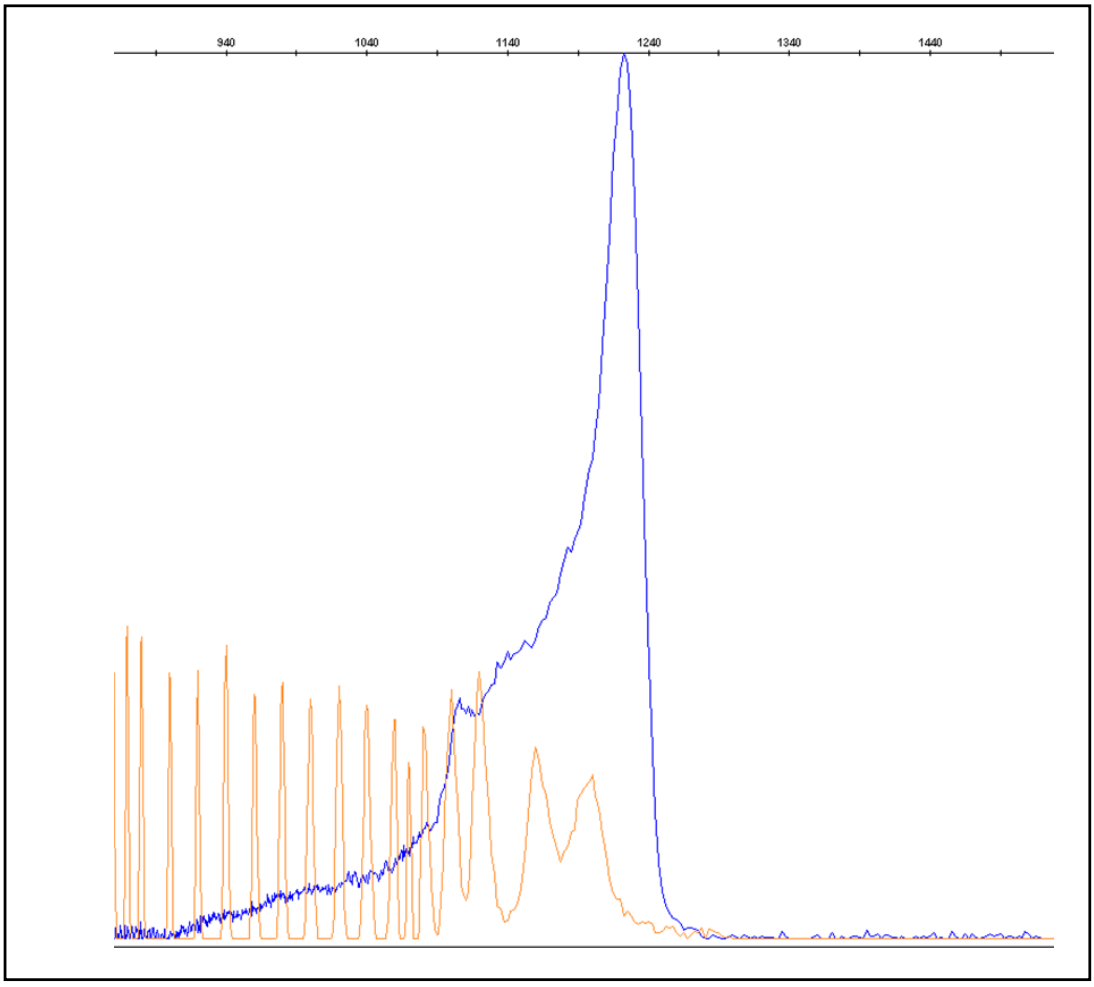
Stem Cell Reports, Volume *7*

Supplemental Information

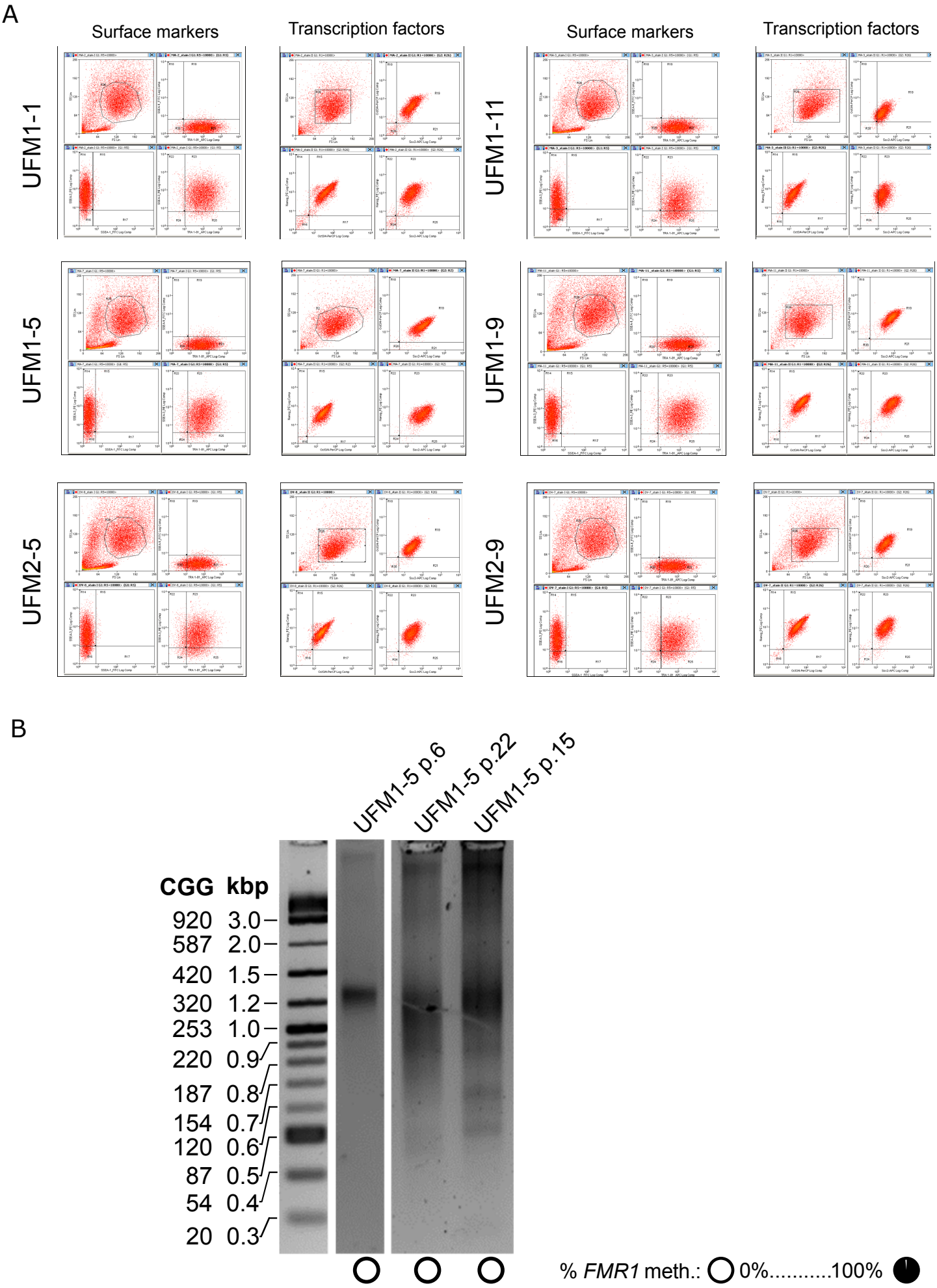
CGG Repeat-Induced *FMR1* Silencing Depends on the Expansion Size in Human iPSCs and Neurons Carrying Unmethylated Full Mutations

Urszula Brykczynska, Eline Pecho-Vrieseling, Anke Thiemeyer, Jessica Klein, Isabelle Fruh, Thierry Doll, Carole Manneville, Sascha Fuchs, Mariavittoria Iazeolla, Martin Beibel, Guglielmo Roma, Ulrike Naumann, Nicholas Kelley, Edward J. Oakeley, Matthias Mueller, Baltazar Gomez-Mancilla, Marc Bühler, Elisabetta Tabolacci, Pietro Chiurazzi, Giovanni Neri, Tewis Bouwmeester, Francesco Paolo Di Giorgio, and Barna D. Fodor

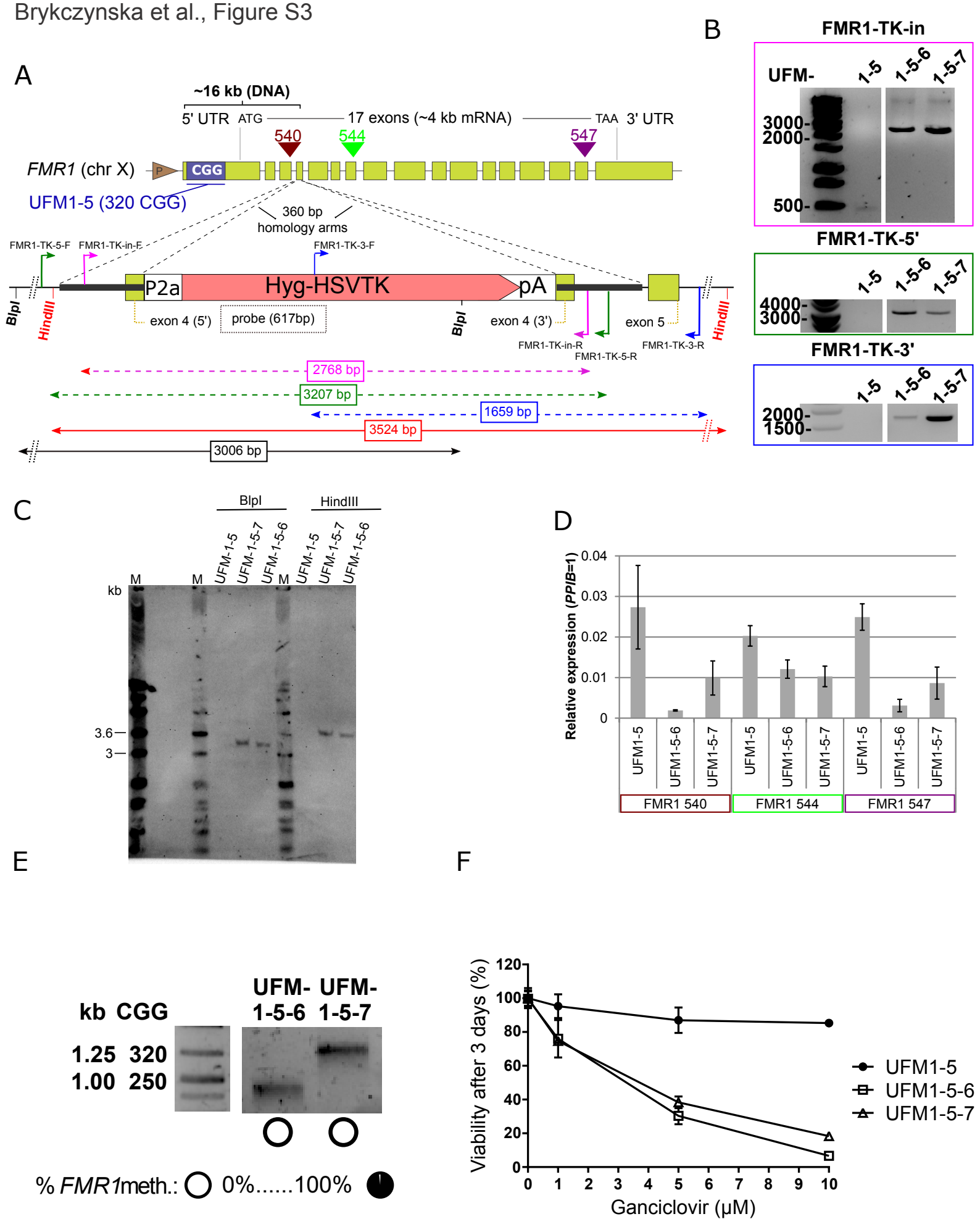
Brykczynska et al., Figure S1



Brykczynska et al., Figure S2

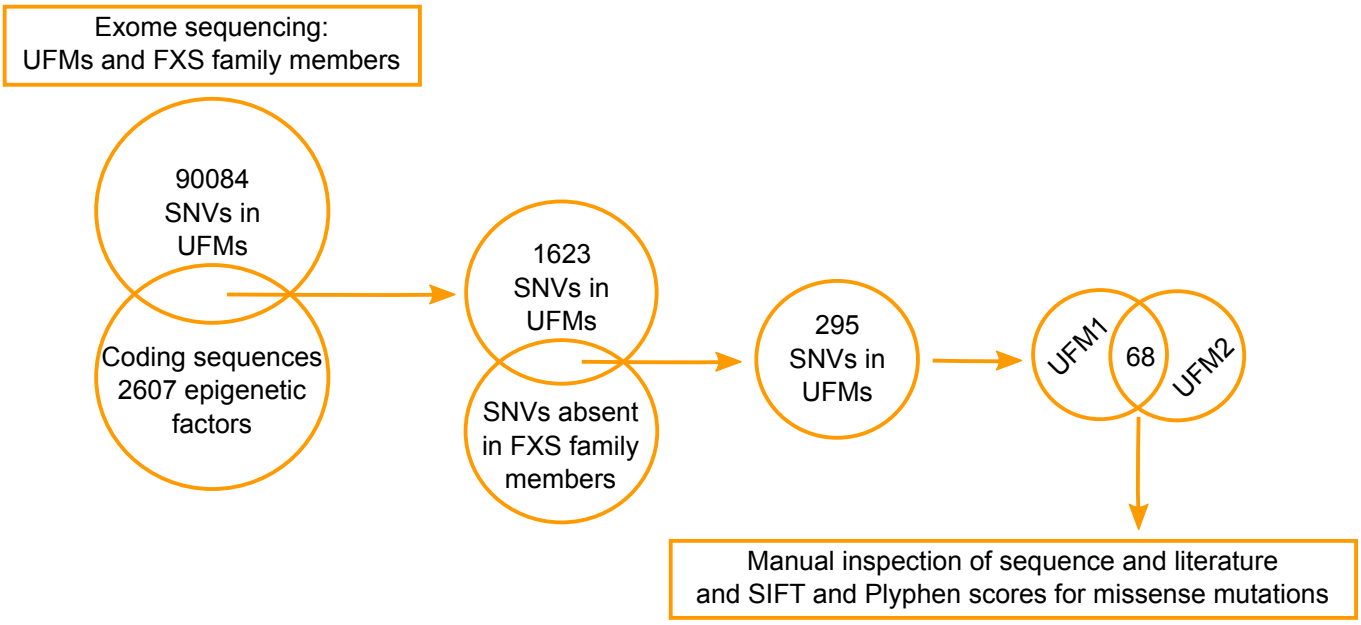


Brykczynska et al., Figure S3

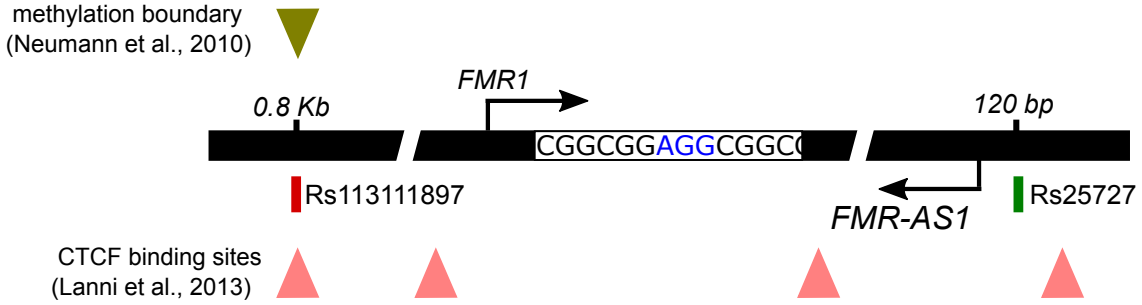


Brykczynska et al., Figure S4

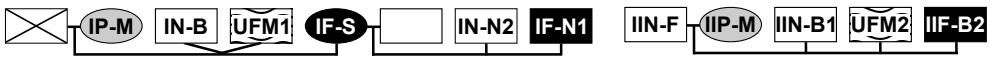
A



B



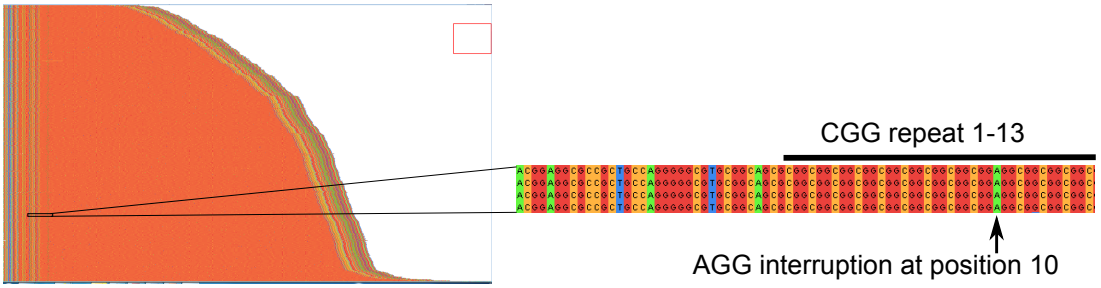
C



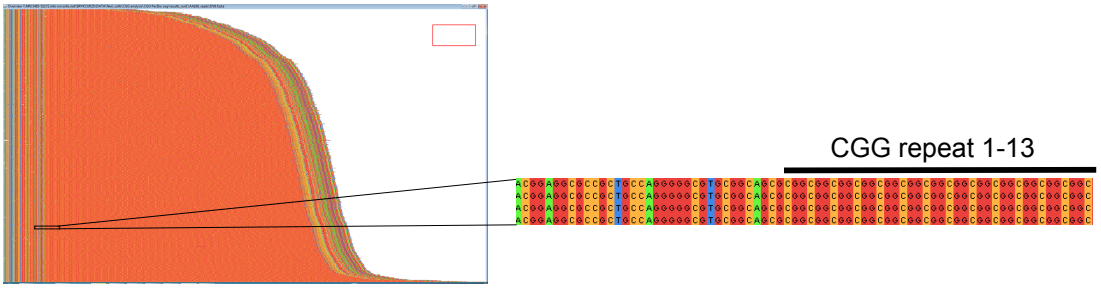
Rs113111897	G/A	G	A	G/A	G	A	G	G/G	G	G	G
Rs25727	NA	T	T	NA	NA	T	NA	NA	T	C	C
AGG position	9/19	9/19	10*	8	5/14	10*	9/19	9/19	9/19	0*	0

D

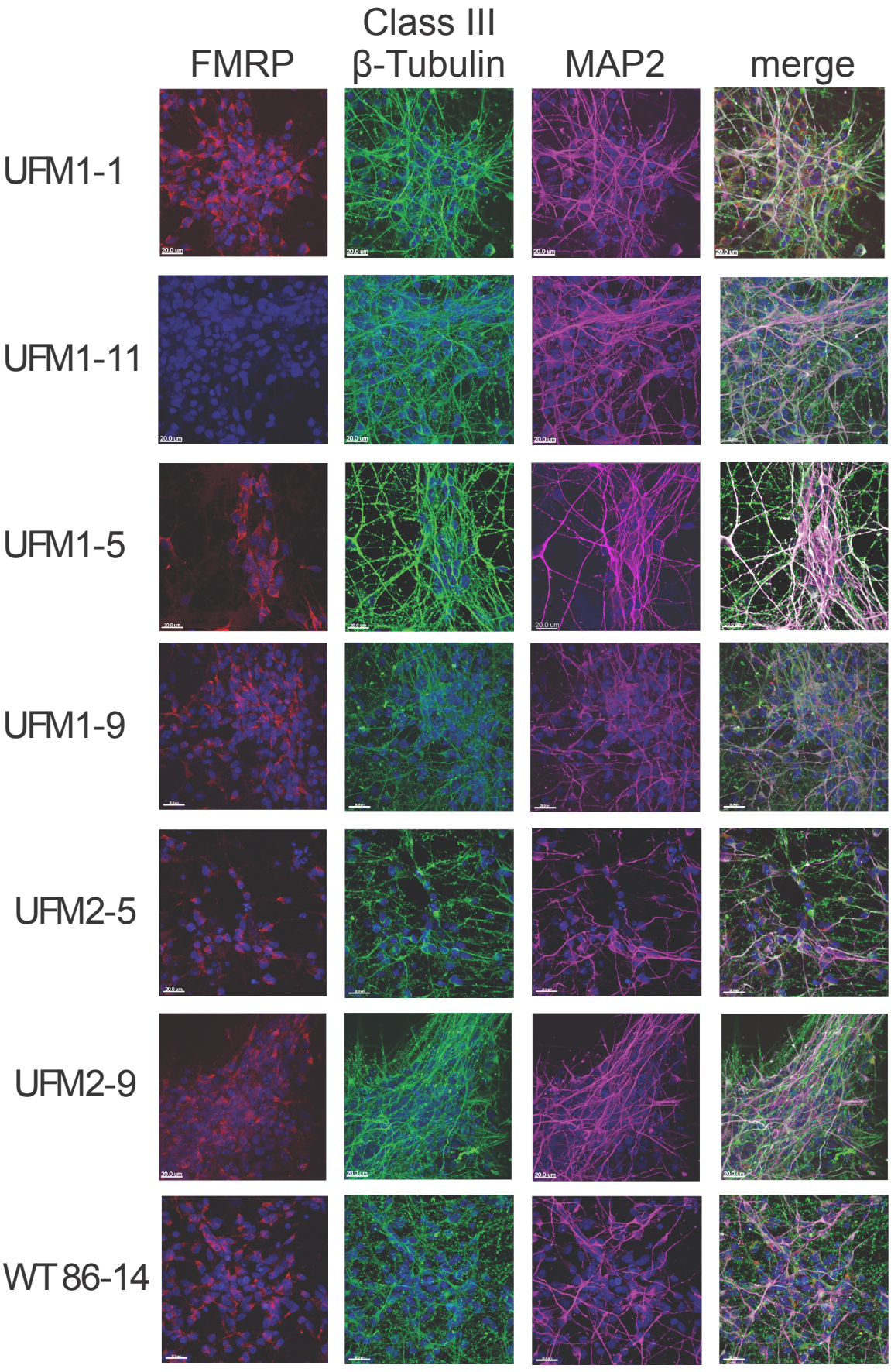
UFM1-5 529 forward sequences



UFM2-7 742 forward sequences



Brykczynska et al., Figure S5



Supplemental Figure Legends

Figure S1. Related to Figure 1

Analysis of CGG expansion in UFM2

The electropherogram of the CGG triplet sequence (in blue) displays an expansion in the range of full mutation of around 280-330 triplets. The orange peaks are those of the 1200 LIZ dye size standard, the numbers in the upper part of the panel indicate the size of the run in bp.

Figure S2. Related to Figure 2 and Figure 3

Characterization of iPSC derived cells

A) Fluorescent-activated cell sorting (FACS) analysis of iPSC lines generated from T-cells of UFM individuals. Data for representative clones that were used in further experiments is shown. As expected majority of the cells are double positive for pluripotency surface markers SSEA-3 and TRA1-81 and negative for differentiation marker SSEA-1, as well as positive for pluripotency associated transcription factors NANOG, OCT4 and SOX2.

B) Stability of CGG repeat size and *FMR1* hypomethylation upon passaging of UFM iPSCs. PCR analysis of the CGG repeat size using Asuragen AmplideX FMR1 PCR kit followed by agarose gel electrophoresis for UFM1-5 iPSC clone at passage 6 and following two independent passaging to passage 22 and 15. Black circles under each lane indicate mean percentage of DNA methylation across 22 CpGs of the *FMR1* promoter based on data in Table S1. Certain instability of the CGG repeat is observed upon passaging but the unmethylated status of the *FMR1* promoter is not changed.

Figure S3. Related to Figure 4

Generation and characterization of the *FMR1* knock-in iPSC lines

A) Schematic representation of the *FMR1* knock-in strategy. Location of PCR primers used for validation of the correct integration is indicated with arrows and sizes of the expected PCR products between the arrows below. Similarly restriction sites of enzymes used for Southern blot analysis are indicated and sizes of the expected genomic fragments between the arrows below. Location of TaqMan probes used for evaluation of expression of *FMR1* after knock-in is indicated with triangles.

B) Validation of the correct cassette integration in two selected clones UFM1-5-6 and UFM1-5-7using PCR primers indicated in the scheme in panel (a) with sequences given in Supplementary Experimental Procedures. PCR primers FMR1-TK-in are spanning the homology arms. Cassette integration results in increase in the PCR product size from 495 bp in parental line UFM1-5 to 2768 bp in knock-in lines. PCR primers FMR1-TK-5 and FMR1-TK-3 are spanning the cassette and the 5' and 3' genomic regions respectively resulting in the 3207 bp product for FMR1-TK-5 and in 1659 bp products for FMR1-TK-3 when the cassette is integrated.

C) Southern blot confirming single integration of HyTK cassette into the genome. gDNA digested with BlpI and HindIII is loaded for parental line UFM1-5 and two selected clones UFM1-5-6 and UFM1-5-7. As expected single band is observed for two clones and no signal for the parental line. Band sizes correspond to the genomic DNA fragments of the *FMR1* locus with integrated HyTK cassette as indicated in panel A.

D) Expression of *FMR1* before (UFM1-5) and after the cassette integration (UFM1-5-6 and UFM1-5-7) based on TaqMan experiment with 3 probes located along the FMR1 mRNA as indicated in the scheme in panel A. Both knock-in clones express *FMR1* although at the reduced levels comparing to the parental line. Data are presented as a mean of three independent biological replicates. Error bars represent standard deviation.

E) CGG repeat size and *FMR1* promoter methylation in two knock-in clones UFM1-5-6 and UFM1-5-7. Black circles under each lane indicate mean percentage of DNA methylation across 22 CpGs of the *FMR1* promoter for a given clone based on data in Table S1.

F) Survival of parental line UFM1-5 and knock-in clones UFM1-5-6 and UFM1-5-7 with increasing amount of ganciclovir evaluated by Presto Blue fluorescence measurement (Thermo Fisher) after 3 days of treatment. Both subclones expressing Thymidine Kinase show lethality in 5-10 μM ganciclovir range in contrast to the parental line. Data are presented as a mean of three independent biological replicates. Error bars represent standard deviation.

Figure S4. Related to Figure 4

A) Filtering criteria for identification of a causative genetic variant common to two UFMs. To identify a common genetic cause of the UFM phenotype we performed exome-sequencing of UFM individuals and their family members. Strict filtering criteria were applied due to the small sample size. We only considered mutations affecting coding sequences of proteins broadly implicated in epigenetic processes. Further we filtered out the mutations present in FXS family members and considered mutations affecting the same gene in both UFMs. Among the 68 shortlisted candidates we haven't identified a mutation with obvious disrupting effect on the epigenetic machinery. This is in line with the fact that UFM cells retain the ability to silence *FMR1*. Furthermore, a dramatic impairment of the core epigenetic machinery would likely have strong consequences on the phenotype of UFM individuals.

B-D) Sequence analysis of *FMR1* and *FMR1-AS1* promoters and CGG repeat tract

B) Schematic representation of the *FMR1* locus with indicated SNPs detected in UFM individuals. Putative CTCF binding sites are indicated based on (Lanni et al., 2013) and the DNA methylation boundary based on (Naumann et al., 2009).

C) Presence and position of AGG interruptions in the CGG repeat tract and presence of SNPs in UFM individuals and their family members. AGG position was determined using Asuragen PCR kit and capillary electrophoresis allowing for analysis of the first 200 3'CGG repeats. For UFM1, UFM2 and F-N1 AGG (indicated by *) position was determined by PacBio single molecule sequencing (SMRT) allowing for evaluation of the entire repeat tract. The AGG position is given in number of CGG repeat units from the 5' end.

D) Representative images of multiple sequence alignment of CGG repeat amplicons sequenced with SMRT technology. 529 forward sequences of iPSC clone UFM1-5 and 742 forward sequences of UFM2-7 were aligned. AGG interruption at position 10 in UFM1-7 is highlighted. In all additional sequenced clones from UFM1 (UFM1-6, UFM1-9, UFM1-11 (silent FMR1)) and F-N1 (F-N1-3) the AGG interruption was detected at this position. No interruptions of the repeat tract were detected in sequenced UFM2 clones (UFM2-5 and UFM2-7).

Figure S5. Related to Figure 5

Characterization of iPSC derived neurons

FMRP (red), Class III β Tubulin (green), MAP2 (red) DAPI (blue) staining of iPSC derived neurons after 60 days of differentiation. Presence of FMRP (protein product of *FMR1*) is detected in all lines except UFM1-11 with 480 CGG repeats and silenced *FMR1*. MAP2 and -Class III β Tubulin staining indicates that at day 60 iPS cells were differentiated into post-mitotic neurons. Scale bar = 20μm

Supplemental Experimental Procedures

CGG repeat length analysis using PCR and gel electrophoresis

Genomic DNA from PBMCs, iPSCs and neurons was extracted using DNeasy kit (Qiagen). CGG repeat number and position of AGG was analyzed by PCR amplification using AmplideX FMR1 PCR kit (Asuragen) with with gel

electrophoresis as described previously (Filipovic-Sadic et al., 2010). Southern blot analysis of the CGG repeat size of UFM2 was performed as previously described (Tabolacci et al., 2008).

CGG repeat analysis using PCR and capillary electrophoresis in Fig. S1

The analysis of CGG-tract expansion in UFM2 peripheral blood lymphocytes' DNA was performed using FAM-fluorescent primers (CG-rich-F: TCA GGC GCT CAG CTC CGT TTC GGT TTC A and CG-rich-R: AAG CGC CAT TGG AGC CCC GCA CTT CC) and GC-rich PCR system (Roche, 12140306001). The PCR product was then separated on capillary electrophoresis on 3130 Genetic Analyzer (Life technologies) together with a 1200 LIZ dye size standard (4379950, Life technologies). The result was analyzed with Sequencing Analysis v5.2 software (Life technologies).

DNA methylation of *FMR1* promoter using bisulfite pyrosequencing

Genomic DNA from PBMCs, iPSCs and neurons was extracted using DNeasy kit (Qiagen). Bisulfite treatment and pyrosequencing analysis of the 22 CpGs of the *FMR1* promoter was performed by EpigenDx according to standard procedures with a unique set of primers that were developed by EpigenDx (assays ADS1451FS1 and ADS1451FS2).

Quantitative RT PCR for *FMR1*

RNA from iPSCs and neurons was extracted using RNeasy kit (Qiagen). Reverse transcription and amplification were performed using TaqMan Fast Virus 1-step reagent (Thermo Fisher) and Viaa7 real-time PCR System (Thermo Fisher). Pre-developed TaqMan assay was used for FMR1 (Hs00924547_m1) and PPIB (Hs00168719_m1) which was used for normalization of the *FMR1* signal.

iPSCs derivation

PBMCs were purified from blood samples using Lymphoprep density gradient (Axis-Shield). T cells were isolated from PBMCs using EasySep Human T Cell Enrichment Kit (Stem Cell technologies) according to kit instructions and activated using anti-CD3 antibody OKT-3 and IL-2 cytokine (Cedarlane) as previously described (Tsoukas et al., 1985). Activated T cells (or fibroblasts in case of GM09497) were transduced with Sendai viruses expressing Oct3/4, Sox2, Klf4, c-Myc (Takahashi et al., 2007) using CytoTune-iPS Reprogramming Kit (Invitrogen) according to kit instructions. 3 days after transduction cells were transferred onto mouse embryonic fibroblasts (CF1, Millipore) and cultured in iPSC medium containing 80% knockout Dulbecco's modified eagle medium (DMEM)/F12 (Gibco), 20% knockout serum replacement (Gibco), 10 ng/ml bFGF, 1 mM glutamax (Gibco), 0.1 mM b-mercaptoethanol (Gibco) and 1% non-essential amino acid solution (Gibco). After 3-4 weeks colonies with stem cell morphology were selected and manually passaged 5 times. After establishment, iPSCs were transferred into feeder free conditions on matrigel (Corning) and maintained in mTeSR medium (Stem Cell Technologies) with Pen/Strep supplement (Gibco). The absence of reprogramming vectors was confirmed using RT-PCR for Sendai virus expressed Oct3/4, Sox2, Klf4 and c-Myc as described in CytoTune-iPS Reprogramming Kit (Invitrogen). IPSC clones were analyzed between passage 5 and 10 for expression of pluripotency markers by Score card assay (Invitrogen) and for surface markers SSEA-3, TRA1-81 and SSEA-1 (negative marker), as well as pluripotency associated transcription factors NANOG, OCT4 and SOX2, using fluorochrome conjugated antibodies (BD Biosciences) and fluorescence activated cell sorting (FACS) using standard procedures.

Cell culture

iPSCs were cultured on matrigel in Nutristem (Biological Industries) or mTeSR1 (Stem Cell Technologies) medium, supplemented with Pen/Strep (Gibco). Cells were dissociated with TrypLE (Gibco) every 3-4 days, plated at the density of 10000/cm^2 for Nutristem or 20000/cm^2 for mTeSR in the presence of 10 μM ROCK inhibitor. The medium was replaced every day (mTeSR1) or every other day (Nutristem). Stable knock in clones with the Hyg-Tk cassette (UFM1-5-6 and UFM1-5-7) were maintained in Nutristem supplemented with 25 μg/ml hygromycin if not indicated differently. For positive-negative selection cells were plated at 10000-15000/cm^2 density and from next day on medium was supplemented with ganciclovir (10 μM) and with no hygromycin. For the first 3 days of

selection medium was changed every day and plates washed with PBS. Surviving subclones were picked and expanded in the presence of 10 μM ganciclovir (except the two first passages) and with no hygromycin. For the hygromycin selection of ganciclovir resistant clone UFM1-5-7-1, the clone was cultured without ganciclovir for 3 weeks and further with indicated concentrations of hygromycin for 7 days. Surviving colonies were picked and expanded in the presence of hygromycin.

Hygromycin-HSV-Thymidine Kinase knock-in into *FMR1*

To create an endogenous in frame fusion between a positive-negative selection marker and *FMR1* we designed a CRISPR mediated homologous recombination strategy targeting exon 4 of *FMR1* (Fig.4 and Fig. S3). The donor plasmid carried a P2a peptide-Hygromycin-HSV Thymidine Kinase-stop codon-SV40 polyA (Hyg-TK) cassette flanked by 360 bp homology arms. The Cas9-gRNA plasmid was a derivative of pU6-gRNA-SPycas9-2acherry (Smurnyy et al., 2014) and was generated by replacing the CMV promoter with the CBh promoter and removing the mCherry reporter (Yi Yang personal communication). The Cas9-gRNA plasmid was further modified by inserting a designed (http://crispr.mit.edu/) oligo encoding the 5'-ttagctaaccaccaacagca-3' gRNA sequence targeting *FMR1*. Cells were co-nucleofected with the donor and *FMR1* targeting Cas9-gRNA plasmid using Amaxa nucleofector (program B-016) with the Human Stem Cell Nucleofector Kit 1 (VPH-5012, Lonza). 72h after co-nucleofection cells were cultured in presence of 25 μg/ml hygromycin for 9 to 12 days. Single colonies were picked, subcloned and expanded. Selected subclones were further characterized to confirm the correct cassette integration by PCR (Fig. S3B). Primer sequences were FMR1-TK-in-F: 5'-TCAGAGCACTAATTATTGCTG-3'; FMR1-TK-in-R:5'-AAATTAGATATTACTTCAAAATAGA-3';FMR1-TK-5-F:5'GAGTATCCCTGTCTCTCTGTCCA-3'; FMR1-TK-5-R:5'-CAGAAACATACTGAACACATAGTG-3';FMR1-TK-3-F:5'-ACTCAACTGCTCGTAGCCCT -3' and FMR1-TK-3-R: 5'-GCAAACCAAACCATTTTTGC-3'. Two subclones UFM1-7-6 and UFM1-7-7 were selected for further experiments based on correct cassette integration, CGG repeat length, *FMR1* promoter methylation and mRNA expression (Fig. S3 B,D,E). In these two clones insertions of the HyTK cassette outside the *FMR1* locus was ruled out by Southern blot and hybridization on gDNA digested with BlpI and HindIII using the parental line as control. The DIG labeled probe was generated by PCR (Roche) using the primers oP3_for3: 5' ACGAAGTTGCCAACATTTTCTT 3' and oP3a_rev: 5' GGCATCCCCGGCACTAATCT 3' internal to the cassette (Fig. SC).

Neuronal differentiation iPSCs

iPSCs were differentiated into post mitotic neurons using dual SMAD inhibition protocol as previously described (Chambers et al., 2009) with modifications. Briefly, neuronal precursor cells (NPCs) were derived by seeding iPSCs into 96-well ultralow attachment plate (Costar) in neuronal induction medium (DMEM/F12 medium with GlutamaX (Gibco) supplemented with 20% Knockout-Serum Replacement (Gibco), 0.1mM MEM non-essential amino acids (Gibco), 0.1mM 2-Mercaptoethanol (Gibco), Pen/Strep supplement (Gibco), 10 ng/ml hFGF (Gibco), 10 μM SB 431542 (Stemgent) and 1 μM LDN 193189 (Stemgent)) with 10 μM ROCK inhibitor, at density 10 000 cells/well. The formed EBs were transferred after 3 days onto a matrigel coted plates and cultured in induction medium for additional 7 days. At day 10 of NPC derivation non-neuronal cells appearing at the edges of the attached EBs were manually scraped off, NPCs were detached (TrypLE, Gibco) and transferred onto matrigel coted dishes in proliferation medium (DMEM/F12 with Glutamax, supplemented with B27 and N2 (Gibco), Pen/Strep (Gibco), 10 ng/ml hEGF (Gibco), 10 ng/ml hFGF (Gibco)) with 10 μM ROCK inhibitor. NPCs were expanded in proliferation medium. For neuronal differentiation NPCs were seeded on matrigel plates with density: 250 000 cells/cm^2 and cultured in differentiation medium (Neurobasal Medium (Gibco), supplemented with B27 and N2 (Gibco), Pen/Strep supplement (Gibco), 10 ng/ml BDNF (R&D Systems), 10 ng/ml GDNF (R&D Systems), 10 ng/ml hNT3 (R&D Systems)) for the indicated number of days (indicated as number of differentiation days plus 10 days of NPC derivation). For immunofluorescence analysis NPCs were seeded and differentiated on glass slides coated with matrigel.

Differentiation of NPCs in organotypic mouse brain slices

NPCs derived from iPSCs were differentiated on organotypic mouse brain slices (OTBS) using a previously described method (Pecho-Vrieseling et al., 2014). Briefly, 400 μm thick coronal sections of brains of 5 days old pups were prepared using a vibratome (Leica). Slices containing striatum, cortex and hippocampus were selected and plated on Millicell inserts (Millipore). Slice cultures were maintained at 35 °C in a 5% CO_2/ 95% air atmosphere with a relative humidity 95% in OTBS medium (50% (vol/vol) MEM, 25 mM HEPES, 25% (vol/vol) HBSS, 25%

(vol/vol) heat-inactivated horse serum, 2 mM glutamine, 1 ml of penicillin/streptomycin solution and 0.044% (vol/vol) NaHCO3 adjusted to pH 7.2). IPSCs derived NPCs were labeled with GFP using CAG-GFP expression vector (Tchorz et al., 2012) nucleofected with Amaxa using basic nucleofector kit for primary neurons (Lonza). After nucleofection cells were resuspended in Neurobasal medium (Gibco) with 10 µM ROCK inhibitor at a concentration of 2×10^3 cells per µl and 0.5 µl of the cell suspension was injected into each brain slice hemisphere (30–45° angle), prepared 2 days before, using a Hamilton syringe (Hamilton). Co-cultures were further maintained as described above for slice cultures.

Immunostaining and image analysis

Cells and oragnotypic brain slices were fixed with 4% paraformaldehyde for 5min at 35°C for brain slices and at room temperature for cells, permeabilized and blocked in 0.1% (vol/vol) Triton-X100 (Sigma), 1% bovine serum albumin (Sigma) in PBS (blocking buffer). Staining was performed overnight at 4°C with the following primary antibodies diluted in blocking buffer: rabbit anti-Class III β-Tubulin (PRB-435P, BioLegend), chicken anti-MAP2 (ab5392, Abcam), mouse anti-FMRP (Sc-101048, Santa Cruz), rabbit anti-ubiquitin clone 10H4L21 (701339 Thermo Fisher). Appropriate fluorescence-labeled secondary antibodies were applied for 2h at room temperature, and nuclei were counterstained with 4′,6-diamidino-2-phenylindole (Sigma). Slides were mounted with ProLong Gold (Invitrogen). Fluorescence was imaged on a LSM 700 confocal microscope (Zeiss). Pictures were analyzed and snapshots were taken using Imaris software (version 7.6.5, Bitplane). Ubiquitin and FMRP dots in GFP-positive h-neurons were analyzed using the 3D crop function. The first and last z-planes showing GFP staining were discarded to ensure that ubiquitin and FMRP dots counted were co-localized within GFP-positive h-neurons. We restricted counting of ubiquitin and FMRP dots to dots ≥ 0.5µm, in order to avoid counting of the background staining.

***FMR1* and *FMR1-AS1* promoter sequencing**

FMR1 promoter region (chrX: 147910465-147911958, genome build hg38) was amplified from genomic DNA from primary patient samples (buccal swap or PBMC as indicated in the text) using following primers: FMR1-F1: GGCAGCTATAAGCACGGTGT with FMR1-R2: CCGGAAGTGAAACCGAAAC and FMR1-F6: TCAGCCCTATTGGGTTCTTG with FMR1-R6: AAGGGACATGGATTGAGTCG. FMR1-AS1 promoter region (chrX: 147922071-147922799, genome build hg38) was amplified using following primers: FMR1-AS1-F: CCAGTTTGAGTGCTTTTCAGG with FMR1-AS1-R: ATTTGCAGCCTGCTTTTGAT. PCR amplification was performed using Phusion High-Fidelity PCR master mix (Thermo Scientific) according to the producers' recommendations. Sanger sequencing of the PCR product was performed in both directions with the same primers as for PCR using BigDye Terminator v3.1 kit (Thermo Fisher) and DNA Analyzer 3730xl (Thermo Fisher).

CGG repeat sequencing

CGG repeat in the 5' UTR of *FMR1* and flanking sequences (127bp 5' and 110bp 3' from the repeat) was amplified from genomic DNA of iPSC clones with discrete repeat sizes using AmplideX FMR1 PCR kit (Asuragen) with non FAM modified primers FMR1-F: TCAGGCGCTCAGCTCCGTTTCGGTTTCA with FMR1-R: AAGCGCCATTGGAGCCCCGCACTTCC according to the kit instructions. PCR products were purified with DNA clean & concentrator kit (Zymo Research) and SMRTbell adapters were ligated to them. They were then loaded by diffusion loading onto a Pacific Biosciences RSII sequencer and sequenced using P6-C4 chemistry. Because of the short size of the CGG repeat sequences, we were able to make high quality read-of-insert intramolecular consensuses which were then used for subsequent analysis. The consensus sequences were generated using the cloud-based DNA Nexus computer provider running their cloud app implementation of the PacBio SMRT Analysis v2.3 "read of insert" tool. 500 – 2000 reads containing the full amplicon (CGG repeat and flanking sequences) were obtained from each sample. Alignment of sequences was performed manually using Jalview multiple alignment visualization software (Waterhouse et al., 2009). The alignments were manually inspected for potential alterations in the CGG repeat tract and in the flanking sequences.

Exome sequencing

Exome sequencing was performed for two UFM individuals and their family members indicated in Fig.1 (11 individuals). Primary patient material (buccal swap or PBMC as indicated in the text) was used for genomic DNA

extraction using DNeasy kit (Qiagen). Coding sequences were captured using Agilent SureSelectXT protocol according to manufacturer's instructions. The resulting sequencing libraries were then multiplexed and sequenced on an Illumina HiSeq2500 instrument using TruSeq chemistry with a read length of 2x 76bp. The data is accessible to the FXS research community at the European Genome-Phenome Archive (EGA) (Lappalainen et al., 2015), under accession number EGAS00001001737.

Variant calling followed a standard workflow of Genome Analysis Toolkit (GATK-1.6–11) (McKenna et al., 2010). Sequencing reads were mapped to the human reference genome (GATK b37 bundle version human_g1k_v37_decoy of hg19; GRCh37) using Burrows-Wheeler Aligner (bwa version 0.5.5) (Li and Durbin, 2009) in a paired end mode. Polymerase chain reaction duplicates were removed using picard-tools-1.69 (http://picard.sourceforge.net). Realignment and recalibration were performed using GATK version 1.6-11-g3b2fab9 modules RealignerTargetCreator, IndelRealigner, CountCovariates and TableRecalibration. Reference Indels and SNPs were taken from GATK b37 bundle files Mills_and_1000G_gold_standard.indels.b37.vcf and dbsnp_135.b37.vcf. Variant calling was then performed using GATKlite version 2.3-9-gdcdccbb module UnifiedGenotyper simultaneously for all eleven individuals and restricted to genomic positions covered by the SureSelect kit (bed file provided by the vendor). SNPs and Indels were called in two separate runs. The standard quality parameters stand_call_conf and stand_emit_conf were set to thresholds 30.0 and 10.0 respectively. Low quality variant calls were removed using the GATKlite version 2.3-9-gdcdccbb module VariantFiltration based on quality parameters with separate filtering steps for SNPs and Indels. Filtering rules for exclusion of SNPs were QD < 2, MQ < 40, FS > 60, HaplotypeScore > 13, MQRankSum < -12.5 and ReadPosRankSum < -8. Filtering rules for exclusion of Indels were QD < 2.0, FS > 200 and ReadPosRankSum < -20. Positions with low coverage (less than 66 reads), low quality (QUAL < 30) and long runs of homozygosity (HRun > 5) were excluded. SNPs overlapping with Indels were masked. Clusters of SNPs defined by a clusterSize of 3 with a clusterWindowSize of 10 were also excluded. Annotation of variants was performed using Variant Effect Predictor (VEP) version 2.7 and Ensembl data base version 69 (Cunningham et al., 2015). Variant call vcf files were exported to tabular format using GATKlite version 2.3-9-gdcdccbb module VariantsToTable and further processed wit R version 2.15.2. Final results (variant calls and annotation) were integrated for data mining using Spotfire. (Tibco Software Inc.).

90084 variants (SNPs and indels) were identified in one or both UFMs. A list of 2607 genes broadly implicated in epigenetic processes was assembled using internal databases, publications and online resources ((Bartke et al., 2010; Fodor et al., 2010; Gao et al., 2012; Hamperl and Cimprich, 2014; Hu et al., 2010; Spruijt et al., 2013), http://sciencepark.mdanderson.org/labs/wood/DNA_Repair_Genes). Transcription factors predicted to bind *FMR1* promoter based on search with Genomatix MatIspector software were included in the list (www.genomatix.com). Full list of candidate genes is provided in Table S2. Variants were filtered for ones affecting the coding sequence of these genes (frameshift, initiator codon change, STOP codon change, missense, in-frame sequence change), resulting in 1623 variants. Variants present in at least one FXS individual were excluded (Variants homozygous in UFM and heterozygous in FXS were kept), resulting in 295 variants present in at least one UFM. Further, 68 variants affecting the same gene (same or different variant) in both UFM individuals were selected and are provided in Table S3. These variants were further evaluated for the impact on the protein function using manual sequence inspection, literature, and SIFT and Polyphen scores for missense mutations.

Supplemental References

Bartke, T., Vermeulen, M., Xhemalce, B., Robson, S.C., Mann, M., and Kouzarides, T. (2010). Nucleosome-interacting proteins regulated by DNA and histone methylation. Cell *143*, 470-484.

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

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., *et al.* (2015). Ensembl 2015. Nucleic Acids Res *43*, D662-669.

Filipovic-Sadic, S., Sah, S., Chen, L., Krosting, J., Sekinger, E., Zhang, W., Hagerman, P.J., Stenzel, T.T., Hadd, A.G., Latham, G.J., *et al.* (2010). A novel FMR1 PCR method for the routine detection of low abundance expanded alleles and full mutations in fragile X syndrome. Clin Chem *56*, 399-408.

Fodor, B.D., Shukeir, N., Reuter, G., and Jenuwein, T. (2010). Mammalian Su(var) genes in chromatin control. Annu Rev Cell Dev Biol *26*, 471-501.

Gao, Z., Zhang, J., Bonasio, R., Strino, F., Sawai, A., Parisi, F., Kluger, Y., and Reinberg, D. (2012). PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. Mol Cell *45*, 344-356.

Hamperl, S., and Cimprich, K.A. (2014). The contribution of co-transcriptional RNA:DNA hybrid structures to DNA damage and genome instability. DNA Repair (Amst) *19*, 84-94.

Hu, X.V., Rodrigues, T.M., Tao, H., Baker, R.K., Miraglia, L., Orth, A.P., Lyons, G.E., Schultz, P.G., and Wu, X. (2010). Identification of RING finger protein 4 (RNF4) as a modulator of DNA demethylation through a functional genomics screen. Proc Natl Acad Sci U S A *107*, 15087-15092.

Lanni, S., Goracci, M., Borrelli, L., Mancano, G., Chiurazzi, P., Moscato, U., Ferre, F., Helmer-Citterich, M., Tabolacci, E., and Neri, G. (2013). Role of CTCF protein in regulating FMR1 locus transcription. PLoS Genet *9*, e1003601.

Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J.D., Ur-Rehman, S., Saunders, G., Kandasamy, J., Caccamo, M., Leinonen, R., *et al.* (2015). The European Genome-phenome Archive of human data consented for biomedical research. Nat Genet *47*, 692-695.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res *20*, 1297-1303.

Naumann, A., Hochstein, N., Weber, S., Fanning, E., and Doerfler, W. (2009). A distinct DNA-methylation boundary in the 5'- upstream sequence of the FMR1 promoter binds nuclear proteins and is lost in fragile X syndrome. Am J Hum Genet *85*, 606-616.

Smurnyy, Y., Cai, M., Wu, H., McWhinnie, E., Tallarico, J.A., Yang, Y., and Feng, Y. (2014). DNA sequencing and CRISPR-Cas9 gene editing for target validation in mammalian cells. Nat Chem Biol *10*, 623-625.

Spruijt, C.G., Gnerlich, F., Smits, A.H., Pfaffeneder, T., Jansen, P.W., Bauer, C., Munzel, M., Wagner, M., Muller, M., Khan, F., *et al.* (2013). Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. Cell *152*, 1146-1159.

Tsoukas, C.D., Landgraf, B., Bentin, J., Valentine, M., Lotz, M., Vaughan, J.H., and Carson, D.A. (1985). Activation of resting T lymphocytes by anti-CD3 (T3) antibodies in the absence of monocytes. J Immunol *135*, 1719-1723.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189-1191.