Non-coding region variants upstream of *MEF2C* cause severe developmental disorder through three distinct loss-of-function mechanisms

Caroline F. Wright[1], Nicholas M. Quaife[2,3], Laura Ramos-Hernández[4], Petr Danecek[5], Matteo P. Ferla[6], Kaitlin E. Samocha[5], Joanna Kaplanis[5], Eugene J. Gardner[5], Ruth Y. Eberhardt[5], Katherine R. Chao[7,8], Konrad J. Karczewski[7,8], Joannella Morales[9], Giuseppe Gallone[5†], Meena Balasubramanian[10,11], Siddharth Banka[12,13], Lianne Gompertz[12], Bronwyn Kerr[13], Amelia Kirby[14], Sally A. Lynch[15], Jenny E.V. Morton[16], Hailey Pinz[17], Francis H. Sansbury[18], Helen Stewart[19], Britton D. Zuccarelli[20], Genomics England Research Consortium, Stuart A. Cook[2], Jenny C. Taylor[6], Jane Juusola[21], Kyle Retterer[21], Helen V. Firth[5,22], Matthew E. Hurles[5], Enrique Lara-Pezzi[4,23], Paul J.R. Barton[2,3] and Nicola Whiffin[5,8,24*]

[1]Institute of Biomedical and Clinical Science, University of Exeter Medical School, Royal Devon & Exeter Hospital, Exeter, EX2 5DW, UK

[2]National Heart & Lung Institute and MRC London Institute of Medical Sciences, Imperial College London, London, W12 0NN, UK

[3]Cardiovascular Research Centre, Royal Brompton & Harefield Hospitals NHS Trust, London, SW3 6NP, UK

[4]Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), 28029 Madrid, Spain

[5]Human Genetics Programme, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1RQ, UK

[6]National Institute for Health Research Oxford Biomedical Research Centre, Wellcome Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK

[7]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA

[8]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

[9]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, CB10 1SD, UK

[10]Sheffield Clinical Genetics Service, Sheffield Children's NHS Foundation Trust, Sheffield, S10 2TH, UK

[11]Academic Unit of Child Health, Department of Oncology & Metabolism, University of Sheffield, Sheffield, S10 2TH, UK

[12]Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester University Hospitals NHS Foundation Trust, Health Innovation Manchester, Manchester, M13 9WL, UK

[13]Division of Evolution and Genomic Sciences, School of Biological Sciences, University of Manchester, Oxford Road, Manchester, M13 9PL, UK

[14]Department of Pediatrics, Wake Forest School of Medicine, Winston-Salem, NC 27101, USA

[15]UCD Academic Centre on Rare Diseases, School of Medicine and Medical Sciences, University College Dublin, and Clinical Genetics, Temple Street Children's University Hospital, Dublin, D01 XD99, Ireland

[16]West Midlands Regional Clinical Genetics Service and Birmingham Health Partners, Birmingham Women's and Children's Hospitals NHS Foundation Trust, Birmingham, B4 6NH, UK

[17]Department of Pediatrics, Saint Louis University School of Medicine, Saint Louis, MO 63104, USA

[18]All Wales Medical Genomics Service, NHS Wales Cardiff and Vale University Health Board, Institute of Medical Genetics, University Hospital of Wales, Cardiff, CF14 4AY, UK

[19]Oxford Centre for Genomic Medicine, Oxford University Hospitals NHS Foundation Trust, Oxford, OX3 7LE, UK

[20]Department of Neurology, University of Kansas School of Medicine-Salina Campus, Salina, KS 67401, USA

[21]GeneDx, Gaithersburg, MD 20877, USA

[22]East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge, CB2 0QQ, UK

[23]CIBER de enfermedades CardioVasculares (CIBERCV), 28029 Madrid, Spain

[24]Wellcome Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK

†Current address: Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, 14195 Berlin, Germany

*Correspondence should be addressed to nwhiffin@well.ox.ac.uk

Abstract

Clinical genetic testing of protein-coding regions identifies a likely causative variant in only around half of developmental disorder (DD) cases. The contribution of regulatory variation in non-coding regions to rare disease, including DD, remains very poorly understood. We screened 9,858 probands from the Deciphering Developmental Disorders (DDD) study for *de novo* mutations in the 5'untranslated regions (5'UTRs) of dominant haploinsufficient DD genes. We identified four single nucleotide variants and two copy number variants upstream of *MEF2C* in a total of 10 individual probands. We developed multiple bespoke and orthogonal experimental approaches to demonstrate that these variants cause DD through three distinct loss-of-function mechanisms, disrupting transcription, translation, and/or protein function. These non-coding region variants represent 23% of likely diagnoses identified in *MEF2C* in the DDD cohort, but these would all be missed in standard clinical genetics approaches. Nonetheless, these variants are readily detectable in exome sequence data, with 30.7% of 5'UTR bases across all genes well covered in the DDD dataset. Our analyses show that non-coding variants upstream of known disease-causing genes are an important cause of severe disease and demonstrate that analysing 5'UTRs can increase diagnostic yield. We also show how non-coding variants can help inform both the disease-causing mechanism underlying protein-coding variants, and dosage tolerance of the gene.

Introduction

The importance of non-coding regulatory variation in common diseases and traits has long been appreciated, however, the contribution of non-coding variation to rare disease remains poorly understood[1–4]. Consequently, current clinical testing approaches for rare disease focus almost exclusively on regions of the genome that code directly for protein, within which we are able to relatively accurately estimate the effect of any individual variant. Using this approach, however, disease-causing variants are only identified in around 36% of individuals

with developmental disorders (DD)[5] using exome sequencing, with a further 15-20% diagnosed through chromosomal microarrays[6]. In previous work, we assessed the role of *de novo* mutations (DNMs) in distal regulatory elements, and estimated that 1-3% of undiagnosed DD cases carry pathogenic DNMs in these regions[1].

Untranslated regions (UTRs) at the 5' and 3' end of genes present a unique opportunity to expand genetic testing outside of protein coding regions given they have important regulatory roles in controlling both the amount and location of mRNA in the cell, and the rate at which it is translated into protein[7,8]. Crucially, we also know the genes/proteins that these regions regulate. Given that UTRs account for around the same genomic footprint as protein-coding exons, they have substantial potential to harbour novel Mendelian diagnoses[9,10]. UTRs are, however, not regularly included in exome sequence capture regions, and are excluded in most analysis pipelines. This is primarily due to a lack of guidance on how to determine when UTR variants are likely to be pathogenic.

Recently, we demonstrated that variants creating upstream start codons (uAUGs) in 5'UTRs are under strong negative selection, and are an important cause of Mendelian diseases, including neurofibromatosis and Van der Woude syndrome[11,12]. Initiation of translation at a newly created uAUG can decrease translation of the downstream coding sequence (CDS). The strength of negative selection acting on uAUG-creating variants varies depending on both the match of the sequence surrounding the uAUG to the Kozak consensus, which is known to regulate the likelihood that translation is initiated[13,14], and the nature of the upstream open reading frame (uORF) that is created. Variants that result in ORFs which overlap the CDS have a larger impact on CDS translation and hence are more deleterious[11,15].

Here, we screened 9,858 probands from the Deciphering Developmental Disorders (DDD)[5] study for DNMs in the 5'UTRs of known dominant DD genes. We uncover novel likely

disease-causing variants that are entirely non-coding and show how these variants cause disease through three distinct loss-of-function mechanisms. We further show how disease-causing missense variants in MEF2C [MIM:600662] are clustered at the N-terminus and likely also cause loss-of-function by disrupting binding of MEF2C protein to DNA. Finally, we analyse the coverage across all UTRs in the DDD exome sequencing dataset to demonstrate how these regions can be readily screened in existing datasets to increase diagnostic yield and glean insight into disease causing mechanisms.

Materials and Methods

Recruitment, sample collection and clinical data

The DDD Study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South REC, and GEN/284/12 granted by the Republic of Ireland REC). Individuals with severe, undiagnosed developmental disorders and their parents were recruited and systematically phenotyped by the 24 Regional Genetics Services within the United Kingdom (UK) National Health Service and the Republic of Ireland. Saliva samples were collected from probands and parents, and DNA extracted as previously described[16]; blood-extracted DNA was also collected for probands where available. Clinical data (growth measurements, family history, developmental milestones, etc.) were collected using a standard restricted-term questionnaire within DECIPHER[17]. Informed consent was obtained for all participants.

Genetic data

Array-CGH analysis was performed using 2 x 1M probe custom designed microarrays (Agilent; Amadid No.s 031220/031221) as described previously[16]. Exome sequencing was performed using Illumina HiSeq (75-base paired-end sequencing) with SureSelect baits (Agilent Human All-Exon V3 Plus and V5 Plus with custom ELID C0338371) and variants

were called and annotated as described previously[16]. We used DeNovoGear[18] (version 0.54) to detect likely DNMs from trio exome BAM files and Ensembl Variant Effect Predictor[19] was used to annotate predicted consequences. The data are available under managed access from the European Genome-phenome Archive (Study ID EGAS00001000775), and likely diagnostic variants are available open access in DECIPHER.

Defining a gene-set of interest

We limited our analysis to 359 DDG2P[20] genes with a confirmed or probable role in developmental disorders and with a dominant (including X-linked dominant) loss-of-function disease mechanism (downloaded from https://www.ebi.ac.uk/gene2phenotype/downloads on 21st July 2020; Table S1).

Identifying uAUG-creating variants in DDD

We defined high-confidence DNMs in DD as previously[21], using the following criteria: minor allele frequency < 0.01 in our cohort and reference databases, depth in the child > 7, depth in both parents > 5, Fisher strand bias p-value > 10^{-3}, and a posterior probability of being a DNM from DeNovoGear > 0.00781[18]. Additionally, we filtered out DNMs with some evidence of an alternative allele in one of the parents and indels with a low variant allele fraction (<30% of the reads support the alternative) that had a minor allele frequency > 0. We cross-referenced this list of high-confidence DNMs with a list of all possible uAUG-creating SNVs from previous work[11]. We also assessed any small insertions and deletions that could form uAUGs.

The strength of the Kozak consensus surrounding each uAUG was assessed as described previously[11]. Specifically, we assessed the positions at −3 and +3 relative to the A of the

AUG, requiring both the −3 base to be either A or G and the +3 to be G for an annotation of 'Strong'. if only one of these conditions was true, the strength was deemed to be 'Moderate' and if neither was the case 'Weak'.

Defining the 5'UTR of MEF2C

We used the MANE Select transcript ENST00000504921.7 for which the 5'UTR was defined using CAGE data from the FANTOM5 project[22] , RNA-seq supported intron data from the Intropolis resource[23], and exon level expression from the GTEx project[24] . The Matched Annotation from the NCBI and EMBL-EBI (MANE) is a collaborative project that aims to define a representative transcript (MANE Select) for each protein-coding locus across the genome. The MANE set perfectly aligns to the GRCh38 reference assembly and includes pairs of 100% identical RefSeq and Ensembl/GENCODE transcripts (https://www.ncbi.nlm.nih.gov/refseq/MANE/[25]). The 5'UTR of *MEF2C* was therefore defined as two exons: chr5:88178772-88179001 and chr5:88119606-88119747 on GRCh37, or chr5:88882955-88883184 and chr5:88823789-88823930 on GRCh38.

Searching for MEF2C 5'UTR variants in external datasets

We queried the regions corresponding to the *MEF2C* 5'UTR for DNMs in (1) a set of 18,789 DD trios sequenced by the genetic testing company GeneDx[5], (2) 13,949 rare disease trios from the main programme v9 release of the Genomics England 100,000 Genomes Project[26] (https://cnfl.extge.co.uk/display/GERE/De+novo+variant+research+dataset), and (3) variants in the v3.0 dataset of the Genome Aggregation Database (gnomAD)[27].

Assessing 5'UTR coverage

Regions corresponding to 5'UTRs were extracted from the .gff file from the MANE project v0.91 (ftp://ftp.ncbi.nlm.nih.gov/refseq/MANE/MANE_human/release_0.91/ ; MANE Select transcripts). For each base, we calculated the mean coverage across 1,000 randomly selected samples from DDD. A mean coverage of >10x was used to call a base 'covered'. Analysis was limited to genes with a defined MANE Select transcript. For our DD haploinsufficient genes this was 345/359 genes (96.1%).

To identify all possible uAUG-creating variants in DD haploinsufficient genes, we extracted the 5'UTR sequence from the MANE rna.fna file and used the UTRannotator[28] to find all possible uAUG-creating sites and annotate their consequence.

Functional validation of variants creating out-of-frame ORFs (oORFs): by MEF2C 5'UTR-luciferase translation assay

Expression constructs: WT and variant MEF2C 5'UTRs were cloned directly upstream of Gaussia luciferase (GLuc) in the pEZX-GA02 backbone (Labomics) and sequenced to confirm integrity. Secreted alkaline phosphatase (SEAP) was expressed on the same construct for normalisation of transfection efficiency.

Cull culture, transfection and analysis: HEK293T cells were purchased from ATCC and cultured in Dulbecco's Modified Eagle Medium (glutamine+, pyruvate+) supplemented with 10% foetal bovine serum and 1% penicillin/streptomycin. Cells were transfected with MEF2C 5'UTR-luciferase constructs using Lipofectamine 3000, following manufacturer's protocols. After 24h, culture medium was sampled and GLuc and SEAP were simultaneously quantified using the Secrete-Pair Dual Luminescence assay (Genecopoeia). Fifteen technical replicates were performed across three independent experiments.

qPCR: RNA was purified from cells using phenol-chloroform extraction and the Qiagen RNeasy Miniprep kit. RNA quantity was normalised and cDNA generated using IV VILO reverse transcriptase following manufacturer's protocols. Quantitative PCR was performed using SYBR green master mix on a Quantstudio 7 Real-time PCR system and results normalised to co-amplified GAPDH. The following primers were used: GLUC F: 5' CTGTCTGATCTGCCTGTCCC 3', GLUC R: 5' GGACTCTTTGTCGCCTTCGT 3', SEAP F: 5' ACCTTCATAGCGCACGTCAT 3' and SEAP R: 5' TCTAGAGTAACCCGGGTGCG 3', GAPDH F: 5' GGAGTCAACGGATTTGGTCG 3', GAPDH R: ATCGCCCCACTTGATTTTGG 3'.

Kozak mutagenesis: The kozak context of the c.-103G>A MEF2C 5'UTR-luciferase construct was modified using the Quikchange II mutagenesis kit, following manufacturers protocols. The following PAGE-purified mutagenesis primers were used: F: 5'CTCCTTCTTCAGCATTTTCACAGCTCAGTTCCCAA 3', R: 5' TTGGGAACTGAGCTGTGAAAATGCTGAAGAAGGAG 3'. Constructs were fully sequenced to verify mutation and construct integrity in each case

Functional validation of CDS-elongating variants: by MEF2 binding site-luciferase transactivation assay

Expression and reporter constructs: WT and variant *MEF2C* 5' UTR+CDS oligos were cloned into the pReceiver-M02 expression construct (Labomics) and sequenced to confirm integrity. For normalisation of transfection efficiency, cells were co-transfected with pRL-Renilla. A desMEF2-luciferase reporter construct was used to quantify the transactivational efficiency of each MEF2C expression construct, and consisted of three copies of a high-affinity MEF2 binding site[29], linked to an hsp68 minimal promoter in pGL3 (Promega)[30].

Cell culture and transfection: HL1 cardiomyocytes were cultured in Claycomb medium, supplemented with 2 mM L-glutamine, 10% FBS and 100 g/ml Penicillin/Streptomycin. Culture surfaces were pre-treated with gelatin/fibronectin. Cells were co-transfected with 1) desMEF2-luciferase reporter construct, 2) pRL-Renilla transfection control, and 3) expression construct of either: i) empty pcDNA3.1 (negative control), ii) WT MEF2C 5' UTR+CDS, iii) MEF2C -26C>T, or iv) MEF2C -8C>T. Transfection was with Lipofectamine 2000, following manufacturers protocols. 48h after transfection, firefly and Renilla Luciferases were quantified by the Promega Dual-Luciferase Reporter Assay System. Eighteen technical replicates were performed across three independent experiments.

Western blot: HL1 cells were lysed in RIPA buffer in the presence of protease and phosphatase inhibitors (04693159001 and 04906845001, Roche Diagnostics). Lysates were separated on SDS-PAGE gels and transferred to PVDF membranes, which were blocked with 3% skimmed milk in TBS. The primary antibody was anti-MEF2C (ab211493, Abcam), and the secondary antibody was anti-mouse P0447 from Dako. The membrane was developed using ECL reagent (AC2204, Azure Biosystems) and intensity of the bands quantified using ImageJ software.

Statistical analysis for all assays: Data were analysed for statistical significance using 1-way ANOVA followed by Tukey's post-test, using GraphPad Prism 8.0.

CNV calling

Four CNV detection algorithms (XHMM[31], CONVEX[16], CLAMMS[32] and CANOES[33]) were used to ascertain CNVs from exome data, followed by a random forest machine learning approach to integrate and filter the results (manuscript in preparation).

Layered H3K4me3 data (to visualise active promoter regions) was downloaded from the UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables) for GN12878 as a representative cell line and plotted alongside the identified CNVs in Figure S1.

Modelling missense disruption to DNA-binding

We collated a set of missense variants identified in *MEF2C* in DD cases comprising all *de novo* variants from trios in DDD and GeneDx published previously[5], and variants from ClinVar either flagged as being identified as *de novo*, or with functional evidence (Table S3).

As a comparator, we used missense variants from gnomAD v2.1.1[27]. Given that there are only three variants in the N-terminal region of *MEF2C* in gnomAD, but the sequence of the N-terminal region is near identical across the four MEF2 proteins (Figure S4), we used missense variants from all four genes (*MEF2A-D*; Table S4).

Based on structures of the N-terminal MADS-box of *MEF2A* homodimer(1egw, 3kov and 6byy, residues 1-92) bound to its DNA consensus sequence[34], we categorised residues into one of four categories: (1) in N-terminal random coil and in contact with the DNA (2) in N-terminal alpha-helix pointing towards the DNA; (3) in N-terminal alpha-helix pointing away from the DNA; or (4) distal to the DNA contact surface (Table S5). We used a two-sided Fisher's exact test to assess for an enrichment of variants in contact or pointing towards the DNA helix in DD cases (Table S6).

The Swissmodel threaded model of MEF2C based upon PDB:6BYY (89% identity)[35,36] was energy minimised using Pyrosetta[37] with 15 FastRelax cycles[38] against the electron density of PDB:6BYY and 5 unconstrained. The DNA was extended on both ends due to the proximity of R15. Mutations were introduced and the 10 Å neighbourhood was energy minimised. Gibbs free energy was calculated using the Rosetta ref2015 scorefunction[39].

Gibbs free energy of binding was calculated by pulling away the DNA and repacking sidechains and, in the case of residues in the N-terminal loop, thoroughly energy minimising the backbone of the loop as this is highly flexible when unbound. N-terminal extensions were made using the RemodelMover[40] with residues 2-5 also remodelled as determined by preliminary test. Closest distance of each residue to the DNA was calculated with the Python PyMOL module. Code used for this analysis can be found at: https://github.com/matteoferla/MEF2C_analysis. This interactive page was made in MichelaNGLo[41].

All missense variants are annotated with respect to the Ensembl canonical transcript ENST00000340208.5.

Calculating regional missense constraint and de novo enrichment

We determined regional missense constraint by (1) extracting observed variant counts from the 125,748 samples in gnomAD v2.1.1, (2) calculating the expected variant count per transcript, and (3) applying a likelihood ratio test to search for significant breaks that split a transcript into two or more sections of variable missense constraint.

Observed missense variants were extracted from the gnomAD exomes Hail Table (version 2.1.1) as described previously[27], using the following criteria:

- Annotated as a missense change in a canonical transcript of a protein-coding gene in Gencode v19 by Variant Effect Predictor (VEP, version 85)
- Median coverage greater than zero in the gnomAD exomes data
- Passed variant filters
- Adjusted allele count of at least one and an allele frequency less than 0.1% in the gnomAD exomes

To calculate the expected variant count, we extended methods described previously[27] to compute the proportion of expected missense variation per base. Briefly, we annotated each possible substitution with local sequence context, methylation level (for CpGs), and associated mutation rate from the table computed in Karczewski *et al*.[27] We aggregated these mutation rates across the transcript and calibrated models based on CpG status and median coverage. To determine the expected variants for a given section of the transcript, we calculated the fraction of the overall the mutation rate represented by the section and multiplied it by the aggregated expected variant count for the full transcript.

We defined missense constraint by extending the methods from Samocha *et al*.[42] We employed a likelihood ratio test to compare the null model (transcript has no regional variability in missense constraint) with the alternative model (transcript has evidence of regional variability in missense constraint). We required a χ^2 value above a threshold of 10.8 to determine significance for each breakpoint, and in the case of multiple breakpoints, retained the breakpoint with the maximum χ^2. This approach defined a single breakpoint in the *MEF2C* canonical transcript at chr5:88057138 (GRCh37).

To evaluate the enrichment of DNMs in the transcript when removing the N-terminal section, we determined the probability of a missense mutation in that region and then compared the observed number of DNMs (n=3) with the expected count in 28,641 individuals using a Poisson test. Specifically, we took the probability of a missense mutation (mu_mis) as provided in the gnomAD v2.1 constraint files for *MEF2C* and adjusted it for the fraction of mutability represented in the latter section of the gene (~79.5%).

Results

Identifying de novo 5'UTR variants in DD cases

To investigate the contribution of uAUG-creating variants to severe DD cases, we analysed 29,523 high-confidence DNMs identified in exome sequencing data from 9,858 parent-offspring trios in the DDD study[5]. Although the majority of DNMs identified are coding, as expected with exome sequencing data, many non-coding variants are also detectable, particularly near exon boundaries. Given that uAUG-creating variants that decrease CDS translation would only be expected to be deleterious in genes that are dosage sensitive, we restricted our analysis to the 5'UTRs of 359 known haploinsufficient developmental disorder genes from the curated DDG2P database[20] (Table S1).

We identified five unique uAUG-creating *de novo* single nucleotide variants (SNVs) in five unrelated probands upstream of two different genes. All of these variants are absent from the Genome Aggregation Database (gnomAD) population reference dataset (both v2.1.1 and v3.0)[27]. Notably, four of the five variants were found in the 5'UTR of *MEF2C* in probands with phenotypes consistent with *MEF2C* haploinsufficiency (Table 1; [MIM: 613443])[43]. Two of these DNMs create uAUGs out-of-frame with the *MEF2C* CDS, which are expected to reduce downstream protein translation, whilst the other two create uAUGs in-frame with the CDS, which are expected to elongate the protein (Figure 1). The fifth variant was located in a strong Kozak consensus upstream of *STXBP1* (ENST00000373302.8:c.-26C>G), creating an uAUG out-of-frame with the *STXBP1* CDS; the phenotype of the proband with this variant is consistent with *STXBP1* haploinsufficiency[44], including global developmental delay, microcephaly, and delayed speech and language development.

Given the identification of multiple uAUG-creating *de novo* SNVs in *MEF2C* in the DDD study, we subsequently queried high-confidence DNMs identified in 18,789 trios with DD that were exome sequenced by GeneDx[5] for additional *MEF2C* DNMs. We uncovered three additional *de novo* occurrences of two of the uAUG-creating variants observed in the DDD study. In addition, we identified a further *de novo* occurrence of one of these variants in a DD proband in the UK 100,000 Genomes Project[26](Table 1).

In a separate analysis, we analysed copy number variants (CNVs) identified in the DDD study using exome sequencing data and identified five de novo CNVs overlapping *MEF2C* (Figure S1). Two of these CNVs (each found in a single additional proband) overlap the 5'UTR of *MEF2C* without impacting any of the coding exons (Table 1). These two non-coding CNVs delete the first exon of the *MEF2C* 5'UTR and >40kb of immediately upstream sequence (294kb and 97kb, respectively), removing the entire promoter (as defined by the Ensembl regulatory build[45] and H3K4me3 peaks from ENCODE[46]) and likely abolishing transcription of this allele (Figure S1). There are no large deletions (>600bps) in this upstream region in the gnomAD structural variant dataset (v2.1)[47]. Both coding *MEF2C* disruptions and non-coding deletions further upstream of *MEF2C* that are predicted to disrupt enhancer function have been identified in DD patients previously[48,49].

De novo 5'UTR variants cause phenotypes consistent with MEF2C haploinsufficiency

We collated all available clinical data for the ten probands with *MEF2C* 5'UTR *de novo* variants and in each case the observed phenotype is consistent with previously reported *MEF2C* haploinsufficiency[50,51] (Table S2). Specifically, of the nine individuals for which detailed phenotypic information was available, the following features were noted: global developmental delay (9/9) with delayed or absent speech (9/9), seizures (8/9), hypotonia (5/9) and stereotypies (2/9). These probands had no other likely disease-causing variants in the coding sequence of *MEF2C*, or in any other known DD genes following exome sequencing.

uAUG-creating SNVs cause loss-of-function by reducing translation or disrupting protein function

The four uAUG-creating SNVs identified in *MEF2C* result in two different downstream effects. We used two distinct experimental approaches to evaluate the impact of i) out-of-frame uAUG-creating variants on downstream translation and ii) CDS-elongating variants on MEF2C-dependent transactivation.

Two of the variants (c.-66A>T and c.-103G>A), each found in a single proband, create uAUGs that are out-of-frame with the coding sequence (CDS), creating an overlapping ORF (oORF) that terminates 128 bases after the canonical start site (Figure 1b). Using a translation assay, with wild-type or mutant 5'UTR sequence cloned upstream of a luciferase reporter gene, we show that both variants result in a significant decrease in translational efficiency (Figure 2a; Figure S2a). The amount by which translation is reduced appears to be dependent on the uAUG match to the Kozak consensus sequence, consistent with previous observations[11]. The c.-103G>A variant, which creates an uAUG with a weak Kozak consensus, results in only a moderate decrease in luciferase expression, and the proband with this variant displays a milder phenotype on clinical review. To validate that this difference in effect is indeed due to the differing Kozak strengths, in the c.-103G>A translation assay, we mutated a single base to alter the oORF start context to a moderate Kozak consensus match (see methods). This modification resulted in significantly decreased translational efficiency compared to the unmodified c.-103G>A variant, to a level equivalent to the c.-66A>T variant (Figure S3). The individual carrying the c.-103G>A does not have any other 5'UTR variants that could similarly modify the variant's effect. These data suggest that MEF2C is sensitive to even partial loss-of-function.

The other two variants (c.-8C>T and c.-26C>T) are both observed recurrently *de novo*, each in three unrelated probands (Table 1). Both variants create uAUGs that are in-frame with the CDS, resulting in N-terminal extensions of three and nine amino acids respectively (Figure 1c). MEF2C is a transcription factor, and critical to its function is the DNA-binding domain located at the extreme N-terminal region[52]. Although no structure is available for the MEF2C

protein, numerous crystal and NMR structures of the N-terminal DNA-binding domain of human MEF2A are available, which is 96% identical in sequence to MEF2C. These structures show clearly that the extreme N-terminus of the protein is in direct contact with DNA[34,53], and that the first few residues bind directly into the minor groove (Figure 3). We assayed MEF2C-dependent transactivation using MEF2C expression constructs with wild-type and mutant 5'UTR sequences. These data demonstrate significantly reduced activation of target gene transcription from the variants (Figure 2b; Figure S2b and c), compared to wild-type MEF2C. Once again, the strength of the effect is dependent on the uAUG context, with the c.-8C>T variant that creates a strong Kozak consensus having a larger effect, almost abolishing transactivation activity.

We looked in the gnomAD dataset[27] for uAUG-creating variants that might have similar impacts. Across the exome (v2.1.1) and genome (v3.0) sequencing datasets, there are only two uAUG-creating variants in the *MEF2C* 5'UTR. Crucially, neither of these fall into the proximal 5'UTR exon and neither create ORFs overlapping the CDS. In both instances, the uAUGs are created into weak Kozak-consensus contexts, and they have in-frame stop codons after 6bps (allele count = 6) and 57bps (allele count = 1) respectively (Table 1; Figure 1d). These variants would therefore not be expected to have substantial, if any, effect on *MEF2C* translation.

Pathogenic de novo missense variants likely cause loss-of-function of MEF2C through disrupting DNA-binding

Whilst the major recognised mechanism through which pathogenic variants in *MEF2C* lead to severe developmental phenotypes is loss-of-function, *de novo* missense variants are also significantly enriched in DD trios (P=1.3x10^{-14})[5] and multiple pathogenic missense variants are reported in ClinVar[54]. These variants are almost exclusively found at the extreme N-terminus of the protein (Table S3), in the DNA-binding region, which is also highly

constrained for missense variants in gnomAD (obs/exp=0.069; calculated on 125,748 exome sequenced samples in v2.1.1; Figure 3a). We hypothesised that these pathogenic missense variants are also causing loss-of-function by disrupting DNA-binding of MEF2C as has been demonstrated for random disruptions to the N-terminal region[52] and two patient variants[49] previously. Using the structure of the N-terminal MEF2A homodimer bound to DNA, we modelled the location of pathogenic missense variants in MEF2C, as well as missense variants in gnomAD v2.1.1 across all members of the myocyte enhancer factor 2 protein family (MEF2A-D; 84% N-terminal domain sequence identity; Table S4; Figure S4), and saw a significant enrichment of pathogenic variants interacting directly with DNA via both the N-terminal loop and DNA-binding helix (Fisher's P=2.6x10^{-5}, Figure 3b; Tables S5 & S6). We further calculated the change in Gibbs free energy (ΔΔG) of both the protein-DNA interaction and the complex stability for each missense change. Variants found in DD cases have significantly increased ΔΔG scores compared to gnomAD variants (Wilcoxon P=2.7x10^{-4}; Figure 3c) and are significantly closer to the bound DNA (Wilcoxon P=1.5x10^{-5}; Figure 3d; Table S7). Together, these data suggest that disease-causing missense variants in *MEF2C* act through a loss-of-function mechanism, as has been experimentally demonstrated for two patient variants previously[49]. Indeed, excluding the N-terminal DNA-binding domain, the remainder of *MEF2C* shows much weaker constraint against missense variants in gnomAD (obs/exp=0.41), and only nominal enrichment for *de novo* missense variants in DD cases (P=0.041).

Disease-causing 5'UTR variants can be detected in exome sequencing data

Given our ability to identify 5'UTR variants in *MEF2C*, we investigated the extent to which these regions are captured across all genes in the exome sequencing dataset from the DDD study. We find that 30.7% of all gene 5'UTR bases and 20.4% of 5'UTR bases of our DD haploinsufficient genes (average of 73 bps per gene; n=345 with MANEv0.91 transcripts) are covered at a mean coverage threshold of >10x. The average length of 5'UTRs in DD

haploinsufficient genes is 356 bps (Figure 4a), with 42.0% containing multiple exons (Figure 4b). As expected, 5'UTR coverage decays as distance from the CDS increases (Figure 4c), with distal exons very poorly covered (6.7% of bases >10x). In comparison, a much lower proportion of 3'UTR bases (6.0%) are covered at >10x, which is unsurprising given that 3'UTRs are much longer than 5'UTRs, at an average of 2,652 bps for our DD haploinsufficient genes.

To determine the proportion of all possible uAUG-creating variants that are sufficiently covered in the DDD exome sequence data, we computationally identified 3,962 possible uAUG-creating variants in DD haploinsufficient genes that would create out-of-frame overlapping ORFs (n=2,782) or CDS-elongations (n=1,180). Of these, 42.4% are sequenced at >10x coverage across the DDD study dataset (40.2% of out-of-frame and 47.6% of CDS-elongating). However, we would not expect CDS-elongating variants to cause a loss-of-function for the majority of genes. Rather, we expect this to be limited to genes with important functional domains at the extreme N-terminus that would be adversely affected by the addition of extra N-terminal amino acids, either through disrupting binding or altering protein structure. Based on Pfam domain predictions, only three of the proteins encoded by our 359 DD haploinsufficient genes, including MEF2C, have DNA-binding domains that start within 10 bps of the N-terminus (Figure 4d); the other two (*ZNF750* and *SIM1*) encode an N-terminal zinc-finger and basic helix-loop-helix, respectively, and although no structures are available, these bind DNA via specific motifs that are unlikely to include the extreme N-terminal residues.

Discussion

Here, we have identified six unique non-coding, pathogenic DNMs in *MEF2C* in ten individuals with severe developmental disorders (six in the DDD study, three in a cohort from GeneDx, and one in the UK 100,000 Genomes Project). These variants act via three distinct

loss-of-function mechanisms at different stages of expression regulation: (1) two large deletions remove the promoter and part of the 5'UTR and are predicted to abolish normal transcription of *MEF2C*; (2) two SNVs create out-of-frame uAUGs and reduce normal translation of the *MEF2C* coding sequence; and (3) two SNVs create in-frame uAUGs that elongate the *MEF2C* coding sequence, disrupting binding of the MEF2C protein to DNA and reducing subsequent transactivation of gene-expression. We also identified a single uAUG-creating variant in *STXBP1* in a proband whose phenotype was consistent with STXBP1 haploinsufficiency. This variant is predicted to create an out-of-frame oORF into a strong Kozak consensus, thus decreasing normal *STXBP1* translation (as ribosomes first encounter, and begin to translate from this new uAUG), leading to reduced levels of STXBP1 protein.

These observations demonstrate the importance of screening 5'UTRs of known disease genes in individuals that remain genetically undiagnosed. We have previously identified 20 probands with diagnostic DNMs (15 SNVs and 5 CNVs) impacting *MEF2C* protein-coding regions in the 9,858 family trios analysed in the DDD study. The six additional non-coding DNMs described here (4 SNVs and 2 CNVs) therefore comprise 23% of diagnoses impacting *MEF2C* in this cohort.

Our data show that 5'UTR variants can be identified in existing datasets that were primarily designed to capture coding sequences, with 30.7% of 5'UTR bases having sufficient (>10x) coverage in exome sequencing data from the DDD study. However, exome sequencing data is likely to only identify UTR variants that are proximal to the first and last exons of genes, and whole genome or expanded panel sequencing will be required to assay distal or poorly covered UTRs. Furthermore, given their large size, 3'UTRs are particularly poorly covered in exome sequencing datasets. There are examples of disease-causing variants within 3'UTRs, including those impacting polyA signals and microRNA binding[9,55–57], which will not be detected using these methodologies but that could increase diagnostic yield.

Although we screened DNMs in the 5'UTRs of a set of 359 known haploinsufficient DD genes, four of the five identified *de novo* uAUG-creating variants were found in *MEF2C*. This enrichment in a single gene is likely due to a combination of factors (Figure S5). Firstly, *MEF2C* has a proximal 5'UTR exon that is very well covered in the DDD exome sequencing data. Secondly, this 5'UTR exon contains a large number of sites where a variant could create an uAUG, with only two DD haploinsufficient genes having more well-covered possible uAUG-creating sites. Thirdly, unlike the other genes with well-covered possible uAUG-creating sites, *MEF2C* haploinsufficiency is a recurrent cause of DD within the DDD study (Figure S5). Finally, due to the direct interaction of the extreme N-terminus of MEF2C with DNA, CDS-elongating variants are also likely to be pathogenic, which is unlikely to be the case in the vast majority of other haploinsufficient DD genes. As a result, *MEF2C* may be unusual in its potential for pathogenic mutations in the 5'UTR and similarly large increases in diagnostic yield are unlikely across most DD haploinsufficient genes. Nethertheless the enrichment of uAUG-creating variants in *MEF2C* is striking: only 14 of 426 possible variants create uAUGs (at 142 5'UTR bases that are well-covered in the DDD study exome sequencing data), yet all four DNMs observed in the DDD study in the *MEF2C* 5'UTR are uAUG-creating (binomial P=1.2x10^{-6}).

In our functional data, we see a difference in the size of variant effects dependent on the strength of the Kozak consensus surrounding the newly created uAUG. The Kozak sequence is known to influence the likelihood of a ribosome initiating translation at any given AUG as it scans along the 5'UTR from the 5' cap[13]. Our four uAUG-creating variants each generate a new uORF that overlaps the coding sequence. Ribosomes that initiate translation at these uAUGs will not be available to translate from the wild-type coding start site (which itself has a strong Kozak consensus), resulting in reduced translation of the CDS. The stronger the Kozak consensus around the uAUG, the greater this effect will be.

As we extend our analyses to detect non-coding variants, we caution that interpretation of UTR variants still remains a critical challenge. Every 5'UTR has a unique combination of regulatory elements tightly regulating RNA stability and protein expression[58,59], and the impact of any variant will vary with the gene-specific context. Functional validation of identified variants will therefore be crucial to prove (or reject) causality. Some variants may have only a partial regulatory effect, but these variants can nonetheless be harnessed to assess the extent to which perturbation of protein levels or function is tolerated, potentially leading to reduced expressivity and/or lower penetrance. In the case of *MEF2C*, our results suggest that even partial reductions in protein expression lead to severe disease.

Finally, we note how the mechanism of action of non-coding variants can inform the mechanisms underlying protein-coding variants. Identification and characterisation of the effect of the CDS-elongating *MEF2C* variants led us to analyse the domain structure of MEF2C protein and confirm that all the currently identified missense variants likely also act via disrupting DNA-binding, leading to a loss-of-function.

In conclusion, our results further highlight the important contribution of non-coding regulatory variants to rare disease and underscore the huge promise of large whole-genome sequencing datasets to both find new diagnoses and further our understanding of regulatory disease mechanisms.

Supplementary Data

Supplementary data include five figures and seven tables. Also included is the Genomics England Research Consortium author list.

Declaration of Interests

K.J.K. is a consultant for Vor Biopharma. J.J. and K.R. are employees of GeneDx, Inc. K.R. holds shares in Opko Health, Inc. B.D.Z. is a member of the speakers bureau for Biogen, Neurelis, and Supernus. S.A.C. is co-founder and shareholder of Enleofen Bio Pte Ltd. M.E.H. is co-founder, shareholder, consultant, and non-executive director of Congenica Ltd. All other authors declare no competing interests.

Acknowledgements

NW is currently supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 220134/Z/20/Z). Initial work was completed whilst NW was supported by a Rosetrees and Stoneygate Imperial College Research Fellowship. NQ is supported by the Imperial College Academic Health Science Centre. This work is additionally supported by The Rosetrees Trust (Grant Number H5R01320), the Wellcome Trust [WT200990/Z/16/Z, WT200990/A/16/Z], Fondation Leducq [16 CVD 03], the National Institute for Health Research (NIHR) Imperial College Biomedical Research Centre, the Cardiovascular Research Centre, Royal Brompton & Harefield NHS Trust, and the NIHR Oxford Biomedical Research Centre Programme. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. The DDD study presents independent research commissioned by the Health Innovation Challenge Fund (Grant Number HICF-1009-003) a parallel funding partnership between the Wellcome Trust and the Department of Health, and the Wellcome Trust Sanger Institute (Grant Number WT098051). See Nature 2015;519:223-8 or www.ddduk.org/access.html for full acknowledgement. This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes

Project uses data provided by patients and collected by the National Health Service as part of their care and support. A CC BY or equivalent licence is applied to the VoR in accordance with Wellcome open access conditions.

Data and Code Availability

The DDD study data are available under managed access from the European Genome-phenome Archive (Study ID EGAS00001000775), and likely diagnostic variants are available open access in DECIPHER. Code used for modelling case and population variants on the MEF2C protein structure can be found here: https://github.com/matteoferla/MEF2C_analysis

Web resources

Online Mendelian Inheritance in Man (http://www.omim.org)

Gene-2-phenotype (https://www.ebi.ac.uk/gene2phenotype/)

References

1. Short, P.J., McRae, J.F., Gallone, G., Sifrim, A., Won, H., Geschwind, D.H., Wright, C.F., Firth, H.V., FitzPatrick, D.R., Barrett, J.C., et al. (2018). De novo mutations in regulatory elements in neurodevelopmental disorders. Nature *555*, 611–616.

2. Spielmann, M., and Mundlos, S. (2016). Looking beyond the genes: the role of non-coding variants in human disease. Hum. Mol. Genet. *25*, R157–R165.

3. An, J.-Y., Lin, K., Zhu, L., Werling, D.M., Dong, S., Brand, H., Wang, H.Z., Zhao, X., Schwartz, G.B., Collins, R.L., et al. (2018). Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science *362*,.

4. Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith, J.D., Harrell, T.M., McMillin, M.J., Wiszniewski, W., Gambin, T., et al. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am. J. Hum. Genet. *97*, 199–215.

5. Kaplanis, J., Samocha, K.E., Wiel, L., Zhang, Z., Arvai, K.J., Eberhardt, R.Y., Gallone, G., Lelieveld, S.H., Martin, H.C., McRae, J.F., et al. (2020). Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature *586*, 757–762.

6. Srivastava, S., Love-Nichols, J.A., Dies, K.A., Ledbetter, D.H., Martin, C.L., Chung, W.K., Firth, H.V., Frazier, T., Hansen, R.L., Prock, L., et al. (2019). Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. Genet. Med. *21*, 2413–2421.

7. Mignone, F., Gissi, C., Liuni, S., and Pesole, G. (2002). Untranslated regions of mRNAs. Genome Biol *3*, reviews0004.1.

8. Mortimer, S.A., Kidwell, M.A., and Doudna, J.A. (2014). Insights into RNA structure and function from genome-wide studies. Nature Reviews Genetics *15*, 469–479.

9. Wanke, K.A., Devanna, P., and Vernes, S.C. (2018). Understanding Neurodevelopmental Disorders: The Promise of Regulatory Variation in the 3'UTRome. Biol. Psychiatry *83*, 548–557.

10. Chatterjee, S., and Pal, J.K. (2009). Role of 5′- and 3′-untranslated regions of mRNAs in human diseases. Biology of the Cell *101*, 251–262.

11. Whiffin, N., Genome Aggregation Database Production Team, Karczewski, K.J., Zhang, X., Chothani, S., Smith, M.J., Gareth Evans, D., Roberts, A.M., Quaife, N.M., Schafer, S., et al. (2020). Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. Nature Communications *11*,.

12. de Lima, R.L.L.F., Hoper, S.A., Ghassibe, M., Cooper, M.E., Rorick, N.K., Kondo, S., Katz, L., Marazita, M.L., Compton, J., Bale, S., et al. (2009). Prevalence and nonrandom

distribution of exonic mutations in interferon regulatory factor 6 in 307 families with Van der Woude syndrome and 37 families with popliteal pterygium syndrome. Genet. Med. *11*, 241–247.

13. Kozak, M. (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res. *15*, 8125–8148.

14. Noderer, W.L., Flockhart, R.J., Bhaduri, A., Diaz de Arce, A.J., Zhang, J., Khavari, P.A., and Wang, C.L. (2014). Quantitative analysis of mammalian translation initiation sites by FACS-seq. Mol. Syst. Biol. *10*, 748.

15. Sample, P.J., Wang, B., Reid, D.W., Presnyak, V., McFadyen, I.J., Morris, D.R., and Seelig, G. (2019). Human 5' UTR design and variant effect prediction from a massively parallel translation assay. Nat. Biotechnol. *37*, 803–809.

16. Deciphering Developmental Disorders Study (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature *519*, 223–228.

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

18. Ramu, A., Noordam, M.J., Schwartz, R.S., Wuster, A., Hurles, M.E., Cartwright, R.A., and Conrad, D.F. (2013). DeNovoGear: de novo indel and point mutation discovery and phasing. Nat. Methods *10*, 985–987.

19. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

20. Thormann, A., Halachev, M., McLaren, W., Moore, D.J., Svinti, V., Campbell, A., Kerr, S.M., Tischkowitz, M., Hunt, S.E., Dunlop, M.G., et al. (2019). Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. Nat. Commun. *10*, 2373.

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

22. (dgt), T.F.C.A.T.R.P.A.C., and The FANTOM Consortium and the RIKEN PMI and CLST (DGT) (2014). A promoter-level mammalian expression atlas. Nature *507*, 462–470.

23. Nellore, A., Jaffe, A.E., Fortin, J.-P., Alquicira-Hernández, J., Collado-Torres, L., Wang, S., Phillips, R.A., III, Karbhari, N., Hansen, K.D., Langmead, B., et al. (2016). Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. Genome Biol. *17*, 266.

24. Consortium, T.G., and The GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.

25. Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., et al. (2020). Ensembl 2020. Nucleic Acids Res. *48*, D682–D688.

26. Caulfield, M., Davies, J., Dennys, M., Elbahy, L., Fowler, T., Hill, S., Hubbard, T., Jostins, L., Maltby, N., Mahon-Pearson, J., et al. (2019). The National Genomics Research and Healthcare Knowledgebase (figshare).

27. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

28. Zhang, X., Wakeling, M., Ware, J., and Whiffin, N. (2020). Annotating high-impact 5'untranslated region variants with the UTRannotator. Bioinformatics.

29. Naya, F.J., Wu, C., Richardson, J.A., Overbeek, P., and Olson, E.N. (1999). Transcriptional activity of MEF2 during mouse embryogenesis monitored with a MEF2-dependent transgene. Development *126*, 2045–2052.

30. Wu, H., Rothermel, B., Kanatous, S., Rosenberg, P., Naya, F.J., Shelton, J.M., Hutcheson, K.A., DiMaio, J.M., Olson, E.N., Bassel-Duby, R., et al. (2001). Activation of MEF2 by muscle activity is mediated through a calcineurin-dependent pathway. EMBO J. *20*, 6414–6423.

31. Fromer, M., Moran, J.L., Chambert, K., Banks, E., Bergen, S.E., Ruderfer, D.M., Handsaker, R.E., McCarroll, S.A., O'Donovan, M.C., Owen, M.J., et al. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am. J. Hum. Genet. *91*, 597–607.

32. Packer, J.S., Maxwell, E.K., O'Dushlaine, C., Lopez, A.E., Dewey, F.E., Chernomorsky, R., Baras, A., Overton, J.D., Habegger, L., and Reid, J.G. (2016). CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. Bioinformatics *32*, 133–135.

33. Backenroth, D., Homsy, J., Murillo, L.R., Glessner, J., Lin, E., Brueckner, M., Lifton, R., Goldmuntz, E., Chung, W.K., and Shen, Y. (2014). CANOES: detecting rare copy number variants from whole exome sequencing data. Nucleic Acids Res. *42*, e97.

34. Santelli, E., and Richmond, T.J. (2000). Crystal structure of MEF2A core bound to DNA at 1.5 A resolution. J. Mol. Biol. *297*, 437–449.

35. Bienert, S., Waterhouse, A., de Beer, T.A.P., Tauriello, G., Studer, G., Bordoli, L., and Schwede, T. (2017). The SWISS-MODEL Repository-new features and functionality. Nucleic Acids Res. *45*, D313–D319.

36. Lei, X., Kou, Y., Fu, Y., Rajashekar, N., Shi, H., Wu, F., Xu, J., Luo, Y., and Chen, L. (2018). The Cancer Mutation D83V Induces an α-Helix to β-Strand Conformation Switch in MEF2B. J. Mol. Biol. *430*, 1157–1172.

37. Chaudhury, S., Lyskov, S., and Gray, J.J. (2010). PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics *26*, 689–691.

38. Conway, P., Tyka, M.D., DiMaio, F., Konerding, D.E., and Baker, D. (2014). Relaxation of backbone bond geometry improves protein energy landscape modeling. Protein Sci. *23*, 47–55.

39. Alford, R.F., Leaver-Fay, A., Jeliazkov, J.R., O'Meara, M.J., DiMaio, F.P., Park, H., Shapovalov, M.V., Renfrew, P.D., Mulligan, V.K., Kappel, K., et al. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J. Chem. Theory Comput. *13*, 3031–3048.

40. Huang, P.-S., Ban, Y.-E.A., Richter, F., Andre, I., Vernon, R., Schief, W.R., and Baker, D. (2011). RosettaRemodel: a generalized framework for flexible backbone protein design. PLoS One *6*, e24109.

41. Ferla, M.P., Pagnamenta, A.T., Damerell, D., Taylor, J.C., and Marsden, B.D. (2020). MichelaNglo: sculpting protein views on web pages without coding. Bioinformatics *36*, 3268–3270.

42. Samocha, K.E., Kosmicki, J.A., Karczewski, K.J., O'Donnell-Luria, A.H., Pierce-Hoffman, E., MacArthur, D.G., Neale, B.M., and Daly, M.J. (2017). Regional missense constraint improves variant deleteriousness prediction (bioRxiv).

43. Le Meur, N., Holder-Espinasse, M., Jaillard, S., Goldenberg, A., Joriot, S., Amati-Bonneau, P., Guichet, A., Barth, M., Charollais, A., Journel, H., et al. (2010). MEF2C haploinsufficiency caused by either microdeletion of the 5q14.3 region or mutation is responsible for severe mental retardation with stereotypic movements, epilepsy and/or cerebral malformations. J. Med. Genet. *47*, 22–29.

44. Suri, M., Evers, J.M.G., Laskowski, R.A., O'Brien, S., Baker, K., Clayton-Smith, J., Dabir, T., Josifova, D., Joss, S., Kerr, B., et al. (2017). Protein structure and phenotypic analysis of pathogenic and population missense variants in. Mol Genet Genomic Med *5*, 495–507.

45. Zerbino, D.R., Wilder, S.P., Johnson, N., Juettemann, T., and Flicek, P.R. (2015). The ensembl regulatory build. Genome Biol. *16*, 56.

46. ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710.

47. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature *581*, 444–451.

48. D'haene, E., Bar-Yaacov, R., Bariah, I., Vantomme, L., Van Loo, S., Cobos, F.A., Verboom, K., Eshel, R., Alatawna, R., Menten, B., et al. (2019). A neuronal enhancer network upstream of MEF2C is compromised in patients with Rett-like characteristics. Hum. Mol. Genet. *28*, 818–827.

49. Zweier, M., Gregor, A., Zweier, C., Engels, H., Sticht, H., Wohlleber, E., Bijlsma, E.K., Holder, S.E., Zenker, M., Rossier, E., et al. (2010). Mutations in MEF2C from the 5q14.3q15 microdeletion syndrome region are a frequent cause of severe mental retardation and diminish MECP2 and CDKL5 expression. Hum. Mutat. *31*, 722–733.

50. Bienvenu, T., Diebold, B., Chelly, J., and Isidor, B. (2013). Refining the phenotype associated with MEF2C point mutations. Neurogenetics *14*, 71–75.

51. Vrečar, I., Innes, J., Jones, E.A., Kingston, H., Reardon, W., Kerr, B., Clayton-Smith, J., and Douzgou, S. (2017). Further Clinical Delineation of the MEF2C Haploinsufficiency Syndrome: Report on New Cases and Literature Review of Severe Neurodevelopmental Disorders Presenting with Seizures, Absent Speech, and Involuntary Movements. J. Pediatr. Genet. *6*, 129–141.

52. Molkentin, J.D., Black, B.L., Martin, J.F., and Olson, E.N. (1996). Mutational analysis of the DNA binding, dimerization, and transcriptional activation domains of MEF2C. Mol. Cell. Biol. *16*, 2627–2636.

53. Potthoff, M.J., and Olson, E.N. (2007). MEF2: a central regulator of diverse developmental programs. Development *134*, 4131–4140.

54. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart,

J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. *44*, D862–D868.

55. Bennett, C.L., Brunkow, M.E., Ramsdell, F., O'Briant, K.C., Zhu, Q., Fuleihan, R.L., Shigeoka, A.O., Ochs, H.D., and Chance, P.F. (2001). A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA-->AAUGAA) leads to the IPEX syndrome. Immunogenetics *53*, 435–439.

56. Devanna, P., Chen, X.S., Ho, J., Gajewski, D., Smith, S.D., Gialluisi, A., Francks, C., Fisher, S.E., Newbury, D.F., and Vernes, S.C. (2018). Next-gen sequencing identifies non-coding variation disrupting miRNA-binding sites in neurological disorders. Molecular Psychiatry *23*, 1375–1384.

57. Reamon-Buettner, S.M., Cho, S.-H., and Borlak, J. (2007). Mutations in the 3'-untranslated region of GATA4 as molecular hotspots for congenital heart disease (CHD). BMC Med. Genet. *8*, 38.

58. Araujo, P.R., Yoon, K., Ko, D., Smith, A.D., Qiao, M., Suresh, U., Burns, S.C., and Penalva, L.O.F. (2012). Before It Gets Started: Regulating Translation at the 5′ UTR. Comparative and Functional Genomics *2012*, 1–8.

59. Leppek, K., Das, R., and Barna, M. (2018). Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. Nature Reviews Molecular Cell Biology *19*, 158–174.

Figure legends

Figure 1: Schematic of the wild-type *MEF2C* gene (a) and the position and effect of uAUG-creating variants identified as *de novo* in developmental disorder cases (b and c) and in gnomAD population controls (d). The two 5'UTR exons are shown as light grey boxes, separated by an intron shown as a thinner broken grey line. Upstream open reading frames (uORFs) already present in the sequence are shown in green. Variant positions are represented by arrows. New ORFs created by the variants are shown as blue boxes. (b) Two case variants create novel ORFs that overlap the coding sequence (CDS) out-of-frame (oORF-creating). If translation initiates at the uAUG, the ribosome will not translate the CDS. (c) Two recurrent case variants create uAUGs in-frame with the CDS. If translation initiates at this uAUG, an elongated protein will be translated. (d) Two variants identified in gnomAD create uORFs far upstream of the CDS which would not be predicted to disrupt translation of the normal protein.

Figure 2: uAUG-creating variants decrease translation of MEF2C (a) or transactivation of target genes (b). (a) *MEF2C* 5' UTR out-of-frame overlapping ORF (oORF)-creating variants c.-103G>A and c.-66A>T (Figure 1b) reduce downstream luciferase expression relative to wild-type (WT) 5' UTR in a translation reporter assay. Reduction is stronger for c.-66A>T (moderate uAUG Kozak context) than for c.-103G>A (weak Kozak context). (b) Overexpression of MEF2C with the WT 5' UTR/CDS induces expression of luciferase from a MEF2C-dependent enhancer-luciferase reporter construct, relative to an empty pcDNA3.1 construct negative control. The *MEF2C* N-terminus-extending variants c.-26C>T (9 amino acids) and c.-8C>T (3 amino acids; Figure 1c) both reduce transactivation. For (a) and (b) bars are coloured by Kozak consensus: yellow=weak; orange=moderate; red=strong. Luciferase expression was normalised for transfection efficiency.

Figure 3: (a) The N-terminal region of *MEF2C* is highly constrained for missense variants in gnomAD (obs/exp=0.069), with much lower constraint across the rest of the protein

(obs/exp=0.41). This region of high constraint correlates with the location of the majority of *de novo* missense variants identified in DD cases (red circles), while gnomAD variants are mostly outside of this N-terminal region (grey circles). (b) The N-terminal portion of the MEF2C dimer [1-92], modelled using structures of the human MEF2A dimer which is 96% identical in sequence to MEF2C, bound directly to its consensus DNA sequence. Side chains of amino acids with pathogenic *de novo* missense variants from DDD, GeneDx and ClinVar are shown in yellow, with gnomAD *MEF2C* missense variants in grey. Most pathogenic missense variants either protrude directly into the DNA or are located in the DNA-binding helix. In particular, the terminal amine (Gly2, top inset) along with Arg3 (bottom inset) act as reader-heads for nucleobase specificity, which is likely disrupted in the N-terminal extension variants (middle inset). All pathogenic and gnomAD variants can be viewed in our interactive protein structure browser here: https://michelanglo.sgc.ox.ac.uk/r/mef2c. (c-d) Missense variants from DD cases (DDD, GeneDx and ClinVar) are significantly more disruptive to the interaction with DNA as measured by ΔΔG values (c) and closer to the bound DNA molecule (d) than *MEF2A-D* variants in gnomAD (see online methods).

Figure 4: 5'UTRs of DD haploinsufficient genes (red) are longer (a), and a higher proportion have multiple exons (b) compared to 5'UTRs of all genes (light grey), and other DD genes (dark grey). Mean lengths for each gene set in (a) are shown as dotted lines. (c) The coverage of 5'UTRs decays rapidly with distance from the CDS (x-axis truncated at 1000 bps). Note that these figures were calculated using exome sequence data from the DDD study and may vary between different exome capture designs. (d) The position of DNA-binding domains (including homeodomains, zinc-fingers, and specific DNA-binding domains) in DD haploinsufficient genes with respect to the N-terminus of the protein; MEF2C is one of three proteins with a DNA-binding domain that starts within 10 bps of the N-terminus.

Tables

variant (GRCh37)	cDNA description (ENST00000504921.7)	variant effect	deletion size	kozak strength	proband ID(s)	proband count	gnomAD v3 AC
uUAG-creating de novo variants discovered in probands with DD:							
chr5:88119671 T>A	c.-66A>T	out-of-frame oORF created	-	moderate	1	1	-
chr5:88119708 C>T	c.-103G>A	out-of-frame oORF created	-	weak	2	1	-
chr5:88119613 G>A	c.-8C>T	CDS-elongating	-	strong	3,4,5	3	-
chr5:88119631 G>A	c.-26C>T	CDS-elongating	-	moderate	6,7,8	3	-
uAUG-creating variant present in gnomAD:							
chr5:88883052 G>A	c.-240C>T	uORF created	-	weak	-	0	1
chr5:88883059 G>A	c.-247C>T	uORF created	-	weak	-	0	6
-							
chr5:88133089-88427361 del	-	promoter and partial 5'UTR deletion	294kb	-	9	1	-
chr5:88123099-88220350 del	-	promoter and partial 5'UTR deletion	97kb	-	10	1	-

Table 1: Details of *MEF2C* uAUG-creating and upstream deletion variants discussed in this work. Shown are the four uAUG SNVs identified in DDD, uAUG SNVs observed in gnomAD v3.0, and non-coding CNVs found upstream of *MEF2C* in DDD. oORF = overlapping ORF; uORF = upstream ORF; AC = allele count. Proband IDs refer to those used in Table S2.