary Table 4**).

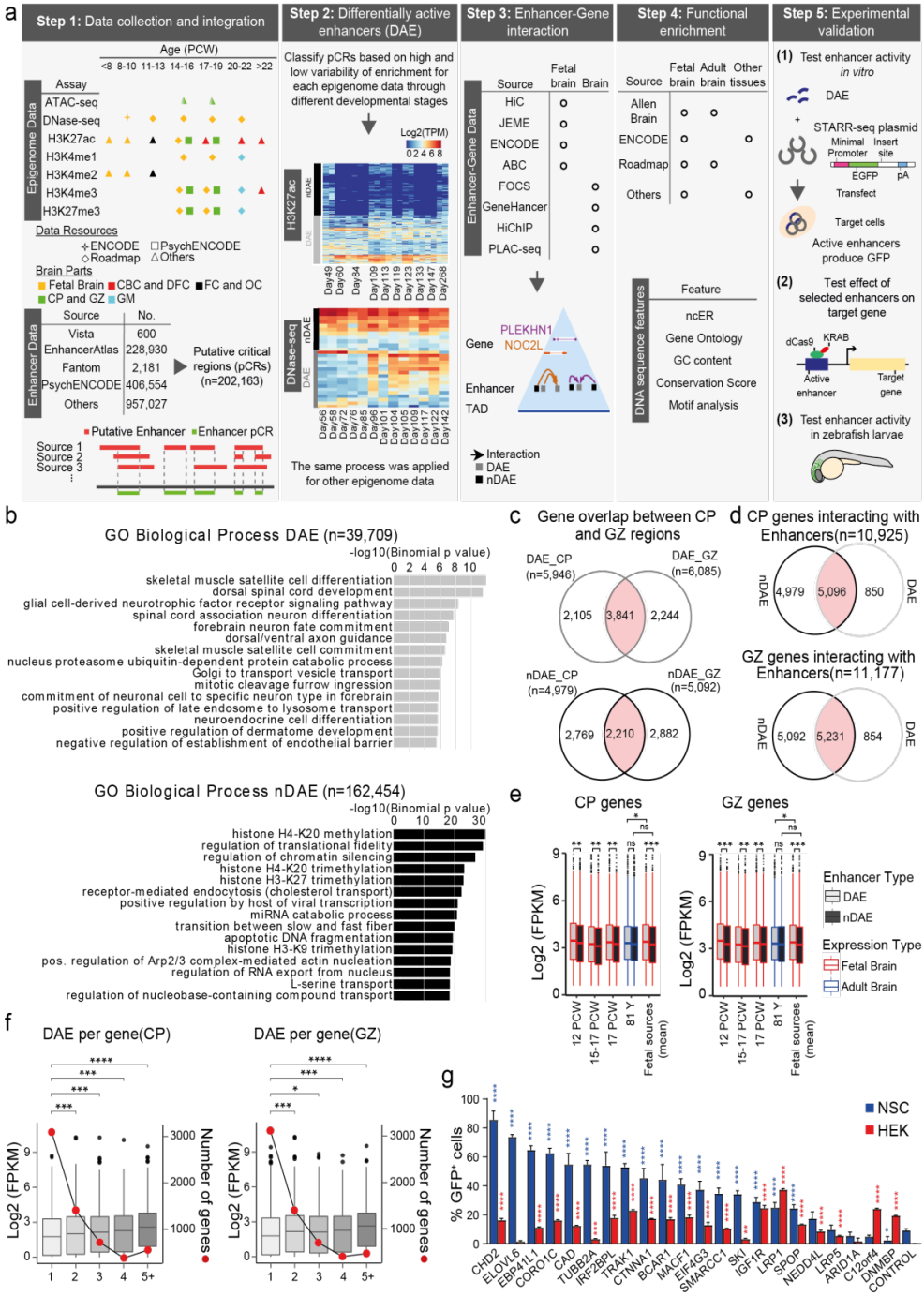
To have a more specific view about the genes regulated by these pCRs, we next linked

DAEs and nDAEs to their target genes, using different resources, which either link enhancer to gene promoters by direct chromatin interaction as determined by chromatin conformation capture techniques (HiC[26], HiChIP[44], PLAC-seq[45]) or by predicting enhancer-gene interactions using statistical models and correlation between gene expression, omics data and epigenome features (JEME, FOCS, GeneHancer, ENCODE, Activity-by-contact (ABC) method)[40-43,46] (**Supplementary Table 5**). Since only a limited number of interactions between DAEs or nDAEs and target genes identified by these different methods were supported by >2 of the available resources (**Supplementary Figure 3A**), and as most interactions and target genes were predicted by the HiC data (**Supplementary Figure 3B)**, we focused on these HiC predicted target gene interactions for the remainder of the analysis. These HiC data were generated from post conceptional week (PCW) 17-18 human brains[26], and were available for the germinal zone (GZ) (containing primarily mitotically active neural progenitors), and the cortical and subcortical plate (CP) (consisting primarily of post-mitotic and migrating neurons). Enhancer-promoter interactions derived from these HiC data do not exclude the fact that the identified DAEs and nDAEs interact with or regulate other genes at other developmental time points or in other cell types, for which at this moment no specific enhancer-promoter predictions are available.

Enhancer-promoter interactions derived from these HiC data do not exclude the fact that the identified DAEs and nDAEs interact with or regulate other genes at other developmental time points or in other cell types, for which at this moment no specific enhancer-promoter predictions are available.

Taking only those enhancer-promoter interactions that occurred in the same topological associated domain (TAD) into account, we found that from all DAEs, 6,858 and 6,883 for CP and GZ, respectively, interacted with promoters of protein coding genes or lincRNAs, of which the majority of interactions occur with protein coding genes. Similarly, 27,004 and 27,161 nDAEs interacted with target genes in CP and GZ, respectively, with a similar distribution between protein-coding and lincRNAs (**Supplementary Figure 3C, D**).

In total, DAEs interacted with 5,946 and 6,085 protein coding and lincRNA genes in CP and GZ, respectively, of which 3,841 genes were shared between both CP and GZ (**Figure 1C**). The majority of these genes (86%) also had interactions with nDAEs (**Figure 1D**). We next integrated available gene expression data from fetal and adult brain (**Supplementary Table 6**), and found that genes that interacted with a


a
Step 1: Data collection and integration
Step 2: Differentially active enhancers (DAE)
Step 3: Enhancer-Gene interaction
Step 4: Functional enrichment
Step 5: Experimental validation
Age (PCW)
Assay
ATAC-seq
DNase-seq
H3K27ac
H3K4me1
H3K4me2
H3K4me3
H3K27me3
Epigenome Data
Data Resources
ENCODE
Roadmap
PsychENCODE
Others
Brain Parts
Fetal Brain
CBC and DFC
FC and OC
CP and GZ
GM
Enhancer Data
Source
No.
Vista 600
EnhancerAtlas 228,930
Fantom 2,181
PsychENCODE 406,554
Others 957,027
Putative critical regions (pCRs) (n=202,163)
Putative Enhancer
Enhancer pCR
Source 1
Source 2
Source 3
Classify pCRs based on high and low variability of enrichment for each epigenome data through different developmental stages
Log2(TPM)
H3K27ac
nDAE
DAE
DNase-seq
The same process was applied for other epigenome data
Enhancer-Gene Data
HiC
JEME
ENCODE
ABC
FOCS
GeneHancer
HiChIP
PLAC-seq
PLEKHN1
NOC2L
Gene
Enhancer
TAD
Interaction
DAE
nDAE
Fetal brain
Brain
Adult brain
Other tissues
Allen Brain
Roadmap
DNA sequence features
Feature
ncER
Gene Ontology
GC content
Conservation Score
Motif analysis
(1) Test enhancer activity in vitro
DAE
STARR-seq plasmid
Minimal Promoter
Insert site
EGFP
pA
Transfect
Target cells
Active enhancers produce GFP
(2) Test effect of selected enhancers on target gene
dCas9
KRAB
Active enhancer
Target gene
(3) Test enhancer activity in zebrafish larvae
b
GO Biological Process DAE (n=39,709)
-log10(Binomial p value)
skeletal muscle satellite cell differentiation
dorsal spinal cord development
glial cell-derived neurotrophic factor receptor signaling pathway
spinal cord association neuron differentiation
forebrain neuron fate commitment
dorsal/ventral axon guidance
skeletal muscle satellite cell commitment
nucleus proteasome ubiquitin-dependent protein catabolic process
Golgi to transport vesicle transport
mitotic cleavage furrow ingression
commitment of neuronal cell to specific neuron type in forebrain
positive regulation of late endosome to lysosome transport
neuroendocrine cell differentiation
positive regulation of dermatome development
negative regulation of establishment of endothelial barrier
GO Biological Process nDAE (n=162,454)
-log10(Binomial p value)
histone H4-K20 methylation
regulation of translational fidelity
regulation of chromatin silencing
histone H4-K20 trimethylation
histone H3-K27 trimethylation
receptor-mediated endocytosis (cholesterol transport)
positive regulation by host of viral transcription
miRNA catabolic process
transition between slow and fast fiber
apoptotic DNA fragmentation
histone H3-K9 trimethylation
pos. regulation of Arp2/3 complex-mediated actin nucleation
regulation of RNA export from nucleus
L-serine transport
regulation of nucleobase-containing compound transport
c
Gene overlap between CP and GZ regions
DAE_CP (n=5,946)
DAE_GZ (n=6,085)
2,105
3,841
2,244
nDAE_CP (n=4,979)
nDAE_GZ (n=5,092)
2,769
2,210
2,882
d
CP genes interacting with Enhancers(n=10,925)
nDAE
DAE
4,979
5,096
850
GZ genes interacting with Enhancers(n=11,177)
5,092
5,231
854
e
CP genes
GZ genes
Log2 (FPKM)
12 PCW
15-17 PCW
17 PCW
81 Y
Fetal sources (mean)
Enhancer Type
DAE
nDAE
Expression Type
Fetal Brain
Adult Brain
f
DAE per gene(CP)
DAE per gene(GZ)
Log2 (FPKM)
Number of genes
g
% GFP+ cells
NSC
HEK
CHD2
ELOVL6
EBP41L1
CORO1C
CAD
TUBB2A
IRF2BPL
TRAK1
CTNNA1
BCAR1
MACF1
EIF4G3
SMARCC1
SKI
IGF1R
LRP1
SPOP
NEDD4L
LRP5
ARID1A
C12orf4
DNMBP
CONTROL

DAE had a significantly higher gene expression compared to those genes not interacting with a DAE, at various regions and stages of fetal development but not in adult brain (12 PCW: CP genes *p-value*=0.002671, GZ genes *p-value*=5.111e-05; 15-17 PCW: CP genes *p-value* =0.003251, GZ genes *p-value*=0.003813; 17 PCW: CP genes *p-value* =0.002533, GZ genes *p-value*=0.001813); 81 years: CP genes *p-value* =0.1377, GZ genes *p-value*= 0.2641; fetal sources mean: CP genes *p-value* =0.0002696, GZ genes *p-value*=0.00046; DAE fetal sources mean vs DAE 81 years: CP genes *p-value* =0.04744, GZ genes *p-value*=0.01525; nDAE fetal sources mean vs nDAE 81 years: CP genes *p-value* =0.781, GZ genes *p-value*=0.4904, wilcox.test) (**Figure 1E, Supplementary Figure 3E**). Similar observations were made when using the alternative, not HiC-based enhancer-promoter predictions (**Supplementary Figure 3F)**. In line with earlier findings[93], we find that the more enhancers a gene is interacting with, the higher the gene expression is, and this was also true for the DAEs (**Figure 1F**). A recent study determined gene expression trajectories in the dorsolateral prefrontal cortex during pre- and postnatal development. This study identified constant, rising and falling genes, that showed respectively similar, increased or decreased gene expression levels upon development[57]. In line with the earlier gene expression findings, we found that the odds ratio between DAE and nDAE linked genes (as determined by HiC) was significantly higher (odds ratio=1.183, *p*-value=0.0008 for GZ; odds ratio=1.198, *p*-value=0.0004

Figure 1. Integrative analysis of brain enhancers during fetal development. A) Various steps taken in the integrative analysis of this study. See text for details. **B)** Functional enrichment analysis using GREAT[17], for DAEs (upper panel, n = 39,709) and nDAEs (lower panel, n = 162,454), determined using whole genome as a background. X-axis reports the − Log10 *p value* as determined by GREAT. **C)** Venn diagram showing the overlap between DAEs (upper panel) and nDAEs (lower panel) interacting with protein-coding and lincRNA genes in CP (left) and GZ (right). **D)** Venn diagram showing the overlap between interactions of protein-coding and lincRNA genes with nDAEs (left) and DAEs (right), for protein-coding and lincRNA genes in CP (upper panel) and GZ (lower panel). **E)** Box plots showing gene expression levels as determined by RNA-seq, for genes that interact by HiC with DAEs (light gray) or nDAEs (dark gray) in CP (left) and GZ (right), for fetal (red) or adult (blue) brain samples. Boxes are interquartile range (IQR); line is median; and whiskers extend to 1.5 the IQR. PCW, postconceptional week. FPKM, fragments per kilobase of transcript per million mapped reads. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; ns, not significant (wilcox.test). Data obtained from: 12 PCW, Yan et al[18]; 15-17 PCW, De la Torre-Ubieta et al[19]; 17 PCW, Roadmap[10]; 81 years, Roadmap[10]; mean of fetal sources is the mean expression of the first three fetal samples. **F)** Box plots showing RNA-seq gene expression for genes interacting with 1, 2, 3, 4, or 5 or more DAEs in CP (left) and GZ (right). Left y-axis shows gene expression (log2 FPKM), right y-axis, and line plot shows the number of genes per DAE group. * $p < 0.05$; ***$p < 0.001$; ****$p < 0.0001$ (wilcox.test). RNA-seq data from Allen human brain atlas[20]. **G)** Bar plot showing the percentage of GFP+ cells in NSCs (blue) and HEK cells (red), from cell transfection experiments with an enhancer reporter plasmid for 22 tested enhancers and an empty plasmid control. Plotted is the percentage of GFP+ in cells co-transfected with an mCherry expressing plasmid, to correct for transfection efficiency. Bars show the average from two independent experiments, with each enhancer tested each in duplicate. Error bars represent standard deviation. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$ (one-way ANOVA test followed by multiple comparison test (Fisher's LSD test).

for CP, Fisher's exact test) for falling genes, that showed higher gene expression levels in prenatal RNA-seq samples (**Supplementary Figure 3G,H**).

Finally, to validate that DAEs can function as enhancers, we selected 22 DAEs linked to genes that are expressed in human neural stem cells (NSCs), cloned them in an enhancer reporter plasmid[88] and tested their enhancer activity in cell transfection experiments. Upon transfection in NSCs, 18 out of 22 tested sequences showed significantly increased percentage of GFP+ cells compared to control (normalized for transfection efficiency using an mCherry spiked-in control), confirming enhancer activity (**Figure 1G**). Transfecting the same plasmids in non-neural HEK cells showed less pronounced activity. This indicates that 81.8% of the tested DAEs had a measurable enhancer activity using this assay in an *in vitro* neural cell type. Of note, this does not exclude activity of those 4 DAEs that do not show enhancer activity in NSCs, in other cell types during fetal brain development.

We conclude that an integrative data analysis of virtually all previously reported brain enhancers identifies a set of DAEs which are associated with a brain developmental gene ontology, increased gene expression in fetal brain and display enhancer activity *in vitro*.

Multi-gene interacting enhancers regulate genes implicated in multiple cellular processes and have distinguishing sequence characteristics

In order to understand the biological function of DAEs and nDAEs in more detail, we further characterized these two groups. When determining the number of genes that each DAE is interacting with, we found that the majority of DAEs interact with 1 or 2 genes; but, in addition, a considerable fraction of DAEs also interact with more than 2 genes (19.7 % for CP, 19.4% for GZ) (**Figure 2A**), and the same was found for nDAEs (**Figure 2B**). When comparing the enrichment of biological processes for the genes that interact by HiC with these multi-gene interacting DAEs using Enrichr, we found that these genes were enriched for broader developmental and metabolic processes. However, genes that interact with DAEs that only regulate single genes were enriched for more specific brain related terms, such as "neuron differentiation" and "neuron migration" (**Supplementary Table 7**). Similar results were obtained using GREAT and Metascape analysis, where multi-gene interacting DAEs for example were enriched in mouse phenotypes associated with "early lethality", whereas DAEs associated with only a single gene were enriched for "regulation of neural precursor cell proliferation" (**Supplementary Table 7**).

We next asked whether DAEs that regulate single or multiple genes could have distinguishing DNA sequence characteristics, which could support their presumed distinct functional roles. To answer this, we focused on scores that provide some weight based on the underlying sequences: non-coding essential regulation (ncER) score[47], GC content[48] and phastcons score[48]. The ncER scores were recently established using a machine learning model[47], taking functional, mutational and structural features into account, including sequence constraint in the human population, and provides a score where 0 is non-essential, and 1 is putative-essential. We observe that DAEs that interact with 3 or more genes have a significantly higher ncER percentile compared to DAEs that interact with only 1 gene (**Figure 2C**). This might reflect their biological function regulating multiple genes, resulting in a higher tendency to be constraint. A similar trend was observed for GC content, where DAEs interacting with more than one gene had a significantly higher GC content, whereas for the phastcons score, an indicator of multi-species conservation, differences were not significant (**Figure 2C**). Similar observations were made for nDAEs (**Figure 2D**). Higher GC content has also been observed in more broadly active enhancers in the immune system[94] and might be explained by binding of broadly active transcription factors (TFs) to GC-rich motifs[95].

Sequence characteristics distinguish DAEs from nDAEs

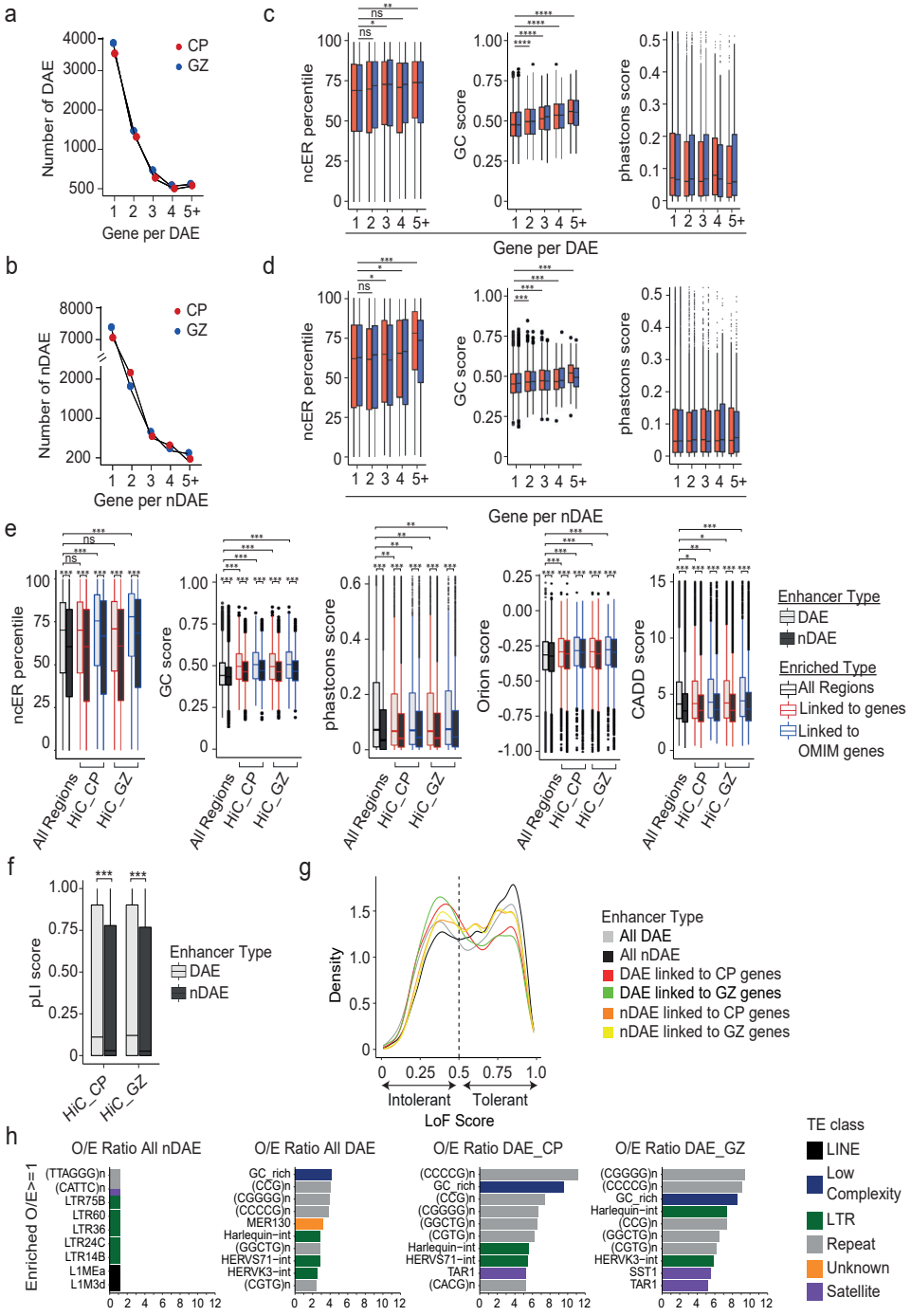
Given the differences in gene ontology between DAE and nDAE linked genes (**Figure 1B**) and the differences in ncER score and CG content between enhancers that regulate single versus multiple genes (**Figure 2C, D**), we next asked whether there are differences between these scores in DAEs and nDAEs, and whether any potential difference would be influenced by gene interactions that these regulatory elements have. We observed a significantly higher ncER percentile, CG content and phastcons score when comparing all DAEs to nDAEs (**Figure 2E**). Interestingly, some of these scores further increased, when only considering those DAEs and nDAEs that interact with target genes (as determined by HiC). This increased even further when only considering interacting target genes that are associated with known *Online Mendelian Inheritance in Man* (OMIM) phenotypes. Similar observations were made when using the Orion[49] and CADD scores[50] (**Figure 2E**) that similarly take depletion of variation in the human population and likelihood of deleteriousness of a given nucleotide based on integration of various annotations into account, respectively. Again, DAEs scored significantly higher for Orion and CADD scores than nDAEs, emphasizing the potentially biological important role of DAEs during brain development. Genes that are essential in humans are generally depleted of loss-of-

function alleles, and this is reflected by a higher probability of loss-of-function intolerance (pLI) score[52]. When we plotted the median pLI of genes linked to DAEs, or to nDAEs, genes linked to DAEs scored significantly higher (**Figure 2F**). Finally, a recent study determined loss-of-function tolerance scores for non-coding sequences, by using machine learning and structural variants from whole genome sequencing, including homozygous enhancer deletions[51]. Using this analysis, we observed that DAEs were more likely to be intolerant to loss-of-function, whereas nDAEs were more often tolerant to loss-of-function (**Figure 2G**). Again, when only considering those interactions linked to known target genes, scores further improved, in favor of DAEs.

We and others previously showed that functional enhancers can be enriched for transposable elements (TEs), some of which can be human specific[62, 96-98]. We thus asked whether DAEs and nDAEs showed a similar TE enrichment, and whether any TEs could distinguish both groups (**Figure 2H, Supplementary Table 8**). nDAEs showed a small enrichment for various LTR-containing TEs (e.g. LTR75B, LTR60, LTR36). Compared to nDAEs, DAEs were mainly enriched for CG rich repeat sequences, and a number of LTR repeats, such as Harlequin-int, HERVS71-int and HERVK3-int. Enrichment of the latter LTR repeats was not seen when only considering gene-interacting DAEs. The MER130 repeat family was previously shown to be enriched near critical genes for the development of the mouse neocortex and suggested to be co-opted for developmental enhancers of these genes[99]. Interestingly, MER130 repeats were enriched in all DAEs, but this enrichment was lost when only assessing DAEs that interact with genes, which made it difficult to further investigate the role of MER130 in human brain regulation. Compared to our previous findings in human embryonic stem cells (ESCs)[62], the overall TE enrichment in enhancers in brain was markedly different, with none of the TEs enriched in active enhancers in ESC showing enrichment at brain enhancers. This could indicate that different TEs co-opted into the regulatory landscape acquired different tissue specific roles during evolution.

Together this indicates that by investigating unbiased variability in epigenome marks over putative brain enhancers across developmental time points, DAEs and nDAEs can be identified which are associated with different gene ontologies, show different enrichments, have different sequence characteristics, and are distinctively linked to disease relevant genes.

Figure 2


a
Number of DAE
CP
GZ
Gene per DAE
b
Number of nDAE
Gene per nDAE
c
ncER percentile
GC score
phastcons score
Gene per DAE
d
e
phastcons score
Orion score
CADD score
All Regions
HiC_CP
HiC_GZ
Enhancer Type
DAE
nDAE
Enriched Type
All Regions
Linked to genes
Linked to OMIM genes
f
pLI score
g
Density
Intolerant
Tolerant
LoF Score
All DAE
All nDAE
DAE linked to CP genes
DAE linked to GZ genes
nDAE linked to CP genes
nDAE linked to GZ genes
h
O/E Ratio All nDAE
O/E Ratio All DAE
O/E Ratio DAE_CP
O/E Ratio DAE_GZ
Enriched O/E>=1
TE class
LINE
Low Complexity
LTR
Repeat
Unknown
Satellite

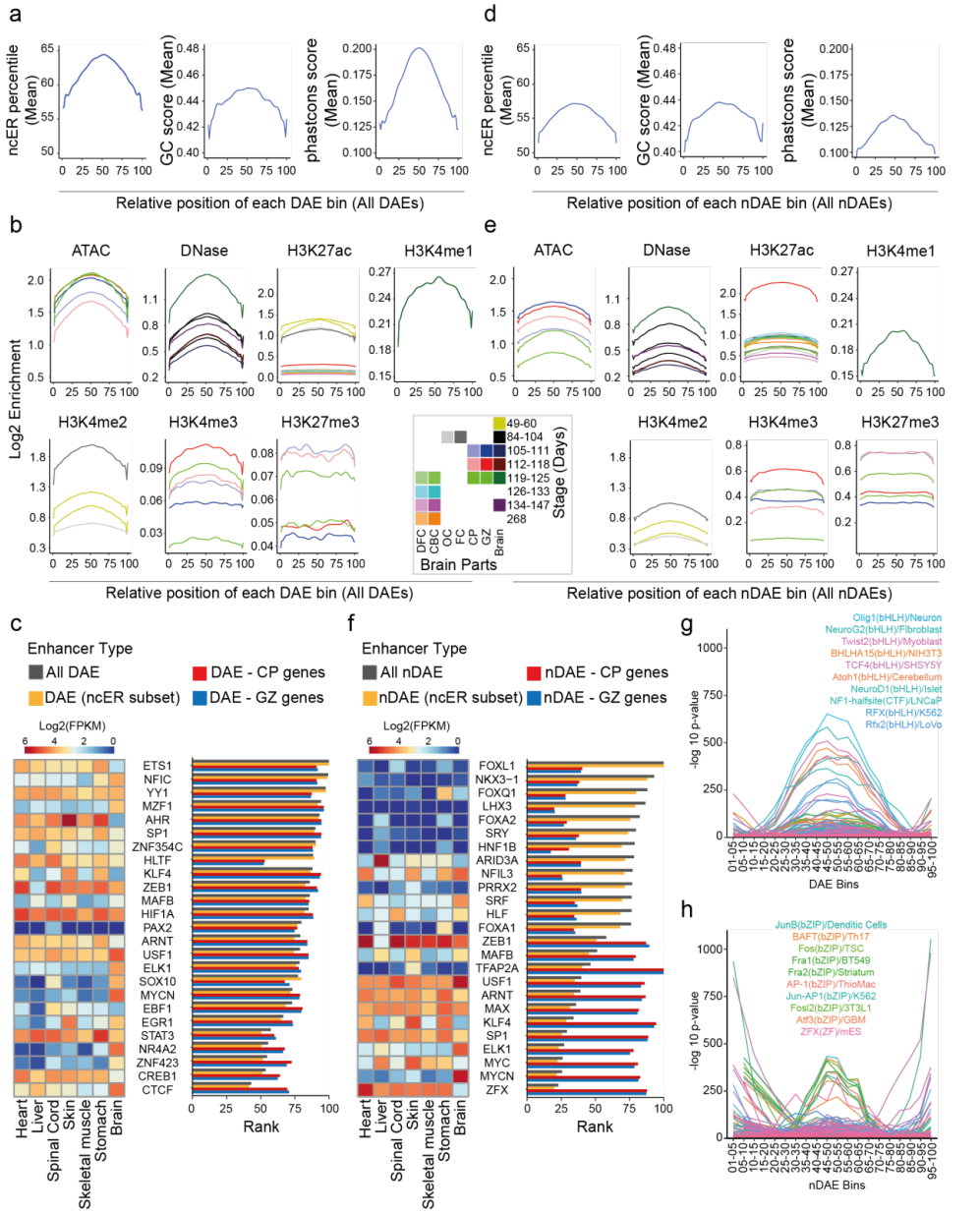
DAEs and nDAEs are enriched for distinct transcription factor binding sites

As the merging of pCRs and subsequent variability calling identified DAEs with distinct sequence characteristics, we next wondered whether we could further zoom in into each of the DAEs, to identify functional relevant nucleotides. To this end, we again made use of the ncER, CG and phastcons scores, assuming that the functional relevant nucleotides in each DAE might be those that have higher scores. As the identified DAEs varied in size between 50 and 1000 bps, we first split up each DAE into 10 bp bins, and assigned the median ncER, CG and phastcons scores to each bin. To be able to compare the score distribution within each bin between all DAEs, we re-scaled each DAE to a relative bin position from 1-100 (see Methods for details). Strikingly, the mean of ncER, CG and phastcons scores were highest between bins 40-60 (**Figure 3A**). To exclude that this was an artefact from the bin-rescaling, we plotted the mean distribution for the same scores also for DAEs that had an identical length and found similar results (**Supplementary Figure 4A**). We next calculated the number of reads from all epigenome data sets and plotted the log2 enrichment over the same relative DAE bin positions. We found that ATAC-seq, DNase-seq, H3K27ac, H3K4me1 and H3K4me2 signals (all associated with enhancers) again were most enriched between bins 40-60, whereas signal for H3K4me3 (of which high levels are associated with promoters and lower levels are found at enhancers) and H3K27me3 (associated with repressed chromatin) showed a broader distribution (**Figure 3B**), and this holds true for all developmental time points assessed.

Figure 2. Distinct sequence characteristics between DAEs and nDAEs. A) Line graph showing the number of protein-coding and lincRNA genes (1, 2, 3, 4, or 5 or more) that each DAE is interacting with, and the number of DAEs per category, for CP (red) and GZ (blue). **B)** As A), but here for nDAEs. **C)** Box plots showing the median ncER percentile (left)[47], GC content score (middle)[48] or phastcons score (right)[48] for DAEs-CP (red) and DAEs-GZ (blue) that interact with 1, 2, 3, 4, or 5 or more protein-coding and lincRNA genes. Boxes are IQR; line is median; and whiskers extend to 1.5 the IQR. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$; ns, not significant (wilcox.test). **D)** As C), but here for nDAEs. **E)** Box plots, showing from left to right ncER percentile[47], GC content score[48], phastcons score[48], Orion score[49], and CADD score[50], for all DAEs (light gray) and nDAEs (dark gray), or for those DAEs and nDAEs that are interacting in CP or GZ with protein-coding and lincRNA genes (red) or genes with a known OMIM phenotype (blue). Boxes are IQR; line is median; and whiskers extend to 1.5 the IQR. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; ns, not significant (wilcox.test). **F)** Box plot showing the pLI score[52] of genes interacting with DAEs (light gray) and nDAEs (dark gray) in CP or GZ. Boxes are IQR; line is median; and whiskers extend to 1.5 the IQR. *** $p < 0.001$; (wilcox.test). **G)** Kernel density plot showing the distribution of loss-of-function tolerance scores for non-coding sequences[51] for all DAEs (light gray), all nDAEs (dark gray), DAEs linked to protein-coding and lincRNA genes in CP (red), DAEs linked to protein-coding and lincRNA genes in GZ (green), nDAEs linked to protein-coding and lincRNA genes in CP (orange), and nDAEs linked to protein-coding and lincRNA genes in GZ (yellow). **H)** Bar chart showing the most enriched transposable elements (TEs) overlapping with from left to right all nDAEs, all DAEs, DAEs interacting with protein-coding and lincRNA genes in CP, and DAEs interacting with protein-coding genes in GZ. Plotted is a ratio between the observed (O) number of TEs over the expected (E). Different classes of TE are indicated with different colors as indicated.

This suggests that on average the center of the DAEs most likely contains the functional relevant sequences, and given the increased chromatin accessibility at those locations, this could indicate binding of functionally relevant TFs in these central regions.

To investigate this further, we first performed TF enrichment analysis using *Locus Overlap Analysis* (LOLA)[60], on both full length DAEs, as well as on only the central DAE parts between bin 40-60 (ncER subset). LOLA performs enrichment analysis based on genomic regions and tests the overlap of the query regions with a core reference database assembled from public data, including amongst others ChIP-seq data from CODEX[100]. We found a similar enrichment of TF binding sites between full length and central parts of DAEs (**Figure 3C, Supplementary Table 8**), and between all DAEs and those interacting with target genes in CP and GZ. The most enriched TFs at DAEs according to LOLA, included well-known TFs with essential roles for brain development. This includes amongst others ETS1, a widely studied TF with functions in different biological systems which was previously shown to be necessary for radial glia formation in vertebrates[101] and FGF-dependent patterning of anterior-posterior compartments in the central nervous system of *Ciona* (a marine invertebrate that is a well-suited model to study cell fate specification in chordates)[102]; YY1, a crucial TF which is involved in both gene activation and repression[103], mediating enhancer-promoter interactions[104] and of which mutations cause a neurodevelopmental disorder[105]; and CTCF, a master regulator of chromatin structure, of which *de novo* mutations cause intellectual disability[106]. We next repeated the same analysis for nDAEs (**Figure 3D-F, Supplementary Table 8**). Similar to our observations for DAEs, nDAEs had higher ncERs, CG content and conservation at the central part, with those regions being enriched for enhancer associated histone marks, but showed less variability over time. When performing TF enrichment analysis using LOLA, we observed a different and less specific set of TFs enriched at nDAEs compared to DAEs. Also, enrichment was lower at those nDAEs that were interacting with target genes. Again similar enrichment was found in the central part compared to the whole nDAEs. Enriched TFs for nDAEs included amongst others FOXL1, a transcriptional repressor that regulates central nervous system development[107]; the LIM homeodomain TF LHX3, that is essential for pituitary and nervous system development[108, 109]; and FOXA2, which plays a role in midbrain dopaminergic neurons[110, 111] (**Figure 3F**). Shared TFs enriched both at DAEs and nDAEs included SP1, loss of which in astrocytes impacts on neurons in the cortex and hippocampus of mice[112]; MAFB, a basic leucine zipper TF that plays a role in hindbrain development[113-115] and postnatal


a
ncER percentile (Mean)
GC score (Mean)
phastcons score (Mean)
Relative position of each DAE bin (All DAEs)
d
Relative position of each nDAE bin (All nDAEs)
b
ATAC
DNase
H3K27ac
H3K4me1
H3K4me2
H3K4me3
H3K27me3
Log2 Enrichment
Stage (Days)
Brain Parts
e
c
Enhancer Type
All DAE
DAE - CP genes
DAE (ncER subset)
DAE - GZ genes
Log2(FPKM)
Rank
f
All nDAE
nDAE - CP genes
nDAE (ncER subset)
nDAE - GZ genes
g
-log 10 p-value
DAE Bins
h
nDAE Bins

brain development[116, 117]; and ZEB1 which is required for neuronal differentiation[118, 119].

As LOLA analysis considers a single shared base pair being sufficient for regions to count as overlapping, this analysis could not distinguish well between TFs specifically enriched at the central part of DAEs and nDAEs relative to the flanking regions. We therefore further investigated which TFs motifs were specifically enriched at the central parts versus other parts of DAEs and nDAEs, using motif enrichment analysis with HOMER[61], a motif discovery algorithm, which identifies regulatory elements that are specifically enriched in the query set relative to background. We first split the 100 relative bins into 20 groups of 5 consecutive bins each and determined the significantly enriched TF motifs ($p \leq 0.01$) for each of these 20 bin groups (**Supplementary Table 8)**. Amongst the enriched motifs, we found back, amongst others, the motifs for the TFs enriched using the LOLA analysis, validating these findings (**Supplementary Table 8**). When plotting the number of significant motifs ($p \leq 0.01$) per bin group and the number of target sequences with those motifs, we found that bins located in the central part of both DAEs and nDAEs had both the highest numbers of significant TF motifs and the highest number of target sequences (**Supplementary Figure 4B,C**). As most enriched motifs were found in multiple bins, and there can be multiple TF bindings sites of the same TF within the same enhancer, we next focused on only those TF motifs which were not equally enriched in all 20 bin groups (n= 251 for DAEs and n=218 for nDAEs). For both DAEs and nDAEs, we again found most motif enrichment in the central enhancer part, with DAEs

Figure 3. DAEs and nDAEs are enriched for distinct transcription factor binding sites. A) Line plot showing the distribution of the mean ncER percentile (left)[47], GC content score (middle)[48], and phastcons score (right)[48] over the relative bin position for all DAEs. **B)** Line plot showing the log2 enrichment for various epigenome features as indicated, over the relative bin positions for all DAEs. Different colors indicate different time points of human brain development and different brain regions from which the data were obtained. DFC, dorsal frontal cortex; CBC, cerebellar cortex; OC, occipital cortex; FC, frontal cortex; CP, cortical plate; GZ, germinal zone; Brain, whole brain. Epigenome data used are summarized in Additional file 3: Table S2. **C)** Bar chart showing the relative LOLA enrichment of TFs from JASPAR in all DAEs (light gray), in the central part of all DAEs (ncER subset, orange), in DAEs linked to genes in CP (red) and in DAEs linked to genes in GZ (blue). X-axis displays the rank score (a combination of p value, odds ratio from Fisher's exact test, and the raw number of overlapping regions) from LOLA. The rank was re-scaled between 0 and 100, so that DAEs with a larger TFs enrichment have a higher rank. Also shown is a heatmap showing the RNA-seq expression levels (Log2 FPKM) of the same TFs across various human fetal tissues. RNA-seq data obtained from ENCODE project[7]. **D)** As in A), but here for nDAEs. **E)** As in B), but here for nDAEs. Note the difference in y-axis scale for H3K4me3 and H3K27me3 compared to panel B given the higher enrichment in nDAEs. **F)** As in C), but now for nDAEs. **G)** Line plot showing the distribution of enrichment (− log10 p value as determined by HOMER analysis) across the relative DAE bins, for the 251 TF motifs that were not equally enriched in all 20 bin groups. The most enriched TF motifs are indicated. **H)** As G, but now for 218 TFs that were not equally enriched across the 20 bin groups of all nDAEs.

being more enriched than nDAEs (**Figure 3G,H, Supplementary Figure 4D,E**). Amongst the most enriched TF motifs at the center of DAEs were motifs for the proneural basic helix-loop-helix transcription factors NEUROG2, ATOH1 and NEUROD1, that promote neurogenesis[120-122], OLIG1, a marker of oligodendrocytes[123] that also regulates the neuron-glial switch during earlier embryonic development [124,125], TCF4, that is necessary for neuronal migration and the correct development of the cerebral cortex[126] and loss of which is associated with intellectual disability[127], and NF1, that regulates neuronal and glial differentiation and is causative of neurofibromatosis type 1 when mutant[128] (**Figure 3G)**. Enriched TF motifs at the central part of nDAEs are involved in more ubiquitous processes and include mainly activator protein 1 (AP-1), a heterodimer composed of members of the JUN (including JUNB), FOS (including FOSL2, FRA1, FRA2), ATF (including ATF3, BAFT) and MAF family, that regulates a wide variety of cellular processes in response to a wide range of extracellular cues[129] (**Figure 3H)**.

Together this indicates that on average the central part of brain enhancers (both DAE and nDAEs) contains relevant but partially distinct TF binding sites and might be enriched for functional relevant sequences, which can be further fine-mapped using ncER scores and other sequence characteristics. To test this directly, we selected three DAEs, linked to *IRF2BPL*, *CHD2* and *MACF1*, that showed activity in reporter assays in NSCs (**Figure 1G**), and deleted 10-30 bp of those regions that had the highest ncER scores in those enhancers. Upon transfection of these mutant DAEs, we observed a significantly reduced enhancer activity for *IRF2BPL* and *CHD2*, but not for *MACF1* (**Supplementary Figure 4F**). Deleting regions with a lower ncER score did not affect enhancer activity. Together this indicates that integrative analysis, variability analysis during development and sequence characteristics can identify functional relevant nucleotides in brain enhancers.

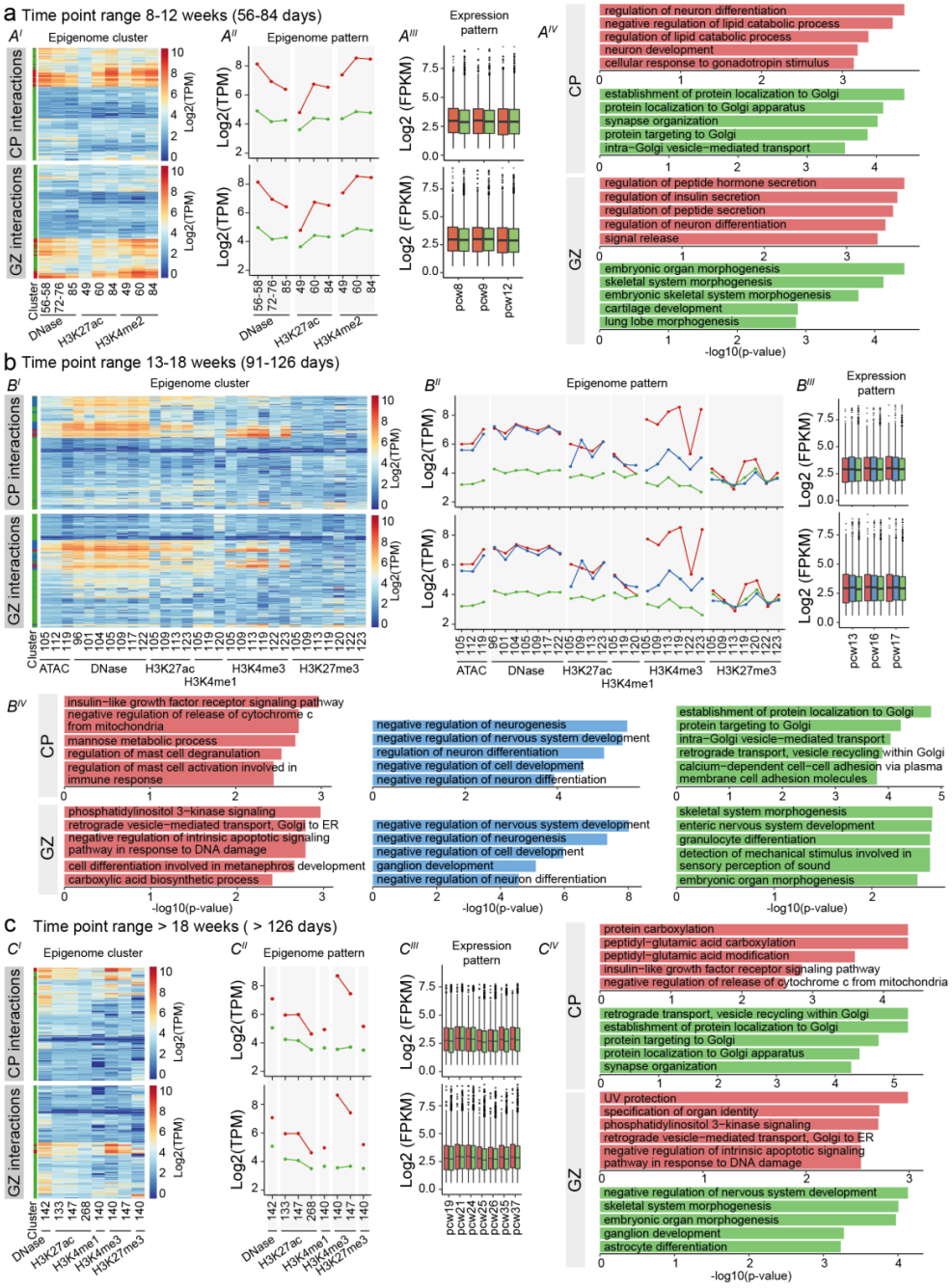
DAEs show temporal epigenome dynamics during human brain development

To further understand the dynamics of enhancer regulation, we subdivided DAEs interacting with genes in GZ and CP by performing clustering analysis on all available epigenome data sets, at different developmental stages (between 8-12 PCW, 13-18 PCW and >18 PCW) (**Figure 4A, Supplementary Table 9**). At 8-12 PCW, we found two clusters for both GZ and CP that showed relatively constant enrichments over time, with the first cluster (red) showing a higher enrichment for all epigenome features available for that developmental stage, com-

pared to the second cluster (green). No statistically significant differences in gene expression levels between genes linked to both clusters were found. Genes associated with cluster 1 DAEs in CP were enriched for gene ontology terms related to neuronal differentiation, whereas cluster 2 was dominated by processes in the Golgi. Likewise, for GZ, genes associated with cluster 1 seemed to be associated with more specific biological functions, whereas processes associated with cluster 2 showed more broad involvements (**Figure 4A, Supplementary Table 9**).

At 13-18 PCW, three clusters emerged in both GZ and CP (**Figure 4B, Supplementary Table 9**). Whereas cluster 3 (green) showed relatively low levels of epigenome marks similar to cluster 2 at 8-12 PCW, cluster 1 (red) and cluster 2 (blue) showed higher epigenome enrichments. Both cluster 1 and 2 had similar levels of H3K27ac, but mainly diverged from each other on the levels of H3K4me3. Cluster 2 was strongly enriched for processes involved in neural system development both in CP and GZ. Gene ontology of genes associated with cluster 1 (red) which showed higher H3K4me3 levels, showed enrichment for insulin-like growth factor receptor signalling and immune cell related processes in CP. Insulin-like growth factors are important for neuronal survival and neurogenesis[130]. As high levels of H3K4me3 have also been found at enhancers in blood cells[131], possibly stabilizing their transcription, it is tempting to speculate that part of this cluster reflects enhancers active in hematopoietic cells from the developing vasculature[132] and microglia (brain tissue macrophages) that are invading the brain at these developmental time points[133]. In GZ, cluster 1 was associated with phosphatidylinositol 3−kinase signaling, which is important for commitment of neural progenitor cells[134, 135].

Finally, at >18 PCWs, we found two clusters of DAEs, of which cluster 1 (red) was marked by higher levels of epigenome marks (**Figure 4C, Supplementary Table 9**). In CP, genes associated with this cluster were enriched for carboxylation processes and insulin-like growth factor receptor signalling. Genes associated with the second cluster (green) were again more enriched for broad developmental processes, including the Golgi system. In GZ, genes associated with cluster 1 (red) were amongst others involved in DNA damage repair. Indeed, alterations in this pathway can lead to reduced proliferation of neural progenitor cells leading to microcephaly[136, 137]. Cluster 2 (green) in GZ was associated with terms related to neurodevelopment and organ morphogenesis. Together, this shows that temporal epigenomic rearrangement in DAEs is reflected in regulating the expression level of genes that are important in developmental and cell type specific processes.


a Time point range 8-12 weeks (56-84 days)
A^I Epigenome cluster
A^II Epigenome pattern
A^III Expression pattern
A^IV
CP interactions
GZ interactions
Log2(TPM)
Log2 (FPKM)
DNase
H3K27ac
H3K4me2
pcw8
pcw9
pcw12
regulation of neuron differentiation
negative regulation of lipid catabolic process
regulation of lipid catabolic process
neuron development
cellular response to gonadotropin stimulus
establishment of protein localization to Golgi
protein localization to Golgi apparatus
synapse organization
protein targeting to Golgi
intra−Golgi vesicle−mediated transport
regulation of peptide hormone secretion
regulation of insulin secretion
regulation of peptide secretion
regulation of neuron differentiation
signal release
embryonic organ morphogenesis
skeletal system morphogenesis
embryonic skeletal system morphogenesis
cartilage development
lung lobe morphogenesis
-log10(p-value)
b Time point range 13-18 weeks (91-126 days)
B^I Epigenome cluster
B^II Epigenome pattern
B^III Expression pattern
ATAC
DNase
H3K27ac
H3K4me3
H3K27me3
H3K4me1
pcw13
pcw16
pcw17
B^IV
CP
GZ
insulin−like growth factor receptor signaling pathway
negative regulation of release of cytochrome c from mitochondria
mannose metabolic process
regulation of mast cell degranulation
regulation of mast cell activation involved in immune response
negative regulation of neurogenesis
negative regulation of nervous system development
regulation of neuron differentiation
negative regulation of cell development
negative regulation of neuron differentiation
establishment of protein localization to Golgi
protein targeting to Golgi
intra−Golgi vesicle−mediated transport
retrograde transport, vesicle recycling within Golgi
calcium−dependent cell−cell adhesion via plasma membrane cell adhesion molecules
phosphatidylinositol 3−kinase signaling
retrograde vesicle−mediated transport, Golgi to ER
negative regulation of intrinsic apoptotic signaling pathway in response to DNA damage
cell differentiation involved in metanephros development
carboxylic acid biosynthetic process
negative regulation of nervous system development
negative regulation of neurogenesis
negative regulation of cell development
ganglion development
negative regulation of neuron differentiation
skeletal system morphogenesis
enteric nervous system development
granulocyte differentiation
detection of mechanical stimulus involved in sensory perception of sound
embryonic organ morphogenesis
-log10(p-value)
c Time point range > 18 weeks (> 126 days)
C^I Epigenome cluster
C^II Epigenome pattern
C^III Expression pattern
C^IV
H3K4me1
H3K4me3
H3K27me3
pcw19
pcw21
pcw24
pcw25
pcw26
pcw35
pcw37
protein carboxylation
peptidyl−glutamic acid carboxylation
peptidyl−glutamic acid modification
insulin−like growth factor receptor signaling pathway
negative regulation of release of cytochrome c from mitochondria
retrograde transport, vesicle recycling within Golgi
establishment of protein localization to Golgi
protein targeting to Golgi
protein localization to Golgi apparatus
synapse organization
UV protection
specification of organ identity
phosphatidylinositol 3−kinase signaling
retrograde vesicle−mediated transport, Golgi to ER
negative regulation of intrinsic apoptotic signaling pathway in response to DNA damage
negative regulation of nervous system development
skeletal system morphogenesis
embryonic organ morphogenesis
ganglion development
astrocyte differentiation
-log10(p-value)

Cell type specificity of DAEs and nDAEs and their dynamics in adult brain

To further investigate cell type specificity of DAEs and nDAEs, we performed two additional analyses. First, we compared DAEs and nDAEs to recently identified cell-type specific regulatory elements. A recent study used scATAC-seq to generate a human cell atlas of fetal chromatin accessibility spanning 15 organs, including fetal brain[85]. When overlapping DAEs and nDAEs to the most specific chromatin accessibility peaks per cell type, we found 7,753 DAEs and 7,946 nDAEs that overlapped with these cell type specific chromatin accessibility peaks, including those found in several types of neurons and astrocytes (**Supplementary Figure 5A,B**). This indicates that our bulk analysis re-identifies cell type specific chromatin accessibility peaks which might therefore present cell type-specific enhancers, and it also shows that the bulk analysis identifies additional enhancers that are not captured by the single cell chromatin accessibility profiles.

We next investigated how these cell type specific enhancer might behave over time. Two recent studies determined cell type specific regulatory elements from postnatal brain with a reasonable overlap between both studies (**Supplementary Figure 5C,D**), by either isolating cell type specific bulk populations from brain followed by ATAC-seq and ChIP-seq for H3K27ac and H3K4me3[45], or by performing scATAC-seq[44]. Comparing the DAEs and nDAEs to these cell type specific regulatory elements, showed as expected that only a fraction of DAEs and nDAEs from the fetal brain analysis showed an overlap with the cell type specific regulatory elements derived from postnatal samples. Amongst those, we found overlap with cell type specific regulatory elements from neurons, oligodendrocytes, astrocytes and microglia (**Supplementary Figure 5E,F**). This indicates that despite determined from an integrative analysis of bulk samples derived during fetal brain development, a fraction of DAEs and nDAEs can be linked to cell-type specific regulatory elements which are likely to also have roles in postnatal brain. In contrast, other DAEs and nDAEs are likely having fetal specific functions.

Figure 4. **Clustering of DAEs unravels temporal dynamics of brain gene regulation. A)** Heatmap displaying all available epigenome features for PCW 8-12, across all DAEs interacting with protein-coding genes in CP (upper heatmap) and GZ (lower heatmap) (AI). K-means clustering analysis of epigenome features (AII) identifies two clusters, cluster 1 (red) and cluster 2 (green). Level of enrichment is indicated on the y-axis in Log2 TPM. Box plots (AIII) shows RNA-seq gene expression of protein-coding genes regulated by the DAEs from each cluster (Expression pattern), for available data from PCW 8, 9, and 12[54]. Boxes are IQR; line is median; and whiskers extend to 1.5 the IQR. Gene enrichment analysis for the corresponding genes in each cluster (AIV). X-axis shows the − log 10 (p value) from Enrichr. **B)** As for A), but now for PCW 13–18. **C)** As for A), but now for PCW > 18.

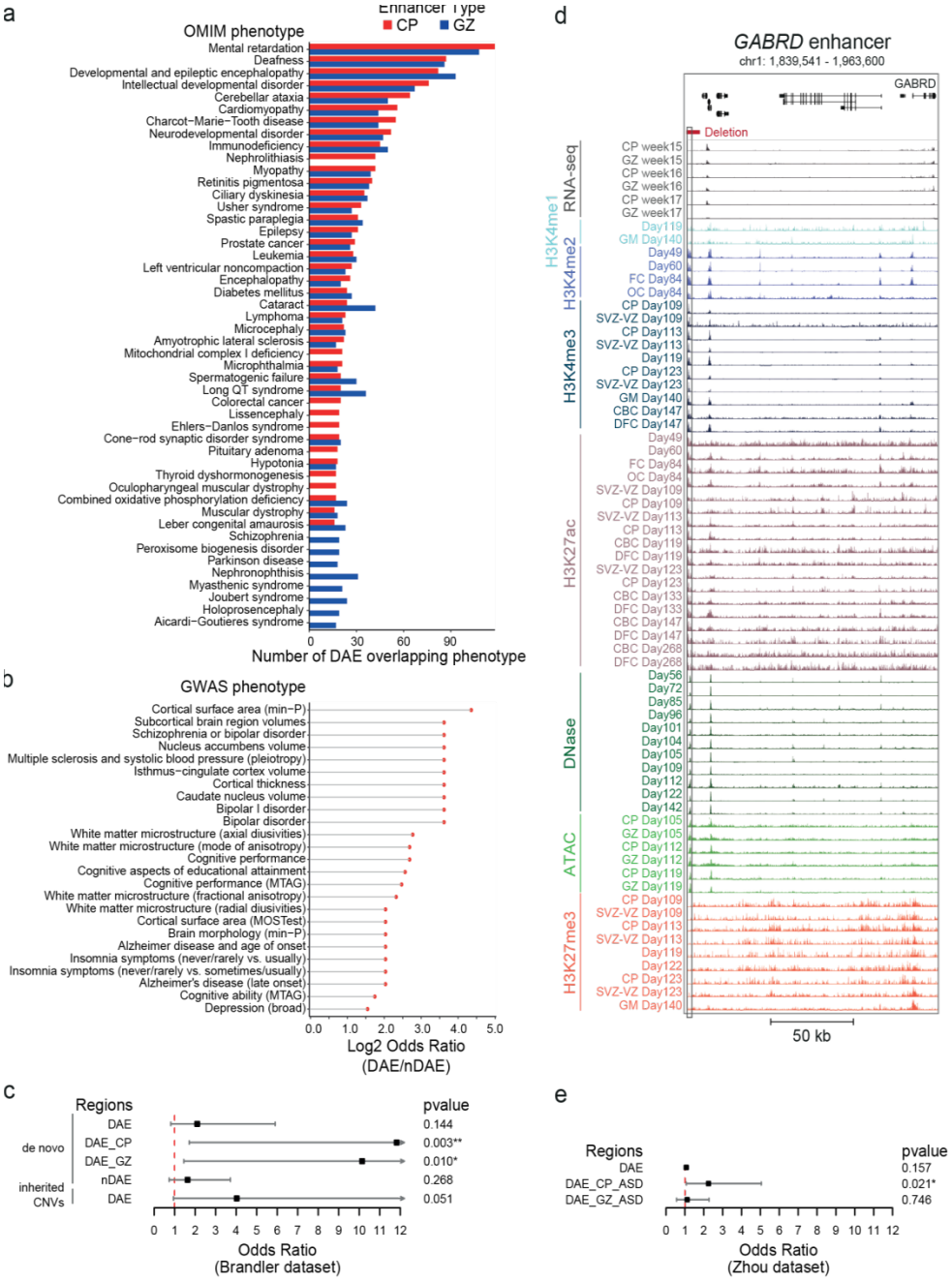
To further investigate the dynamics of DAEs and nDAEs in adult brain, in the second analysis, we compared H3K27ac levels obtained from both fetal and adult samples derived from a single study[86] for all DAEs and nDAEs linked to target genes in GZ and CP by HiC, and performed clustering and gene ontology analysis (**Supplementary Figure 6, Supplementary Table 9**). We found that DAEs that were mainly enriched for H3K27ac in fetal samples, were as expected associated with gene ontology terms related to fetal brain development, including regulation of neuron differentiation. DAEs which also showed H3K27ac enrichment in adult samples were associated with more broad physiological processes.

Together, this shows that part of DAEs and nDAEs can be linked back to cell type specific regulatory elements despite being identified from bulk tissue analysis and that some DAEs and nDAEs are likely to also function in postnatal brain.

DAEs regulate disease relevant genes and are enriched for disease implicated variants

Given our findings that DAEs are associated with genes relevant for brain development, we further investigated which disease relevant genes are regulated by DAEs. We first focused on known disease causing genes retrieved from OMIM. We found that 1,556 OMIM genes are regulated by DAEs (of which 1,165 and 1,166 from the interactions found in GZ and CP, respectively) (**Supplementary Table 10**). Most DAEs are linked to genes involved in mental retardation, developmental and epileptic encephalopathy, and neurodevelopmental disorders (**Figure 5A**). This included genes like *KMT2C*, involved in Kleefstra syndrome (OMIM #617768), and *GRIN2A* of which heterozygous mutations cause epilepsy and speech delay (OMIM #245570). Next to genes, enhancers can also interact with other additional enhancers. Interestingly, the more additional enhancers (DAE and/or nDAE) a DAE was interacting with, the more likely the target gene of this DAE was an OMIM gene (**Supplementary Figure 7A**). This supports recent findings that the number of enhancers linked to a gene reflect its disease pathogenicity[138], and confirms enhancer redundancy for disease relevant genes[139].

We next leveraged published GWAS loci for brain-related traits and disorders (**Supplementary Table 11**). When comparing the odds ratio between DAEs and nDAEs, we found that DAEs were more often enriched for various significant GWAS loci, reflecting a broad variety of both brain developmental processes (e.g. volumes of different anatomical brain regions) and neurodevelopmental disorders (e.g. mental development, autism) (**Figure 5B**). Similar, using LD score regression analysis we


a
Enhancer Type
CP
GZ
OMIM phenotype
Mental retardation
Deafness
Developmental and epileptic encephalopathy
Intellectual developmental disorder
Cerebellar ataxia
Cardiomyopathy
Charcot−Marie−Tooth disease
Neurodevelopmental disorder
Immunodeficiency
Nephrolithiasis
Myopathy
Retinitis pigmentosa
Ciliary dyskinesia
Usher syndrome
Spastic paraplegia
Epilepsy
Prostate cancer
Leukemia
Left ventricular noncompaction
Encephalopathy
Diabetes mellitus
Cataract
Lymphoma
Microcephaly
Amyotrophic lateral sclerosis
Mitochondrial complex I deficiency
Microphthalmia
Spermatogenic failure
Long QT syndrome
Colorectal cancer
Lissencephaly
Ehlers−Danlos syndrome
Cone−rod synaptic disorder syndrome
Pituitary adenoma
Hypotonia
Thyroid dyshormonogenesis
Oculopharyngeal muscular dystrophy
Combined oxidative phosphorylation deficiency
Muscular dystrophy
Leber congenital amaurosis
Schizophrenia
Peroxisome biogenesis disorder
Parkinson disease
Nephronophthisis
Myasthenic syndrome
Joubert syndrome
Holoprosencephaly
Aicardi−Goutieres syndrome
0
30
60
90
Number of DAE overlapping phenotype
b
GWAS phenotype
Cortical surface area (min−P)
Subcortical brain region volumes
Schizophrenia or bipolar disorder
Nucleus accumbens volume
Multiple sclerosis and systolic blood pressure (pleiotropy)
Isthmus−cingulate cortex volume
Cortical thickness
Caudate nucleus volume
Bipolar I disorder
Bipolar disorder
White matter microstructure (axial diusivities)
White matter microstructure (mode of anisotropy)
Cognitive performance
Cognitive aspects of educational attainment
Cognitive performance (MTAG)
White matter microstructure (fractional anisotropy)
White matter microstructure (radial diusivities)
Cortical surface area (MOSTest)
Brain morphology (min−P)
Alzheimer disease and age of onset
Insomnia symptoms (never/rarely vs. usually)
Insomnia symptoms (never/rarely vs. sometimes/usually)
Alzheimer's disease (late onset)
Cognitive ability (MTAG)
Depression (broad)
0.0
1.0
2.0
3.0
4.0
5.0
Log2 Odds Ratio
(DAE/nDAE)
c
Regions
pvalue
de novo
inherited CNVs
DAE
DAE_CP
DAE_GZ
nDAE
DAE
0.144
0.003**
0.010*
0.268
0.051
0
1
2
3
4
5
6
7
8
9
10
11
12
Odds Ratio
(Brandler dataset)
d
GABRD enhancer
chr1: 1,839,541 - 1,963,600
GABRD
Deletion
H3K4me1
RNA-seq
H3K4me2
H3K4me3
H3K27ac
DNase
ATAC
H3K27me3
CP week15
GZ week15
CP week16
GZ week16
CP week17
GZ week17
Day119
GM Day140
Day49
Day60
FC Day84
OC Day84
CP Day109
SVZ-VZ Day109
CP Day113
SVZ-VZ Day113
Day119
CP Day123
SVZ-VZ Day123
GM Day140
CBC Day147
DFC Day147
Day49
Day60
FC Day84
OC Day84
SVZ-VZ Day109
CP Day109
SVZ-VZ Day113
CP Day113
CBC Day119
DFC Day119
SVZ-VZ Day123
CP Day123
CBC Day133
DFC Day133
CBC Day147
DFC Day147
CBC Day268
DFC Day268
Day56
Day72
Day85
Day96
Day101
Day104
Day105
Day109
Day112
Day122
Day142
CP Day105
GZ Day105
CP Day112
GZ Day112
CP Day119
GZ Day119
CP Day109
SVZ-VZ Day109
CP Day113
SVZ-VZ Day113
Day119
Day122
CP Day123
SVZ-VZ Day123
GM Day140
50 kb
e
Regions
pvalue
DAE
DAE_CP_ASD
DAE_GZ_ASD
0.157
0.021*
0.746
0
1
2
3
4
5
6
7
8
9
10
11
12
Odds Ratio
(Zhou dataset)

found enrichment of heritability for variants within DAEs, nDAEs and pCRs, including for the trait "intelligence" (**Supplementary Figure 7B**).

Encouraged by these findings, we next asked whether copy number variants (CNVs) or single nucleotide variants (SNVs) at DAEs could be involved in causing genetic disease. We first leveraged previously published disease implicated CNVs. Brandler *et al* performed WGS in their discovery cohort of individuals affected by an autism spectrum disorder (ASD) and unaffected individuals and reported on 135 *de novo* CNVs (104 deletions, 29 duplications and 2 inversions)[81]. Of these, 25 overlapped a DAE in cases, and 8 in controls (odds ratio=2.10, *p*-value=0.144101). When only considering those CNVs overlapping DAEs linked to target genes, this became 17 in cases and 1 in control for DAEs linked to CP genes (odds ratio=11.83, p=0.003003) and 15 in cases and 1 in control for DAEs linked to GZ genes (odds ratio=10.14, *p*-value=0.010423). For nDAEs, 36 CNVs were found in cases and 15 in controls (odds ratio=1.63, *p*-value=0.267964). However, as not all these CNVs exclusively covered non-coding regions, it cannot be excluded that the observed association is due to disrupted coding genes, rather than involvement of DAEs. We therefore also assessed rare inherited deletions from the same study that did not overlap with coding exons (n=213 in total, 175 in cases and 38 in controls). From these, 32 cases had a deletion covering a DAE, compared to two controls (odds ratio=4.027972, *p*-value=0.05119). Although not significant, this might point to more deletions covering DAEs in ASD individuals but would require a larger sample size to be confirmed (**Figure 5C, Supplementary Table 12**).

In another study, Monlong *et al*[82] reported on CNVs in 198 epilepsy patients detected by WGS. They found an enrichment of rare non-coding CNVs near known epilepsy genes, with the *GABRD* gene showing the strongest and only nominally significant association with 4 non-coding deletions amongst the epilepsy patients. Interestingly, a 4999 bp deletion reported in that study, overlapped with a 386 bp DAE which is located ~110 kb upstream of *GABRD* and which interacts with its promoter (**Figure 5D)**.

Figure 5. Variants in DAEs and nDAEs are associated with human disease. A) Bar graph showing the number of DAEs linked to their target genes in CP and GZ and their most enriched OMIM phenotypes. **B)** Plot showing the top-25 GWAS phenotypes that are enriched in DAEs compared to nDAEs (log2 odds ratio DAE/nDAE). **C)** Line graph showing the odds ratio, confidence interval, and p value for enrichment of CNVs from an ASD cohort at DAEs and nDAEs. CNVs data obtained from Brandler et al[81]. * $p < 0.05$; ** $p < 0.01$ (Fisher's exact test). **D)** Genome browser track showing the regulatory landscape of the GABRD gene. Indicated are a DAE (chr1: 1,840,449-1,840,835) that is interacting with the GABRD promoter, and a deletion (chr1: 1,840,001-1,845,000) that is found in an epilepsy patient (CNET0068) from Monlong et al.[82].* $p < 0.05$ (Fisher's exact test). **E)** Line graph showing the odds ratio, confidence interval, and p value for enrichment of SNV from an ASD cohort at DAEs and nDAEs. SNV data obtained from Zhou et al.[83].

Hence it is possible that deletion of this DAE affects *GABRD* expression, which might be implicated in the phenotype of that individual.

Third, we made use of *de novo* SNVs found in WGS from 1,790 ASD simplex families[83]. We found 932 *de novo* variants that overlapped all DAEs in ASD individuals compared to 829 variants overlapping all DAEs in unaffected individuals (odds ratio=1.07, *p*-value=0.157). We next repeated the analysis with only those DAEs that are interacting with known autism genes from the SFARI Gene database (n=1,003 genes)[84]. We found 26 cases and 11 controls with *de novo* variants in DAEs that interact with autism genes in CP (odds ratio=2.249703, *p*-value=0.021455), whereas for DAEs interacting with autism genes in GZ this was 20 cases and 17 controls (odds ratio=1.11955, *p*-value=0.745628) (**Figure 5E, Supplementary Table 12**). Interestingly, for each of the genes *CIB2*, *FBRSL1*, *PACS2*, *KDM4B* and *MYT1L* we found 2 individuals with autism with *de novo* variants in DAEs interacting with these genes. These variants are either absent or extremely rare in a large control cohort of *gnomAD*[140], possibly pointing to a role in causing the phenotype, although this will require further validation.

Together this indicates that DAEs are linked to disease relevant genes and are enriched for GWAS loci relevant for brain related traits and for variants linked to genetic disorders.

CRISPRi and zebrafish experiments confirm enhancer activity of DAEs regulating genes involved in epileptic encephalopathy

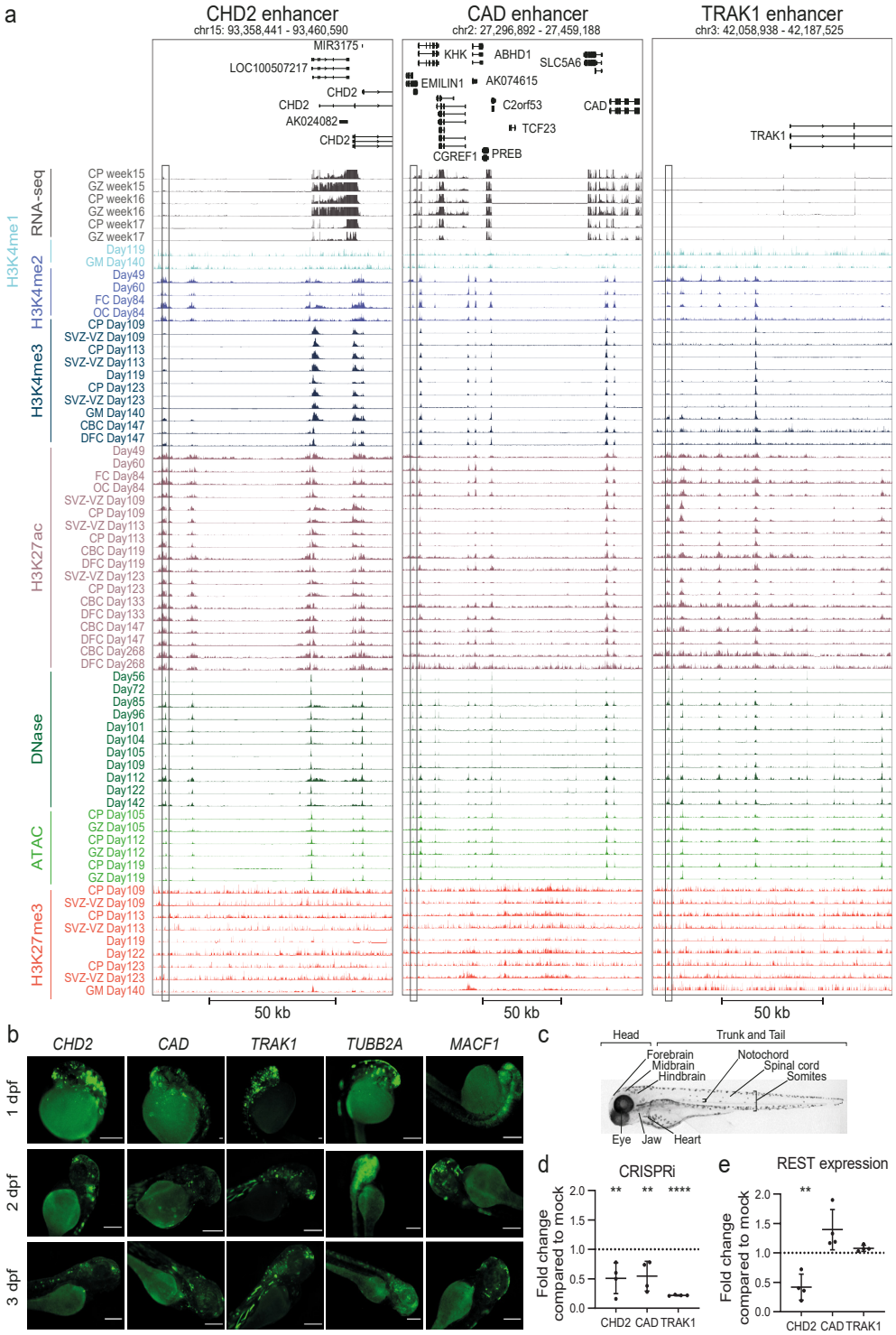
To further substantiate our findings, we validated the biological role of selected enhancers, using *in vivo* zebrafish transgenic reporter assays and CRISPR inhibition in human NSCs by focusing on enhancers linked to disease relevant genes.

CHD2 belongs to the chromodomain helicase DNA binding families of chromatin remodeling proteins, and haplo-insufficiency of this gene has been associated with a developmental and epileptic encephalopathy, presenting with early onset intractable seizures, cognitive regression, intellectual disability and ASD behaviors (OMIM #615369)[141]. Around 80 kb upstream of *CHD2*, we found a DAE that interacts with the *CHD2* promoter (**Figure 6A**). In NSC reporter assays, this region showed strong enhancer activity, and this was less pronounced in non-neural HEK cells (**Figure 1G**). To further study the biological relevance of this region, we first tested enhancer activity *in vivo* using zebrafish transgenesis. Out of the 36 analyzed zebrafish larvae, 61.1% showed GFP-expression in the forebrain at 1 day post fertilization

(dpf), and this increased to 81.8% at 2dpf and 87.9% at 3dpf, indicating enhancer activity (**Figure 6B,C**). Expression was also found in midbrain and hindbrain, at a slightly lower extent, in the eyes, in peripheral neurons and in the spinal cord (**Supplementary Table 13**). GFP expression in the developing zebrafish brain correlated with *in situ* hybridisations of endogenous chd2[142]. To test whether epigenome silencing of this enhancer would affect *CHD2* expression, we performed CRISPR interference (CRISPRi) by targeting dCas9-KRAB-MeCP2 to the enhancer region by co-expression of gRNAs with a GFP fluorescent reporter. Transfection efficiency in these experiments, based on FACS for GFP, was 78-92%, and this resulted in around 50% reduction of *CHD2* expression compared to mock cells transfected solely with dCas9-KRAB-MeCP2 (**Figure 6D**). Interestingly, it was previously shown that silencing of *CHD2* leads to reduced expression of *REST*[143]. In agreement with this, cells with reduced *CHD2* expression upon *CHD2* enhancer silencing showed reduced *REST* expression (**Figure 6E**). This confirms that *CHD2* is under control of the investigated DAE.

Bi-allelic variants in *CAD* cause an early infantile epileptic encephalopathy (OMIM #616457)[144], that is characterized by global developmental delay, loss of skills, therapy refractory epilepsy, brain atrophy, and dyserythropoietic anemia. We found an enhancer located in the third intron of *EMILIN1*, around 135 kb upstream of *CAD*, that interacts with the *CAD* promoter (**Figure 6A**) and which showed strong enhancer reporter activity in NSCs and only limited activity in HEK cells (**Figure 1G**). Targeting this region in NSCs by CRISPRi significantly diminished gene expression of *CAD* to around 50% compared to mock (**Figure 6D**). Similar to *CHD2*, *in vivo* reporter assays in zebrafish recapitulated *in situ* hybridisation results for cad[145]. From the 45 analyzed larvae, GFP expression was found in the forebrain of 88.9% larvae at 1dpf, which remained ~85% at 2 and 3 dpf. Again, GFP expression was observed also in midbrain, hindbrain, eyes, in peripheral neurons, notochord and spinal cord (**Figure 6B, Supplementary Table 13**).

We next focused on an enhancer interacting with *TRAK1*, located ~65 kb upstream of the TSS (**Figure 6A**). *TRAK1* is involved in mitochondrial trafficking, and bi-allelic loss-of-function variants in *TRAK1* are associated with developmental and epileptic encephalopathy (OMIM #618201)[146, 147]. Similar to the *CHD2* enhancer results, the *TRAK1* enhancer showed higher reporter assay activity in NSCs than in HEK cells (**Figure 1G**). Targeting of dCas9-KRAB-MeCP2 to the *TRAK1* enhancer reduced *TRAK1* expression to ~25% residual expression (**Figure 6D**). Interestingly, in the VISTA enhancer browser, another enhancer linked to *TRAK1* (hs2359), ~18 kb upstream


a
CHD2 enhancer
chr15: 93,358,441 - 93,460,590
CAD enhancer
chr2: 27,296,892 - 27,459,188
TRAK1 enhancer
chr3: 42,058,938 - 42,187,525
MIR3175
LOC100507217
CHD2
CHD2
AK024082
CHD2
KHK
ABHD1
SLC5A6
EMILIN1
AK074615
C2orf53
CAD
TCF23
CGREF1
PREB
TRAK1
RNA-seq
H3K4me1
H3K4me2
H3K4me3
H3K27ac
DNase
ATAC
H3K27me3
50 kb
50 kb
50 kb
b
CHD2
CAD
TRAK1
TUBB2A
MACF1
1 dpf
2 dpf
3 dpf
c
Head
Trunk and Tail
Forebrain
Midbrain
Hindbrain
Notochord
Spinal cord
Somites
Eye
Jaw
Heart
d
CRISPRi
Fold change compared to mock
CHD2 CAD TRAK1
e
REST expression
Fold change compared to mock
CHD2 CAD TRAK1

of the TSS, has been reported which did not show enhancer reporter activity in E11.5 mouse embryos. When testing the *TRAK1* enhancer identified here in zebrafish (**Figure 6B**), we found that from 55 larvae, 89.1% showed GFP-expression in the forebrain, as well as in the midbrain (74.5%) and hindbrain (85.5%). The larvae showed decreasing GFP expression in neurons outside of the brain over the different timepoint (83.6% at 1dpf, 65.5% at 2dpf and 67.3% at 3dpf) and increasing expression in both somites (89.1%) and heart (58.2%) at 3dpf, compared to 32.7% and 1.8% at 1dpf larvae, respectively. Moreover, this enhancer was active also in the eye, trunk and tail, notochord and, at 1 dpf, in the spinal cord (**Supplementary Table 13**). Finally, next to these three enhancers, we validated 7 additional enhancers linked to the genes *LRP1*, *LRP5*, *TUBB2A*, *ELOVL6*, *MACF1*, *C12orf4*, and *EBP41L1* using zebrafish reporter assays, and could confirm enhancer activity for all of them with >60% larvae expressing GFP (**Figure 6B, Supplementary Figure 8, Supplementary Table 13**). These included enhancers linked to the disease genes *MACF1* (OMIM #618325) and *TUBB2A* (OMIM #615763), of which coding pathogenic mutations cause brain malformations[148, 149], and *C12orf4* (OMIM #618221) of which bi-allelic variants cause intellectual disability[150]. Together, this shows that DAEs identified in this integrative analysis show enhancer activity *in vitro* and *in vivo* and regulate, amongst others, genes linked to Mendelian disorders.

Figure 6. CRISPRi and zebrafish experiments validate activity of DAEs regulating genes involved in neurogenetic disorders. A) Genome browser tracks showing enhancers interacting with *CHD2* (left), *CAD* (middle), and *TRAK1* (right). Shown are RNA-seq expression profiles, various histone modifications, and ATAC-seq and DNase profiles for various time points during human fetal brain development, as indicated. The tested DAEs are indicated by the box. **B)** Representative fluorescent images showing GFP expression of transgenic enhancer reporter assays in zebrafish larvae at 1, 2, and 3 dpf. Tested are the enhancers for CHD2, CAD, and TRAK1 (shown in A), and two additional enhancers for MACF1 and TUBB2A. The five tested enhancers induced GFP expression in the head of the larvae, amongst others in the forebrain in 61.1%, 81.8%, and 87.9% larvae for *CHD2*; 88.9%, 85.4%, and 85.7% for *CAD*; 87.1%, 70%, and 88.5% for TRAK1; 81.5%, 85.7%, and 76.2% for *MACF1*; and 87.5%, 100%, and 100% for *TUBB2A*, respectively at 1, 2, and 3 dpf. Also peripheral neuron-specific GFP expression was found, with 0%, 60.6%, and 21.2% for *CHD2*; 68.9%, 24.4%, and 51.4% for *CAD*; 83.6%, 65.5%, and 67.3% for *TRAK1*; 37%, 50%, and 33.3% for *MACF1*; and 50%, 83.3%, and 63.3% for *TUBB2A*, respectively at 1, 2, and 3 dpf. See also Additional file 14: Table S13. Scale bars represent 500 μm. **C)** Bright-field image of a wild type zebrafish larvae at 3 dpf (lateral view), with the anatomical sites that were scored for GFP expression indicated. **D)** qRT-PCR showing reduction of *CHD2*, *CAD*, and *TRAK1* expression in NSCs upon silencing of respective enhancer by dCas9-KRAB-MECP2. Data represent fold change of expression of respective genes compared to mock transfected cells (KRAB-MECP2 plasmid only, no gRNA plasmid). Two independent transfection experiments were performed, each in duplicate. All data points and standard deviation are shown. ** $p < 0.01$; **** $p < 0.0001$ (one-way ANOVA test followed by multiple comparison test (Fisher's LSD test). **E)** qRT-PCR showing reduction of *REST* expression in NSCs upon silencing of CHD2, CAD, or TRAK1 enhancers by dCas9-KRAB-MECP2. Data represent fold change of REST expression compared to mock transfected cells (KRAB-MECP2 plasmid only, no gRNA plasmid). Two independent transfection experiments were performed, each in duplicate. All data points and standard deviation are shown. ** $p < 0.01$ (one-way ANOVA test followed by multiple comparison test (Fisher's LSD test).

Discussion

Understanding the role of NCREs in development and disease still needs a significant effort at multiple levels: starting from identifying and annotating NCREs to investigating their target gene(s) and function. In the past few years, the identification and annotation of NCREs have gained a lot of attention. However, despite these developments, due to their sheer number and complex function, more studies and concerted efforts are needed to understand the role of NCREs in development and disease. Here we performed an integrative analysis of virtually all previously described putative enhancers and epigenome datasets of relevance for human brain development.

Our analysis has allowed us to first identify the intersection between previous studies and identify a list of putative NCREs. This is an important step as the different regions that were identified by previous investigations often have slightly different coordinates, length and quality. Our putative regions are thus the commonality between all the different studies that are conducted *hitherto*, but at the same time keep the originality in each of them. To further specify enhancers that might have a biological relevance, mapping epigenomic data to these putative regions allowed us to identify around 40 thousand enhancers that display epigenomic rearrangement during human brain development. These DAEs have different sequence characteristics compared to non-variable enhancers, are bound by distinct sets of TFs, regulate disease relevant genes and can harbor non-coding variants that are associated with human disease. Furthermore, our integrative analysis identified a large number of enhancers linked to known disease genes and expands on the knowledge of regulation of these genes. For example, *CHD2* expression regulation has so far only been known to be influenced by a highly conserved long non-coding RNA (lncRNA) referred to as CHD2 Adjacent Suppressive Regulatory RNA (CHASERR), which is located in proximity to the *CHD2* TSS, and which represses *Chd2* gene expression *in cis* [151]. It has been hypothesized that targeting CHASERR could be used to increase expression of *CHD2* in haploinsufficient individuals [151], and it will be interesting to explore whether targeting the enhancer region of *CHD2* that we find and validate here could be exploited as an alternative target of such a strategy. Similarly, the regulation by enhancers of other disease implicated genes that we validate here adds to the list of potential targets to find disease causing non-coding variants that disturb this regulation.

An interesting finding of our study is that by starting with putative enhancers and

variability of epigenome features over time during development, we recover DAEs and nDAEs that can be distinguished based on sequence characteristics, such as differences in GC content, the level of sequence constraint, tolerance to loss-of-function and differential profiles of TF binding. Also, these DAEs and nDAEs seem to be associated with distinct developmental processes, and result in differences in gene expression levels. It is tempting to speculate that the distinctive features between these two types of enhancers can be used to uncover key nucleotides responsible for those biological regulatory differences. It seems plausible that disturbing these functionally causative sequences could lead to altered physiology resulting in disease. Our analysis revealing GWAS loci enrichment and the link of DAEs supports this statement. We suggest that our results might help interpreting the effects of SNVs in non-coding sequences, which is at this stage not a trivial task. Our annotated database of DAE and nDAE will be instrumental to prioritize SNVs based on distinct sequence characteristics identified for these elements as well as to provide cues on potentially disturbed developmental processes based on differential temporal activity and regulatory targets of the enhancer in question. This in turn can instruct functional validation and help deciphering pathogenicity of variants. With an increasing number of whole genome sequencing data available, it is expected that more, possibly disease implicated, non-coding variants will be identified, and the need to classify those sequences in benign or pathogenic will only further increase. With more computational pathogenicity prediction tools available, such as the ncER score and outcomes of integrative analyses such as performed here that pinpoint likely functional sequences, it might become possible to further decipher the impact of these SNVs.

In this study, by using an integrative computational analysis of virtually all previously described putative enhancers and epigenome datasets, we identified a comprehensive compendium of likely functional enhancers that are involved human brain development and disease. By applying CRISPRi based silencing and zebrafish enhancer reporter assays, we show that these putative regions possess enhancer characteristics. We foresee that these enhancer sequences will be instrumental in identifying disease causing variants which might explain parts of the missing heritability in the field of clinical genetics.

Experimental procedure

Data visualization

To generate UCSC Genome Browser Tracks, aligned reads were converted to bedgraph using genomeCoverageBed, after which the bedGraphToBigWig tool from the UCSC Genome Browser was used to create a bigwig file[33, 34]. All enhancer regions, Enhancer-Gene interactions and TAD coordinates were uploaded directly as bed files. Other plots were drawn using R packages and **Figures 1-6** and **Supplementary Figures 1-8** were assembled in Adobe Illustrator[35]. **Supplementary Tables 1-14** were exported as Text or Excel files.

Data collection and processing

Collection of putative brain enhancers

To generate a comprehensive set of putative brain enhancers active during fetal brain development, we scrutinized PubMed and various enhancer databases (last assessed: April 2019), including amongst others EnhancerAtlas, the FANTOM5 Project, and the Vista Enhancer database[9, 11, 17-27]. This resulted in 1,595,292 putative enhancers (**Supplementary Table 1**). Enhancers with identical coordinates were deduplicated and the unique regions were used to determine putative critical regions (pCRs), reasoning that overlapping parts of a putative enhancer obtained from different sources might point to functional relevant regions of that putative enhancer. If there is any overlap between coordinates of putative enhancers derived from two or more databases, the pCRs were defined as maximum overlapping regions present in those databases using the BEDtools suite (mergeBed, intersectBed, genomeCoverageBed and groupBy sub-commands) (version 2.30.0)[33]. Putative enhancers that were only present in one of the input sources were also included in the pCRs (**Figure 1A, step 1**), as it cannot be excluded that these putative enhancers are biologically relevant. pCRs with length less than 50 bp and more than 1000 bp were excluded. To avoid any overlap with gene promoters, enhancers located within 2 kb upstream or 1 kb downstream of a transcriptional start site (TSS) (Ensembl GRCh37.p13 Release 102) were excluded using intersectBed. Following this procedure, we identified a total of 202,462 pCRs which were used for downstream analyses. Next, we excluded 299 pCRs that were not covered by sufficient amounts of epigenome data (less than 10 reads in at least two samples (see section on defining DAEs)), resulting in a final number of 202,163 pCRs (**Supplementary Table 3**). GREAT web interface was used (version 4.0.4) (http://great.stanford.edu/public/html/)[36] to visualize enhancer-TSS

distance (with *basal plus extension, proximal 5kb upstream and 1kb downstream, plus distal up to 100kb, including curated regulatory domains, and whole genome (GRCh37/hg19) as background* parameters) (**Supplementary Figure 2B**).

Epigenome data

Epigenome data were collected from the Roadmap Epigenomics Consortium, ENCODE, PsychENCODE and other studies (**Supplementary Table 2**). Epigenome data sets used for integration included histone modifications (H3K27ac, H3K27me3, H3K4me1, H3K4me2, H3K4me3) and chromatin accessibility (ATAC-seq and DNase-seq) from different brain regions and different human developmental stages (**Figure 1A, step 1**). To avoid any possible confounding biases because of the various pipelines used in different studies, we reanalyzed the raw FASTQ files using our analysis pipeline (**Additional File 1: Fig. S1**). First, adaptor contamination was removed using Trim Galore (version 0.6.5 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and trimmed data were aligned to the GRCh37/hg19 human genome using Bowtie2 aligner (version 2.4.2)(with *--very-sensitive* parameter)[37]. Only properly paired and uniquely mapped reads, with mapping quality more than 30 (MAPQ >=30), were kept followed by removing any possible duplicated reads using Picard's MarkDuplicates (version 4.0.1.1) (http://broadinstitute.github.io/picard/). These reads were used to define differentially active enhancers (DAEs).

Defining differentially active enhancers (DAEs)

We assumed that pCRs with high variability in different epigenome data (dynamic epigenomic rearrangement) across different developmental stages are more likely to be functional than other pCRs. To determine this variability, the number of overlapping reads (for each epigenome mark) with pCRs were counted using the multiBamCov sub-command of BEDtools and a matrix was generated that included enhancers as rows and epigenome features as columns. Epigenome features were from different brain regions and developmental stages. 299 pCRs with less than 10 reads were excluded, leaving 202,163 pCRs for this analysis. Subsequently, the raw read count matrix was normalized using TMM-normalization[38]. Since there were different developmental stages (time-point factor) and brain regions (brain part factor) in each epigenome data, a design matrix was generated for each factor separately. A limited number of samples without biological replicates were grouped together with other samples based on high correlation (Pearson correlation; $r > 0.89$). The DAEs were defined based on each design matrix using a generalized linear model and quasi-likelihood F-tests. In order to define the final DAE list, DAEs identified from at least two

epigenome data specific matrices were pooled. In total, this resulted in 39,709 DAEs (FDR adjusted *p-value* < 0.05). The remaining 162,454 pCRs that did not show variability were considered as nDAEs (**Supplementary Table 3**).

Identifying chromatin interactions

Enhancer-Gene interactions

In order to define Enhancer-Gene interaction, published HiC data from 3 human fetal brains, for cortical plate (CP) and germinal zone (GZ) at gestation week 17–18 were used[26]. This data provides 10 kb resolution bins for gene loop interactions and 40 kb resolution for topologically associating domain (TAD). Pre-calculated significant interactions were intersected with pCRs (DAEs and nDAEs) using intersectBed to define gene-enhancer interaction for both CP and GZ separately. Out of the 202,163 pCRs, 41,041 pCRs engaged in 101,366 interactions in CP, and 41,085 pCRs had 100,521 interactions in GZ. Enhancer-gene interactions locating within the same TAD were considered for downstream analyses (almost 80% of all interactions were intra-TAD). We only included protein coding and lincRNA genes in our analysis. To determine enhancer-enhancer interactions in **Additional File 1: Fig. S7A** we also intersected HiC data with pCRs, focusing on interactions between DAEs and both DAEs and nDAEs.

In addition to HiC, we employed other enhancer-gene interaction predictions including JEME (http://yiplab.cse.cuhk.edu.hk/jeme/)[40], ENCODE (https://ernstlab.biolchem.ucla.edu/roadmaplinking/)[41], FOCS (http://acgt.cs.tau.ac.il/focs/download.html)[42], and GeneHancer (downloaded from UCSC table browser; hg19; updated 2019)[43]. These databases apply statistical models on different types of omics data to predict enhancer-gene interactions. We collected fetal brain enhancer-gene predictions from JEME and ENCODE and all brain related enhancer-gene predictions from FOCS and GeneHancer, as the latter two resources do not specify fetal specific interactions. In addition, we used H3K27ac HiChIP derived chromatin interactions from several postnatal brain regions[44] cell type specific chromatin conformation capture data from PLAC-seq experiments in postnatal brain tissue[45] and enhancer-gene interaction predictions generated by the Activity-by-contact (ABC) model (https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction)[46]. We performed the ABC model by fixing the length of pCRs to 500 bps from the center (250 bps from each side). The enhancer activity was then determined considering DNase, and H3K27ac samples, and gene expression data from fetal brain[10] using default settings of the "run.neighborhoods.py" function. The ABC score was calculated by integrat-

ing the fetal HiC data and enhancer activity defined using the default settings of the "predict.py" function and adjusting "--hic_type bedpe", "--hic_resolution 10000" flags and ignoring "--cellType" flag.

Intersections between the pCRs and each of these predictions were considered as enhancer-gene interaction (**Supplementary Table 5**). The coordinates of the HiChIP interactions were lifted over to hg19 before intersecting with pCRs.

Functional enrichment analysis

Enhancer sequence characteristics analysis

To determine whether different DNA sequence features distinguish different enhancer groups and whether there is any association between these features and functional prediction, we considered the following features: (i) the non-coding essential regulation (ncER) score (https://github.com/TelentiLab/ncER_datasets/; updated 06-03-2019)[47]; (ii) GC content, as determined by the GCcontent R packages based on BSgenome. Hsapiens.UCSC.hg19 (version 1.4.3); (iii) conservation score for each enhancer, as derived from the gscores R packages based on phastCons100way. UCSC.hg19 (version 3.7.2)[48]; (iv) Orion scores[49]; (v) CADD scores[50]; (vi) Haploinsufficiency scores[51] and (vii) probability of loss-of-function intolerance (pLI) score[52]. The overlaps between DNA sequence features and enhancer coordinates were defined using intersectBed. As assessed enhancers (e.g. pCRs) varied in length between 50-1000 bp, and the above mentioned scores were given either at the nucleotide level or in certain bins (depending on the given scores from the individual resources), we calculated the mean value for each enhancer and used this in group comparisons. For gene specific scores (e.g. pLI), we plotted the scores of the genes linked to the enhancers. Statistical significant differences between groups were determined using Wilcoxon signed rank test in R.

Gene expression correlation

To compare gene expression levels of enhancer target genes between different groups, various transcriptome data were collected. This included transcriptome data from different brain regions and developmental stages, and also various control data from other fetal tissues from the Roadmap Epigenomics Consortium, ENCODE project, Allen human brain atlas and other studies (**Supplementary Table 6**)[7, 10, 29,53, 54]. Raw data (FASTQ) was quality controlled and adaptors and other contaminants were removed using Trim Galore (version 0.6.5), reads were mapped to the GRCh37/hg19

human genome assembly using STAR aligner (version 2.7)[55], and gene counts were obtained using htseq-count (version 0.12.4)[56]. Gene expression levels were normalized based on fragments per kilobase of transcript per million mapped reads (FPKM). To correlate enhancers to gene expression, enhancer-gene interactions were derived from the HiC data or the alternative enhancer-gene predictions as described above. Gene expression levels were plotted and statistical comparison was performed, between expression levels of subgroups, using Wilcoxon signed rank test in R. We also compared genes linked to DAEs and nDAEs by HiC, to the three trajectory gene groups from BrainVar[57]. For this we first found the overlap between genes interacting with DAEs/nDAEs using HiC-CP/GZ and each of the three trajectory groups (e.g. falling, rising and constant genes). We then determined the odds ratio between DAE and nDAE linked genes for each of the three gene trajectories, and used Fisher's exact test to determine significance.

Gene ontology analysis

For functional enrichment analysis, we used GREAT[36], Enrichr[58] and Metascape[59]. GREAT was used via the web interface (version 4.0.4) (http://great.stanford.edu/public/html/) using the following settings: basal plus extension, proximal 5 kb upstream and 1 kb downstream, plus distal up to 100 kb, including curated regulatory domains, and either whole genome or all pCRs as background, as indicated in the tabs of **Supplementary Table 4**. The -log 10 *p*-value was used to rank GREAT enrichment. Enrichr and Metascape were also used via the web interface (https://maayanlab.cloud/Enrichr/; https://metascape.org/gp/index.html#/main/step1), using the default settings and the whole genome set as background. All outputs of *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 and of Log*P*, enrichment, z score and log(*q* value) for Metascape are reported in **Supplementary Table 4, Supplementary Table 7 and Supplementary Table 9.**

Transcription factor binding enrichment

We used LOLA[60] using default settings to assess binding of known transcription factors to DAEs and nDAEs (**Figure 3**). We used motifs from the JASPAR motif database (using reference genome GRCh37/hg19 and LOLAJaspar database core), to test the TF enrichment in DAEs and nDAEs, using all pCRs as background. The mean rank index (a combination of *p*-value, odds ratio from Fisher's exact test and the raw number of overlapping regions), was used to rank the known motifs. To dis-

play TF enrichment in **Figure 3 C** and **F**, we re-scaled the rank between 0-100 using the rescale () R function. To further identify motifs across the different relative DAE or nDAE bins and distinguish motifs in the central versus peripheral parts of the enhancers (**Figure 3G, H**) we split the 100 relative bins into 20 groups of 5 consecutive bins and performed motif enrichment analysis using HOMER (version 4.11)[61], using function "findMotifsGenome.pl" and all pCRs as background. A *p-value* ≤ 0.01 was considered to select significantly enriched motifs.

Transposable element enrichment

The RepeatMask (GRCh37/hg19, updated 20-02-2020) was downloaded from the UCSC table browser and joined to the pCRs. To determine enrichment of transposable elements in brain enhancers, we followed a strategy previously used when investigating active enhancers in human embryonic stem cells[62]. The number of overlaps of each type of repeat (n_overlaps) with all pCRs (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. DAE, nDAE etc.) in any tested group 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 calculate the observed versus expected ratio (O/E). We considered repeats with O/E < 0.5 as depleted, or O/E > 1 as enriched. For the subsequent data interpretation we only focused on transposable elements that were present multiple times (n_overlap > 15) in all pCRs (**Supplementary Table 8**).

Disease relevance enrichment

The *Online Mendelian inheritance in Man* (OMIM) gene list (updated 28-09-2020) was downloaded using biomaRt R package[63] from Ensembl GRCh37.p13 Release 101. The GWAS catalog (GRCh37/hg19,updated 17-03-2021) was downloaded from the UCSC table browser. The GWAS catalog was manually filtered to keep brain related studies and their variants with *p*-value ≤ 9e-08 (**Supplementary Table 11**). Stratified LD score regression analysis was performed by implementing the full baseline model to calculate enrichment (https://github.com/bulik/ldsc/wiki)[64, 65]. Annotation and LD score files were created using the "make_annot.py" and "ldsc.py" functions, respectively. Partitioning heritability was performed using the "ldsc.py" script considering default parameters with "-- h2" flag. We obtained GWAS summary statistics for several brain-related traits including Alzheimer's disease[66], Anorexia Nervosa[67], Anxiety[68], Attention Deficit Hyperactivity Disorder[69], Autism Spectrum Disorder[70], Bipolar Disorder and Schizophrenia[71], Epilepsy[72], Insomnia[73],

Intelligence[74], Major Depressive Disorder[75], Neuroticism[76, 77], Obsessive compulsive disorder / Tourette syndrome[78], Parkinson's disease[79], and Schizophrenia[80] (**Supplementary Table 11**). *Z-scores* were used to calculate the *p-values* which were corrected for multiple hypothesis testing using the Benjamini-Hochberg method. For CNV analysis, we retrieved pre-processed published data from Brandler *et al* (their supplemental table 9: de_novo_SVs sheet, and their supplemental table 7: Primary CR Trans and Replication CR Trans sheets) [81] and Monlong *et al* (cnvs-PopSV-Epilepsy-198affected-301controls-5kb.tsv.gz file in https://figshare.com/s/20dfdedcc4718e465185)[82]. For SNV analysis of the ASD simplex families, we collected *de novo* variants from supplemental table 1 of Zhou et al [83]. Autism genes were collected from the SFARI Gene database (http://gene.sfari.org/database/human-gene) [84]. The overlap between enhancer regions (DAE and nDAE) and each data set was determined using intersectBed. The odds ratio and *p*-value between DAE and nDAE was calculated using fisher.test () R function. The Haldane–Anscombe correction was used to adjust the odds ratio.

Distribution of features across enhancer bins

To investigate the distribution of enrichment of different features (ncER score, GC content, phastCons score and epigenome data) across enhancers, we divided the enhancer regions into 10 bp bins and calculated the relative scores (the median value for ncER score, GC content, phastCons score) and the number of reads (for epigenome data) for each bin. As the enhancers under investigation differed in size between 50-1000 bp, to make enrichments between enhancers comparable, we re-scaled each enhancer bin. To this end, we calculated a relative position between 1-100 for each bin of each enhancer, where 1 is the first bin, and 100 is the last bin of each individual enhancer. We then plotted the distribution of each feature across all these re-scaled enhancer bins.

DAE clustering analysis

The matrix of DAEs was used to determine the pattern of epigenome data through different developmental stages. To determine the optimal clustering algorithm, we used clValid R package which simultaneously compares multiple clustering algorithms (hierarchical, kmeans, model-based, pam and clara). Based on this, the pam algorithm (which is similar to k-means but more robust to noise and outliers) was selected to cluster DAEs using the spearman distance and ward.D^2 method. To define the optimal number of clusters, we used fviz_nbclust and NbClust R packages which compute different indices by bootstrapping (n=1000). The predicted number of clus-

ters were tested using the silhouette R package to examine whether the clustering performed correctly. This approach resulted in 2 clusters for DAEs and epigenome features at 8-12 PCW, 3 clusters for 13-18 PCW and 2 clusters for >18 PCW, for each of CP and GZ, respectively. For each cluster, we determined the gene expression of protein coding genes interacting with the DAEs from each cluster, as obtained from published RNA-seq data sets. Significant differences in expression levels between different clusters were determined using the Wilcoxon signed rank test in R. Also, target genes linked to each cluster were used for functional enrichment analysis using Enrichr[58], as described under gene ontology analysis (**Supplementary Table 9**).

Enhancer cell type specificity and their dynamics in adult brain

To determine cell type specificity of enhancers, we compared DAEs and nDAEs to recently described cell-type specific regulatory elements from two studies on adult brain (obtained from supplementary data Set 4 (data lifted over to hg19) of Corces *et al*[44] and Supplementary Table 5 of Nott *et al*[45] and a study of fetal brain (obtained from supplementary file 4 of Domcke *et al*, specificity scores for top 10000 regions[85]. We used bedtools to intersect DAEs or nDAEs and different cell type specific regulatory elements. For all DAEs and nDAEs linked to target genes in CP and GZ by HiC, we compared dynamics of H3K27ac levels in both fetal and adult samples, using H3K27ac data from Li *et al*[86]. Clustering analysis was performed as described under "DAE clustering analysis" above. Gene ontology analysis for each defined cluster was performed using Enrichr, as described above.

Experimental validation

Cell culture

HEK293 LTV cells (Cell Biolabs) were cultured in DMEM medium (Gibco), supplemented with 10% FBS at 37ºC, 5% CO2. Human neural stem cells (NSCs) (Gibco), were cultured in NSC medium (KnockOut DMEM-F12 (Gibco), 2 mM L-glutamine (Gibco), 20 ng/ml bFGF (Peprotech), 20 ng/ml EGF (Peprotech), 2% StemPro Neural supplement (Gibco), 100U/ml penicillin and 100µg/ml streptomycin), as previously described [87].

Enhancer activity in STARR-seq reporter plasmids

For experimental validation in **Figure 1G**, we randomly selected 22 DAEs that showed interaction with a target gene by HiC, and of which the target gene was expressed in neural stem cells, as indicated from our previously generated RNA-seq

data (GSE137129;[87]). DAEs were amplified from genomic DNA and cloned into the STARR-seq plasmid (kind gift of A.Stark)[88] as previously described[62]. For the additional tested enhancer deletions (**Supplementary Fig. 4**), the obtained STARR-seq plasmids containing *IRF2BPL*, *CHD2* and *MACF1* enhancers were modified by site-directed mutagenesis to remove regions with high or low ncER score. The following regions were deleted: *IRF2BPL* (chr14: 77422484-77422514); *CHD2* (ncER1 chr15: 93363603-93363640, ncER3 chr15: 93363780-93363790); *MACF1* (ncER1 chr1: 39598824-39598844, ncER2 chr1:39598744-39598754). The regions with low ncER score at the 5' and 3' ends (80-100bp) of *IRF2BPL*, *CHD2* and *MACF1* enhancers were excluded by Gibson assembly. Primer sequences are provided in **Supplementary Table 14**. HEK293 and NSC were transfected with STARR-seq plasmid containing enhancer regions using polyethylenimine (PEI, Sigma) or Lipofectamine™ Stem Transfection Reagent (Thermo Scientific) respectively. Spike-in of a pmCherry-N1 plasmid (Clonetech) was used as a transfection control. 24h post transfection cells were collected, stained with Hoechst dye and the enhancer activity was measured by FACS analysis (20,000 cells per sample). GFP-positive cells within the mCherry-positive population were quantified to assess enhancer activity compared to an empty STARR-seq vector. Two independent transfection experiments were performed, each in duplicates. Statistical analysis was performed using a one-way ANOVA test followed by multiple comparison test (Fisher's LSD test). Calculations were conducted in GraphPad Prism (version 8).

dCas9-KRAB-MeCP2 silencing of active enhancers in NSC

We selected DAEs linked to *CHD2*, *CAD* and *TRAK1* and designed for each DAE two targeting gRNAs (primer sequences are given in **Supplementary Table 14**). gRNAs were cloned into a pgRGFP plasmid (Addgene #82695, a kind gift of Allan Mullen)[89]. NSCs were co-transfected with dCas9-KRAB-MeCP2 (Addgene #110824, kind gift of Alejandro Chavez and George Church)[90] and the two gRNAs/DAE and collected for RNA isolation 48h post transfection. Transfection efficiency was estimated by FACS analysis (78-92% GFP-positive cells detected). RNA was isolated using TRI reagent (Sigma) followed by cDNA preparation using iSCRIPT cDNA synthesis kit (BioRad). Fold change in gene expression (ΔΔct method) was evaluated by qPCR (iTaq universal SYBR Green Supermix) (Sigma), performed in CFX96RTS thermal cycler (Bio-Rad), as previously described[87]. TBP expression was used as housekeeping normalization control. Statistical analysis was performed using a one-way ANOVA test followed by multiple comparison test (Fisher's LSD test). Calculations were conducted in GraphPad Prism (version 8).

Zebrafish studies

Zebrafish (Danio rerio) were raised and maintained under standard conditions[91]. Adult and larval fish were kept on a 14h/10h light–dark cycle at 28°C. Larvae were kept in HEPES-buffered E3 medium. Media was refreshed daily and at 24 hpf 0.003% 1-phenyl 2-thiourea (PTU) was added to prevent pigmentation. All zebrafish experiments were performed in compliance with Dutch animal welfare legislation. Selected DAEs used in the *in vitro* experiments were transferred by Gibson assembly between the AscI and PacI site of a E1b-GFP-Tol2 enhancer assay plasmid (a kind gift of Ramon Birnbaum)[92] containing an E1b minimal promoter followed by GFP, using the following transfer primers: Transfer_fw: 5'-AGATGGGCCCTCGGGTAGAGCATGCACCGG-3' and Transfer_rv: 5'-TCGAGAGATCTTAATGGCCGAATTCGTCGA-3'. Constructs were injected into zebrafish embryos using standard procedures, together with Tol2 mRNA to facilitate genomic integration. At least 50 embryos were injected per construct in at least two different injection experiments. GFP expression was observed and annotated at 1, 2 and 3 dpf by a fluorescent Leica M165FC stereomicroscope (**Supplementary Table 13**). Images were analyzed using imageJ (FIJI). An enhancer was considered active when at least 30% of the larvae showed consistent GFP expression.

Availability of data and materials

All primary data used in this study are given in **Supplementary Tables 1, 2 and 6** Some of the primary data that were used to support the findings of this study are available from dbGaP and PsychENCODE, but restrictions apply to the availability of these data, which were used under license for the current study and so are not publicly available (third party data). The source code and all processed data for all analysis performed in this study are available in the repositories https://github.com/syousefi87/Differentially-Active-Enhancers[152], and https://figshare.com/projects/Differentially-Active-Enhancers/122965[153].

Declarations

Ethics approval and consent to participate:

For zebrafish studies, no larvae older than 5 days post fertilization were used. Zebrafish were kept according to guidelines of the EMC animal welfare office and all zebrafish experiments were performed in compliance with Dutch animal welfare legislation.

Funding

RD is supported by a China Scholarship Council (CSC) PhD Fellowship (201906300026) for her PhD studies at the Erasmus Medical Center, Rotterdam, the Netherlands. EM is supported by Netherlands Organisation for Scientific Research (ZonMW Off Road grant). TSB 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. Funding bodies did not have any influence on study design, results and data interpretation or final manuscript.

Authors' contribution

SY performed primary computational analysis, with help of RD. KL, AN, and EP performed validation experiments in cells. EMS, HCvsL and TvH performed zebrafish validation experiments. SY, EM and TSB wrote the manuscript with input from all authors. All authors read and approved the final manuscript. EM and TSB conceived and jointly supervised the work.

Acknowledgements

We would like to dedicate this paper to prof. Robert M.W. Hofstra, head of the department of Clinical Genetics at Erasmus MC, who sadly passed away during the preparation of this work. Ramon Birnbaum (Ben-Gurion University of the Negev, Israel) is acknowledged for sharing zebrafish reporter plasmids, and Raymond Poot (Erasmus MC, The Netherlands) for providing neural stem cells.

We would like to thank the following third parties for providing approved access to the indicated data sets that were used to support the findings presented in this study:

-phs000755.v2.p1: "BrainSpan Atlas of the Human Brain". The datasets used for the analysis described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000406.v2.p1. Submission of the data, phs000406.v2.p1, to dbGaP was provided by Dr. Nenad Sestan. Collection of the data and analysis was supported by grants from the National Institutes of Health (MH089929, MH081896, and MH090047). Additional support was provided by the Kavli Foundation, a James S. McDonnell Foundation Scholar Award, NARSAD, and the Foster-Davis Foundation.

-phs000791.v1.p1: "Roadmap Epigenomics Program - UCSF". Funding support for the NIH Roadmap Epigenomics Program was provided through the NIH Common Fund (Office of Strategic Coordination). Support for collection of datasets and samples was provided by a series of UO1 cooperative agreements with The Broad Institute [1U01ES017155-01], The Ludwig Institute for Cancer Research [1U01ES017166-01], The University of California San Francisco [1U01ES017154-01], and The Uni-

versity of Washington [1U01ES017156-01]. Data analysis and coordination were supported by an agreement with Baylor College of Medicine [1U01DA025956-01]. Assistance with data curation was supplied by GEO, and data access and visualization was supported by the NCBI.

-phs001226.v1.p1: "Regulatory Genomics of Human Embryonic Development". The datasets used for the analysis described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs001226.v1.p1.

-phs001438.v1.p1: "The dynamic landscape of open chromatin during human cortical neurogenesis". Datasets from this study used for the analyses described in this manuscript were generated in the Geschwind laboratory and supported by NIH grants to D.H.G. (5R01MH060233; 5R01MH100027; 3U01MH103339; 1R01MH110927; 1R01MH094714). Datasets were obtained from dbGaP found at http://www.ncbi.nlm.nih.gov/gap through dbGaP study accession numbers phs001438.v1.p1.

-Data access from PsychENCODE was obtained for the following data sets: SynapseID: syn12033248[9]; SynapseID: syn17092080[86]**.** Data were generated as part of the PsychENCODE Consortium supported by: U01MH103339, U01MH103365, U01MH103392, U01MH103340, U01MH103346, R01MH105472, R01MH094714, R01MH105898, R21MH102791, R21MH105881, R21MH103877, and P50MH106934 awarded to: Schahram Akbarian (Icahn School of Medicine at Mount Sinai), Gregory Crawford (Duke), Stella Dracheva (Icahn School of Medicine at Mount Sinai), Peggy Farnham (USC), Mark Gerstein (Yale), Daniel Geschwind (UCLA), Thomas M. Hyde (LIBD), Andrew Jaffe (LIBD), James A. Knowles (USC), Chunyu Liu (UIC), Dalila Pinto (Icahn School of Medicine at Mount Sinai), Nenad Sestan (Yale), Pamela Sklar (Icahn School of Medicine at Mount Sinai), Matthew State (UCSF), Patrick Sullivan (UNC), Flora Vaccarino (Yale), Sherman Weissman (Yale), Kevin White (UChicago) and Peter Zandi (JHU).

References

1 Spitz F, Furlong EE: Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* 2012, 13:613-626.
2 Nord AS, West AE: Neurobiological functions of transcriptional enhancers. *Nat Neurosci* 2020, 23:5-14.
3 D'Haene E, Vergult S: Interpreting the impact of noncoding structural variation in neurodevelopmental disorders. *Genet Med* 2021, 23:34-46.
4 Perenthaler E, Yousefi S, Niggl E, Barakat TS: Beyond the Exome: The Non-coding Genome and Enhancers in Neurodevelopmental Disorders and Malformations of Cortical Development. *Front Cell Neurosci* 2019, 13:352.
5 Carullo NVN, Day JJ: Genomic Enhancers in Brain Health and Disease. *Genes (Basel)* 2019, 10.
6 Chatterjee S, Ahituv N: Gene Regulatory Elements, Major Drivers of Human Disease. *Annu Rev Genomics Hum Genet* 2017, 18:45-63.
7 Consortium EP: An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012, 489:57-74.
8 Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al: Integrative analysis of 111 reference human epigenomes. *Nature* 2015, 518:317-330.
9 Amiri A, Coppola G, Scuderi S, Wu F, Roychowdhury T, Liu F, Pochareddy S, Shin Y, Safi A, Song L, et al: Transcriptome and epigenome landscape of human cortical development modeled in organoids. *Science* 2018, 362.
10 Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, et al: The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* 2010, 28:1045-1048.
11 Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al: An atlas of active enhancers across human cell types and tissues. *Nature* 2014, 507:455-461.
12 Townsley KG, Brennand KJ, Huckins LM: Massively parallel techniques for cataloguing the regulome of the human brain. *Nat Neurosci* 2020, 23:1509-1521.
13 Ryan GE, Farley EK: Functional genomic approaches to elucidate the role of enhancers during development. *Wiley Interdiscip Rev Syst Biol Med* 2020, 12:e1467.
14 Montalbano A, Canver MC, Sanjana NE: High-Throughput Approaches to Pinpoint Function within the Noncoding Genome. *Mol Cell* 2017, 68:44-59.
15 Kleftogiannis D, Kalnis P, Bajic VB: Progress and challenges in bioinformatics approaches for enhancer identification. *Brief Bioinform* 2016, 17:967-979.
16 Rojano E, Seoane P, Ranea JAG, Perkins JR: Regulatory variants: from detection to predicting impact. *Brief Bioinform* 2019, 20:1639-1654.
17 Gao T, Qian J: EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res* 2020, 48:D58-D64.
18 Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, Clarke D, Gu M, Emani P, Yang YT, et al: Comprehensive functional genomic resource and integrative model for the human brain. *Science* 2018, 362.
19 Visel A, Minovitsky S, Dubchak I, Pennacchio LA: VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* 2007, 35:D88-92.
20 Vermunt MW, Reinink P, Korving J, de Bruijn E, Creyghton PM, Basak O, Geeven G, Toonen PW, Lansu N, Meunier C, et al: Large-scale identification of coregulated enhancer networks in the adult human brain. *Cell Rep* 2014, 9:767-779.
21 Valensisi C, Andrus C, Buckberry S, Doni Jayavelu N, Lund RJ, Lister R, Hawkins RD: Epigenomic Landscapes of hESC-Derived Neural Rosettes: Modeling Neural Tube Formation and Diseases. *Cell Rep* 2017, 20:1448-1462.
22 Sun W, Poschmann J, Cruz-Herrera Del Rosario R, Parikshak NN, Hajan HS, Kumar V, Ramasamy R, Belgard TG, Elanggovan B, Wong CCY, et al: Histone Acetylome-wide Association Study of Autism Spectrum Disorder. *Cell* 2016, 167:1385-1397 e1311.
23 Emera D, Yin J, Reilly SK, Gockley J, Noonan JP: Origin and evolution of developmental en-

hancers in the mammalian neocortex. *Proc Natl Acad Sci U S A* 2016, 113:E2617-2626.
24 Vermunt MW, Tan SC, Castelijns B, Geeven G, Reinink P, de Bruijn E, Kondova I, Persengiev S, Netherlands Brain B, Bontrop R, et al: Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. *Nat Neurosci* 2016, 19:494-503.
25 Yao P, Lin P, Gokoolparsadh A, Assareh A, Thang MW, Voineagu I: Coexpression networks identify brain region-specific enhancer RNAs in the human brain. *Nat Neurosci* 2015, 18:1168-1174.
26 Won H, de la Torre-Ubieta L, Stein JL, Parikshak NN, Huang J, Opland CK, Gandal MJ, Sutton GJ, Hormozdiari F, Lu D, et al: Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* 2016, 538:523-527.
27 Capra JA, Erwin GD, McKinsey G, Rubenstein JL, Pollard KS: Many human accelerated regions are developmental enhancers. *Philos Trans R Soc Lond B Biol Sci* 2013, 368:20130025.
28 Reilly SK, Yin J, Ayoub AE, Emera D, Leng J, Cotney J, Sarro R, Rakic P, Noonan JP: Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* 2015, 347:1155-1159.
29 de la Torre-Ubieta L, Stein JL, Won H, Opland CK, Liang D, Lu D, Geschwind DH: The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. *Cell* 2018, 172:289-304 e218.
30 Meuleman W, Muratov A, Rynes E, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F, Teodosiadis A, et al: Index and biological spectrum of human DNase I hypersensitive sites. *Nature* 2020, 584:244-251.
31 Vierstra J, Lazar J, Sandstrom R, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F, Haugen E, et al: Global reference mapping of human transcription factor footprints. *Nature* 2020, 583:729-736.
32 Zhang J, Lee D, Dhiman V, Jiang P, Xu J, McGillivray P, Yang H, Liu J, Meyerson W, Clarke D, et al: An integrative ENCODE resource for cancer genomics. *Nat Commun* 2020, 11:3696.
33 Quinlan AR: BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* 2014, 47:11 12 11-34.
34 Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D: BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* 2010, 26:2204-2207.
35 Golding M: Adobe Illustrator CS5 : for Web and Interactive Design. 2010.
36 McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G: GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* 2010, 28:495-501.
37 Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012, 9:357-359.
38 Robinson MD, Oshlack A: A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 2010, 11:R25.
39 Robinson MD, McCarthy DJ, Smyth GK: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010, 26:139-140.
40 Cao Q, Anyansi C, Hu X, Xu L, Xiong L, Tang W, Mok MTS, Cheng C, Fan X, Gerstein M, et al: Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nat Genet* 2017, 49:1428-1436.
41 Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al: Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 2011, 473:43-49.
42 Hait TA, Amar D, Shamir R, Elkon R: FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map. *Genome Biol* 2018, 19:56.
43 Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M, Safran M, et al: GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* 2017, 2017.
44 Corces MR, Shcherbina A, Kundu S, Gloudemans MJ, Fresard L, Granja JM, Louie BH, Eulalio T, Shams S, Bagdatli ST, et al: Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat Genet* 2020, 52:1158-1168.
45 Nott A, Holtman IR, Coufal NG, Schlachetzki JCM, Yu M, Hu R, Han CZ, Pena M, Xiao J, Wu Y, et al: Brain cell type-specific enhancer-promoter interactome maps and disease-risk association.

Science 2019, 366:1134-1139.
46 Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, Grossman SR, Anyoha R, Doughty BR, Patwardhan TA, et al: Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat Genet* 2019, 51:1664-1669.
47 Wells A, Heckerman D, Torkamani A, Yin L, Sebat J, Ren B, Telenti A, di Iulio J: Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nat Commun* 2019, 10:5241.
48 Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al: Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005, 15:1034-1050.
49 Gussow AB, Copeland BR, Dhindsa RS, Wang Q, Petrovski S, Majoros WH, Allen AS, Goldstein DB: Orion: Detecting regions of the human non-coding genome that are intolerant to variation using population genetics. *PLoS One* 2017, 12:e0181604.
50 Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014, 46:310-315.
51 Xu D, Gokcumen O, Khurana E: Loss-of-function tolerance of enhancers in the human genome. *PLoS Genet* 2020, 16:e1008663.
52 Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al: Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016, 536:285-291.
53 Yan L, Guo H, Hu B, Li R, Yong J, Zhao Y, Zhi X, Fan X, Guo F, Wang X, et al: Epigenomic Landscape of Human Fetal Brain, Heart, and Liver. *J Biol Chem* 2016, 291:4386-4398.
54 Miller JA, Ding SL, Sunkin SM, Smith KA, Ng L, Szafer A, Ebbert A, Riley ZL, Royall JJ, Aiona K, et al: Transcriptional landscape of the prenatal human brain. *Nature* 2014, 508:199-206.
55 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013, 29:15-21.
56 Anders S, Pyl PT, Huber W: HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 2015, 31:166-169.
57 Werling DM, Pochareddy S, Choi J, An JY, Sheppard B, Peng M, Li Z, Dastmalchi C, Santpere G, Sousa AMM, et al: Whole-Genome and RNA Sequencing Reveal Variation and Transcriptomic Coordination in the Developing Human Prefrontal Cortex. *Cell Rep* 2020, 31:107489.
58 Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A: Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 2013, 14:128.
59 Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK: Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* 2019, 10:1523.
60 Sheffield NC, Bock C: LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* 2016, 32:587-589.
61 Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010, 38:576-589.
62 Barakat TS, Halbritter F, Zhang M, Rendeiro AF, Perenthaler E, Bock C, Chambers I: Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell Stem Cell* 2018, 23:276-288 e278.
63 Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W: BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 2005, 21:3439-3440.
64 Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, Anttila V, Xu H, Zang C, Farh K, et al: Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* 2015, 47:1228-1235.
65 Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C, Patterson N, Daly MJ, Price AL, Neale BM: LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015, 47:291-295.

66 Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, Boland A, Vronskaya M, van der Lee SJ, Amlie-Wolf A, et al: Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. *Nat Genet* 2019, 51:414-430.
67 Duncan L, Yilmaz Z, Gaspar H, Walters R, Goldstein J, Anttila V, Bulik-Sullivan B, Ripke S, Eating Disorders Working Group of the Psychiatric Genomics C, Thornton L, et al: Significant Locus and Metabolic Genetic Correlations Revealed in Genome-Wide Association Study of Anorexia Nervosa. *Am J Psychiatry* 2017, 174:850-858.
68 Otowa T, Hek K, Lee M, Byrne EM, Mirza SS, Nivard MG, Bigdeli T, Aggen SH, Adkins D, Wolen A, et al: Meta-analysis of genome-wide association studies of anxiety disorders. *Mol Psychiatry* 2016, 21:1391-1399.
69 Demontis D, Walters RK, Martin J, Mattheisen M, Als TD, Agerbo E, Baldursson G, Belliveau R, Bybjerg-Grauholm J, Baekvad-Hansen M, et al: Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat Genet* 2019, 51:63-75.
70 Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, Pallesen J, Agerbo E, Andreassen OA, Anney R, et al: Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet* 2019, 51:431-444.
71 Bipolar D, Schizophrenia Working Group of the Psychiatric Genomics Consortium. Electronic address drve, Bipolar D, Schizophrenia Working Group of the Psychiatric Genomics C: Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. *Cell* 2018, 173:1705-1715 e1716.
72 International League Against Epilepsy Consortium on Complex Epilepsies. Electronic address e-auea: Genetic determinants of common epilepsies: a meta-analysis of genome-wide association studies. *Lancet Neurol* 2014, 13:893-903.
73 Jansen PR, Watanabe K, Stringer S, Skene N, Bryois J, Hammerschlag AR, de Leeuw CA, Benjamins JS, Munoz-Manchado AB, Nagel M, et al: Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nat Genet* 2019, 51:394-403.
74 Savage JE, Jansen PR, Stringer S, Watanabe K, Bryois J, de Leeuw CA, Nagel M, Awasthi S, Barr PB, Coleman JRI, et al: Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat Genet* 2018, 50:912-919.
75 Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, Adams MJ, Agerbo E, Air TM, Andlauer TMF, et al: Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet* 2018, 50:668-681.
76 Nagel M, Jansen PR, Stringer S, Watanabe K, de Leeuw CA, Bryois J, Savage JE, Hammerschlag AR, Skene NG, Munoz-Manchado AB, et al: Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat Genet* 2018, 50:920-927.
77 Okbay A, Baselmans BM, De Neve JE, Turley P, Nivard MG, Fontana MA, Meddens SF, Linner RK, Rietveld CA, Derringer J, et al: Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat Genet* 2016, 48:624-633.
78 Yu D, Sul JH, Tsetsos F, Nawaz MS, Huang AY, Zelaya I, Illmann C, Osiecki L, Darrow SM, Hirschtritt ME, et al: Interrogating the Genetic Determinants of Tourette's Syndrome and Other Tic Disorders Through Genome-Wide Association Studies. *Am J Psychiatry* 2019, 176:217-227.
79 Nalls MA, Blauwendraat C, Vallerga CL, Heilbron K, Bandres-Ciga S, Chang D, Tan M, Kia DA, Noyce AJ, Xue A, et al: Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol* 2019, 18:1091-1102.
80 Pardinas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N, Legge SE, Bishop S, Cameron D, Hamshere ML, et al: Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet* 2018, 50:381-389.
81 Brandler WM, Antaki D, Gujral M, Kleiber ML, Whitney J, Maile MS, Hong O, Chapman TR, Tan S, Tandon P, et al: Paternally inherited cis-regulatory structural variants are associated with autism. *Science* 2018, 360:327-331.
82 Monlong J, Girard SL, Meloche C, Cadieux-Dion M, Andrade DM, Lafreniere RG, Gravel M,

Spiegelman D, Dionne-Laporte A, Boelman C, et al: Global characterization of copy number variants in epilepsy patients from whole genome sequencing. *PLoS Genet* 2018, 14:e1007285.
83 Zhou J, Park CY, Theesfeld CL, Wong AK, Yuan Y, Scheckel C, Fak JJ, Funk J, Yao K, Tajima Y, et al: Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat Genet* 2019, 51:973-980.
84 Banerjee-Basu S, Packer A: SFARI Gene: an evolving database for the autism research community. *Dis Model Mech* 2010, 3:133-135.
85 Domcke S, Hill AJ, Daza RM, Cao J, O'Day DR, Pliner HA, Aldinger KA, Pokholok D, Zhang F, Milbank JH, et al: A human cell atlas of fetal chromatin accessibility. *Science* 2020, 370.
86 Li M, Santpere G, Imamura Kawasawa Y, Evgrafov OV, Gulden FO, Pochareddy S, Sunkin SM, Li Z, Shin Y, Zhu Y, et al: Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science* 2018, 362.
87 Perenthaler E, Nikoncuk A, Yousefi S, Berdowski WM, Alsagob M, Capo I, van der Linde HC, van den Berg P, Jacobs EH, Putar D, et al: 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* 2020, 139:415-442.
88 Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A: Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 2013, 339:1074-1077.
89 Daneshvar K, Pondick JV, Kim BM, Zhou C, York SR, Macklin JA, Abualteen A, Tan B, Sigova AA, Marcho C, et al: DIGIT Is a Conserved Long Noncoding RNA that Regulates GSC Expression to Control Definitive Endoderm Differentiation of Embryonic Stem Cells. *Cell Rep* 2016, 17:353-365.
90 Yeo NC, Chavez A, Lance-Byrne A, Chan Y, Menn D, Milanova D, Kuo CC, Guo X, Sharma S, Tung A, et al: An enhanced CRISPR repressor for targeted mammalian gene regulation. *Nat Methods* 2018, 15:611-616.
91 Westerfield M: The Zebrafish Book. A Guide for the Laboratory Use of Zebrafish (Danio rerio). 2000.
92 D'Haene E, Bar-Yaacov R, Bariah I, Vantomme L, Van Loo S, Cobos FA, Verboom K, Eshel R, Alatawna R, Menten B, et al: A neuronal enhancer network upstream of MEF2C is compromised in patients with Rett-like characteristics. *Hum Mol Genet* 2019, 28:818-827.
93 Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P: Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat Ecol Evol* 2018, 2:152-163.
94 Lecellier CH, Wasserman WW, Mathelier A: Human Enhancers Harboring Specific Sequence Composition, Activity, and Genome Organization Are Linked to the Immune Response. *Genetics* 2018, 209:1055-1071.
95 Colbran LL, Chen L, Capra JA: Short DNA sequence patterns accurately identify broadly active human enhancers. *BMC Genomics* 2017, 18:536.
96 Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G: Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* 2010, 42:631-634.
97 Jacques PE, Jeyakani J, Bourque G: The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet* 2013, 9:e1003504.
98. Glinsky G, Barakat TS: The evolution of Great Apes has shaped the functional enhancers' landscape in human embryonic stem cells. *Stem Cell Res* 2019, 37:101456.
99. Notwell JH, Chung T, Heavner W, Bejerano G: A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. *Nat Commun* 2015, 6:6644.
100 Sanchez-Castillo M, Ruau D, Wilkinson AC, Ng FS, Hannah R, Diamanti E, Lombard P, Wilson NK, Gottgens B: CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res* 2015, 43:D1117-1123.
101 Kiyota T, Kato A, Kato Y: Ets-1 regulates radial glia formation during vertebrate embryogenesis. *Organogenesis* 2007, 3:93-101.
102 Gainous TB, Wagner E, Levine M: Diverse ETS transcription factors mediate FGF signaling in the Ciona anterior neural plate. *Dev Biol* 2015, 399:218-225.
103 Verheul TCJ, van Hijfte L, Perenthaler E, Barakat TS: The Why of YY1: Mechanisms of Tran-

scriptional Regulation by Yin Yang 1. *Front Cell Dev Biol* 2020, 8:592164.
104 Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ, Cohen MA, Nabet B, Buckley DL, et al: YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* 2017, 171:1573-1588 e1528.
105 Gabriele M, Vulto-van Silfhout AT, Germain PL, Vitriolo A, Kumar R, Douglas E, Haan E, Kosaki K, Takenouchi T, Rauch A, et al: YY1 Haploinsufficiency Causes an Intellectual Disability Syndrome Featuring Transcriptional and Chromatin Dysfunction. *Am J Hum Genet* 2017, 100:907-925.
106 Gregor A, Oti M, Kouwenhoven EN, Hoyer J, Sticht H, Ekici AB, Kjaergaard S, Rauch A, Stunnenberg HG, Uebe S, et al: De novo mutations in the genome organizer CTCF cause intellectual disability. *Am J Hum Genet* 2013, 93:124-131.
107 Nakada C, Satoh S, Tabata Y, Arai K, Watanabe S: Transcriptional repressor foxl1 regulates central nervous system development by suppressing shh expression in zebra fish. *Mol Cell Biol* 2006, 26:7246-7257.
108 Mullen RD, Park S, Rhodes SJ: A distal modular enhancer complex acts to control pituitary- and nervous system-specific expression of the LHX3 regulatory gene. *Mol Endocrinol* 2012, 26:308-319.
109 Savage JJ, Hunter CS, Clark-Sturm SL, Jacob TM, Pfaeffle RW, Rhodes SJ: Mutations in the LHX3 gene cause dysregulation of pituitary and neural target genes that reflect patient phenotypes. *Gene* 2007, 400:44-51.
110 Pristera A, Lin W, Kaufmann AK, Brimblecombe KR, Threlfell S, Dodson PD, Magill PJ, Fernandes C, Cragg SJ, Ang SL: Transcription factors FOXA1 and FOXA2 maintain dopaminergic neuronal properties and control feeding behavior in adult mice. *Proc Natl Acad Sci U S A* 2015, 112:E4929-4938.
111 Stott SR, Metzakopian E, Lin W, Kaestner KH, Hen R, Ang SL: Foxa1 and foxa2 are required for the maintenance of dopaminergic properties in ventral midbrain neurons at late embryonic stages. *J Neurosci* 2013, 33:8022-8034.
112 Hung CY, Hsu TI, Chuang JY, Su TP, Chang WC, Hung JJ: Sp1 in Astrocyte Is Important for Neurite Outgrowth and Synaptogenesis. *Mol Neurobiol* 2020, 57:261-277.
113 Manzanares M, Trainor PA, Nonchev S, Ariza-McNaughton L, Brodie J, Gould A, Marshall H, Morrison A, Kwan CT, Sham MH, et al: The role of kreisler in segmentation during hindbrain development. *Dev Biol* 1999, 211:220-237.
114 Blanchi B, Kelly LM, Viemari JC, Lafon I, Burnet H, Bevengut M, Tillmanns S, Daniel L, Graf T, Hilaire G, Sieweke MH: MafB deficiency causes defective respiratory rhythmogenesis and fatal central apnea at birth. *Nat Neurosci* 2003, 6:1091-1100.
115 Koshida R, Oishi H, Hamada M, Takei Y, Takahashi S: MafB is required for development of the hindbrain choroid plexus. *Biochem Biophys Res Commun* 2017, 483:288-293.
116 Pai EL, Vogt D, Clemente-Perez A, McKinsey GL, Cho FS, Hu JS, Wimer M, Paul A, Fazel Darbandi S, Pla R, et al: Mafb and c-Maf Have Prenatal Compensatory and Postnatal Antagonistic Roles in Cortical Interneuron Fate and Function. *Cell Rep* 2019, 26:1157-1173 e1155.
117 Maimaiti S, Koshida R, Ojima M, Kulathunga K, Oishi H, Takahashi S: Neuron-specific Mafb knockout causes growth retardation accompanied by an impaired growth hormone/insulin-like growth factor I axis. *Exp Anim* 2019, 68:435-442.
118 Wang H, Xiao Z, Zheng J, Wu J, Hu XL, Yang X, Shen Q: ZEB1 Represses Neural Differentiation and Cooperates with CTBP2 to Dynamically Regulate Cell Migration during Neocortex Development. *Cell Rep* 2019, 27:2335-2353 e2336.
119 Jiang Y, Yan L, Xia L, Lu X, Zhu W, Ding D, Du M, Zhang D, Wang H, Hu B: Zinc finger E-box-binding homeobox 1 (ZEB1) is required for neural differentiation of human embryonic stem cells. *J Biol Chem* 2018, 293:19317-19329.
120 Aslanpour S, Han S, Schuurmans C, Kurrasch DM: Neurog2 Acts as a Classical Proneural Gene in the Ventromedial Hypothalamus and Is Required for the Early Phase of Neurogenesis. *J Neurosci* 2020, 40:3549-3563.
121 Mulvaney J, Dabdoub A: Atoh1, an essential transcription factor in neurogenesis and intestinal and inner ear development: function, regulation, and context dependency. *J Assoc Res Otolaryngol* 2012, 13:281-293.

122 Pataskar A, Jung J, Smialowski P, Noack F, Calegari F, Straub T, Tiwari VK: NeuroD1 reprograms chromatin and transcription factor landscapes to induce the neuronal program. *EMBO J* 2016, 35:24-45.
123 Xin M, Yue T, Ma Z, Wu FF, Gow A, Lu QR: Myelinogenesis and axonal recognition by oligodendrocytes in brain are uncoupled in Olig1-null mice. *J Neurosci* 2005, 25:1354-1365.
124 Silbereis JC, Nobuta H, Tsai HH, Heine VM, McKinsey GL, Meijer DH, Howard MA, Petryniak MA, Potter GB, Alberta JA, et al: Olig1 function is required to repress dlx1/2 and interneuron production in Mammalian brain. *Neuron* 2014, 81:574-587.
125 Jakovcevski I, Zecevic N: Olig transcription factors are expressed in oligodendrocyte and neuronal cells in human fetal CNS. *J Neurosci* 2005, 25:10064-10073.
126 Chen T, Wu Q, Zhang Y, Lu T, Yue W, Zhang D: Tcf4 Controls Neuronal Migration of the Cereral Cortex through Regulation of Bmp7. *Front Mol Neurosci* 2016, 9:94.
127 Whalen S, Heron D, Gaillon T, Moldovan O, Rossi M, Devillard F, Giuliano F, Soares G, Mathieu-Dramard M, Afenjar A, et al: Novel comprehensive diagnostic strategy in Pitt-Hopkins syndrome: clinical score and further delineation of the TCF4 mutational spectrum. *Hum Mutat* 2012, 33:64-72.
128 Hegedus B, Dasgupta B, Shin JE, Emnett RJ, Hart-Mahon EK, Elghazi L, Bernal-Mizrachi E, Gutmann DH: Neurofibromatosis-1 regulates neuronal and glial cell differentiation from neuroglial progenitors in vivo by both cAMP- and Ras-dependent mechanisms. *Cell Stem Cell* 2007, 1:443-457.
129 Shaulian E, Karin M: AP-1 as a regulator of cell life and death. *Nat Cell Biol* 2002, 4:E131-136.
130 Werner H, LeRoith D: Insulin and insulin-like growth factor receptors in the brain: physiological and pathological aspects. *Eur Neuropsychopharmacol* 2014, 24:1947-1953.
131 Russ BE, Olshansky M, Li J, Nguyen MLT, Gearing LJ, Nguyen THO, Olson MR, McQuilton HA, Nussing S, Khoury G, et al: Regulation of H3K4me3 at Transcriptional Enhancers Characterizes Acquisition of Virus-Specific CD8(+) T Cell-Lineage-Specific Function. *Cell Rep* 2017, 21:3624-3636.
132 Paredes I, Himmels P, Ruiz de Almodovar C: Neurovascular Communication during CNS Development. *Dev Cell* 2018, 45:10-32.
133 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* 2006, 499:565-582.
134 Zhang X, He X, Li Q, Kong X, Ou Z, Zhang L, Gong Z, Long D, Li J, Zhang M, et al: PI3K/AKT/mTOR Signaling Mediates Valproic Acid-Induced Neuronal Differentiation of Neural Stem Cells through Epigenetic Modifications. *Stem Cell Reports* 2017, 8:1256-1269.
135 Sanchez-Alegria K, Flores-Leon M, Avila-Munoz E, Rodriguez-Corona N, Arias C: PI3K Signaling in Neurons: A Central Node for the Control of Multiple Functions. *Int J Mol Sci* 2018, 19.
136 Jayaraman D, Bae BI, Walsh CA: The Genetics of Primary Microcephaly. *Annu Rev Genomics Hum Genet* 2018, 19:177-200.
137 Oegema R, Barakat TS, Wilke M, Stouffs K, Amrom D, Aronica E, Bahi-Buisson N, Conti V, Fry AE, Geis T, et al: International consensus recommendations on the diagnostic work-up for malformations of cortical development. *Nat Rev Neurol* 2020, 16:618-635.
138 Wang X, Goldstein DB: Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease. *Am J Hum Genet* 2020, 106:215-233.
139 Kvon EZ, Waymack R, Elabd MG, Wunderlich Z: Enhancer redundancy in development and disease. *Nat Rev Genet* 2021.
140 Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al: The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020, 581:434-443.
141 Wilson MM, Henshall DC, Byrne SM, Brennan GP: CHD2-Related CNS Pathologies. *Int J Mol Sci* 2021, 22.
142 Thisse B, Thisse C: Fast Release Clones: A High Throughput Expression Analysis. . *ZFIN Direct Data Submission* 2004.
143 Shen T, Ji F, Yuan Z, Jiao J: CHD2 is Required for Embryonic Neurogenesis in the Developing Cerebral Cortex. *Stem Cells* 2015, 33:1794-1806.
144 Rymen D, Lindhout M, Spanou M, Ashrafzadeh F, Benkel I, Betzler C, Coubes C, Hartmann H,

Kaplan JD, Ballhausen D, et al: Expanding the clinical and genetic spectrum of CAD deficiency: an epileptic encephalopathy treatable with uridine supplementation. *Genet Med* 2020, 22:1589-1597.

145 Thisse B, Pflumio S, Fürthauer M, Loppin B, Heyer V, Degrave A, Woehl R, Lux A, Steffan T, Charbonnier XQ, Thisse C: Expression of the zebrafish genome during embryogenesis (NIH R01 RR15402). *ZFIN Direct Data Submission* 2001.

146 Sagie S, Lerman-Sagie T, Maljevic S, Yosovich K, Detert K, Chung SK, Rees MI, Lerche H, Lev D: Expanding the phenotype of TRAK1 mutations: hyperekplexia and refractory status epilepticus. *Brain* 2018, 141:e55.

147 Barel O, Malicdan MCV, Ben-Zeev B, Kandel J, Pri-Chen H, Stephen J, Castro IG, Metz J, Atawa O, Moshkovitz S, et al: Deleterious variants in TRAK1 disrupt mitochondrial movement and cause fatal encephalopathy. *Brain* 2017, 140:568-581.

148 Dobyns WB, Aldinger KA, Ishak GE, Mirzaa GM, Timms AE, Grout ME, Dremmen MHG, Schot R, Vandervore L, van Slegtenhorst MA, et al: MACF1 Mutations Encoding Highly Conserved Zinc-Binding Residues of the GAR Domain Cause Defects in Neuronal Migration and Axon Guidance. *Am J Hum Genet* 2018, 103:1009-1021.

149 Brock S, Vanderhasselt T, Vermaning S, Keymolen K, Regal L, Romaniello R, Wieczorek D, Storm TM, Schaeferhoff K, Hehr U, et al: Defining the phenotypical spectrum associated with variants in TUBB2A. *J Med Genet* 2021, 58:33-40.

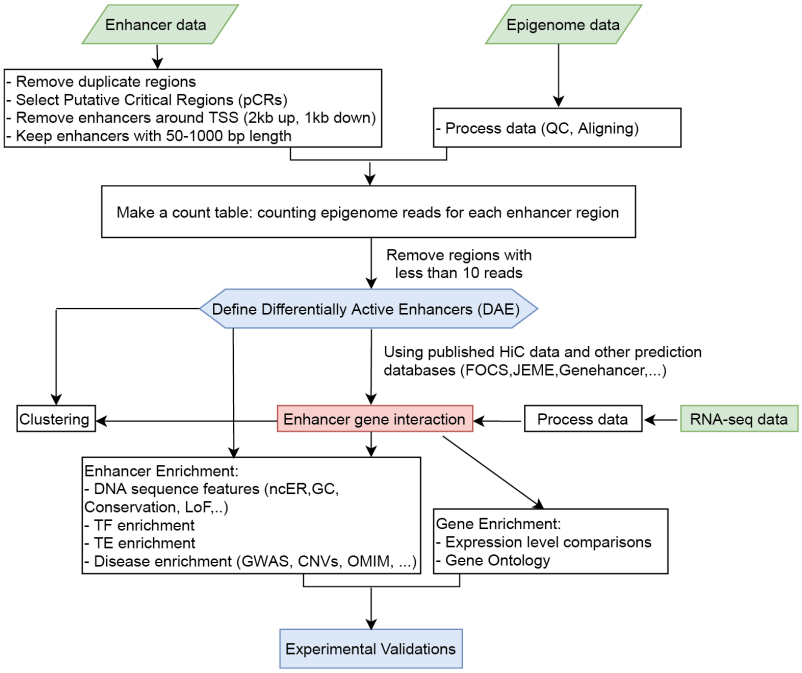
150 Philips AK, Pinelli M, de Bie CI, Mustonen A, Maatta T, Arts HH, Wu K, Roepman R, Moilanen JS, Raza S, et al: Identification of C12orf4 as a gene for autosomal recessive intellectual disability. *Clin Genet* 2017, 91:100-105.

151 Rom A, Melamed L, Gil N, Goldrich MJ, Kadir R, Golan M, Biton I, Perry RB, Ulitsky I: Regulation of CHD2 expression by the Chaserr long noncoding RNA gene is essential for viability. *Nat Commun* 2019, 10:5092.

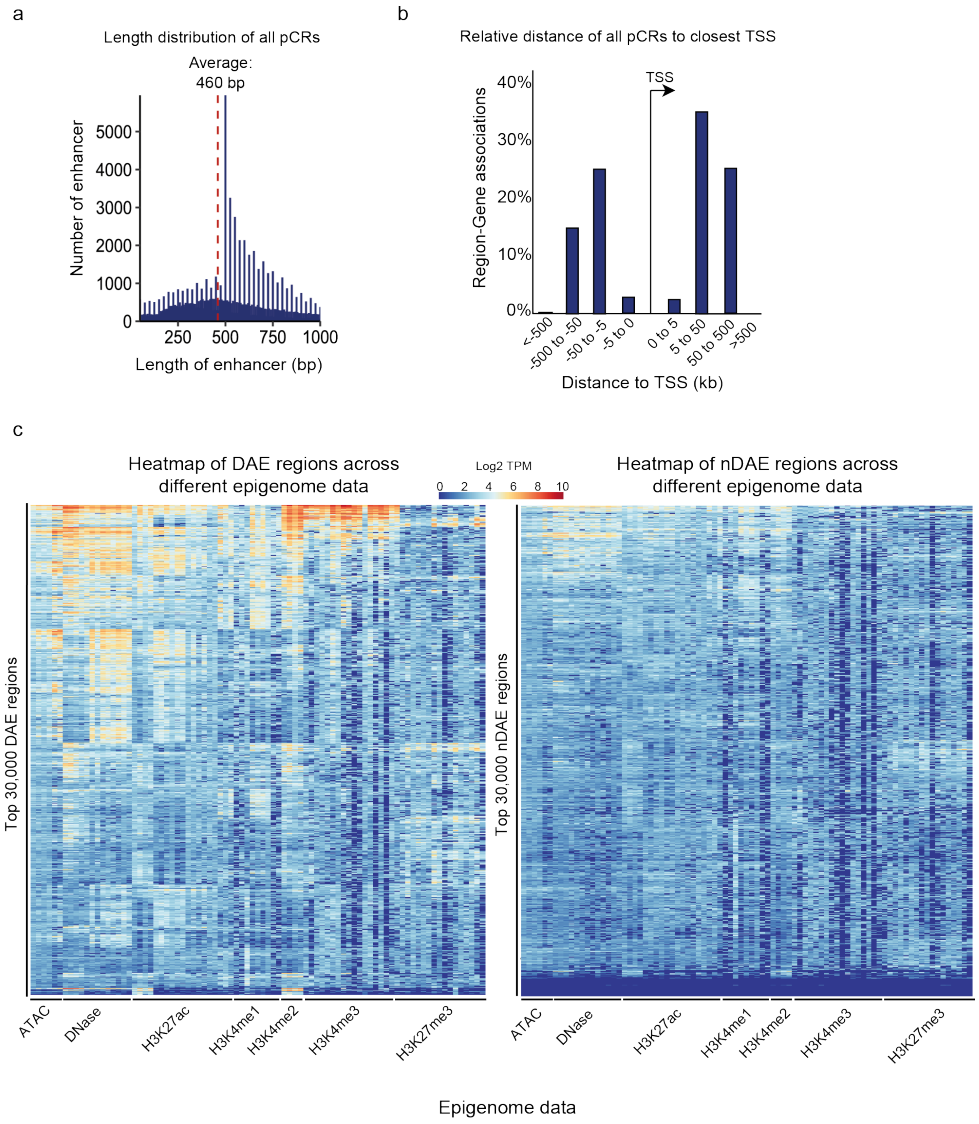
152 Yousefi S: Differentially-Active-Enhancers (GitHub). 2021.

153 Yousefi S: Differentially-Active-Enhancers (figshare). 2021.

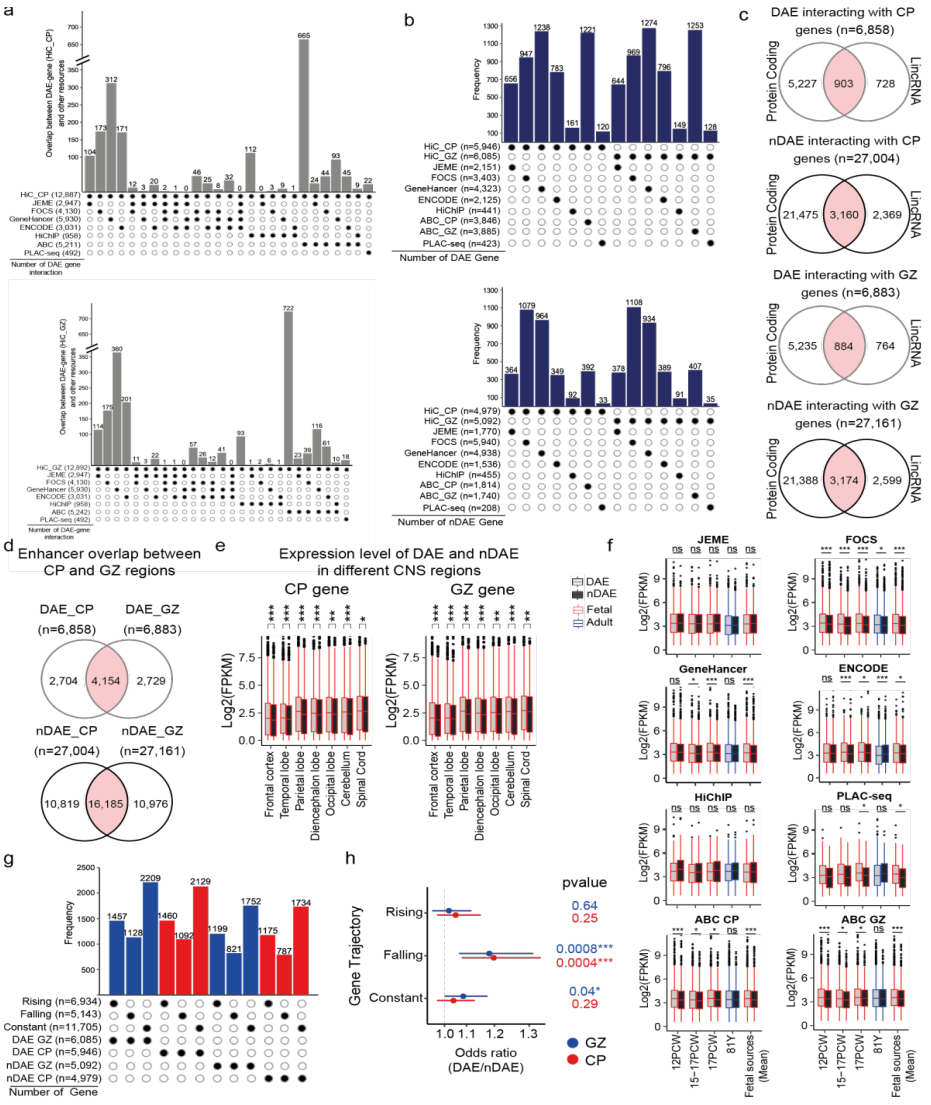
Supplementary Figures



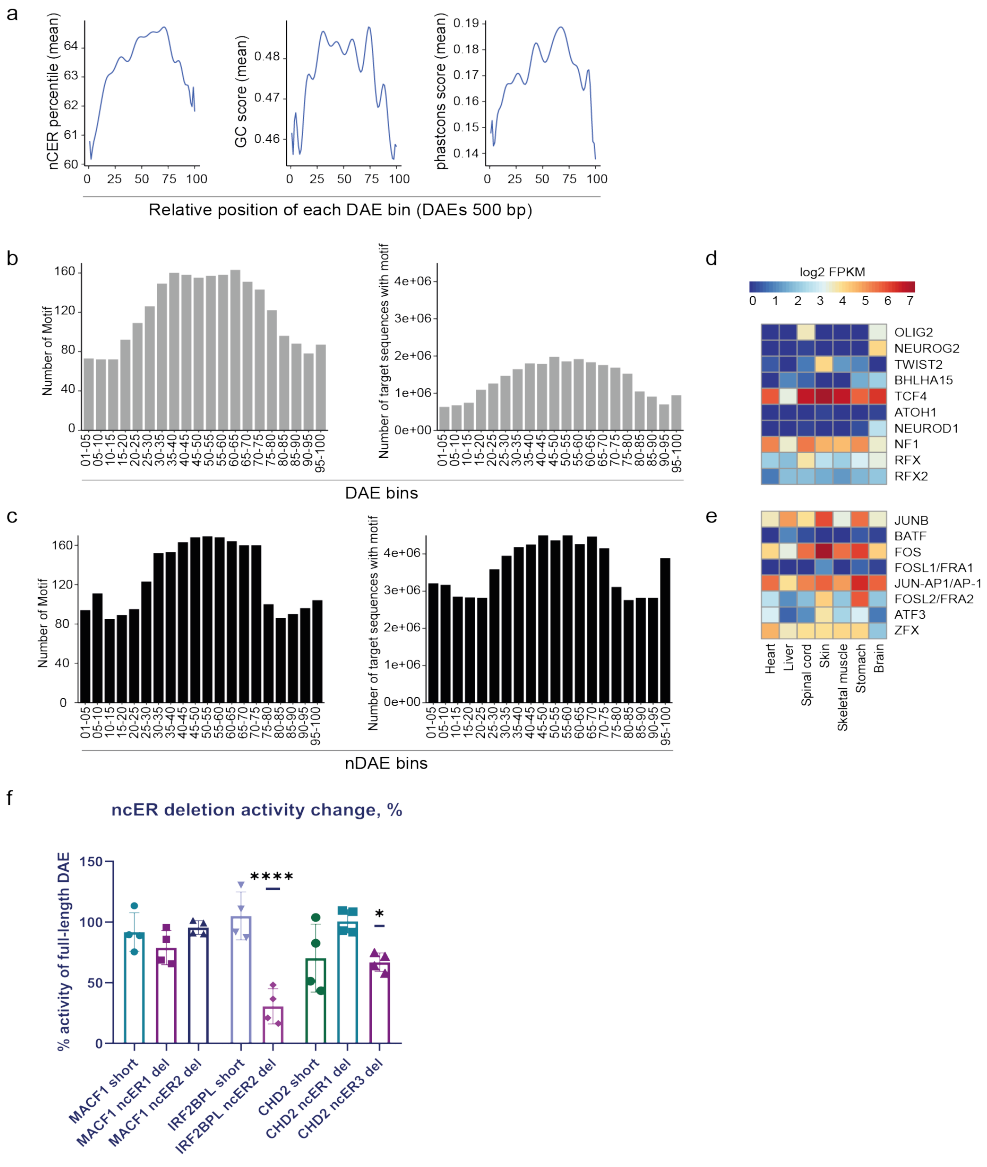
Supplementary Figure 1. Flow chart of integrative data analysis. Overview of the various analysis steps performed in this study. See text and methods for additional details.



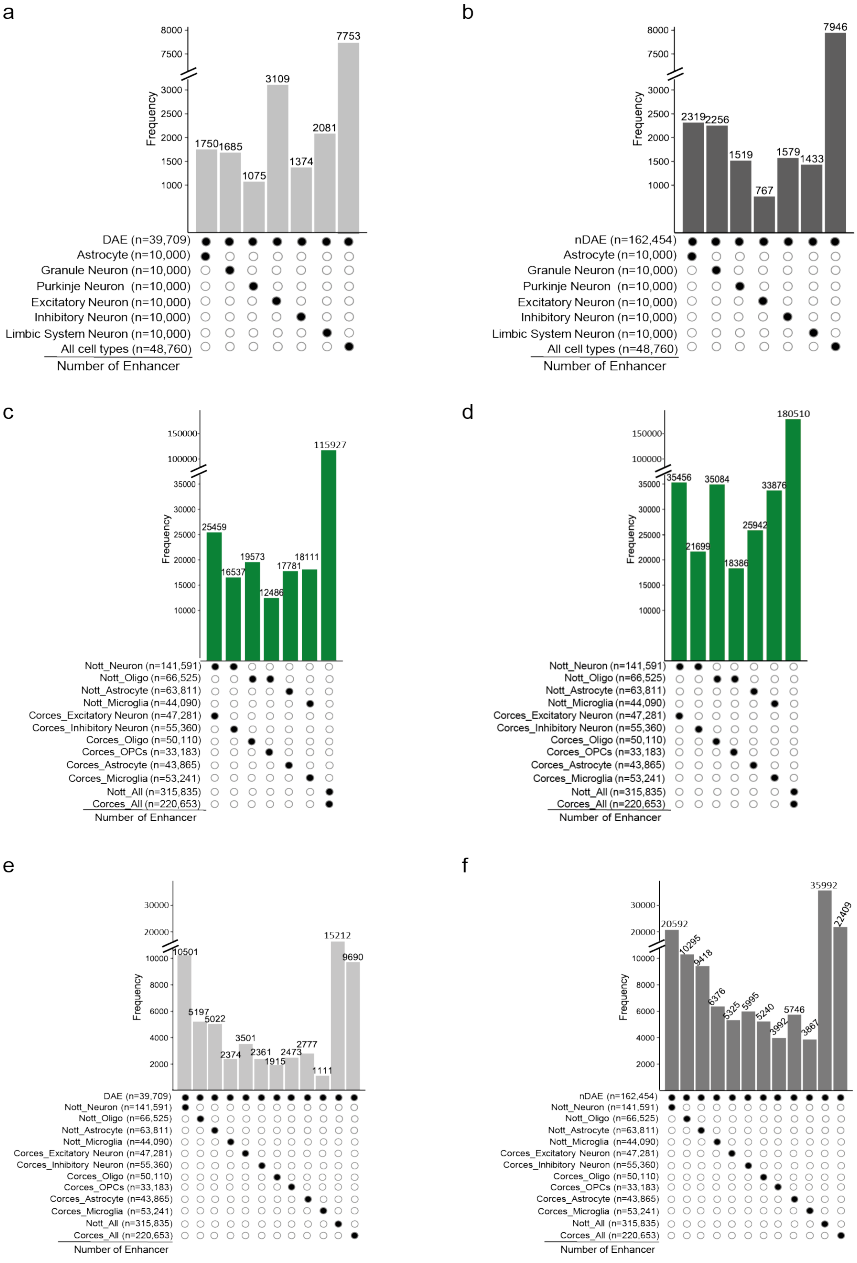
Supplementary Figure 2. Derivation of pCRs and DAEs. A) Density plot showing the size distribution of the 202,163 pCRs in bps. The red dashed line indicates the average length of all pCRs (460 bp). **B)** Relative distribution of all 202,163 pCRs in relation to their closest transcriptional start site. Graph generated using GREAT. **C)** Heatmaps showing variability across all epigenome features for the top 30,000 DAEs (left) and nDAEs (right). Columns represent in total 494 epigenome data sets used for the various types of histone marks and chromatin accessibility, as indicated.


a
b
c
DAE interacting with CP genes (n=6,858)
Protein Coding
LincRNA
5,227
903
728
nDAE interacting with CP genes (n=27,004)
21,475
3,160
2,369
DAE interacting with GZ genes (n=6,883)
5,235
884
764
nDAE interacting with GZ genes (n=27,161)
21,388
3,174
2,599
Number of DAE gene interaction
Number of DAE Gene
Number of nDAE Gene
d
Enhancer overlap between CP and GZ regions
DAE_CP (n=6,858)
DAE_GZ (n=6,883)
2,704
4,154
2,729
nDAE_CP (n=27,004)
nDAE_GZ (n=27,161)
10,819
16,185
10,976
e
Expression level of DAE and nDAE in different CNS regions
CP gene
GZ gene
Log2(FPKM)
Frontal cortex
Temporal lobe
Parietal lobe
Diencephalon lobe
Occipital lobe
Cerebellum
Spinal Cord
f
DAE
nDAE
Fetal
Adult
JEME
FOCS
GeneHancer
ENCODE
HiChIP
PLAC-seq
ABC CP
ABC GZ
12PCW
15~17PCW
17PCW
81Y
Fetal sources (Mean)
g
Frequency
Rising (n=6,934)
Falling (n=5,143)
Constant (n=11,705)
DAE GZ (n=6,085)
DAE CP (n=5,946)
nDAE GZ (n=5,092)
nDAE CP (n=4,979)
Number of Gene
h
Gene Trajectory
Rising
Falling
Constant
Odds ratio (DAE/nDAE)
pvalue
0.64
0.25
0.0008***
0.0004***
0.04*
0.29
GZ
CP

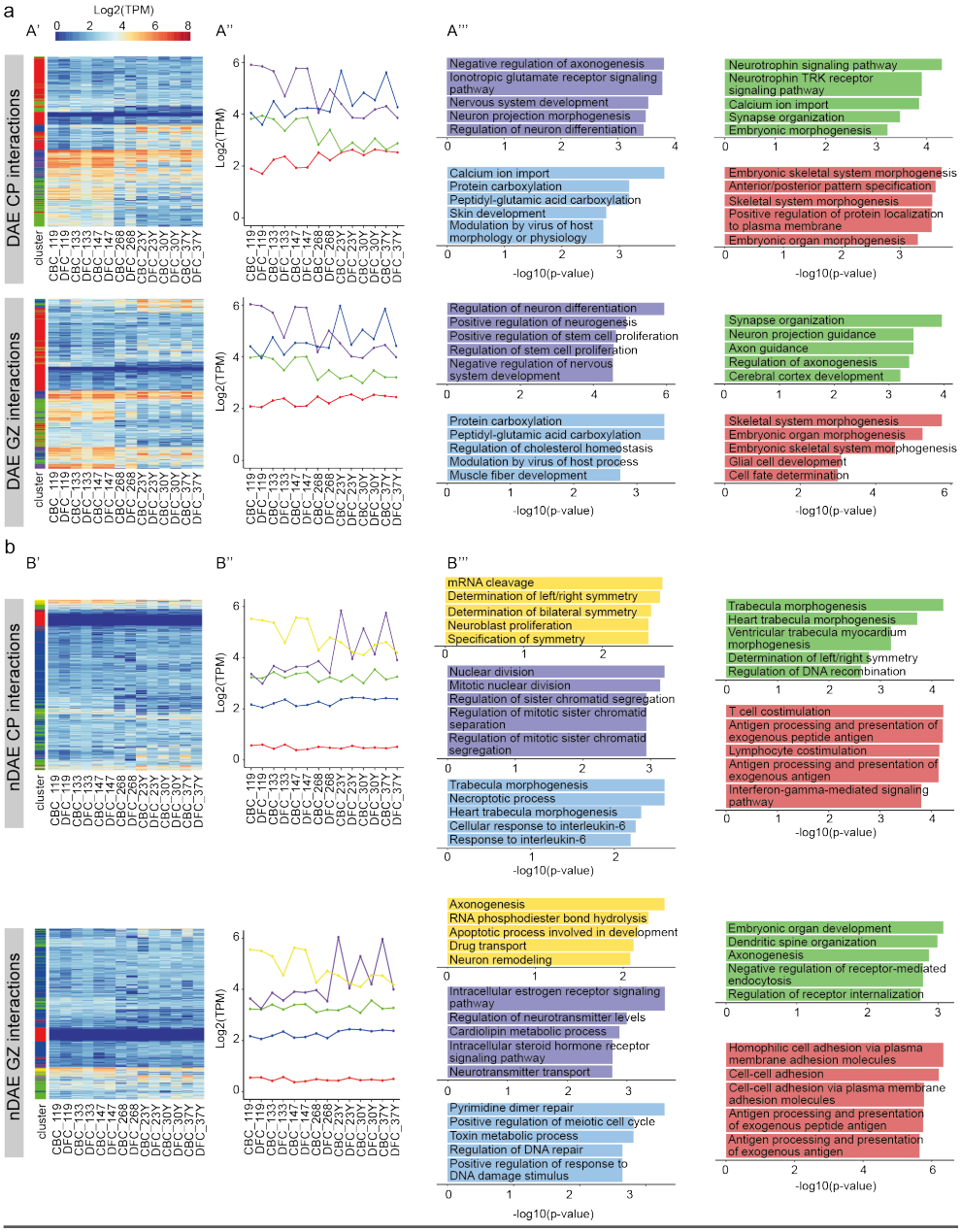
Supplementary Figure 3. Enhancer-Gene predictions and target gene expression. A) Bar chart showing the overlap between predicted enhancer-gene interactions from HiC of CP (upper panel) or GZ (lower panel) and the alternatively used enhancer-prediction methods JEME, ENCODE, FOCS,-GeneHancer, HiChIP, PLAC-seq and ABC model. **B)** Bar chart showing the number of target genes that overlap between the HiC enhancer-gene interaction predictions and the target gene predictions from the alternative methods, for DAEs (upper panel) and nDAEs (lower panel). **C)** Venn diagrams showing the interactions of DAEs (first and third panel) or nDAEs (second and fourth panel) with protein coding genes (left) and lincRNA (right) within the same TAD, for interactions from HiC in CP (first and second panel) or GZ (third and fourth panel). **D)** Venn diagrams showing the overlap between DAEs (upper panel) or nDAEs (lower panel) that interact with genes in CP (left) or GZ (right). **E)** Box plots showing the RNA-seq gene expression levels (in log2 FPKM) of genes linked to DAEs or nDAEs in CP (left) or GZ (right) for different brain regions. Boxes are interquartile range (IQR); line is median; and whiskers extend to 1.5 the IQR. * $p<0.05$; ** $p<0.01$; *** $p<0.001$; (wilcox.test). RNA-seq data obtained from ENCODE project. **F)** Box plots showing gene expression levels as determined by RNA-seq, for genes that interact with DAEs (light gray) or nDAEs (dark gray) as predicted by JEME, FOCS, GeneHancer, ENCODE, HiChIP, PLAC-seq, or the activity-by-contact (ABC) method, as indicated, for either CP or GZ, for fetal (red) or adult (blue) brain samples. Boxes are interquartile range (IQR); line is median; and whiskers extend to 1.5 the IQR. PCW, postconceptional week. FPKM, fragments per kilobase of transcript per million mapped reads. * $p<0.05$; *** $p<0.001$; ns, not significant (wilcox.test). Data obtained from: 12 PCW, Yan et al; 15-17 PCW, De la Torre-Ubieta et al; 17 PCW, Roadmap; 81 years, Roadmap; mean of fetal sources is the mean expression of the first three fetal samples. **G)** Bar plot showing the overlap between rising, falling and constantly expressed genes from BrainVar and DAE and nDAE target genes as predicted by HiC in CP or GZ. **H)** Line plot showing the odds ratio between DAE and nDAE linked genes in CP (red) or GZ (blue) (as determined by HiC), for rising, falling or constant genes from BrainVar. * $p<0.05$; *** $p<0.001$, Fisher's exact test.


a
nCER percentile (mean)
GC score (mean)
phastcons score (mean)
Relative position of each DAE bin (DAEs 500 bp)
b
Number of Motif
Number of target sequences with motif
DAE bins
c
Number of Motif
Number of target sequences with motif
nDAE bins
d
log2 FPKM
OLIG2
NEUROG2
TWIST2
BHLHA15
TCF4
ATOH1
NEUROD1
NF1
RFX
RFX2
e
JUNB
BATF
FOS
FOSL1/FRA1
JUN-AP1/AP-1
FOSL2/FRA2
ATF3
ZFX
Heart
Liver
Spinal cord
Skin
Skeletal muscle
Stomach
Brain
f
ncER deletion activity change, %
% activity of full-length DAE
****
*
MACF1 short
MACF1 ncER1 del
MACF1 ncER2 del
IRF2BPL short
IRF2BPL ncER2 del
CHD2 short
CHD2 ncER1 del
CHD2 ncER3 del

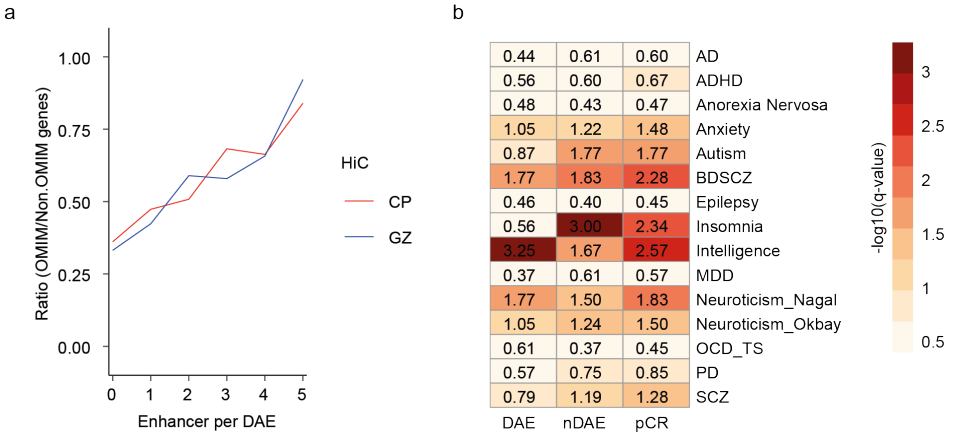
Supplementary Figure 4. **Features and motifs in DAEs and nDAEs. A)** Line plot showing the distribution of the mean ncER percentile (left) , GC content score (middle) and phastcons score (right) over all DAEs that have a size of 500 bp (n= 768). **B)** Bar chart showing the number of significant motifs from HOMER analysis (left) or the total number of target sequences for these motifs, across the 20 relative bin groups for DAEs. **C)** As **B**, but now for nDAEs. **D)** Heatmap showing the RNA-seq expression levels (Log2 FPKM) of the most enriched TFs at the center of DAEs from the HOMER analysis presented in Figure 3G, across various human fetal tissues. RNA-seq data obtained from ENCODE project. **E)** As D), but now for the most enriched TFs at the center of nDAEs from the HOMER analysis reported in Figure 3H. **F)** Effect of ncER deletion on activity of DAEs linked to *IRF2BPL*, *CHD2* and *MACF1*. Percentage of activity of modified DAEs (see methods) compared to the full-length DAE in STARR-seq enhancer reporter experiments is plotted. Two independent transfection experiments were performed, each in duplicate. All data points and standard deviation are shown. * $p<0.05$; **** $p<0.0001$ (one-way ANOVA test followed by multiple comparison test (Fisher's LSD test).


a
Frequency
1750
1685
1075
3109
1374
2081
7753
DAE (n=39,709)
Astrocyte (n=10,000)
Granule Neuron (n=10,000)
Purkinje Neuron (n=10,000)
Excitatory Neuron (n=10,000)
Inhibitory Neuron (n=10,000)
Limbic System Neuron (n=10,000)
All cell types (n=48,760)
Number of Enhancer
b
Frequency
2319
2256
1519
767
1579
1433
7946
nDAE (n=162,454)
Astrocyte (n=10,000)
Granule Neuron (n=10,000)
Purkinje Neuron (n=10,000)
Excitatory Neuron (n=10,000)
Inhibitory Neuron (n=10,000)
Limbic System Neuron (n=10,000)
All cell types (n=48,760)
Number of Enhancer
c
Frequency
25459
16537
19573
12486
17781
18111
115927
Nott_Neuron (n=141,591)
Nott_Oligo (n=66,525)
Nott_Astrocyte (n=63,811)
Nott_Microglia (n=44,090)
Corces_Excitatory Neuron (n=47,281)
Corces_Inhibitory Neuron (n=55,360)
Corces_Oligo (n=50,110)
Corces_OPCs (n=33,183)
Corces_Astrocyte (n=43,865)
Corces_Microglia (n=53,241)
Nott_All (n=315,835)
Corces_All (n=220,653)
Number of Enhancer
d
Frequency
35456
21699
35084
18386
25942
33876
180510
Nott_Neuron (n=141,591)
Nott_Oligo (n=66,525)
Nott_Astrocyte (n=63,811)
Nott_Microglia (n=44,090)
Corces_Excitatory Neuron (n=47,281)
Corces_Inhibitory Neuron (n=55,360)
Corces_Oligo (n=50,110)
Corces_OPCs (n=33,183)
Corces_Astrocyte (n=43,865)
Corces_Microglia (n=53,241)
Nott_All (n=315,835)
Corces_All (n=220,653)
Number of Enhancer
e
Frequency
10501
5197
5022
2374
3501
2361
1915
2473
2777
1111
15212
9690
DAE (n=39,709)
Nott_Neuron (n=141,591)
Nott_Oligo (n=66,525)
Nott_Astrocyte (n=63,811)
Nott_Microglia (n=44,090)
Corces_Excitatory Neuron (n=47,281)
Corces_Inhibitory Neuron (n=55,360)
Corces_Oligo (n=50,110)
Corces_OPCs (n=33,183)
Corces_Astrocyte (n=43,865)
Corces_Microglia (n=53,241)
Nott_All (n=315,835)
Corces_All (n=220,653)
Number of Enhancer
f
Frequency
20592
10295
9418
6576
5325
5995
5240
5692
5746
3867
35992
22409
nDAE (n=162,454)
Nott_Neuron (n=141,591)
Nott_Oligo (n=66,525)
Nott_Astrocyte (n=63,811)
Nott_Microglia (n=44,090)
Corces_Excitatory Neuron (n=47,281)
Corces_Inhibitory Neuron (n=55,360)
Corces_Oligo (n=50,110)
Corces_OPCs (n=33,183)
Corces_Astrocyte (n=43,865)
Corces_Microglia (n=53,241)
Nott_All (n=315,835)
Corces_All (n=220,653)
Number of Enhancer

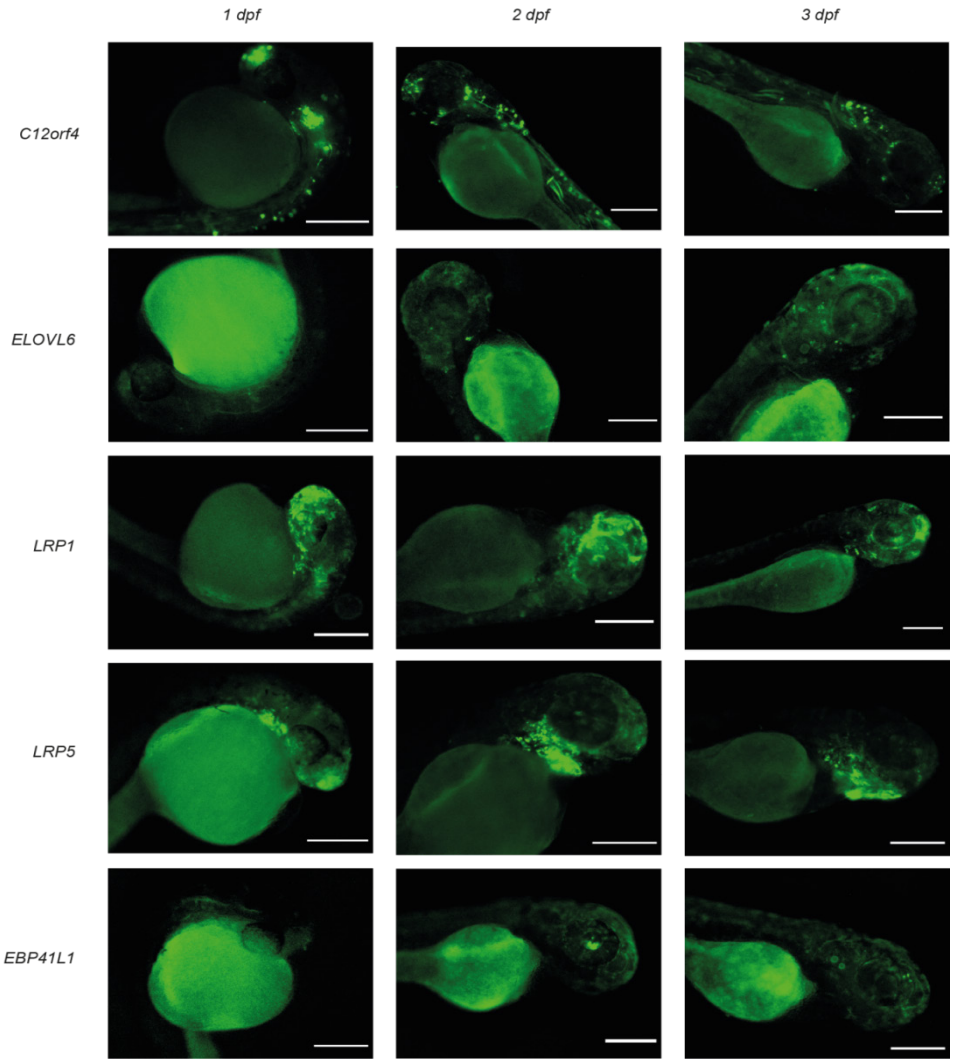
Supplementary Figure 5. Cell type specificity of DAEs and nDAEs. A) Bar chart showing the overlap between DAEs and cell type-specific chromatin accessibility peaks derived from Domcke et al, generated by scATAC-seq on fetal brain. **B)** As **A**, but not nDAEs. **C)** Bar chart showing the overlap between cell type specific putative enhancers from postnatal brain from Nott et al and Corces et al, using the putative enhancers from Nott et al as reference for the intersection. **D)** As **C**, but now using the putative enhancers from Corces et al as reference for the intersection. **E)** Bar chart showing the overlap between DAEs and the postnatal, cell type specific putative enhancers from Nott et al and Corces et al. **F)** As **E**, but now for nDAEs.



Supplementary Figure 6. Dynamics of DAEs and nDAEs in comparison to adult brain. A) Heatmap displaying H3K27ac for pre- and postnatal samples from Li et al, across all DAEs interacting with protein coding genes in CP (upper heatmap) and GZ (lower heatmap) (A^{I}). K means clustering analysis of H3K27ac enrichment (A^{II}) identifies four clusters, depicted in purple, blue, green and red. Level of enrichment is indicated on the y-axis in Log2 TPM. Gene enrichment analysis for the corresponding genes in each cluster (A^{III}). X-axis shows the - log 10 (p-value) from Enrichr. **B)** As **A**, but then for nDAEs. K means clustering identifies 5 different clusters for nDAEs, depicted in yellow, purple, green, blue and red.



Supplementary Figure 7. DAEs and nDAEs in human disease, related to Figure 5. A) Line graph showing the fraction between OMIM divided by nonOMIM genes as a function of the number of enhancers that a DAE is interacting with, for interactions in CP (red) and GZ (blue). The more enhancers a DAE is interacting with, the more likely it is that the target gene of that DAE is a OMIM gene. **B)** Heat map showing the –log10 p-value obtained from LD score regression analysis using relevant publicly available GWAS data for several brain related disorders (see Supplementary Table 11), for DAEs, nDAEs and pCRs. AD, Alzheimer's disease; ADHD, attention-deficit hyperactivity disorder; BDSCZ, bipolar disorder and schizophrenia; MDD, major depressive disorder; OCD_TS, obsessive compulsive disorder / Tourette syndrome; PD, Parkinson's disease; SCZ, schizophrenia.



Supplementary Figure 8. Zebrafish enhancer reporter assay. Panel of additional fluorescent images for validated enhancers, showing GFP expression in zebrafish at 1, 2 and 3 dpf. Scale bars represent 500 µm.

Supplementary Tables

https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-021-00980-1

Supplementary Table 1: List of all putative enhancers collected (also available at

https://figshare.com/projects/Differentially-Active-Enhancers/122965)

Supplementary Table 2: Overview of all epigenome data processed in this study

Supplementary Table 3: List of all pCRs, with DAEs and nDAEs indicated

Supplementary Table 4: Functional enrichment analysis using GREAT for DAEs and nDAEs

Supplementary Table 5: Enhancer-gene predictions

Supplementary Table 6: Overview of all RNA-seq data sets used in this study

Supplementary Table 7: Functional enrichment analysis using GREAT, Enrichr and Metascape for multigene interacting DAEs and nDAEs

Supplementary Table 8: TF and TE enrichment at DAEs and nDAEs

Supplementary Table 9: DAE clusters and associated functional enrichment using Enrichr

Supplementary Table 10: DAEs linked to OMIM genes

Supplementary Table 11: Significant GWAS loci used in this study

Supplementary Table 12: Calculations and p-values for gene variant associations

Supplementary Table 13: Zebrafish quantifications

Supplementary Table 14: Oligonucleotides used in this study

Appendix

- Summary
- Samenvatting
- Curriculum Vita
- List of Publications
- PhD Portfolio

Summary

As sequencing costs decrease, whole genome sequencing (WGS) will be more often routinely employed for diagnostics of patients with presumed genetic diseases in the field of human genetics. These studies provide the opportunity to investigate regions of our genome that currently used routine methods, such as whole exome sequencing (WES) fail to assess, including non-coding regions. Variants in these regions beyond the exome, are excellent candidates in which genetic variants might contribute to the phenomena of missing heritability. Namely, independent of the precise indication for genetic testing, in general a molecular diagnosis is not achieved in around 50% of individuals suspected of a genetic disorder, including neurodevelopmental disorders, and it seems likely that at least part of this missing heritability is caused by alterations of the non-coding genome. Particular interesting parts of that non-coding genome are enhancers, which are non-coding elements that ensure correct spatio-temporal expression of their target genes. An increasing number of studies have linked alterations of such enhancers to human disease, but still, their wide-spread investigation in a clinical setting is hampered. One of the key reasons is that it is still challenging to predict the location and the activity of functional enhancers genome-wide. Given these hurdles, it is even more challenging to interpret the effect of genetic variants in such non-coding regulatory elements. It is thus of crucial relevance to better functionally annotate non-coding regulatory elements as this will greatly facilitate the interpretation of genetic variants identified outside of the exome in WGS studies.

In this thesis, I aimed to solve part of the missing heritability in neurodevelopmental disorders, using computational approaches. Next to the investigations of a novel epilepsy syndrome and investigations aiming to elucidate the regulation of the gene involved, I investigated and prioritized genomic sequences that have implications in gene regulation during the developmental stages of human brain, with the goal to create an atlas of high confidence non-coding regulatory elements that future studies can assess for genetic variants in genetically unexplained individuals suffering from neurodevelopmental disorders that are of suspected genetic origin.

In **chapter 1**, we provided an overview of the current knowledge of the role of the non-coding genome in gene regulation and diseases, with a particular focus on enhancers. We discussed the main techniques currently applied to identify putative and functional enhancers, their target genes, and computational approaches that facilitate future investigations on non-coding causes of genetic diseases.

In **chapter 2A**, we applied WES to identify the cause of a severe epileptic encephalopathy in a child visiting the outpatient clinic of the Clinical Genetics department, identifying a homozygous variant in the essential gene *UGP2*. Subsequently, through international collaboration, we identified in 21 additional individuals presenting with the same phenotype exactly the same variant, establishing *UGP2* as a new cause of developmental epileptic encephalopathy. The *UGP2* gene encodes two different protein isoforms, which only differ by 11 amino acids at the N-terminal, and which do not display any functional differences. The rare variant in the affected individuals results in a tolerable Met12Val missense change of the longer *UGP2* isoform but disrupts the start codon of the shorter isoform, which is predominantly expressed in the brain. Affected individuals therefore become functionally depleted of UGP2 in brain, but still have expression of the functional long isoform in other tissues. Absence of UGP2 leads to alterations in glycogen metabolism, protein glycosylation and increased ER-stress, leading to neuronal dysfunction. This is the first disease caused by the specific absence of a tissue-specific isoform of an essential gene, and our computational analysis shows that a similar mechanism might apply as well to other essential genes with a similar structure of the gene locus.

In **chapter 2B**, we investigated the mechanisms underlying the *UGP2* gene regulation and the switch of isoform expression amongst various cell types. To identify potential *UGP2* regulatory elements and the 3D interactions with the *UGP2* promoter, we combined ChIP-seq with T2C maps, allowing a high-resolution investigation of all interactions around the *UGP2* promoter. A potential *UGP2* regulatory element (pDE4 enhancer) is located adjacent to the *OTX1* gene and shows interactions with the *UGP2* promoter. Also, partially knocking out the pDE4 enhancer sequence provided evidence for potential *UGP2* regulation by this enhancer. Computational motif analysis at pDE4 showed two binding sites of ZNF281, a zinc finger transcription factor which is expressed at a higher level in human embryonic stem cells (ESCs) compared to neural stem cells (NSCs). This ongoing work provides the first glimpse in the mechanisms underlying gene regulation at the *UGP2* locus.

In **chapter 3**, we introduced a pipeline to computationally define likely functional enhancers during different stages of fetal brain developmental. Data integration allowed us to generate a comprehensive list of functional enhancers, integrating virtually all available epigenome data and previously proposed putative enhancers for brain development. We first collected more than 1.6 million putative enhancers from literature and various data bases, and assessed their individual overlap which we hypothesized might identify the real biological relevant sequences, leading to the

identification of around 200 thousand putative critical regions. The epigenome state of these putative critical regions was assessed across ~500 distinct epigenome data during the early stages of human brain development, likely reflecting the activity of those sequences during development. This multi-omics integration analysis defined 39,709 differentially active enhancers (DAEs) with dynamic epigenomic rearrangement during fetal brain development. Many of these DAEs are linked to clinically relevant genes, and functional validation of selected DAEs in cell models and zebrafish confirmed their role in gene regulation. DAEs were subjected to higher sequence constraints in humans, different sequence characteristics, and a distinct transcription factor binding landscapes. Also, DAEs are enriched for GWAS loci for brain-related traits and genetic variation found in individuals with autism spectrum disorder. Together, this work provides a large atlas of regulatory elements that play a role during human brain development, many of which regulate disease relevant genes.

Chapter 4 provides a catalog of active enhancers in NSCs identified using the massively parallel reporter assay ChIP-STARR-seq, a technology previously developed in our laboratory which allows the genome-wide identification of functional enhancers in a genome-wide manner. Using this approach, we could assess the activity of more than 148 thousands genomic regions and identify around 14 thousands highly active enhancers. Using computational studies, we showed correlations between enhancer activity and various characteristics, such as gene expression levels of their target genes, sequence constraint, conservation, and enrichment for transcription factor (TF) motifs and transposable elements (TEs), providing insights into the mechanisms underlying gene regulation in NSCs. Furthermore, testing the same genomic regions in ESCs allowed the assessment of differential enhancer activity in the two cell types. We determined a subset of enhancers with higher activity in ESCs that were surprisingly linked to neural genes with evidence of epigenetic silencing at the endogenous chromatin context in ESCs, suggesting they might be silenced in ESCs but primed for activation at later stages of (neural) development. Interestingly, a small subset of these enhancers are enriched for binding sites of ZNF281, which might indicate that this zinc finger protein also plays a role in priming these enhancers. Together, this study provides a catalogue of functionally validated enhancers in NSCs, and provides novel insights in gene regulatory mechanisms in NSCs.

In **chapter 5**, we provided a graphical user interface to explore the enhancer-related information obtained from previous chapters in a user-friendly manner. This visualization application allows users to explore enhancer activity, enhancer-gene inter-

actions and enhancer-disease associations, and other enhancer-related information during development. Launching of the application will facilitate widespread data access to users that do not have to be trained in bioinformatics and will help facilitating the use of enhancer analysis in a clinical setting.

Finally, **chapter 6** provides a general discussion of the findings of this thesis, highlighting notable results and discusses them in the context of recent literature, with an outlook to future clinical implementation of these findings.

Samenvatting

Naarmate de sequencingkosten dalen, zal whole genome sequencing (WGS) vaker routinematig worden toegepast voor de diagnostiek van patiënten met vermoedelijk genetische ziekten in het veld van de klinische genetica. Deze studies bieden de mogelijkheid om regio's van ons genoom te onderzoeken die momenteel met de huidige technologieën, zoals whole exome sequencing (WES), niet worden geëvalueerd, zoals niet-coderende regio's buiten de eiwitcoderende genen. Varianten in deze regio's buiten het exoom, zijn uitstekende kandidaten waarin genetische varianten zouden kunnen bijdragen tot het fenomeen van ”missing heritability”. Onafhankelijk van de precieze indicatie voor genetische tests wordt namelijk in het algemeen bij ongeveer 50% van de van een genetische aandoening verdachte personen, met inbegrip van neurologische ontwikkelingsstoornissen, geen moleculaire diagnose gesteld, en het lijkt waarschijnlijk dat ten minste een deel van deze missing heritability wordt veroorzaakt door veranderingen in het niet-coderende genoom. Bijzonder interessante delen van dat niet-coderende genoom zijn enhancers, dat zijn niet-coderende elementen die er voor zorgen dat de genen die zij reguleren op het juiste moment aan en uit gaan, ook wel regulatie van correcte spatio-temporele genexpressie genoemd. Een toenemend aantal studies heeft afwijkingen van dergelijke enhancers in verband gebracht met ziekten bij de mens, maar nog steeds wordt het onderzoek hiervan op grote schaal in een klinische setting belemmerd. Een van de belangrijkste redenen hiervoor is dat het nog steeds een uitdaging is om de locatie en de activiteit van functionele enhancers genoombreed te voorspellen. Gezien deze hindernissen is het zelfs nog moeilijker om het effect van genetische varianten in dergelijke niet-coderende regulerende elementen te interpreteren. Het is dus van cruciaal belang om niet-coderende regulatorische elementen beter functioneel te annoteren, aangezien dit de interpretatie van genetische varianten die buiten het exoom in WGS studies worden geïdentificeerd, aanzienlijk zal vergemakkelijken.

In dit proefschrift heb ik getracht een deel van de ontbrekende erfelijkheid in neurologische ontwikkelingsstoornissen op te lossen, gebruik makend van bioinformatische benaderingen. Naast het onderzoek naar een nieuw epilepsie syndroom en onderzoek gericht op het ophelderen van de regulatie van het betrokken gen, heb ik genomische sequenties onderzocht en geprioriteerd die implicaties hebben in genregulatie tijdens de ontwikkelingsstadia van het menselijk brein, met als doel een atlas te creëren van niet-coderende regulatoire elementen met een hoge betrouw-baarheid, die toekomstige studies kunnen gebruiken om genetische varianten op pathogeniciteit te beoordelen in genetisch onverklaarde individuen die lijden aan vermoedelijk

genetische neurologische ontwikkelingsstoornissen.

In **hoofdstuk 1** gaven we een overzicht van de huidige kennis van de rol van het niet-coderende genoom in genregulatie en ziekten, met een bijzondere nadruk op enhancers. We bespraken de belangrijkste technieken die momenteel worden toegepast om mogelijke en functionele enhancers te identificeren, hun doelgenen in kaart te brengen, en bioinformatische benaderingen die toekomstig onderzoek naar niet-coderende oorzaken van genetische ziekten vergemakkelijken.

In **hoofdstuk 2A** hebben we WES toegepast om de oorzaak van een ernstige epileptische encefalopathie vast te stellen bij een kind dat de polikliniek van de afdeling Klinische Genetica bezocht, waarbij we een homozygote variant in het essentiële gen *UGP2* hebben geïdentificeerd. Door internationale samenwerking hebben we vervolgens bij 21 andere personen met hetzelfde fenotype precies dezelfde variant geïdentificeerd, waardoor *UGP2* als een nieuwe oorzaak van epileptische encefalopathie (developmental epileptic encephalopathy) werd vastgesteld. Het *UGP2* gen codeert voor twee verschillende eiwit isovormen, die slechts 11 aminozuren verschillen aan de N-terminal, en die geen functionele verschillen vertonen. De zeldzame variant in de getroffen individuen resulteert in een tolereerbare Met12Val missense verandering van de langere *UGP2* isovorm, maar verstoort het startcodon van de kortere isovorm, die overwegend in de hersenen tot expressie komt. Getroffen individuen hebben hierdoor geen functioneel UGP2 in de hersenen, maar hebben nog expressie van de functionele lange isovorm in andere weefsels. Afwezigheid van UGP2 leidt tot veranderingen in het glycogeenmetabolisme, eiwitglycosylering en verhoogde ER-stress, wat leidt tot neuronale disfunctie. Dit is de eerste ziekte die veroorzaakt wordt door de specifieke afwezigheid van een weefselspecifieke isovorm van een essentieel gen, en onze rekenkundige analyse toont aan dat een zelfde mechanisme ook zou kunnen gelden voor andere essentiële genen met een gelijkaardige structuur van het gen locus.

In **hoofdstuk 2B** hebben we de mechanismen onderzocht die ten grondslag liggen aan de *UGP2* genregulatie en de omschakeling van isovorm expressie tussen verschillende celtypen. Om potentiële *UGP2* regulerende elementen en de 3D interacties met de *UGP2* promoter te identificeren, combineerden we ChIP-seq met T2C kaarten, waardoor een hoge-resolutie onderzoek van alle interacties rond de *UGP2* promoter mogelijk werd. Een potentieel *UGP2* regulerend element (pDE4 enhancer) bevindt zich naast het *OTX1* gen en vertoont interacties met de *UGP2* promoter. Ook het gedeeltelijk uitschakelen van de pDE4 enhancer sequentie leverde bewijs voor

potentiële *UGP2* regulatie. Computermotiefanalyse op pDE4 toonde twee bindingsplaatsen van ZNF281, een zinkvingertranscriptiefactor die op een hoger niveau tot expressie komt in menselijke embryonale stamcellen (ESCs) in vergelijking met neurale stamcellen (NSCs). Dit werk levert een eerste glimp op van de mechanismen die ten grondslag liggen aan genregulatie op de *UGP2* locus.

In **hoofdstuk 3** introduceerden we een analyse pipeline voor het bioinformatisch definiëren van waarschijnlijke functionele enhancers tijdens verschillende stadia van de foetale hersenontwikkeling van de mens. Data-integratie stelde ons in staat om een uitgebreide lijst van functionele enhancers te genereren, waarin vrijwel alle beschikbare epigenoom data en eerder voorgestelde mogelijke enhancers voor hersenontwikkeling zijn geïntegreerd. Eerst verzamelden we meer dan 1.6 miljoen putatieve enhancers uit de literatuur en verschillende databases, en beoordeelden hun individuele overlap, waarvan we veronderstelden dat dit de echte biologisch relevante sequenties zou kunnen identificeren, wat leidde tot de identificatie van ongeveer 200 duizend vermeende kritieke gebieden. De epigenoom toestand van deze mogelijke kritieke regio's werd beoordeeld in ~500 verschillende epigenoom data tijdens de vroege stadia van de menselijke hersenontwikkeling, wat waarschijnlijk de activiteit van deze sequenties tijdens de ontwikkeling weerspiegelt. Deze multi-omics integratie analyse identificeerde 39,709 differentieel actieve enhancers (DAEs) met dynamische epigenomische herschikking tijdens de foetale hersenontwikkeling. Veel van deze DAEs zijn gelinkt aan klinisch relevante genen, en functionele validatie van geselecteerde DAEs in celmodellen en zebravissen bevestigden hun rol in genregulatie. DAEs zijn onderhevig aan hogere sequence constraint bij de mens, hebben verschillende sequentie karakteristieken, en een verschillend transcriptie factor bindingslandschap. Bovendien zijn DAEs verrijkt voor GWAS loci voor hersen-gerelateerde eigenschappen en genetische variatie gevonden bij personen met autisme spectrum stoornis. Samen levert dit werk een grote atlas op van regulatoire elementen die een rol spelen tijdens de ontwikkeling van het menselijk brein, en waarvan vele genen reguleren die relevant zijn voor ziekten.

Hoofdstuk 4 geeft een catalogus van actieve enhancers in NSCs die geïdentificeerd zijn met behulp van de massaal parallelle reporter assay ChIP-STARR-seq, een technologie die eerder in ons laboratorium is ontwikkeld en die het mogelijk maakt om functionele enhancers op genoom-brede wijze te identificeren. Met deze aanpak konden we de activiteit van meer dan 148 duizend genomische regio's bepalen en ongeveer 14 duizend zeer actieve enhancers identificeren. Met behulp van bioinformatische studies toonden we correlaties aan tussen enhancer activiteit en verschillende

karakteristieken, zoals genexpressieniveau van de genen die door deze enhancers gereguleerd worden, sequence constraint en conservatie, en verrijking voor transcriptiefactor (TF) motieven en transposable elementen (TEs), wat samen inzicht verschaft in de mechanismen die ten grondslag liggen aan genregulatie in NSCs. Bovendien maakte het testen van dezelfde genomische regio's in ESCs de beoordeling mogelijk van differentiële enhancer activiteit in de twee celtypes. We vonden een subset van enhancers met hogere activiteit in ESCs die verrassend gelinkt waren aan neurale genen met bewijs van epigenetische silencing in de endogene chromatine context in ESCs. Dit suggereert dat ze mogelijk in ESCs gesilenced zijn maar klaar staan voor activatie in latere stadia van (neurale) ontwikkeling. Interessant is dat een klein deel van deze enhancers verrijkt is met bindingsplaatsen van ZNF281, wat erop zou kunnen wijzen dat dit zinkvingereiwit ook een rol speelt in de priming van deze enhancers. Samengevat biedt deze studie een catalogus van functioneel gevalideerde enhancers in NSCs, en verschaft nieuwe inzichten in genregulerende mechanismen in NSCs.

In **hoofdstuk 5** hebben we een grafische gebruikersinterface gemaakt om de enhancer-gerelateerde informatie, verkregen uit voorgaande hoofdstukken, op een gebruikersvriendelijke manier te verkennen. Deze visualisatie applicatie stelt gebruikers in staat om enhancer-activiteit, enhancer-gen interacties en enhancer-ziekte associaties, en andere enhancer-gerelateerde informatie tijdens de ontwikkeling interactief te verkennen. De lancering van de applicatie zal een wijdverspreide datatoegang vergemakkelijken voor gebruikers die niet hoeven te zijn opgeleid in bioinformatica en zal het gebruik van enhanceranalyse in een klinische setting helpen vergemakkelijken.

Tenslotte geeft **hoofdstuk 6** een algemene discussie van de bevindingen van dit proefschrift, waarbij relevante resultaten worden belicht en besproken in de context van recente literatuur, met een vooruitblik naar toekomstige klinische implementatie van deze bevindingen.

Curriculum Vitae

Personal Information

Name: Soheil Yousefi
Place of birth: Gorgan, Iran
Email: soheil.yousefi01@gmail.com

Professional Experience and Education

PhD research, group of Dr. Tahsin Stefan Barakat in the Department of Clinical Genetics, Erasmus University Medical Center, The Netherlands (2018-2022).

Master of Science in Animal Genetics and Statistics, Gorgan University of Agricultural Sciences and Natural Resources, Iran (2009-2011).

Bachelor of Science in Animal Science, Gorgan University of Agricultural Sciences and Natural Resources, Iran (2005-2008).

Internships

AJA Medical Science University, Iran (2016-2017).

Group of Prof. Dr. Peter-Bram `t Hoen in the Department of Human Genetics, Leiden University Medical Center, The Netherlands (2015-2016).

List of publications

Soheil Yousefi, Eskeatnaf Mulugeta and Tahsin Stefan Barakat. ***EnhancerExplorer:* an interactive graphical user interface application to explore non-coding regulatory elements for brain development in the human genome** (*in preparation*).

Elena Perenthaler, Rutger W.W. Brouwer, Soheil Yousefi, Anita Nikoncuk, Ruizhi Deng, Kristina Lanko, Wilfred F.J. van IJcken and Tahsin Stefan Barakat. **Investigating the chromatin architecture of the UGP2 locus by targeted chromatin conformation capture** (*in preparation*).

Elena Perenthaler*, Soheil Yousefi*, Ruizhi Deng*, Anita Nikoncuk, Wilfred F.J. van Ijcken, Eskeatnaf Mulugeta and Tahsin Stefan Barakat. **Identification of the active enhancer landscape in Neural Stem Cells by ChIP-STARR-seq** (*in preparation*).

Elena Perenthaler, Soheil Yousefi, Ruizhi Deng, Anita Nikoncuk, Kristina Lanko, Wilfred F.J. van IJcken, Eskeatnaf Mulugeta, Jeroen Demmers and Tahsin Stefan Barakat. **YY1 interacts with cell-type specific complexes in embryonic and neural stem cells** (*in preparation*).

Elena Perenthaler*, Ruizhi Deng*, Soheil Yousefi*, Anita Nikoncuk, Tsung Wai Kan, Mariana Pelicano de Almeida, Wilfred F.J. van Ijken, Eskeatnaf Mulugeta and Tahsin Stefan Barakat. **Dissecting the role of YY1 in determining enhancer activity and gene expression** (*in preparation*).

Ruizhi Deng, Eva Medico-Salsench, Anita Nikoncuk, Reshmi Ramakrishnan, Kristina Lanko, Nikolas A. Kühn, Herma C van der Linde, Sarah Lor-Zade, Fatimah Albuainain, Yuwei Shi, Soheil Yousefi, Ivan Capo, Evita Medici van den Herik, Marjon van Slegtenhorst, Rick van Minkelen, Geert Geeven, Monique T. Mulder, George J.G. Ruijter, Dieter Lütjohann, Edwin H. Jacobs, Genomics England Research Consortium, Henry Houlden, Alistair T. Pagnamenta, Kay Metcalfe, Adam Jackson, Siddharth Banka, Lenika De Simone, Abigail Schwaede, Nancy Kuntz, Timothy Blake Palculict, Safdar Abbas, Muhammad Umair, Mohammed AlMuhaizea, Hanan AlQudairy, Maysoon Alsagob, Catarina Pereira, Roberta Trunzo, Vasiliki Karageorgou, Aida M. Bertoli-Avella, Peter Bauer, Arjan Bouman, Lies H. Hoefsloot, Tjakko J. van Ham, Mahmoud Issa, Maha S. Zaki, Joseph G. Gleeson, Rob Willemsen, Namik Kaya, Stefan T. Arold, Reza Maroofian, Leslie E. Sanderson and Tahsin Stefan Baraka. **Loss-of-function of AMFR causes autosomal recessive hereditary spas-**

tic paraplegia by altering lipid metabolism and can be positively modulated by statin treatment in a preclinical model (*Submitted*).

Soheil Yousefi, Ruizhi Deng*, Kristina Lanko*, Eva Medico Salsench*, Anita Nikoncuk*, Herma C van der Linde, Elena Perenthaler, Tjakko J van Ham, Eskeatnaf Mulugeta and Tahsin Stefan Barakat. **Comprehensive multi-omics integration identifies differentially active enhancers during human brain development with clinical relevance.** *Genome Medicine. 2021; 13(1):162*.

Elena Perenthaler, Anita Nikoncuk*, Soheil Yousefi*, Woutje M Berdowski*, Maysoon Alsagob*, Ivan Capo, Herma C van der Linde, Paul van den Berg, Edwin H Jacobs, Darija Putar, Mehrnaz Ghazvini, Eleonora Aronica, Wilfred FJ van IJcken, Walter G de Valk, Evita Medici-van den Herik, Marjon van Slegtenhorst, Lauren Brick, Mariya Kozenko, Jennefer N Kohler, Jonathan A Bernstein, Kristin G Monaghan, Amber Begtrup, Rebecca Torene, Amna Al Futaisi, Fathiya Al Murshedi, Renjith Mani, Faisal Al Azri, Erik-Jan Kamsteeg, Majid Mojarrad, Atieh Eslahi, Zaynab Khazaei, Fateme Massinaei Darmiyan, Mohammad Doosti, Ehsan Ghayoor Karimiani, Jana Vandrovcova, Faisal Zafar, Nuzhat Rana, Krishna K Kandaswamy, Jozef Hertecant, Peter Bauer, Stephanie Efthymiou, Henry Houlden, Aida M Bertoli-Avella, Reza Maroofian, Kyle Retterer, Alice S Brooks, Tjakko J van Ham and Tahsin Stefan Barakat. **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 Neuropathologica. 2020; 139(3):415-442*.

Elena Perenthaler*, Soheil Yousefi*, Eva Niggl* and Tahsin Stefan Barakat. Beyond the Exome: **The Non-coding Genome and Enhancers in Neurodevelopmental Disorders and Malformations of Cortical Development.** *Frontiers in Cellular Neuroscience. 2019; 13:352*.

Soheil Yousefi, Tooba Abbassi-Daloii, Mojtaba Tahmoorespur and Mohammad Hadi Sekhavati. **Nanoparticle or conventional adjuvants: which one improves immune response against Brucellosis?** *Iranian Journal of Basic Medical Sciences. 2019; 22:360-366*.

Soheil Yousefi, Tooba Abbassi-Daloii, Thirsa Kraaijenbrink, Martijn Vermaat, Hailiang Mei, Peter van‘t Hof, Maarten van Iterson, DariaV. Zhernakova, Annique Claringbould, Lude Franke, Leen M. ‘t Hart, Roderick C. Slieker, Amber van der Heijden, Peter de Knijff, BIOS consortium and Peter A.C.’t Hoen. **A SNP panel for**

identification of DNA and RNA specimens. *BMC Genomics. 2018;19:90*.

Soheil Yousefi, Tooba Abbassi-Daloii, Mohammad Hadi Sekhavati and Mojtaba Tahmoorespur. **Evaluation of immune responses induced by polymeric OMP25-BLS Brucella antigen**. *Microbial Pathogenesis. 2018; 115:50-56*.

Soheil Yousefi*, Tooba Abbassi-Daloii*, Mohammad Hadi Sekhavati and Mojtaba Tahmoorespur. **Impact of heat shock protein 60KD in combination with outer membrane proteins on immune response against Brucella melitensis**. *APMIS. 2018; 126: 65–75*.

Tooba Abbassi-Daloii, Soheil Yousefi, Eleonora de Klerk, Laurens Grossouw, Muhammad Riaz, Peter A.C.'t Hoen and Vered Raz. **An alanine expanded PAB-PN1 causes increased utilization of intronic polyadenylation sites**. *npj Aging and Mechanisms of Disease. 2017; 3:6*.

Mohammad Hadi Sekhavati, Reza Majidzadeh Heravi, Mojtaba Tahmoorespur, Soheil Yousefi, Tooba Abbassi-Daloii and Rahabe Akbari. **Cloning, molecular analysis and epitopics prediction of a new chaperone GroEL Brucella melitensis antigen.** *Iranian Journal of Basic Medical Sciences. 2015; 18:499-505*.

* These authors contributed equally

PhD Portfolio

Courses	Year
The Workshop UCSC Genome Browser	2018
R2 Workshop	2018
Scientific Integrity	2019
Next Generation Sequencing Data Analysis	2019
Single Cell Analysis	2019
Python Programming	2019
Biomedical Writing for PhD candidate	2021
SNP Course: SNP and human Diseases	2021
Survival Analysis	2021

Seminars and Workshops	
26th MGC Workshop	2019
27th MGC Workshop	2021
28th MGC Workshop	2022
Clinical Genetics Meeting	2018-2022
Clinical Genetics Seminars	2018-2022
Journal Clubs PhD Students	2018-2022
Cell Biology Meeting	2018-2022
Cell Biology Seminars	2018-2022
Sophia Research Days	2018-2022

(Inter)national Conferences	
MGC Symposium	2019,2022
European Society of Human Genetics	2020
European Society of Human Genetics	2022
Oral Presentation, 27th MGC Workshop	2021
Oral Presentation, 28th MGC Workshop	2022
Oral Presentation, The UK and Dutch Clinical Genetics joint meeting	2022
Poster Presentation, Biomedical Science PhD Day	2022

Acknowledgements

Writing the last chapter of my dissertation feels like both a relief and a sorrow. But it is time to begin a new journey in my life and to thank everyone who has supported me throughout these years. Without them, this dissertation would have been incomplete.

My deepest gratitude goes to my co-promotors Dr. Tahsin Stefan Barakat and Dr. Eskeatnaf Mulugeta. Dear Stefan, I remember the time when I emailed you to join your group as a bioinformatician. I thank you for giving me this opportunity and supporting me on my way to PhD and helping me to grow as a scientist. Dear Eskeww, thank you for your support and helpful discussions we had about my projects.

I would like to thank my late promotor Prof.dr. Robert Hofstra for giving me confidence and making me happier after each meeting where we talked about my project and my plans. You are greatly missed. I would also like to extend my gratitude to my promotor, Prof.dr. Ype Elgersma, who supported me on my way to my PhD and in writing my thesis.

I would also like to thank my dissertation committee, Prof.dr. Peter-Bram ‘t Hoen, Prof.dr. Joost Gribnau, Dr. Annelies de Klein, Prof.dr. Tjitske Kleefstra and Prof.dr. Ruud Delwel. I thank you for your time and effort in reading my work. It is an honor to have you on my committee.

Dear Anita and dear Eva M., thank you for being my paranymphs. I am so glad that you will stand by my side. A very special thanks also to Elena, Anita and Ruizhi. Dear Elena and dear Anita, I cannot thank you enough for all the effort you put into the wet lab parts of this book. Dear Ruizhi, thank you so much for your help. I enjoyed working with you and talking about different topics. I also want to thank all the other members of the group, Kristina, Sarah, Yuwei, Aliya, Fatima, Leslie, Rachel, Mashiro and Varun. You are the best colleagues I could have. Thank you so much for your contribution and for the great time we had. Dear friends, I consider myself lucky to have worked with you, and without your help, this book would not have been finished.

I would also like to thank Geert and Walter for their bioinformatics support and the useful discussions we had.

I thank all my wonderful colleagues and friends in the Department of Clinical Genetics and 9th floor. It has been a great experience to work in this department. Adriana, Alessandro, Almira, Ana, Atze, Bert, Bianca, Chantal, Charlotte, Christina, Claudia, Daphne, Darija, Douglas, Erik, Erwin, Esmay, Eva V, Eva N, Fabio, Federico, Fenne, Frank, Gerben, Grazia, Guido, Herma, Jeannette, Jonathan, Jordy, Katherine, Kirke, Kyra, Laura K, Laura V, Lies-Anne, Maria, Mariana, Marjoleine, Martyna, Max, Mehrnaz, Michiel, Mike, Monica, Naomi, Natasha, Niko, Nina , Nynke, Pablo, Pim, Rodrigo, Rianne, Rob, Rob V, Roy, Saif, Shamiram, Stijn, Tjakko, Thomas, Tom, Valerie, Vincenzo, Wilfred, Wim, Wim Q, Woutje, Yu-Ying and the other past and present colleagues thank you for the good time and cooperation.

Finally, no words can express my gratitude to my parents who made me who I am and for their unconditional support and kindness. Last but not least, it was my wife and closest friend Tooba who inspired, motivated and encouraged me throughout all these years of fun, stress and hard work. Dear Tooba, I owe this success to your support and love. Thank you for standing by me through all the challenges.