De novo mutations in fetal brain specific enhancers play a significant role in severe intellectual disability

Matias G De Vas[1], Myles G Garstang[2,3], Shweta S Joshi[1], Tahir Khan[2], Goutham Atla[1,4,5], David Parry[6], David Moore[7], Ines Cebola[1], Shuchen Zhang[8], Wei Cui[8], Anne K Lampe[7], Wayne W Lam[7], David R FitzPatrick[6], Madapura M Pradeepa[2,3*] and Santosh S Atanur[1,9*]

1 - Department of Medicine, Faculty of Medicine, Imperial College London, London, W12 0NN

2 - Blizard institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, E1 2AT

3 - School of Biological Sciences, University of Essex, Colchester, CO4 3SQ

4- CIBER de Diabetes y Enfermedades Metabólicas Asociadas, Spain

5 - Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain

6 - MRC Human Genetics Unit, University of Edinburgh, EH4 2XU

7 - South-East Scotland Regional Genetics Service, Western General Hospital, Edinburgh EH4 2XU

8 - Institute of Reproductive and Developmental Biology, Faculty of Medicine, Imperial College London, London, W12 0NN

9 - Previous institute: Centre for Genomic and Experimental Medicine, University of Edinburgh, Edinburgh EH4 2XU

* Corresponding authors

Correspondence to Santosh Atanur satanur@ic.ac.uk

or Madapura Pradeepa email: p.m.madapura@qmul.ac.uk

Abstract

The genetic aetiology of a large proportion of intellectual disability (ID) cases still remains undiagnosed as *de novo* mutations (DNMs) in protein coding regions of the genome explain only 35-40% of the cases[1,2]. We sequenced whole genomes of 70 individuals, including 24 ID probands, to identify potentially pathogenic DNMs at distal *cis*-regulatory elements, as they may explain some of these genetically undiagnosed ID cases. In our cohort, DNMs were significantly enriched in fetal brain specific enhancers that were intolerant to mutations within the human population. The majority of these enhancer DNMs showed a significant effect on enhancer activity compared to the reference alleles when tested using luciferase reporter assays. Furthermore, CRISPR interference of *CSMD1* enhancer resulted in overexpression of the neurogenesis gene *CSMD1* suggesting that the disruption of a transcriptional repressor binding site due to an enhancer DNM could be a potential cause of ID in one of the patients. Taken together, our results demonstrate that DNMs in tissue specific enhancers play an important role in the aetiology of ID.

Introduction

Intellectual disability (ID) is a neurodevelopmental disorder that is characterized by limitations in both intellectual functioning and adaptive behavior[1]. The worldwide prevalence of ID is thought to be 1%[3], which provides a significant medical and social challenge to our society. The clinical presentation of ID tends to be highly heterogeneous and often occurs along with congenital malformations or other neurodevelopmental disorders such as epilepsy and autism[1]. The methods to establish aetiological diagnosis in patients with ID have been significantly improved in recent years[4]. The genetics is known to play a significant role in the aetiology of ID[1]. With the introduction of next generation sequencing (NGS) technologies, over the past decade, great progress has been made in identifying genetic causes of ID. It has been shown that *de novo* protein truncating mutations and *de novo* copy number variations play a crucial role in the aetiology of ID[2,5]. Large numbers of genes have been implicated in ID and related neurodevelopmental disorders[1], however, mutations in coding regions of the genome only explain up to 35-40% of ID cases[2]. The genetic aetiology of a large proportion of ID cases still remains unidentified. *De novo* mutations (DNM) in non-coding regions of the genome, specifically distal *cis* regulatory elements (CRE), could explain some of the cases in which no causal pathogenic coding mutation has been identified. However, to date, limited efforts have been made to understand contribution of enhancer mutations in ID.

Previous studies have implicated CRE mutations in monogenic developmental disorders including preaxial polydactyly (*SHH*)[6,7], Pierre Robin syndrome (*SOX9*)[8], congenital heart disease (*TBX5*)[9] and pancreatic agenesis (*PTF1A*)[10]. In addition, sequencing of evolutionarily ultra-conserved regions in developmental disorder probands found enrichment of DNMs only in fetal brain enhancers[11]. On the contrary, large-scale whole genome sequencing of autism spectrum disorder (ASD) patients along with unaffected parents and unaffected siblings (quad) has found no significant association between any non-coding classes other than promoter regions with ASD[12]. In the case of coding mutations, only protein truncating mutations, which make up a relatively small fraction of all coding mutations, show significant enrichment in neurodevelopmental disorders[12,13]. Unlike protein coding regions, it is challenging to distinguish between potentially pathogenic mutations and benign mutations in distal CREs due to the absence of codon structure.

The brain is the most complex organ in the body. During brain development, various lineage specific genes need to be expressed in the right amount, at the right time,

and at the right location. The spatio-temporal expression of genes is regulated by distal CREs. It has been shown that the disease-associated variants are enriched in diseases relevant tissues[14]. Thus, DNMs in fetal brain specific (FBS) enhancers that regulate expression of essential brain developmental genes could affect brain morphogenesis with severe functional consequences. Furthermore, advanced human cognition has been attributed to human brain enhancers that are gained during evolutionary expansion and elaboration of the human cerebral cortex[15]. It is therefore possible that DNMs in human gained (HG) enhancers could have a significant impact on human cognition.

In the present study, we performed whole genome sequencing of 70 individuals, including 24 ID probands and their unaffected parents. To understand the role of distal *cis*-regulatory mutations, we specifically focused on FBS and HG enhancers. We performed an extensive experimental validation and showed that the majority of enhancer DNMs tested in this study lead to significant loss or gain of enhancer activity. Perturbation studies of putative *CSMD1* enhancer lead to a striking transcriptional phenotype suggesting a contribution of *CSMD1* enhancer DNMs in neurodevelopmental disorders.

Results

Whole genome sequencing and identification of *de novo* mutations

We performed whole genome sequencing (WGS) of 70 individuals including 24 probands with severe intellectual disability (ID) and their unaffected parents at an average genome wide depth of 37X (supplementary table 1). Our cohort includes 22 trios and one quad family with two affected probands. We identified on average 4.08 million genomic variants per individual that include 3.36 million single nucleotide variants (SNVs) and 0.72 million short indels (supplementary table 1). We focused our analysis mainly on *de novo* mutations (DNMs), as it has been shown that DNMs contribute significantly to neurodevelopmental disorders[5,16]. We identified a total of 1,261 DNMs in 21 trios after excluding one trio from the analysis due to excessive number of DNMs. An average of 60 high quality DNMs per proband were identified, which includes 55.2 SNVs and 4.8 indels per proband (supplementary table 2). The number of DNMs identified in this study is similar to the number of DNMs identified per proband in previous WGS studies on neurodevelopmental disorders[2,12,17]. It has been shown that *de novo* copy number variants (CNVs) play a significant role in severe ID[2]. We identified a total of 45 *de novo* CNVs (2.14 per proband; supplementary table 3).

Protein coding *de novo* mutations and copy number variants

The role of protein truncating mutations in ID is well established; hence we first looked at DNMs located in protein coding regions of the genome. A total of 23 DNMs were located in the protein coding regions of the genome (average 1.1 DNMs per proband). Of the 23 coding mutations, 15 were non-synonymous coding mutations or protein truncating mutations. In six ID probands, we identified various types of pathogenic mutations in the genes *KAT6A*, *TUBA1A*, *KIF1A*, *NRXN1* and *PNKP* all of them previously implicated in ID[2,18]. The mutation in *KAT6A* gene resulted in a premature stop codon while genes *TUBA1A* and *KIF1A* showed non-synonymous coding mutations that were reported to be pathogenic in ClinVar[19] (supplementary table 4). One *de novo* CNV resulted in partial deletion of known ID gene *NRXN1*. A family with two affected siblings was analyzed for the presence of recessive variants. We identified a homozygous 17bp insertion in the gene *PNKP* (supplementary table 4) in both siblings. This insertion has been reported as pathogenic in ClinVar[19].

***De novo* mutations in *cis*-regulatory elements (CRE)**

In our severe ID cohort, we did not identify pathogenic coding DNMs in 17 ID cases (~70%); hence we decided to investigate potentially pathogenic mutations in distal *cis*-regulatory regions (CRE) of the genome. To increase the statistical power, we included 30 previously published severe ID samples in which no pathogenic protein coding DNMs have been found using WGS[2]. We hypothesized that the DNMs in fetal brain specific distal CRE could perturb expression levels of genes that are essential for brain development, leading to ID. We identified 27,420 fetal brain specific enhancers using epigenomic roadmap data[14](see Methods). In addition, we included 8,996 human gained enhancers that have been shown to be active during cerebral corticogenesis[15].

A total of 83 DNMs (an average of 1.77 DNMs per proband) were located within fetal brain specific or human gained enhancers. To test the enrichment of observed number of DNMs in fetal brain specific enhancers and human gained enhancers we used previously defined framework for interpreting DNMs[20]. In short, the model determines mutability of a given base by taking into consideration one nucleotide on each side (trinucleotide context). We used DNMs identified in healthy individuals in genome of the Netherlands (GoNL)[21] to test the suitability of the mutational model for estimating expected number of DNMs in enhancer regions. In healthy individuals the observed number of DNMs (n=252) was almost equal to expected number of DNMs

(n=250) in fetal brain specific enhancers (*P*=0.53; Figure1), suggesting that the mutational model fits well to determine expected number of DNMs within tissue specific enhancers. Next, we investigated the enrichment of observed number of DNMs over expected in fetal brain specific enhancers and human gain enhancers within the ID cohort. As a control we used multiple fetal and adult tissue specific enhancers as well as a set of randomly selected, sequence composition matched quiescent regions. The fetal brain specific enhancers and human gained enhancers showed enrichment for DNMs in our ID cohort as compared to the expected number of mutations, however enrichment was statistically not significant (*P* = 0.10; Figure1). However, fetal brain specific enhancers showed significant enrichment when compared to fetal lung specific enhancers as a control (*P* = 0.038). Almost all the non-brain tissue specific enhancers and control quiescent regions showed depletion of DNMs in ID patients. Our results demonstrate that the enhancers from disease relevant tissues are specifically enriched for DNMs in this disease cohort.

Recurrent DNMs in fetal brain specific or human gained enhancers

Due to the smaller sample size, we were not able to identify individual enhancers with recurrent DNMs hence we investigated whether clusters of enhancers that regulate the same gene show recurrent DNMs. To identify clusters of enhancers that target the same gene, we used the following approaches: neuronal progenitor cells (NPC) promoter capture Hi-C (PCHi-C) data[22]; correlation of H3K27ac signal at promoters and enhancers across multiple tissues; and promoter-enhancer correlation using chromHMM segmentation data. If any enhancer remained unassigned after application of these approaches, we assigned the closest fetal brain expressed gene as a target gene of the enhancer. For all approaches, we restricted our search space to brain topologically associated domains (TADs)[23] as the majority of enhancer-promoter interactions happen within TADs[24]. The enhancer clusters associated with three genes, *CSMD1*, *OLFM1* and *POU3F3* showed recurrent mutations within our cohort (supplementary figure 1). The presence of three enhancer clusters with recurrent DNMs was significantly higher than expected (*P* = 0.016).

Biological properties of genes associated with DNM-containing enhancers

To investigate whether enhancer associated genes have previously been implicated in ID or related neurodevelopmental disorders, we used three gene sets; known ID genes[2,18], genes implicated in neurodevelopmental disorders in DDD project[16] and autism risk genes (SFARI genes)[25]. This provided us with a unique set of 1,868

genes previously implicated in neurodevelopmental disorders. The genes associated with DNM-containing enhancers were enriched for known neurodevelopmental disorder genes (25 genes, P = 0.025, supplementary table 5). Furthermore, the putative target genes of DNM-containing enhancers were involved in nervous system development (P = 7.4x10^{-4}, supplementary table 6) and were predominantly expressed in the prefrontal cortex (P =6.5 x10^{-3} supplementary table 7), the region of the brain that has been implicated in social and cognitive behavior, personality expression, and decision-making.

The potential functional effect of enhancer mutations may be through altered expression target genes. Recently, it has been shown that the majority of known severe haploinsufficient human disease genes are intolerant to loss of function (LoF) mutations[26]. We compared the putative target genes of DNM-containing enhancers with the recently compiled list of genes that are intolerant to LoF mutations (pLI >=0.9)[26]. We found that a significantly higher proportion of enhancer DNM target genes were intolerant to LoF mutations than expected (P = 4.2x10^{-5}, supplementary table 8). Taken together, this analysis suggests that haploinsufficiency of fetal brain specific genes involved in nervous system development due to CRE mutations could lead to severe ID.

Population constraints in fetal brain specific enhancers and human gain enhancers

It has been shown that single nucleotide changes in enhancer regions can lead to severe developmental defects[10]. On the contrary, recent studies suggest that functional redundancy of enhancers of developmental genes reduce the likelihood of severe functional consequences of enhancer mutations[27]. We reasoned that enhancers that are essential for human cognitive development must be intolerant to mutations within the human population. Hence we investigated whether fetal brain specific enhancers themselves are intolerant to mutations. The recently developed context dependent tolerance score (CDTS) provides sequence constrains across human population in non-coding regions of the genome at 10bp resolution[28]. The human gain enhancers as well as fetal brain specific enhancers showed significant enrichment (P < 2.2x10^{-16} and P < 2.2x10^{-16} respectively) for constrained genomic regions (CDTS <=30, figure 2A) when compared to sequence composition matched control regions. This finding suggests that enhancers that regulate spatio-temporal expression of genes during brain development tend to be intolerant to mutations. In our ID cohort, 25 enhancer DNMs were located within constrained regions (CDTS

score <= 30) of the genome and the majority (13 out of 25) of them were associated with at least one loss of function intolerant (pLI >= 0.9) gene (figure 2C). We found marginal enrichment of DNMs in constrained enhancers (CDTS<=30) as compared to control quiescent regions (*P*=0.072; Figure 2B). However, when we restricted the analysis to highly constrained regions (CDTS<=20), we found significant excess of DNMs (*P*=0.027; Figure 2B) in fetal brain specific enhancers as compared to control regions. An increased burden of DNMs in ID patients in enhancers that are intolerant to mutations within the human population suggests that DNMs in population-constrained regions are more likely to be functional, specifically if they are associated with loss of function mutation intolerant genes.

Effect of DNMs on transcription factor binding sites

Enhancers regulate gene expression through the binding of sequence-specific transcription factors (TFs) at TF binding sites (TFBS) within the enhancer[29]. DNMs that alter the sequence of putative TFBS or create putative TFBS within the enhancer region could have a significant impact on target gene expression. Of the 82 *de novo* SNVs, 32 were predicted to alter putative TFBS affinity strongly, either by destroying or creating TFBS (supplementary table 9).

Functional validation of DNMs using luciferase reporter assays

DNMs in enhancers could have severe functional consequences, specifically if they regulate intolerant genes. To test the functional impact of regulatory mutations on enhancer activity, we investigated 11 enhancer DNMs (supplementary table 10) using luciferase reporter assays in a neuroblastoma cell line SH-SY5Y. Of the 11 DNM containing enhancers, 10 showed at least one allele (wild type or mutant) with significantly higher activity than the negative control, suggesting that these putative enhancers are indeed active enhancers in this neuronal cell line. Of 10 active enhancers, nine showed allele specific activity with five showing loss of activity and four showing gain of activity due to DNMs (Figure 3). The majority of the DNMs that showed allele-specific enhancer activity altered the core base of the TF motif with position specific weight >= 0.95 (supplementary table 9). Both the DNMs in the putative *CSMD1* enhancer cluster showed gain of activity compared to the wild type allele, while both *OLFM1* enhancer DNMs showed loss of activity due to DNMs. These results demonstrate that the DNMs in fetal brain specific enhancers or human gain enhancers that significantly alter TF binding affinity are more likely to be functional.

Functional validation of *CSMD1* enhancer DNM using CRISPR interference

We identified recurrent DNMs in the *CSMD1* enhancer cluster with two DNMs (chr8: 2177122C>T and chr8: 2411360T>C) from two unrelated ID probands (fam6 and fam3 respectively). Both the probands showed developmental delay and both were overgrown with high birth weights (above the 91st centile) and remain large. Both *CSDM1* enhancer DNMs resulted in gain of enhancer activity (Figure 3). The *CSMD1* gene is involved in neurogenesis[30] and has been shown to be associated with schizophrenia[31]. Motif analysis predicted that the DNM chr8:2411360T>C from fam3 disrupts the binding site for the transcriptional repressor *TCF7L1*[32,33] (Figure 4A). *TCF7L1* is known to inhibit premature neurogenesis[34]. To decipher the functional effect of DNM chr8:2411360T>C (fam3) located in putative transcriptional enhancer of *CSMD1* (2.44 Mb away from TSS of *CSMD1*), we performed lentiviral CRISPR interference (CRISPRi) by recruiting dCas9 fused with the KRAB domain to the *CSMD1* enhancer (fam3) in the neuronal precursor cell line – Lund human mesencephalic (LUHMES). Given that the expression of *CSMD1* was not detectable in LUHMES cells (Figure 4D), probably due to repression by *TCF7L1*, we differentiated these cells into neurons. Differentiated neurons with CRISPRi of *CSMD1* enhancer showed significantly higher expression of *CSMD1* than control cells (P=0.004; Figure 4B). The KRAB domain is known to repress transcription through heterochromatin spreading[35,36] or simply by steric interference of endogenous regulatory components[37]. In this case it is possible that the transcriptional repressor *TCF7L1* may not be able to bind at the TFBS due to heterochromatinisation or obstruction of the *CSMD1* enhancer leading to overexpression of *CSMD1*.

To investigate the detailed molecular phenotype of CRISPRi at CSMD1, we performed RNAseq upon differentiation of neuronal precursors to neurons. RNAseq data shows significant up-regulation of CSMD1 in CRISPRi inactivated neurons as compared to controls. In addition, the genes *MYH3* (myosin heavy chain 3), expressed exclusively during embryonic development, and *PCDHGA11* (Protocadherin Gamma-A11), which plays a significant role in establishment of cell-cell connections in the brain, showed a significantly strong down-regulation (Figure 4C). Up-regulation of *CSMD1* and down-regulation of *MYH3* and *PCDHGA11* suggest that *CSMD1* regulatory DNM chr8:2411360T>C may affect the formation of axons and the establishment of neuronal connections, which might be a potential cause of ID in the fam3 proband.

We observed that *CSMD1* was not expressed in LUHMES cells while *TCF7L1* was highly expressed (Figure 4D). Inversely, in differentiated neurons, *CSMD1* showed a high expression while a striking down-regulation was observed for TCF7L1 compared to the undifferentiated state (Figure 4D). Analyzing gene expression of *CSMD1* and *TCF7L1* across multiple tissues using GTEx data[38], we found that the expression pattern of both of these genes were almost mutually exclusive. In tissues where *TCF7L1* was expressed at high levels, *CSMD1* did not show any expression. On the contrary, *CSMD1* showed expression only in the tissues where *TCF7L1* showed relatively lower levels of expression (Figure 4E), strengthening our assumption that TCF7L1 represses *CSMD1* in LUHMES cells. Taken together, our analysis shows that the DNM in the putative *CSMD1* enhancer might affect the binding of the transcriptional repressor *TCF7L1*, leading to premature expression of *CSMD1*. This could be the potential mechanism by which enhancer DNM leads to ID in this patient.

Next, we explored occurrence of DNMs in the *CSMD1* enhancer cluster in large-scale WGS studies on autism spectrum disorders (ASD). The Simons Simplex Collection (SSC) has sequenced WGS of 1,902 ASD families (quad)[12], while the MSSNG database has compiled WGS of 2,281 ASD trios[17]. Within the SSC cohort, four DNMs from unrelated ASD cases were located in the *CSMD1* enhancer cluster, while unaffected siblings did not contain any DNM in *CSMD1* enhancer cluster. Additionally, we found four ASD patients from the MSSNG database harboring DNMs in the *CSMD1* enhancer cluster. Presence of a total of eight DNMs in ASD cohort suggests that DNMs in *CSMD1* enhancer cluster tend to be highly penetrant in neurodevelopmental disorders.

Discussion

This study provides strong clues about the role of disease relevant tissue specific enhancers in severe ID as we see a marginal enrichment of DNMs in fetal brain specific enhancers. However, an absence of ways to distinguish potentially pathogenic mutations from benign regulatory mutations makes it challenging to determine the true burden of pathogenic mutations in CREs. Aggregation of a minority of the pathogenic mutations with the majority of benign regulatory mutations nullifies any signal from pathogenic mutations in the disease cohort[12,13].

The large majority of the DNMs tested in this study showed allele-specific activity. The majority of validated DNMs alter core bases of the motif with position weight

>=0.95 in the position specific weight matrix. This reiterates the fact that DNMs that alter core bases of the TF binding motif strongly alter TF binding affinity, thus are more likely to be functional[39]. Furthermore, enrichment of DNMs in the enhancers that are intolerant to mutations within human populations suggests that the DNMs in essential enhancers are more likely to be functional. This study shows that the tissue specificity of enhancers, population constraint and changes in TF binding affinity are some of the key factors that could determine the pathogenicity of the enhancer DNMs. To confirm these findings and to identify additional properties of potentially pathogenic regulatory DNMs, functional validation of a large number of enhancer DNMs from large-scale WGS studies[13,17], using high throughput assays such as massively parallel reporter assays[40] or BiT-STARR-seq[41], is required. Furthermore, to understand the role of regulatory mutations in ID and other neurodevelopmental disorders it is important to understand the functional genomic architecture of brain tissue at various developmental stages[42]. Improved annotation of functional enhancers using epigenomic profiling methods, together with genome-wide reporter assays and accuracy in identifying enhancer-promoter interactions using promoter capture Hi-C[43] and Hi-ChIP[44], would significantly improve identification of potentially pathogenic of regulatory mutations.

We showed that a DNM in a transcriptional repressor binding site leads to overexpression of *CSMD1*, a gene involved in neurogenesis. The majority of previously reported pathogenic regulatory mutations in monogenic disorders have been shown to abolish enhancer activity[6–10]. At least to our knowledge, mutations in a repressor-binding site that lead to a monogenic disorder have not yet been documented. However, such reports can be found in common diseases and cancer[45,46]. The obesity risk allele associated with *FTO* was shown to disturb a conserved motif for the ARID5B repressor, leading to *IRX3* and *IRX5* overexpression during early adipocyte differentiation[46]. Similarly, prostate cancer risk SNPs shown to disrupt a repressive loop, leading to overexpression of *HOXA13* and *HOTTIP*[45].

Taken together, our results show that tissue-specific regulatory mutations play a significant role in severe ID. Incorporating regulatory regions in the clinical diagnostic panel would significantly improve the diagnostic yield of ID and other neurodevelopmental disorders. Furthermore, this work demonstrates the importance of comprehensive analysis of WGS, followed by functional assays, to understand novel mechanisms through which regulatory DNMs lead to neurodevelopmental disorders.

Materials and methods

Selection criteria of intellectual disability patients for this study and ethical approval

The inclusion criteria for this study were that the affected individuals had a severe undiagnosed developmental or early onset pediatric neurological disorder and that samples were available from both unaffected parents. Written consent was obtained from each patient family using a UK multicenter research ethics approved research protocol (Scottish MREC 05/MRE00/74).

Sequencing and quality control

Whole genome sequencing was performed on the Illumina X10 at Edinburgh Genomics. Genomic DNA (gDNA) samples were evaluated for quantity and quality using an AATI, Fragment Analyzer and the DNF-487 Standard Sensitivity Genomic DNA Analysis Kit. Next Generation sequencing libraries were prepared using Illumina SeqLab specific TruSeq Nano High Throughput library preparation kits in conjunction with the Hamilton MicroLab STAR and Clarity LIMS X Edition. The gDNA samples were normalized to the concentration and volume required for the Illumina TruSeq Nano library preparation kits then sheared to a 450bp mean insert size using a Covaris LE220 focused-ultrasonicator. The inserts were ligated with blunt ended, A-tailed, size selected, TruSeq adapters and enriched using 8 cycles of PCR amplification. The libraries were evaluated for mean peak size and quantity using the Caliper GX Touch with a HT DNA 1k/12K/HI SENS LabChip and HT DNA HI SENS Reagent Kit. The libraries were normalised to 5nM using the GX data and the actual concentration was established using a Roche LightCycler 480 and a Kapa Illumina Library Quantification kit and Standards. The libraries were normalised, denatured, and pooled in eights for clustering and sequencing using a Hamilton MicroLab STAR with Genologics Clarity LIMS X Edition. Libraries were clustered onto HiSeqX Flow cell v2.5 on cBot2s and the clustered flow cell was transferred to a HiSeqX for sequencing using a HiSeqX Ten Reagent kit v2.5.

Alignment and variant calling

The de-multiplexing was performed using bcl2fastq (2.17.1.14) allowing 1 mismatch when assigning reads to a barcodes. Adapters were trimmed during the de-multiplexing process. Raw reads were aligned to the human reference genome (build GRCh38) using the Burrows-Wheeler Aligner (BWA) mem (0.7.13)[47]. The duplicated

fragments were marked using samblaster (0.1.22)[48]. The local indel realignment and base quality recalibration was performed using Genome Analysis Toolkit (GATK; 3.4-0-g7e26428)[49–51]. For each genome SNVs and indels were identified using GATK (3.4-0-g7e26428) HaplotypeCaller[52] creating a gvcf file for each genome. The gvcf files of all the individuals from the same family were merged together and re-genotyped using GATK GenotypeGVCFs producing single VCF file per family.

Variant filtering

Variant Quality Score Recalibration pipeline from GATK[49–51] was used to filter out sequencing and data processing artifacts (potentially false positive SNV calls) from true SNV and indel calls. First step was to create a Gaussian mixture model by looking at the distribution of annotation values of each input variant call set that match with the HapMap 3 sites and Omni 2.5M SNP chip array polymorphic sites, using GATK VariantRecalibrator. Then, VariantRecalibrator applies this adaptive error model to both known and novel variants discovered in the call set of interest to evaluate the probability that each call is real. Next, variants were filtered using GATK ApplyRecalibration such that final variant call set contains all the variants with 0.99 or higher probability to be a true variant call.

De novo mutations (DNM) calling and filtering

The de novo mutations (DNMs) were called using GATK Genotype Refinement workflow. First, genotype posteriors were calculated using sample pedigree information and the allele accounts from 1000 genome sequence data as a prior. Next, the posterior probabilities were calculated at each variant site for each sample of the trio. Genotypes with genotype quality (GQ) < 20 based on the posteriors are filtered out. All the sites at which both the parents genotype was homozygous reference (0/0) and child's genotype was heterozygous (0/1), with GQs >= 20 for each sample of the trio, were annotated as the high confidence DNMs. Only high confident DNMs that were novel or had minor allele frequency less than 0.01 in 1000 genome were selected for further analysis.

DNM annotations

DNM annotations were performed using Annovar[53]. To access DNM location with respect to genes, refseq, ENSEMBL and USCS annotations were used. To determine allele frequencies, 1000 genome, dbSNP, Exac and GnomAD databases were used. To determine pathogenicity of coding DNMs, annotations were performed

with CADD, DANN, EIGAN, FATHMM and GERP++ pathogenicity prediction scores. In addition, we determined whether any coding DNM has been reported in ClinVar database as a pathogenic mutation.

Structural variant detection and filtering

To detect structural variants (SV), we used four complimentary SV callers: BreakDancer[54], Manta[55], CNVnator[56] and CNVkit[57]. The BreakDancer and Manta use discordant paired end and split reads to detect deletions, insertions, inversions and translocations, while CNVnator and CNVkit detect copy number variations (deletions and duplications) based on read-depth information. The consensus SV calls were derived using MetaSV[58]. The MetaSV is the integrative SV caller, which merges SV calls from multiple orthogonal SV callers to detect SV with high accuracy. We selected SVs that were called by at least two independent SV callers out of four.

To detect *de-novo* SV, we used SV2[59]. SV2 is a machine-learning algorithm for genotyping deletions and duplications from paired-end whole genome sequencing data. In *de novo* mode SV2 uses trio information to detect *de novo* SVs at high accuracy.

Fetal brain specific enhancers and other tissue specific enhancers

Roadmap Epigenomic Project[14] chromHMM segmentations across 127 tissues and cell type were used to define brain specific enhancers. All the genic (intronic) and intergenic enhancers ("6_EnhG and 7_Enh) from male (E081) and female fetal brain (E082) samples were extracted. Genome-wide chromHMM chromatin state classification was performed in rolling 200bp windows. All the consecutive 200bp windows assigned as an enhancer in fetal brain were merged to get enhancer boundaries. A score was assigned to each enhancer based on the total number of 200pb windows covered by each enhancer. Next, for each fetal brain enhancer, we counted the number of 200bp segments assigned as an enhancer in the remaining 125 tissues and cell types. This provided enhancer scores across 127 tissues and cell types for all fetal brain enhancers. To identify fetal brain specific enhancers, Z scores were calculated for each fetal brain enhancer using the enhancer scores. Z scores were calculated independently for the male and female fetal brain enhancers. Independent Z score cutoffs were used for both male and female fetal brain enhancers such that approximately 35% of enhancers were selected. Furthermore, we intersected these enhancers with DNAse-seq data from male and female fetal

brain respectively. The fetal brain specific enhancers that overlap with DNAse hypersensitivity sites were selected. Next, the male and female fetal brain specific enhancers were merged together to get final set of 27,420 fetal brain specific enhancers. We used similar approach to identify tissue specific enhancers for selected fetal and adult non-brain tissues.

Human gain enhancers

Human gain enhancers published previously by Reilly et al[15] were downloaded from Gene Expression Omnibus (GEO) using accession number GSE63649.

***De novo* mutations from healthy individuals**

We downloaded *de novo* mutations identified in healthy individual in genomes of the Netherland (GoNL) study[21] from GoNL website (http://www.nlgenome.nl/). A total of 11,020 DNMs have been identified in 258 individuals in GoNL, that is 42.71 DNMs per individuals. In our ID cohort we identified 62.53 DNMs per individual. To avoid overestimation of enrichment we normalised number of DNMs per individual to our cohort my multiplying it by normalizing factor of 1.46.

Fetal brain specific genes

Roadmap Epigenomic Project[14] gene expression (RNA-seq) data from 57 tissues was used to identify fetal brain specific genes. Fetal brain specific genes were identified using only female fetal brain gene expression data, as RNA-seq data was available only for female fetal brain. For each gene, Z scores were calculated using RPKM values across 57 tissues. The genes with Z score greater than two were considered as the brain specific genes.

***De novo* mutation enrichment analysis**

The expected number of *de novo* mutations (DNMs) in fetal brain specific enhancers and human gain enhancers was estimated using the previously defined framework for *de novo* mutations[20]. The framework for the null mutation model is based on tri-nucleotide context where the second base is mutated. Using this framework, the probability of mutation for each enhancer was estimated based on the DNA sequence of the enhancer. Probability of mutation of all the enhancers within the enhancer set (fetal brain specific enhancers and human gain enhancers) was summed to estimate the probability of mutation for the entire enhancer set. The probability of mutation for fetal brain specific enhancers and human gain enhancers

was estimated separately. To estimate the expected number of DNMs, the probability of mutation for each enhancer set was multiplied by the cohort size (n=47). To estimate the significance of observed number of DNMs over expected number, Poisson distribution probabilities were invoked using R function ppois.

DNM effect on transcription factor binding

The R bioconductor package motifbreakR[60] was used to estimate the effect of DNM on transcription factor binding. The motifbreakR works with position probability matrices (PPM) for transcription factors (TF). MotifbreakR was run using three different TF databases: viz. homer, encodemotif and hocomoco. To avoid false TF binding site predictions, either with reference allele or with alternate allele, stringent threshold of 0.95 was used for motif prediction. DNMs that show a strong effect on transcription factor binding, as predicted by motifbreakR, were selected for further analysis.

Prediction of target genes of enhancers

Three different methods were used to predict the potential target genes of enhancers.

Chromosome conformation capture (Hi-C) comprehensively detects chromatin interactions in the nucleus; however, it is challenging to identify individual promoter-enhancer interactions using Hi-C due to the complexity of the data. In contrast, promoter capture Hi-C (PCHi-C) specifically identifies promoter-enhancer interactions as it uses sequence capture to enrich the interactions involving promoters of annotated genes[43]. The significant interactions between promoters and enhancers identified using PCHi-C in neuronal progenitor cells[22] were used to assign target genes to the DNM containing enhancers. The enhancers were overlapped with the PCHi-C HindIII fragments. If an overlap was found between an enhancer and the PCHi-C HindIII fragment, the significantly interacting regions (PCHi-C HindIII fragments representing promoters of the genes) of the PCHi-C HindIII fragment were extracted to assign genes to the enhancers.

For an enhancer to interact with a promoter, both promoter and enhancer need to be active in specific cells at a specific stage. To identify promoter-enhancer interactions, all the active promoters in fetal brain (as defined by chromHMM segmentation) were extracted. Promoter-enhancer interactions occur within topologically associated domains (TAD), hence, promoters that were located within the same TAD as that of a

DNM containing enhancer were used for analysis.

For each enhancer and promoter, H3K27ac counts were extracted from all tissues for which H3K27ac data was available in the Roadmap Epigenomic Project[14] ChIP-seq dataset. For fetal brain, H3K27ac ChIP-seq data published by Reilly *et al*[15] was used because H3K27ac ChIP-seq data was not available in Roadmap Epigenomic Project ChIP-seq dataset for fetal brain. The Spearman rank correlation coefficient (Spearman's rho) was calculated between each enhancer-promoter pair within the TAD using Scipy stats.spearmanr function from Python. The permutation test was performed to identify the significance of the correlation. The counts were randomly shuffled, independently for enhancers and promoters, 1000 times to calculate an adjusted *P* value. The interactions with an adjusted *P* value less than 0.01 were considered as a significant interaction between the enhancer and promoter.

Finally, if any enhancer remained unassigned to a gene using these approaches, they were assigned to fetal brain expressed genes within the TAD. A gene with an expression level more than or equal to 1 TPM in the Roadmap Epigenomic Project fetal brain RNA-seq data was considered to be expressed in the fetal brain.

Gene enrichment analysis

To test if enhancer associated genes were enriched for genes previously implicated in neurodevelopmental disorders, three different gene sets were used. 1) Intellectual disability (ID) gene list published in the review by Vissers *et al*[1] was downloaded from Nature website (https://media.nature.com/original/nature-assets/nrg/journal/v17/n1/extref/nrg3999-s1.pdf). 2) We compiled all the genes implicated in neurodevelopmental disorders in Deciphering Developmental Disorder (DDD) project[16]. 3) All the genes implicated in autism spectrum disorder were downloaded from SFARI browser (https://gene.sfari.org/). Significance of enrichment was tested using hypergeometric test in R.

Gene ontology enrichment and tissue enrichment analysis was performed using web-based tool Enricher (http://amp.pharm.mssm.edu/Enrichr/).

Probability of loss of function intolerance (pLI) scores for each gene was downloaded from Exome Aggregation Consortium (ExAC) browser (http://exac.broadinstitute.org/). Significance of enrichment was tested using hypergeometric test in R.

Cell culture

LUHMES cells were purchased from ATCC (CRL-2927). Cells were cultured as previously described (Scholz et al, JNC 2011). Briefly, cells were attached on pre-coated multi-well plates (50 ug/mL poly-L-ornithine and 1 ug/mL fibronectin - Sigma) and grown in proliferation medium: Advanced DMEM/F-12 plus N-2 Supplement (Thermo Fisher), 2 mM L-glutamine (Sigma) and 40 ng/mL recombinant basic fibroblast growth factor (R&D Systems). Cells were differentiated into neurons for 7 days using standard differentiation medium: Advanced DMEM/F-12 plus N-2 Supplement, 2 mM L-glutamine, 1 mM dibutyryl cAMP (Sigma), 1 ug/mL tetracycline (Sigma) and 2 ng/mL recombinant human GDNF (R&D Systems).

Neuroblastoma cell line (SH-SY5Y) was maintained in DMEM/F12 media (Gibco), 1% penicillin-streptomycin, 10 % fetal bovine serum and 2 mM L-glutamine.

Dual luciferase enhancer assays

Enhancer and control regions (500-600 bp) were amplified from human genomic DNA from HEK293T cells using Q5 High-Fidelity Polymerase (NEB). Amplified fragments were cloned into pGL4.23 plasmid (Promega), which consists of a minimal promoter and the firefly luciferase reporter gene. These regions were mutagenized in order to introduce the *de novo* mutations of interest using the Q5 Site-Directed Mutagenesis kit (NEB) using non overlapping primers. pGL4.23 plasmids containing putative enhancer DNA were sequence verified and transfected, together with a Renilla luciferase expressing vector (pRL-TK Promega) into SHSY-5Y cells using Lipofectamine 3000 (Invitrogen) following manufacturer's protocol. Firefly and Renilla luciferase activity was measured 24 hours after transfection using the Dual-Luciferase Reporter Assay System (Cat. number E1910, Promega) as per the manufacturer's instructions. Primers used to amplify genomic DNA and for mutagenesis are provided in supplementary table 11.

CRISPR interference lentivirus preparation and transduction

293FT cells were transfected with 3rd generation system lentiviral plasmids (pMDLg/pRRE, pRSV-Rev and pMD2.G; Addgene #12251, 12253 and 12259, respectively) to generate viral particles using the PEIpro reagent (Polyplus-Transfection). An additional plasmid consisting of KRAB-dCAS9 (Addgene 118155) was co-transfected with the previous ones. These vectors were used to perform CRISPRi Culture medium was changed the following day of transfection (Day 1) and

virus was collected 48 hours after this (Day 3). Lentiviral particles were concentrated ~100 times using Lenti-X Concentrator (Takara). LUHMES cells were seeded in a 12-well-plate format and immediately transduced with 30 ul of virus concentrate. Culture medium was changed the following day and drug selection started 3 days after transduction and continued until control un-transduced cells were completely dead. 4.5 ug/mL of blasticidin (KRAB-dCAS9) treatment was used for LUHMES selection. List of primers used for sgRNA cloning for CRISPRi and CRISPRa is provided in supplementary table 12

Genomic DNA extraction

Genomic DNA was extracted by a modified version of the salting-out method. Briefly, cells were lysed in Lysis Buffer (100 mM Tris-HCl pH 8.5; 5 mM EDTA; 200 mM NaCl; 0.2 % SDS) plus 4 U/mL of Proteinase K (Thermo Fisher) for at least 2 hours at 55ºC with agitation. Then, 0.4x volumes of 5 M NaCl were added to the mixture and centrifuged at max. speed for 10 min. DNA in the supernatant was precipitated with 1x volume of isopropanol. After centrifugation, the pellet was washed with 70% ethanol and air dried for half an hour. DNA was resuspended in water and incubated for at least one hour at 37ºC with agitation.

RNA isolation, cDNA synthesis and RT-qPCR

RNA was extracted using the RNeasy Mini Kit (QIAGEN) and cDNA was produced by employing the Transcriptor First Strand cDNA Synthesis Kit (Roche) following the manufacturer´s procedures. RT-qPCR reactions were performed with SYBR Green Master Mix (Thermo Fisher) and run on an Applied Biosystems QuantStudio 12K Flex Real-Time PCR machine. Relative gene expression values were calculated with the –ΔCt method, using TATA-Box Binding Protein (TBP) as a house-keeping gene for normalization. Primers used for qPCR are provided in supplementary table 13.

Statistical analysis

All luciferase experiments and gene quantification using qPCR were done in biological triplicates. The significance level was calculated using two-tailed t-test.

RNA-seq and data analysis

RNA-seq libraries using RNA extracted from control LUHMES cells, LUHMES cells with CSMD1 enhancer CRISPRi, control differentiated neurons and CRISPRi differentiated neurons were generated in triplicate. RNA-seq libraries were

sequenced on one lane of Illumina Hi-seq 4000 with 75bp paired end sequencing.

Sequencing data for RNA-Seq samples are adapter trimmed using Fastp and mapped against a reference transcriptome using splice aware aligner STAR v 2.6.1[61]. We generate raw counts per gene using the FeatureCounts tool (v1.6.3)[62]. The differential expression analysis was performed using DEseq2[63].

Acknowledgment

We thank the families of the affected children for their time and support for the research. This work has been funded by grants from the Wellcome Trust Institute Strategic Support and National Institute for Health Research (NIHR) Imperial Biomedical Research Centre, Institute for Translational Medicine and Therapeutics (P70888) obtained by SSA. MMP's work is funded by the Wellcome Trust Research seed award and funding from the School of Medicine and Dentistry, QMUL. We thank Mrs Sophie Shi for contributing to reagent generation.

Author contributions

Conceptualization, S.S.A.; Acquiring funding, S.S.A.; Patient recruitment and sample collection, A.K.L, W.W.L. and D.R.F.; Bioinformatics analysis -Variant calling, S.S.J. and D.P.; Bioinformatics analysis: Enhancer-promoter interactions, G.A. and S.S.A.; Bioinformatics analysis: All other: S.S.A.; Experiments – Sanger sequencing: D.M.; Experiments – Luciferase: M.G.G., T.K., M.G.D., I.C., S.Z. and W.C.; Experiments – CRISPR: M.G.D.; Writing – Original Draft: S.S.A.; Writing – Review & Editing, M.M.P., M.G.D., I.C. and M.G.G.; Supervision, D.R.F., M.M.P. and S.S.A.

Reference

1. Vissers, L. E. L. M., Gilissen, C. & Veltman, J. A. Genetic studies in intellectual disability and related disorders. *Nat. Rev. Genet.* **17,** 9–18 (2015).
2. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511,** 344–347 (2014).
3. Maulik, P. K., Mascarenhas, M. N., Mathers, C. D., Dua, T. & Saxena, S. Prevalence of intellectual disability: A meta-analysis of population-based studies. *Res. Dev. Disabil.* **32,** 419–436 (2011).
4. Kvarnung, M. & Nordgren, A. in *Advances in Experimental Medicine and Biology* **1031,** 39–54 (Springer, Cham, 2017).
5. Vissers, L. E. L. M. *et al.* A de novo paradigm for mental retardation. *Nat. Genet.* **42,** 1109–1112 (2010).
6. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12,** 1725–1735 (2003).
7. Lettice, L. A., Devenney, P., De Angelis, C. & Hill, R. E. The Conserved Sonic Hedgehog Limb Enhancer Consists of Discrete Functional Elements that Regulate Precise Spatial Expression. *Cell Rep.* **20,** 1396–1408 (2017).
8. Benko, S. *et al.* Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nat. Genet.* **41,** 359–364 (2009).
9. Smemo, S. *et al.* Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum. Mol. Genet.* **21,** 3255–3263 (2012).
10. Weedon, M. N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat. Genet.* **46,** 61–64 (2014).
11. Short, P. J. *et al.* De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555,** 611–616 (2018).
12. An, J.-Y. *et al.* Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362,** eaat6576 (2018).
13. Werling, D. M. *et al.* An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.* **50,** 727–736 (2018).
14. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
15. Reilly, S. K. *et al.* Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science (80-.).* **347,** (2015).
16. McRae, J. F. *et al.* Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542,** 433–438 (2017).
17. C Yuen, R. K. *et al.* Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neurosci.* **20,** 602–611 (2017).
18. Vissers, L. E. L. M., Gilissen, C. & Veltman, J. A. Genetic studies in intellectual disability and related disorders. *Nature Reviews Genetics* **17,** 9–18 (2016).
19. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42,** D980-5 (2014).
20. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46,** 944–950 (2014).
21. Francioli, L. C. *et al.* Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46,** 818–825 (2014).
22. Freire-Pritchett, P. *et al.* Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *Elife* **6,** e21926 (2017).
23. Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538,** 523–527 (2016).
24. 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 (2016).
25. Abrahams, B. S. *et al.* SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol. Autism* **4,** 36 (2013).
26. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
27. Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554,** 239–243 (2018).
28. di Iulio, J. *et al.* The human noncoding genome defined by genetic diversity. *Nat. Genet.* **50,** 333–337 (2018).
29. Zhao, J., Li, D., Seo, J., Allen, A. S. & Gordân, R. Quantifying the Impact of Non-coding Variants on Transcription Factor-DNA Binding. *Res. Comput. Mol. Biol. ... Annu. Int. Conf. RECOMB ... proceedings. RECOMB (Conference 2005-)* **10229,** 336–352 (2017).
30. Molenaar, J. J. *et al.* Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* **483,** 589–593 (2012).
31. Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **43,** 969–976 (2011).
32. Kim, C.-H. *et al.* Repressor activity of Headless/Tcf3 is essential for vertebrate head formation. *Nature* **407,** 913–916 (2000).
33. Ramakrishnan, A.-B. & Cadigan, K. M. Wnt target genes and where to find them. *F1000Research* **6,** (2017).
34. Gribble, S. L., Kim, H.-S., Bonner, J., Wang, X. & Dorsky, R. I. Tcf3 inhibits spinal cord neurogenesis by regulating sox4a expression. *Development* **136,** 781–789 (2009).
35. Groner, A. C. *et al.* KRAB–Zinc Finger Proteins and KAP1 Can Mediate Long-Range Transcriptional Repression through Heterochromatin Spreading. *PLoS Genet.* **6,** e1000869 (2010).
36. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154,** 442–51 (2013).
37. O'Geen, H. *et al.* dCas9-based epigenome editing suggests acquisition of histone methylation is not sufficient for target gene repression. *Nucleic Acids Res.* **45,** 9901–9916 (2017).
38. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–585 (2013).
39. Grossman, S. R. *et al.* Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc. Natl. Acad. Sci. U. S. A.* **114,** E1291–E1300 (2017).
40. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30,** 271–277 (2012).
41. Kalita, C. A. *et al.* High-throughput characterization of genetic effects on DNA-protein binding and gene transcription. *Genome Res.* **28,** 1701–1708 (2018).
42. Sullivan, P. F. & Geschwind, D. H. Defining the Genetic, Genomic, Cellular, and Diagnostic Architectures of Psychiatric Disorders. *Cell* **177,** 162–183 (2019).
43. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167,** 1369–1384.e19 (2016).
44. Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13,** 919–922 (2016).
45. Luo, Z., Rhie, S. K., Lay, F. D. & Farnham, P. J. A Prostate Cancer Risk Element Functions as a Repressive Loop that Regulates HOXA13. *Cell Rep.* **21,** 1411–1417 (2017).
46. Claussnitzer, M. *et al. FTO* Obesity Variant Circuitry and Adipocyte Browning

in Humans. *N. Engl. J. Med.* **373,** 895–907 (2015).
47. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
48. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30,** 2503–5 (2014).
49. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–303 (2010).
50. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).
51. Van der Auwera, G. A. *et al.* in *Current Protocols in Bioinformatics* 11.10.1-11.10.33 (John Wiley & Sons, Inc., 2013). doi:10.1002/0471250953.bi1110s43
52. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 (2017). doi:10.1101/201178
53. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164–e164 (2010).
54. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6,** 677–81 (2009).
55. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32,** 1220–1222 (2016).
56. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21,** 974–984 (2011).
57. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLOS Comput. Biol.* **12,** e1004873 (2016).
58. Mohiyuddin, M. *et al.* MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics* **31,** 2741–4 (2015).
59. Antaki, D., Brandler, W. M. & Sebat, J. SV2: accurate structural variation genotyping and de novo mutation detection from whole genomes. *Bioinformatics* **34,** 1774–1777 (2018).
60. Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. *motifbreakR* : an R/Bioconductor package for predicting variant effects at transcription factor binding sites: Fig. 1. *Bioinformatics* **31,** btv470 (2015).
61. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
62. Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41,** e108 (2013).
63. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).

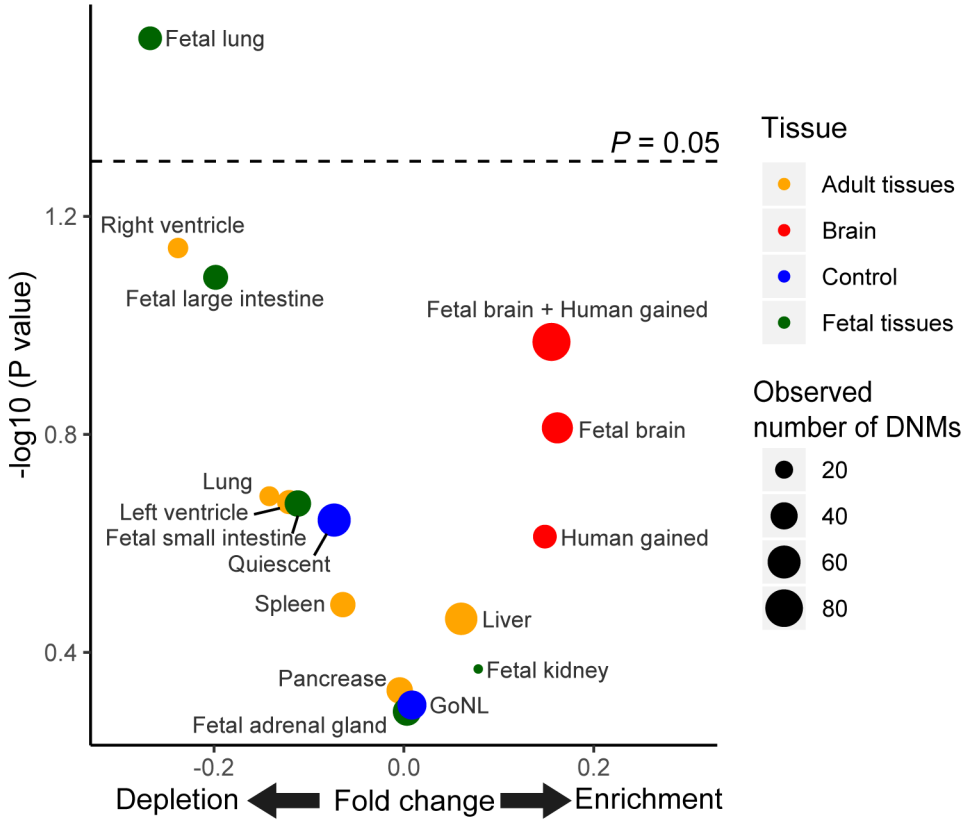


Figure 1: Enrichment of *de novo* mutations (DNMs) in enhancers. Enrichment or depletion of observed number of DNMs over expected number across multiple tissues and controls. The x-axis scale is centred around zero which indicates an equal number of observed, and expected number, of DNMs. Positive numbers of x-axis indicate enrichment and negative numbers indicate depletion of DNMs. The red dots indicate enhancers from brain tissue (fetal brain specific enhancers and human gain enhancers). Green dots indicate tissue specific enhancers from fetal tissues other than brain while orange dots indicate tissue specific enhancers from adult tissues. Two controls; sequence matched quiescent regions, and DNMs from healthy individuals from Genomes of The Netherland (GoNL) study overlapping fetal brain specific enhancers, are represented as blue dots. The size of the dot indicate observed number of DNMs. Black dotted line indicates *P*-value threshold of 0.05.

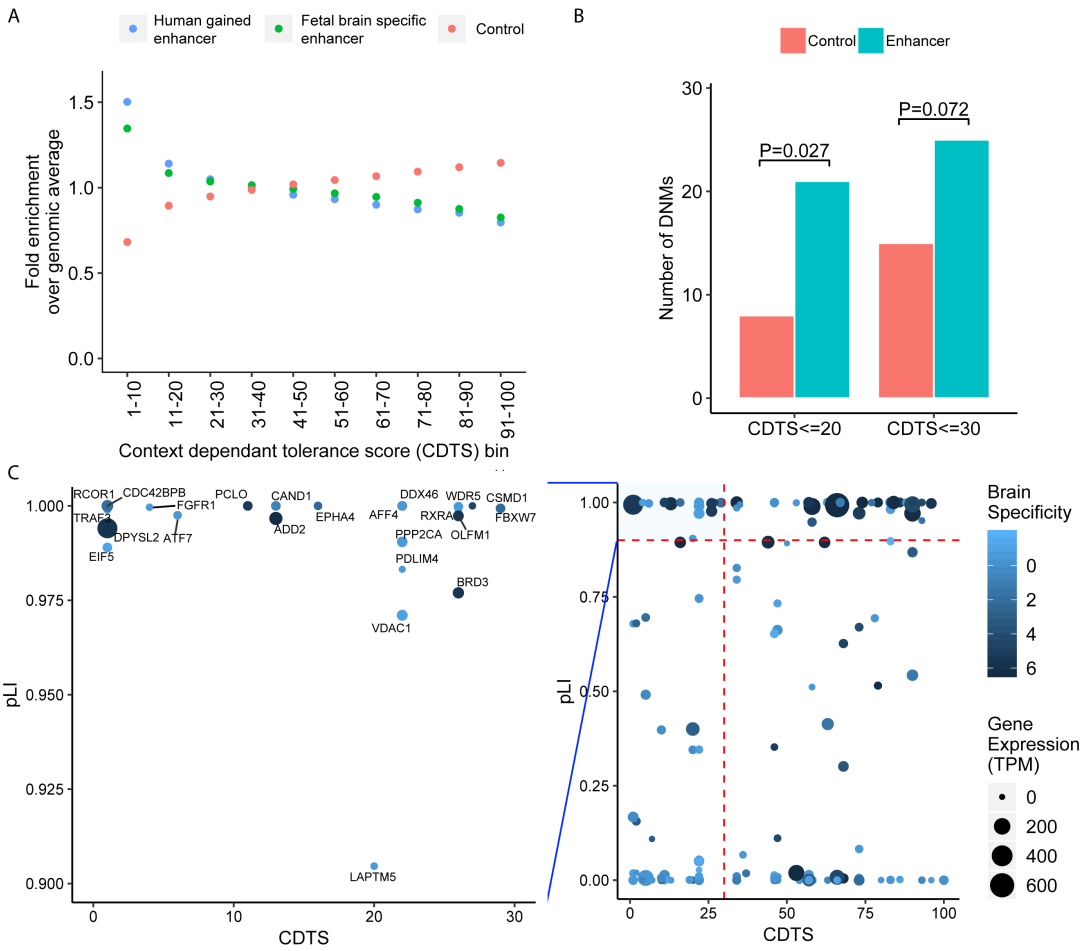


Figure 2: Population constraint at regulatory *do novo* mutations. A) Enrichment of context dependent tolerance score (CDTS) across fetal brain specific and human gain enhancers. X-axis indicates bins of CDTS score. Y-axis indicates enrichment of CDTS relative to whole genome control. The blue dots indicate human gained enhancers; green dots indicate fetal brain specific enhancers while orange dots indicate sequence matched control regions. B) Number of de novo mutations (DNMs) constrained enhancer (turquoise) and sequence matched control regions (orange) C) X-axis indicates CDTS score of the enhancer and Y-axis indicates probability of intolerance to loss of function mutation (pLI) score. Each dot indicates target genes of DNM-containing enhancer. Size of the dot indicates gene expression level in fetal brain while colour indicates fetal brain specificity of the gene. The enhancers with CDTS score less than or equal to 30 with target gene pLI score greater than equal to 0.9 are highlighted.

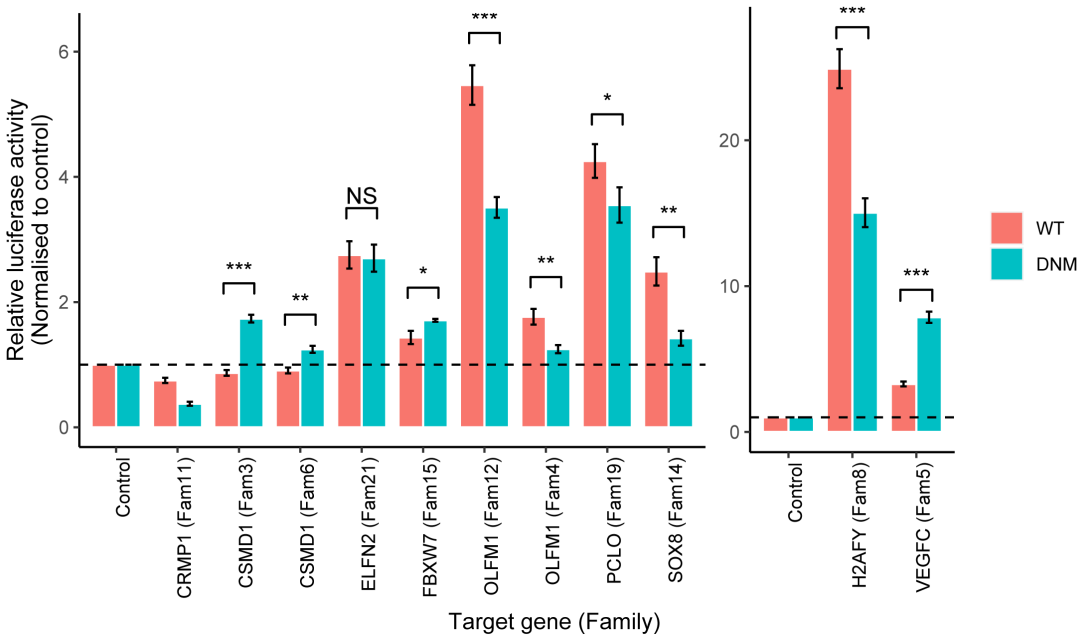


Figure 3: Effect of DNM on enhancer activity. Duel luciferase reporter assay of wild type (reference) and the mutant (DNM) allele. X-axis indicates putative target genes of the enhancer, while the family IDs are shown in brackets. Y-axis indicates relative luciferase activity normalised to empty plasmid. The error bars indicate standard error of means of three biological replicates. The significance level was calculated using two-tailed t-test. *** Indicates p-value <=0.001, ** indicate p-value between 0.01 and 0,001 while * indicates p-value between 0.01 to 0.05.

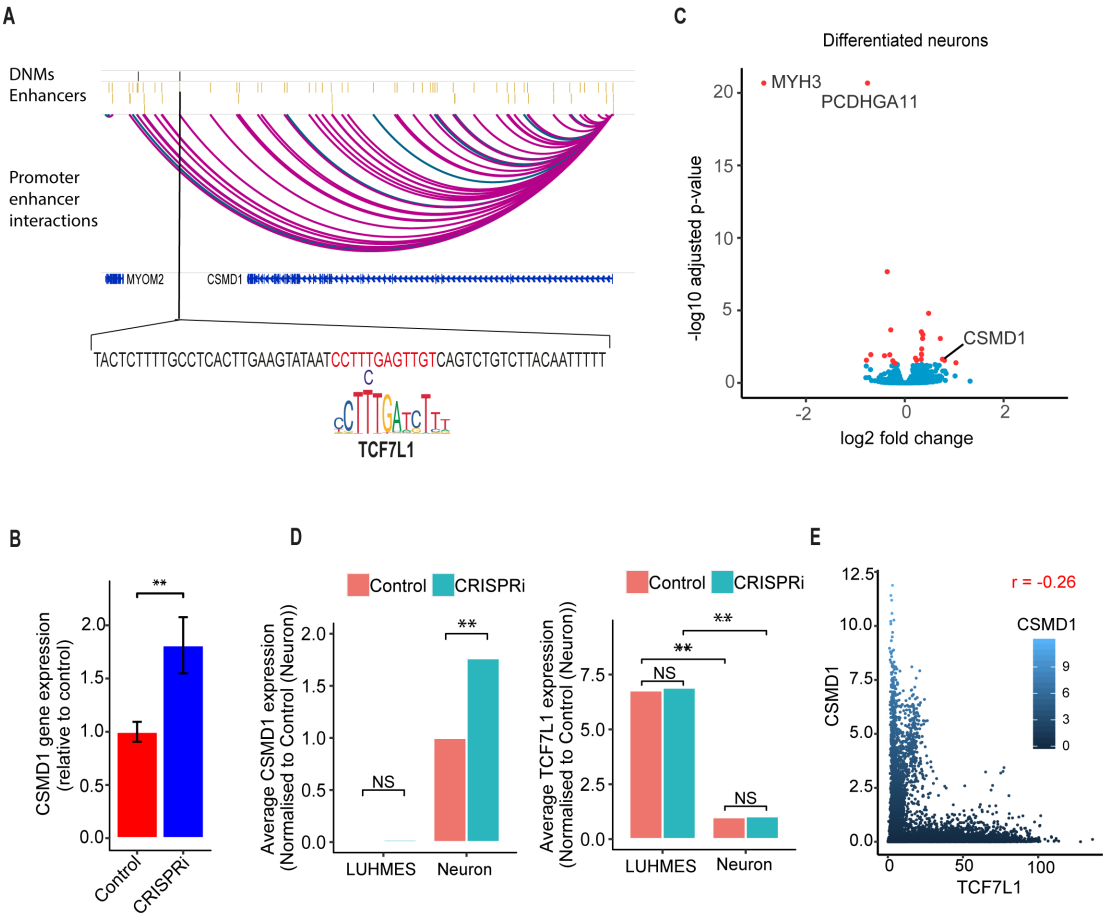
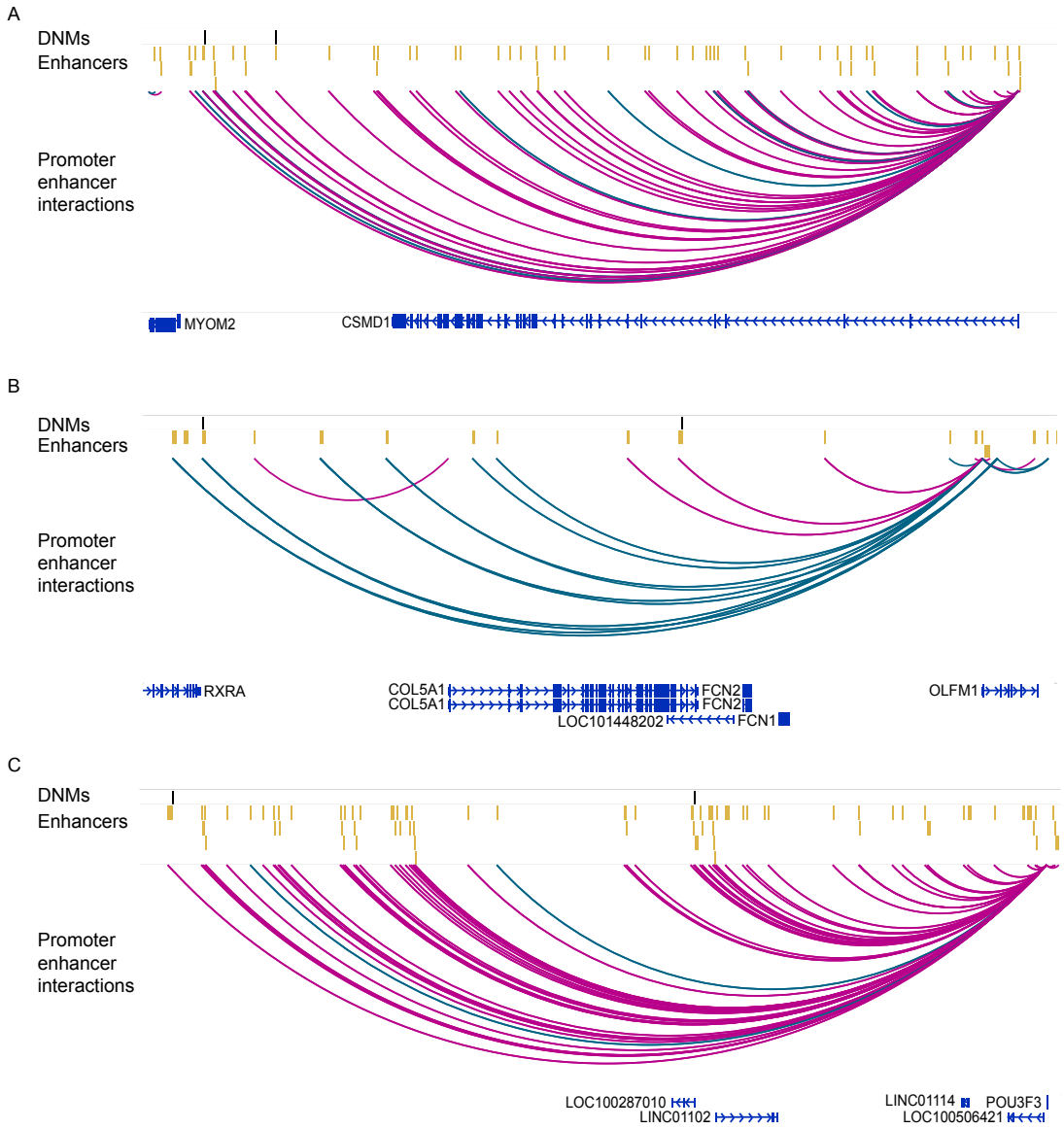


Figure 4: *Do novo* mutation in CSMD1 enhancer lead premature activation of CSMD1. A) Enhancer-promoter interactions in *CSMD1* enhancer cluster. Pink arcs represent fetal brain specific enhancer *CSMD1* promoter interactions while green arcs represent human gain enhancer-promoter interactions. *De novo* mutation in *CSMD1* enhancer (T>C) alters the core base of the *TCF7L1* transcription factor binding site. B) *CSMD1* expression measured using qPCR. CSMD1 shows significantly higher expression in cells containing KRAB-dCas9 (blue bar) as compared to control wild type cells (red bars). The error bars indicate the standard error of means of three replicates. The significance level was calculated using a two-tailed t-test. C) Volcano plot depicting differentially expressed genes between *CSMD1* enhancer CRISPRi and the control in differentiated neuron using RNA-seq. The red dots indicate significantly differentially expressed genes. D) Gene expression of *CSMD1* and *TCF7L1* genes in LUHMES cells and differentiated neurons, as determined by RNA-seq. E) Dot plot indicating gene expression pattern of *CSMD1* and *TCF7L1* across multiple tissues. Gene expression data was obtained from GTEx database (https://gtexportal.org/home/)



Supplementary figure 1: Recurrent de novo mutations (DNMs): A) Recurrent DNMs in CSMD1 enhancer cluster. B) Recurrent DNMs in OLFM1 enhancer cluster. C) Recurrent DNMs in POU3F3 enhancer cluster. Black lines indicate DNMs while yellow bars indicate enhancers. Pink arcs represent fetal brain specific enhancer-promoter interactions while green arcs represent human gain enhancer-promoter interactions.

List of supplementary tables

Supplementary table 1: Mean coverage, number of SNVs and indels in 70 whole genome sequences that include 24 intellectual disability patients

Supplementary table 2: Number of *de novo* mutations per ID proband

Supplementary table 3: *De novo* copy number variants identified in ID probands

Supplementary table 4: Potentially pathogenic variants in protein coding regions of the genome

Supplementary table 5: Enrichment of known ID genes in genes associated with DNM containing enhancers

Supplementary table 6: Gene ontology enrichment analysis of genes associated with DNM containing enhancers

Supplementary table 7: Tissues in which genes associated with DNM containing enhancers are highly expressed in Human gene atlas

Supplementary table 8: Enrichment of loss of function intolerant (pLI >=0.9) genes in genes associated with DNM containing enhancers

Supplementary table 9: Effect of regulatory DNMs on transcription factor binding site

Supplementary table 10: List of DNMs tested using dual luciferase reporter assay

Supplementary table 11: List of primers used to amplify genomic DNA and for site directed mutagenesis for dual luciferase reporter assay

Supplementary table 12: List of primers used for sgRNA cloning for CRISPRi

Supplementary table 13: List of primers used to quantify gene expression by qPCR