Human Genetics
https://doi.org/10.1007/s00439-023-02626-1

ORIGINAL INVESTIGATION

Regulation potential of transcribed simple repeated sequences in developing neurons

Tek Hong Chung[1] · Anna Zhuravskaya[1] · Eugene V. Makeyev[1]

Received: 19 July 2023 / Accepted: 28 November 2023

Abstract

Simple repeated sequences (SRSs), defined as tandem iterations of microsatellite- to satellite-sized DNA units, occupy a substantial part of the human genome. Some of these elements are known to be transcribed in the context of repeat expansion disorders. Mounting evidence suggests that the transcription of SRSs may also contribute to normal cellular functions. Here, we used genome-wide bioinformatics approaches to systematically examine SRS transcriptional activity in cells undergoing neuronal differentiation. We identified thousands of long noncoding RNAs containing >200-nucleotide-long SRSs (SRS-lncRNAs), with hundreds of these transcripts significantly upregulated in the neural lineage. We show that SRS-lncRNAs often originate from telomere-proximal regions and that they have a strong potential to form multivalent contacts with a wide range of RNA-binding proteins. Our analyses also uncovered a cluster of neurally upregulated SRS-lncRNAs encoded in a centromere-proximal part of chromosome 9, which underwent an evolutionarily recent segmental duplication. Using a newly established in vitro system for rapid neuronal differentiation of induced pluripotent stem cells, we demonstrate that at least some of the bioinformatically predicted SRS-lncRNAs, including those encoded in the segmentally duplicated part of chromosome 9, indeed increase their expression in developing neurons to readily detectable levels. These and other lines of evidence suggest that many SRSs may be expressed in a cell type and developmental stage-specific manner, providing a valuable resource for further studies focused on the functional consequences of SRS-lncRNAs in the normal development of the human brain, as well as in the context of neurodevelopmental disorders.

Introduction

Different types of repeats occupy >50% of the ~3 billion nucleotide-long human genome, while only ~1.5% of its capacity is used to encode proteins (Lander et al. 2001; Nurk et al. 2022). The repeated sequences have been traditionally classified as interspersed and tandem repeats, based on their relative positions and the mechanisms driving their expansion. Interspersed repeats often arise through the "selfish" propagation of transposable and retrotransposable elements and their associated sequences (Goodier and Kazazian 2008; Gorbunova et al. 2021). Tandem repeats range from duplications of gene- or gene cluster-sized units to head-to-tail iterations of shorter DNA sequences. Gene duplications can enhance protein activity by increasing the protein production rate or contribute to the processes of evolutionary innovation and speciation (Kent et al. 2003; Lynch and Conery 2000; Nei and Rooney 2005). A notable example of tandem duplication of non-protein-coding genes is provided by the arrays of 47S/45S ribosomal RNA (rRNA) repeats, which are required to sustain the high levels of ribosome biogenesis in the cell (McStay 2023; Nemeth and Grummt 2018).

The biological functions of shorter tandem repeats are less well understood. Depending on the length of their repeated units, this group is often classified into microsatellites (≤9-bp units), minisatellites (10–60-bp units), and satellites (>60-bp units; Wright and Todd 2023). These boundaries may differ depending on the study and the organism, and alternative terms such as short tandem repeats and simple sequence repeats are often used as synonyms for microsatellites. To avoid confusion, we will collectively refer to all tandem repeats with microsatellite-, minisatellite-, or satellite-sized units as "simple repeated sequences" (SRSs).

Defined in this way, SRSs account for ~7.5% of the total human genome length (Nurk et al. 2022). These repeats, particularly satellites, tend to be enriched in the transcriptionally repressed heterochromatin and associated with

✉ Eugene V. Makeyev
eugene.makeyev@kcl.ac.uk

[1] Centre for Developmental Neurobiology, New Hunt's House, King's College London, London SE1 1UL, UK

Published online: 28 December 2023

genome maintenance functions (Altemose 2022; Ugarkovic et al. 2022). However, at least some SRSs are known to be expressed, giving rise to biologically active transcripts (Ninomiya and Hirose 2020; Trigiante et al. 2021). Since individual SRSs contain multiple repeated units, transcripts containing these sequences can form multivalent contacts with cognate proteins or nucleic acid sequences. This in turn may allow SRS-containing transcripts to act as regulators of cellular RNA metabolism or/and contribute to the assembly of large ribonucleoprotein complexes and membraneless compartments.

Many SRSs are genetically unstable and frequently increase or decrease the number of repeated units as a result of replication or/and recombination errors. The genetic expansion of microsatellites in transcribed genomic regions is known to give rise to toxic RNAs in several neuromuscular and neurodegenerative disorders (Baud et al. 2022; Ciesiolka et al. 2017; Fujino and Nagai 2022; Schwartz et al. 2021). For example, the type 1 myotonic dystrophy (DM1) is caused by the expansion of the CTG trinucleotide repeat in the 3′ untranslated region (3′UTR) of DMPK gene (Meola and Cardani 2015; Sznajder and Swanson 2019; Yum et al. 2017). The aberrant DMPK transcripts containing >50 and, occasionally, up to thousands of CUG units can form a hairpin structure stabilized by the G-C base-pairing.

Furthermore, expanded CUGs comprise numerous YGCY motifs (CUGCUG) recognized by the muscleblind-like (MBNL) RNA-binding proteins (RBPs; Mankodi et al. 2001; Miller et al. 2000). This allows for multivalent MBNL binding to the mutant DMPK transcripts and the assembly of pathological ribonucleoprotein foci in the nucleus. The sequestration of MBNL proteins and potentially other RBPs results in dysregulation of pre-mRNA splicing and cleavage/polyadenylation, as well as mRNA stability (Goodwin et al. 2015; Jiang et al. 2004; Masuda et al. 2012).

Pathological RNA foci and MBNL sequestration are also observed in the type 2 myotonic dystrophy (DM2), which is caused by the expansion of the CCTG repeats in intron 1 of the CNBP/ZNF9 gene (Liquori et al. 2001; Mankodi et al. 2001; Sznajder and Swanson 2019). Other repeat expansion disorders have been shown to involve the production of toxic RNAs interacting with multiple copies of different RBPs. For example, the expansion of the CAG-repeat in Huntington's disease (Nalavade et al. 2013), the GGGGCC -repeat in amyotrophic lateral sclerosis and frontotemporal dementia (Balendra and Isaacs 2018), and the CGG-repeat in Fragile X-Associated Tremor/Ataxia Syndrome (Cid-Samper et al. 2018; Sellier et al. 2010), have been shown to promote the formation of RNA foci and sequester different protein factors.

The biomedical relevance of SRSs is not limited to repeat expansion disorders. We have previously identified a UC repeat-enriched long noncoding RNA (lncRNA), PNCTR, produced by RNA polymerase I (Pol I) from an intergenic spacer separating tandem copies of 47S rRNA genes (Yap et al. 2018). We showed that PNCTR is often upregulated in transformed cells, driving the assembly of the cancer-specific perinucleolar compartment (PNC). The UC-rich sequence elements within PNCTR enable multivalent binding of the RBP PTBP1, sequestering it within the PNC. This in turn antagonizes the splicing regulation function of PTBP1 and promotes cancer cell survival (Yap et al. 2018). Using a proximity labeling approach, we have recently shown that PNCTR and the PNC are associated with a host of additional proteins involved in nucleic acid metabolism (Yap et al. 2022a, b).

SRS transcription can also contribute to normal cellular and organismal processes. A classic example is provided by the lncRNA XIST required for X chromosome inactivation (XCI) in female mammalian cells (Patrat et al. 2020). XIST contains several conserved SRS-like elements, referred to as A-, F-, B-, C-, D-, and E-repeats, which may facilitate its function through multivalent RBP recruitment (Almeida et al. 2017; Chu et al. 2015; Lu et al. 2020; McHugh et al. 2015; Pintacuda et al. 2017; Sakaguchi et al. 2016). Another example of a lncRNA expressed in healthy cells is the telomeric repeat-containing RNA TERRA. This lncRNA enriched in UUAGGG repeats plays a crucial role in telomere maintenance by recruiting a specific set of proteins and forming R-loop structures with the telomeric DNA (Arnoult et al. 2012; Deng et al. 2009; Graf et al. 2017; Mei et al. 2021; Porro et al. 2014; Silva et al. 2021). Knockdown of TERRA can lead to telomeric defects, chromosome abnormality, and cell death (Barral and Dejardin 2020; Deng et al. 2009).

Similar to telomeres, centromeric and pericentromeric DNA is enriched in SRSs. Transcripts containing centromeric satellite repeats show a considerable sequence diversity across species but play a conserved role in the assembly of the kinetochore and the centromere passenger complex (CPC; Leclerc and Kitagawa 2021; Perea-Resa and Blower 2018). One common type of centromeric repeats in human is known as alpha satellites, which are characterized by a consensus sequence unit of 171 bp in length (McNulty and Sullivan 2018). Transcripts containing alpha satellites, and possibly other centromeric SRSs, have been shown to associate with the CPC components Aurora B and INCENP, the kinetochore component CENP-C, and the centromeric histone 3 variant CENP-A (Blower 2016; Ideue et al. 2014; McNulty et al. 2017; Quenet and Dalal 2014; Wong et al. 2007).

Knockdown of alpha satellite RNA downregulates these centromeric proteins, resulting in mitotic defects and cell cycle arrest (McNulty et al. 2017; Quenet and Dalal 2014; Wong et al. 2007). Additionally, alpha satellite RNA can associate with SUV39H1, which deposits the repressive heterochromatin marks and recruits the heterochromatin protein

1α (HP1α; Johnson et al. 2017). Therefore, transcribed alpha satellites can act as scaffolds, facilitating the recruitment of proteins involved in centromeric and pericentromeric chromatin maintenance and chromosome segregation.

Not all satellite-containing RNAs are expressed constitutively. The transcription of the pericentromeric human satellite III (HSATIII) repeats is induced under heat shock and cytotoxic stress conditions (Hussong et al. 2017). The resultant GGAAT repeat-rich HSATIII transcripts drive the assembly of membraneless nuclear stress bodies (nSBs; Biamonti and Caceres 2009). Examples of proteins localizing to the nSBs include HSF1, Pol II, Scaffold Attachment Factor B (SAFB), SAFB-Like Transcription Modulator (SLTM), Nuclear Receptor Coactivator 5 (NCOA5), hnRNP proteins M, A1, H1, and serine/arginine-rich (SR) proteins such as SRSF1, SRSF7, and SRSF9 (Aly et al. 2019; Denegri et al. 2001; Jolly et al. 2004; Metz et al. 2004; Weighardt et al. 1999). Another nSB component, CDC-like kinase 1 (CLK1), has been shown to mediate SR protein phosphorylation during the recovery from stress (Ninomiya et al. 2020). Phosphorylation of SRSF9 promotes the intron retention in numerous mRNAs, causing their accumulation in the nucleus and dampening the expression of the corresponding genes (Ninomiya et al. 2020).

The above examples indicate that the normal biological functions of SRS-containing transcripts have been extensively studied in proliferating cells. On the other hand, the effects of abnormally expanded SRSs are better understood in diseases affecting differentiated cells, including neurons. The possible roles of transcribed SRSs in healthy neurons remain largely unexplored, partly due to the challenges associated with analyzing repeat-rich sequences. To address this limitation, we systematically identified SRS-containing lncRNA (SRS-lncRNA) candidates upregulated during neuronal development by mining RNA-sequencing (RNA-seq) data. We further validated our bioinformatics predictions using a newly established system for rapid neuronal differentiation of human induced pluripotent stem cells (iPSCs). Our work provides a valuable resource for further studies focused on the roles of repeat transcription in the development and function of the nervous system.

Methods

Culturing human induced pluripotent stem cells (iPSCs)

Healthy donor-derived iPSCs (HipSci, cat# HPSI0314i-cuhk_1) were maintained in a humidified incubator at 37 °C and 5% CO_2 using Essential 8 (Thermo Fisher Scientific, cat# A1517001) or Essential 8 flex (Thermo Fisher Scientific, cat# A2858501) media supplemented with 100 units/ml PenStrep (Thermo Fisher Scientific, cat# 15140122). Essential 8 medium was changed every day, while Essential 8 flex was refreshed every other day. Tissue culture (TC) plates used for culturing iPSCs (typically Thermo Fisher Scientific cat# 140675) were coated by incubating them with 1 μg/cm^2 of vitronectin (Thermo Fisher Scientific, cat# A14700) at room temperature for 1 h. For normal passaging, iPSC colonies were incubated with 0.5 mM EDTA (Thermo Fisher Scientific, cat# 15575020) in DPBS (no calcium, no magnesium; Thermo Fisher Scientific, cat# 14190094) at room temperature for 4–8 min and then gently dissociated by trituration in the growth medium.

DNA constructs

The pZT-C13-L1 and pZT-C13-R1 constructs encoding the left and right TALENs specific to the *CLYBL* safe-harbor locus were a gift from Jizhong Zou (Addgene plasmid #62196 and #62197; Cerbini et al. 2015). The *CLYBL*-specific homology directed repair construct pUCM-CLYBL-hNIL was a gift from Michael Ward (Addgene plasmid #105841; Fernandopulle et al. 2018). We modified the pUCM-CLYBL-hNIL backbone to act as an acceptor locus in the recombination-mediated cassette exchange (RMCE) protocol for the high-efficiency integration of transgenes of interest (Iacovino et al. 2011). We used standard molecular cloning techniques and restriction and modification enzymes from New England Biolabs to substitute the *hNIL* fragment with a *Lox2272*- and *LoxP*-flanked *Cre* recombinase gene linked to the *ΔNeoR* and *PuroR* markers. The map of the resultant pML630 plasmid is provided Supplementary Data S1. The mouse *Ngn2*-encoding RMCE knock-in plasmid pML156 was generated as described previously (Zhuravskaya et al. 2023).

Generating TRE-Ngn2 iPSCs

We prepared TRE-Ngn2 iPSCs expressing an *Ngn2* transgene from a Dox-inducible promoter using a two-step approach previously used for mouse ESCs (Iacovino et al. 2011). In our case, the first step involved knocking in the rtTA-2Lox-Cre cassette encoded by pML630 into the *CLYBL* safe-harbor locus, and the second step, high-efficiency RMCE substituting the *Cre* coding sequence with the pML156-encoded *Ngn2*.

In the first step, a ~20%-confluent wild-type iPSC culture in a 12-well plate (Corning, cat# 3513) containing 1 ml/well of Essential 8 supplemented with 10 μM ROCK inhibitor (Merck, cat# Y0503) was co-transfected with the pZT-C13-L1 and pZT-C13-R1 plasmids (Cerbini et al. 2015) and the pML630 construct mixed in the 1:1:8 ratio, respectively. We combined 2 μg of the plasmid mixture with 2 μl of Lipofectamine Stem Transfection Reagent (Thermo Fisher

Scientific, cat# STEM00008) and 100 µl of Opti-MEM I (Thermo Fisher Scientific, cat# 31985070), as recommended. The resultant transfection mixture was then added drop-wise to the cells. The medium was refreshed on the next day. To select knock-in clones, 0.25 µg/ml puromycin was added to the medium 2 days post transfection and gradually increased to 0.75 µg/ml by day 12. Puromycin-resistant colonies were picked 12 days post transfection, expanded, and genotyped by PCR using the MLO3670/MLO3671 and MLO3686/MLO1631 primer pairs (Table S1). We also confirmed that the clones express *Cre* in a Dox-inducible manner by reverse transcription (RT) qPCR using human *GAPDH* as the "housekeeping" control (see Table S1 for primer sequences).

In the second step, the rtTA-2Lox-Cre knock-in cells were pre-treated overnight with 2 µg/ml doxycycline (Dox; Sigma, cat# D9891) to activate *Cre* expression. The cells were then transfected with a mixture containing 1 µg of pML156, 2 µl of the Lipofectamine Stem Transfection Reagent, and 100 µl of Opti-MEM I, as described above. To select RMCE-positive clones, 25 µg/ml G418/geneticin (Sigma, cat# 10,131,019) was added to the medium 2 days post transfection and gradually increased to 60 µg/ml by day 12. G418-resistent colonies were picked 2 weeks post-transfection, expanded and genotyped by PCR with the MLO1295/MLO1296 primers (Table S1). We also confirmed the Dox-inducible expression of the *Ngn2* transgene by RT-qPCR with the same primer pair and using human *GAPDH* as the as the "housekeeping" control. Three TRE-Ngn2 iPSC clones selected in this manner were used for the Dox-induced differentiation experiments described below.

Dox-induced neuronal differentiation

We adapted the neurogenin 2-based neuronal differentiation protocol from (Fernandopulle et al. 2018) using plates and coverslips coated for 1 h at 37 °C with Geltrex (Thermo Fisher Scientific, cat# A1413302) diluted to 1% with DMEM/F12 (Thermo Fisher Scientific, cat# 31330038). On differentiation day 0, TRE-Ngn2 iPSCs were dissociated with Accutase (Thermo Fisher Scientific, cat# 00-4555-56) at 37 °C for 5 min and triturated to obtain a single-cell suspension. Cells were then plated onto Geltrex-coated surfaces at 1.5×10^5 cells/cm^2 in the induction medium (DMEM/F12 supplemented with 1× N-2 (Thermo Fisher Scientific, cat# 17,502,048), 1× non-essential amino acids (Thermo Fisher Scientific, cat# 11140035), 1× L-glutamine (Thermo Fisher Scientific, cat# 25030-024), 10 µM ROCK inhibitor, and 2 µg/ml Dox). The medium was replaced on differentiation days 1 and 2, omitting the ROCK inhibitor.

On day 2, we coated fresh 6-well plates (Thermo Fisher Scientific, cat# 140675) or/and 12-well plates with 18-mm coverslips (VWR, cat# 631-1580) with 100 µg/ml poly-L-ornithine (PLO; Merck, cat# A-004-C) and left the plates at 37 °C overnight. On day 3, differentiating cultures were dissociated with Accutase at 37 °C for 5 min and triturated to form a single-cell suspension. Cells were then plated on the PLO-coated plates or coverslips in the maturation medium, which consisted of Neurobasal A (Thermo Fisher Scientific, cat# 10888022), 1× B-27 (Thermo Fisher Scientific, cat# 17504044), 10 µg/ml BDNF (Miltenyi Biotec, cat# 130–093-811), 10 µg/ml NT-3 (Miltenyi Biotec, cat# 130-093-973), and 1 µg/ml laminin (Merck, cat# L2020). Dox was omitted from the medium beginning from day 3 and half-medium changes were carried out twice a week until day 14.

Immunofluorescence

Cells grown on 18-mm coverslips were washed briefly in PBS, fixed with 4% formaldehyde (Thermo Fisher Scientific, cat# 28908), washed three times with PBS, permeabilized with 0.1% Triton X-100 in PBS for 10 min, and washed three times with PBS. Fixed and permeabilized cells were then blocked in PBS containing 0.5% BSA and 0.2% Tween-20 (IF blocking buffer) for 30 min at room temperature and incubated with anti-NGN2 (Cell Signaling Technology, cat# 13144; 1:250 dilution) and/or anti-MAP2 (Biolegend, cat# 822501; 1:2000 dilution) antibodies in the IF blocking buffer at 4 °C overnight. The samples were washed once with 0.2% Tween-20 in PBS and twice with PBS, and incubated with Alexa Fluor-conjugated secondary antibodies in IF blocking buffers for 1 h at room temperature. This was followed by one wash with 0.2% Tween-20 in PBS and two washes with PBS. The coverslips were then counterstained with 0.5 µg/ml DAPI in PBS for 3 min and mounted with ProLong Gold antifade reagent (Thermo Fisher Scientific, cat# P36934). Images were taken using a ZEISS Axio Observer Z1 Inverted Microscope with a LD Plan-Neofluar 40x/0.6 Corr Ph 2 M27 objective.

RNA fluorescence in situ hybridization (RNA-FISH)

In some experiments, immunofluorescently stained cells were post-fixed with 4% formaldehyde for 15 min at room temperature, washed three times with PBS, and transferred to the pre-hybridization buffer containing 10% formamide (Thermo Fisher Scientific, cat# 15515026) and 2× SSC (Thermo Fisher Scientific, cat# AM9763). The samples were then hybridized at 37 °C overnight with a 125-nM mixture of RNA target-specific oligonucleotide probes (Table S1) labeled with digoxigenin (Sigma-Aldrich, cat# 03353575910), which were dissolved in the hybridization buffer containing 10% formamide, 2×SSC, and 10% dextran sulfate (Sigma-Aldrich, cat# D8906). Following hybridization, the samples were washed for 30 min in 2× SSC and

10% formamide at 37 °C, and 15 min in 1× SSC at room temperature. Subsequently, the samples were incubated with mouse anti-digoxigenin antibody (Jackson Laboratories, cat# 200-002-156, 1:500) in 4× SSC and 0.8% BSA for 1 h at 37 °C. Following this incubation, additional washes were performed with 4× SSC, 4× SSC and 0.1% Triton X-100, and 4× SSC at room temperature for 10 min each. The samples were then incubated with Alexa Fluor-647-conjugated anti-mouse secondary antibodies (Thermo Fisher Scientific, cat# A31571, 1:300) in 4× SSC and 0.8% BSA for 1 h at room temperature, followed by washes with 4× SSC, 4× SSC and 0.1% Triton X-100, and 4× SSC at room temperature for 10 min each. Finally, the samples were counterstained with 0.5 µg/ml DAPI in PBS for 3 min and mounted with ProLong Gold antifade reagent. Images were captured using the same ZEISS system as described above, switching to an alpha Plan-Apochromat 100x/1.46 Oil DIC (UV) M27 objective.

PCR-based assays

PCR genotyping was carried out using the PCRBIO HS Taq Mix Red kit (PCR Biosystems, cat# PB10.23-02), as recommended by the manufacturer. To perform RT-qPCR assays, total RNAs were extracted with the EZ-10 DNAaway RNA Mini-Preps kit (Bio Basic, cat# BS88136) as recommended, and reverse transcribed in a 20-µl format. Prior to reverse transcription, traces of genomic DNA were removed by treating 600 ng of total RNA with 1 µl RQ1 RNase-Free DNase (Promega, cat# M6101) in 9 µl reaction additionally containing RQ1 DNase buffer and 1 µl of murine RNase inhibitor (New England Biolabs, cat# M0314L) at 37 °C for 45 min. RQ1 DNase was inactivated by adding 1 µl of Stop Solution from the RQ1 kit and incubating the solution at 65 °C for 10 min. The reaction was placed on ice, supplemented with 1 µl of 100 µM random decamer primers, incubated at 70 °C for 10 min, and returned to ice. The reaction was then supplemented with 1× SuperScript IV reaction buffer, 10 mM DTT, 0.5 mM each of the four dNTPs, 1 µl murine RNase inhibitor, 1 µl of SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific, cat# 18090200), and nuclease-free water (Thermo Fisher Scientific, cat# AM9939) added to the final volume of 20 µl. In RT-negative controls, reverse transcriptase was substituted with an equal volume of nuclease-free water. cDNA was synthesized by incubating the RT mixtures at 50 °C for 40 min, followed by a 10-min 70 °C heat inactivation step. The samples were finally diluted tenfold by adding 180 µl of nuclease-free water and stored at −80 °C until needed. Quantitative PCR (qPCR) was performed using qPCRBIO SyGreen Mix Lo-ROX (PCR Biosystems, cat# PB20.11-51), 100 nM of each primer and 5 µl of diluted cDNA in 20 µl reactions. Reactions were run on a LightCycler 96 Instrument (Roche). The *YWHAZ* gene was used as the "housekeeping" control. Primers used for PCR genotyping and RT-qPCR are listed in Table S1.

Gapmer transfection experiments

Day-7 differentiating TRE-Ngn2 cultures grown in the 12-well plate format were transfected with 50 pmol/well of an XLOC_312995-specific (gmXLOC_312995; Qiagen, cat# 339511 LG00797536-DDA; 5′-G*T*A*G*T*A*G*G*T*G*G*T*C*T*T*T; the asterisks indicate phosphorothioate bonds; the positions of the LNA modifications are not provided by the manufacturer) or a negative control gapmer (gmControl; Qiagen, cat# 339516 LG00000002-DFA; 5′-A*A*C*A*C*G*T*C*T*A*T*A*C*G*C) mixed with 2 µl of Lipofectamine 3000 (Thermo Fisher Scientific, cat# L3000001), according to the manufacturer's protocol. Total cellular RNAs were extracted 48 h post-transfection and analyzed by RT-qPCR, as described above.

Bioinformatics analyses

A flow chart summarizing our bioinformatic analyses is shown in Fig. 1A. RNA-seq data for the nuclear and cytoplasmic fractions of human embryonic stem cells (ESCs), neural progenitor cells (NPCs), and differentiation day-14 and day-50 neurons were downloaded from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100007; Blair et al. 2017). RNA-seq data for H1 human ESCs (https://doi.org/10.17989/ENCSR895ZTB and https://doi.org/10.17989/ENCSR712GOC) and in-vitro differentiated neurons (https://doi.org/10.17989/ENCSR877FRY and https://doi.org/10.17989/ENCSR877FRY), astrocytes (https://doi.org/10.17989/ENCSR129VBC), endothelial cells (https://doi.org/10.17989/ENCSR429EGC), and embryonic endodermal cells (https://doi.org/10.17989/ENCSR559HWG) were form Encode (https://www.encodeproject.org/). RNA-seq data for human cortical organoids were from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE106245 (Fiddes et al. 2018), and autism spectrum disorder and control iPSC-derived neurons, from ASD https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE124308 (DeRosa et al. 2018). HITS-CLIP data for neuronal ELAVL proteins were from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53699 (Scheckel et al. 2016). Quality control and adaptor trimming were performed using Trimmomatic (v0.39; Bolger et al. 2014) as follows:

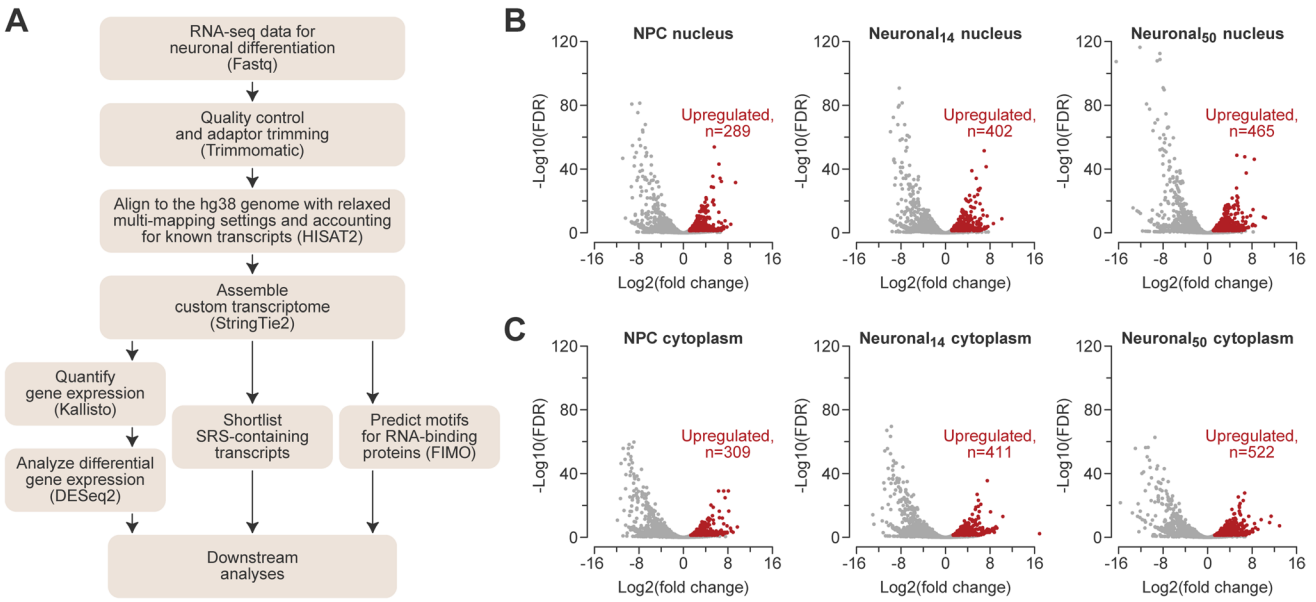


Fig. 1 Systematic discovery of SRS-lncRNAs expressed during neuronal differentiation. **A** The pipeline used to identify regulated SRS-containing transcripts. Volcano plots for SRS-lncRNA expression changes in **B** the nuclear and **C** the cytoplasmic fractions at the NPC and day-14 ($Neuronal_{14}$) and day-50 ($Neuronal_{50}$) neuronal differentiation stages, compared to ESC. Significantly upregulated SRS-lncRNAs are highlighted in red (colour figure online)

```
# Paired-end data
java -jar trimmomatic-0.39.jar PE -threads <threads> \
<input_1.fastq.gz> <input_2.fastq.gz> \
<paired_output_1.fastq.gz> <unpaired_output_1.fastq.gz> \
<paired_output_2.fastq.gz> <unpaired_output_2.fastq.gz> \
LEADING:10 TRAILING:10 SLIDINGWINDOW:5:20 MINLEN:86 \
ILLUMINACLIP:<adaptor.fa>:2:30:10:2:true

#Single-end data
java -jar trimmomatic-0.39.jar SE -threads <threads> \
<input.fastq.gz> \
<output.fastq.gz> \
LEADING:10 TRAILING:10 SLIDINGWINDOW:5:20 MINLEN:86 \
ILLUMINACLIP:<adaptor.fa>:2:30:10
```

Trimmed FASTQ files were aligned to the GRCh38 genome assembly using HISAT2 (v2.2.1; Pertea et al. 2016), allowing for RNA-seq read multi-mapping.

```
#List of known splice sites
python hisat2_extract_splice_sites.py <Gencode V32 annotation GTF file> >
<list_of_known_splice_sites.txt>

#Paired-end data
hisat2 -p <threads> -k 100 --fr --rna-strandness RF --dta-cufflinks --no-unal --no-mixed \
--no-discordant --known-splicesite-infile <list_of_known_splice_sites.txt> \
-x <hg38 genome> \
-1 <trimmomatic paired_output_1.fastq.gz> \
-2 <trimmomatic paired_output_2.fastq.gz> \
-S <output.sam>

#Single-end data
hisat2 -p <threads> -k 100 --rna-strandness R --dta-cufflinks --no-unal \
--no-discordant --known-splicesite-infile <list_of_known_splice_sites.txt> \
-x <hg38 genome> \
-U <trimmomatic output.fastq.gz> \
-S <output.sam>
```

After converting the output SAM files into indexed BAM files using Samtools (v1.11; Li et al. 2009), sample-specific transcriptomes were assembled using StringTie2 (v2.1.4; Pertea et al. 2016) and then merged by Cuffmerge (Trapnell et al. 2010).

```
stringtie -p <threads> --rf -m 200 -M 1 -c 1 -s 2 -f 0.05 \
-G <Gencode V32 annotation GTF file> -o <output.gtf> <input.bam>

cuffmerge -p <threads> --min-isoform-fraction 0.05 -o <output directory> \
-g <Gencode V32 annotation GTF file> <list_of_gtfs.txt>
```

The transcripts were quantified using Kallisto (v0.44.0; Bray et al. 2016) as follows:

```
#Paired-end data
kallisto quant -i <transcriptome index> --bias --single-overhang --rf-stranded -t <threads> \
-g <cuffmerge_output.gtf> \
-c <chromosome_length.txt> \
-o <output directory> \
<trimmomatic paired_output_1.fastq.gz> <trimmomatic paired_output_2.fastq.gz>

#Single-end data
kallisto quant -i <transcriptome index> --bias --single-overhang --rf-stranded --single \
-l 200 -s 50 -t <threads> \
-g <cuffmerge_output.gtf> \
-c <chromosome_length.txt> \
-o <output directory> \
<trimmomatic output.fastq.gz>
```

Differential gene expression analyses were performed in DESeq2 using Wald's test (Love et al. 2014). Genes with $\log_2$(fold change) > 1 and BH-adjusted *P* value (FDR) < 0.05 were considered significantly upregulated, whereas genes with $\log_2$(fold change) < −1 and FDR < 0.05 were considered downregulated. Low-abundance transcripts expressed at <0.1 transcripts per million (TPM) in the nuclear fraction in all samples were excluded from subsequent analyses.

The transcript_type tag "protein_coding" from the Cuffmerge GTF file output was used to define transcripts and their corresponding genes as protein-coding. The remaining transcripts/genes were considered non-coding. To shortlist SRS-containing lncRNAs, Simple Repeat track data were downloaded using the UCSC Genome Browser table browser tool selecting the Simple Repeat track for the GRCh38/hg38 genome assembly. Only repeated sequences of >200 bp in length and comprising ≥3 repeated units were included in downstream analyses. Transcripts consisting entirely of repetitious sequences were filtered out because of the inherent difficulty in quantifying their abundance. We used the same strategy to categorize protein-coding genes as containing or lacking SRSs in the transcribed region in the analyses illustrated in Fig. 3.

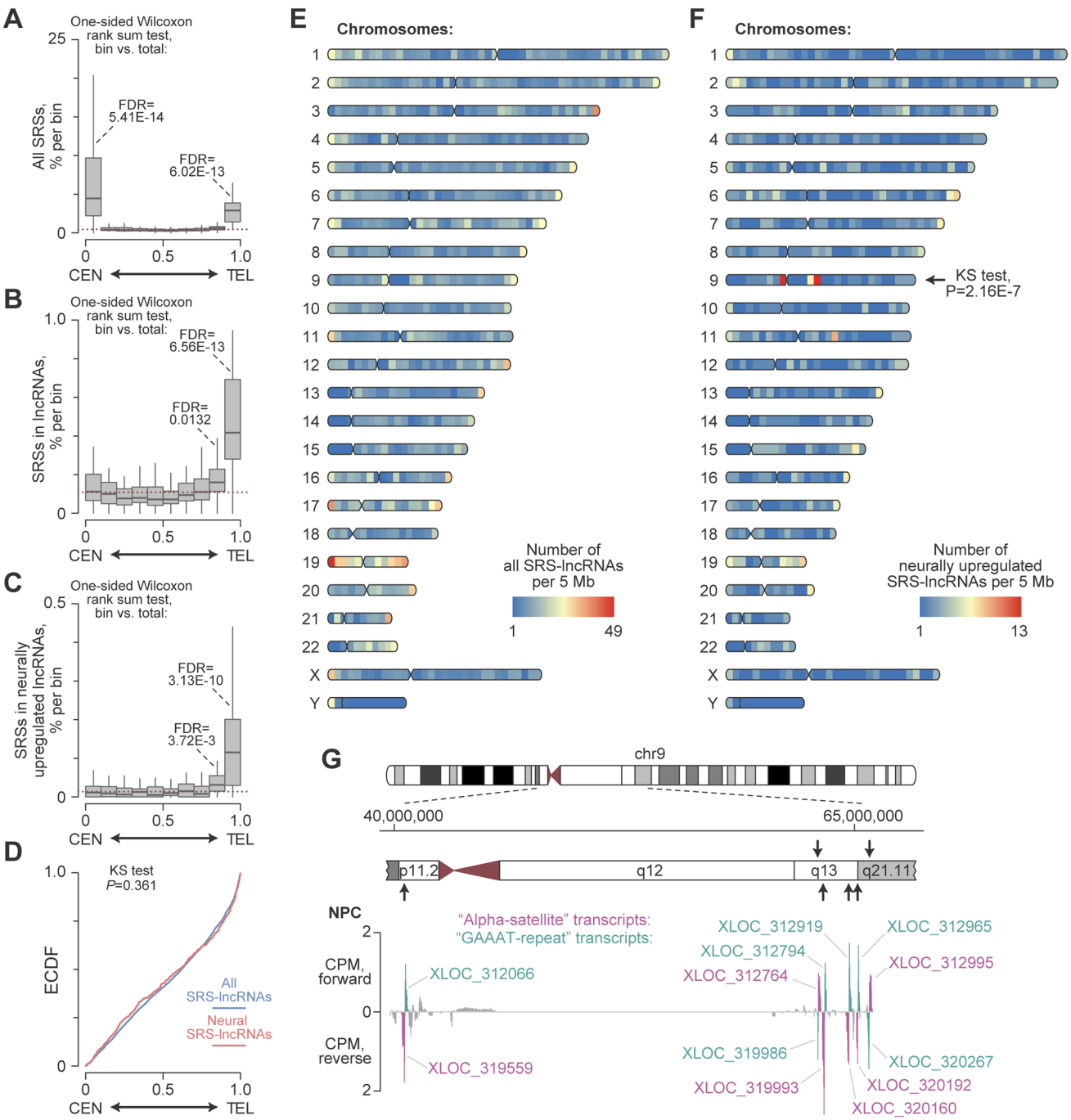
To identify RBP-specific sequence motifs, we used FIMO (Bailey et al. 2015) with the RBP motif data from the CISBP-RNA database (Ray et al. 2013):

```
fimo --text --no-qvalue --norc --thresh 0.001 --motif-pseudo 0.1 \
--max-stored-scores 100000000 \
--bgfile <markov1.b file> \
-oc <output directory> \
<motif file> \
<gene fasta file>
```

To process nELAVL HITS-CLIP data (Scheckel et al. 2016), sequencing reads were aligned to the hg38 genome, and the CLIP peaks were identified using the CLIP Tool Kit (CTK; Shah et al. 2017). Briefly, BAM files were first converted to the BED format and pooled:

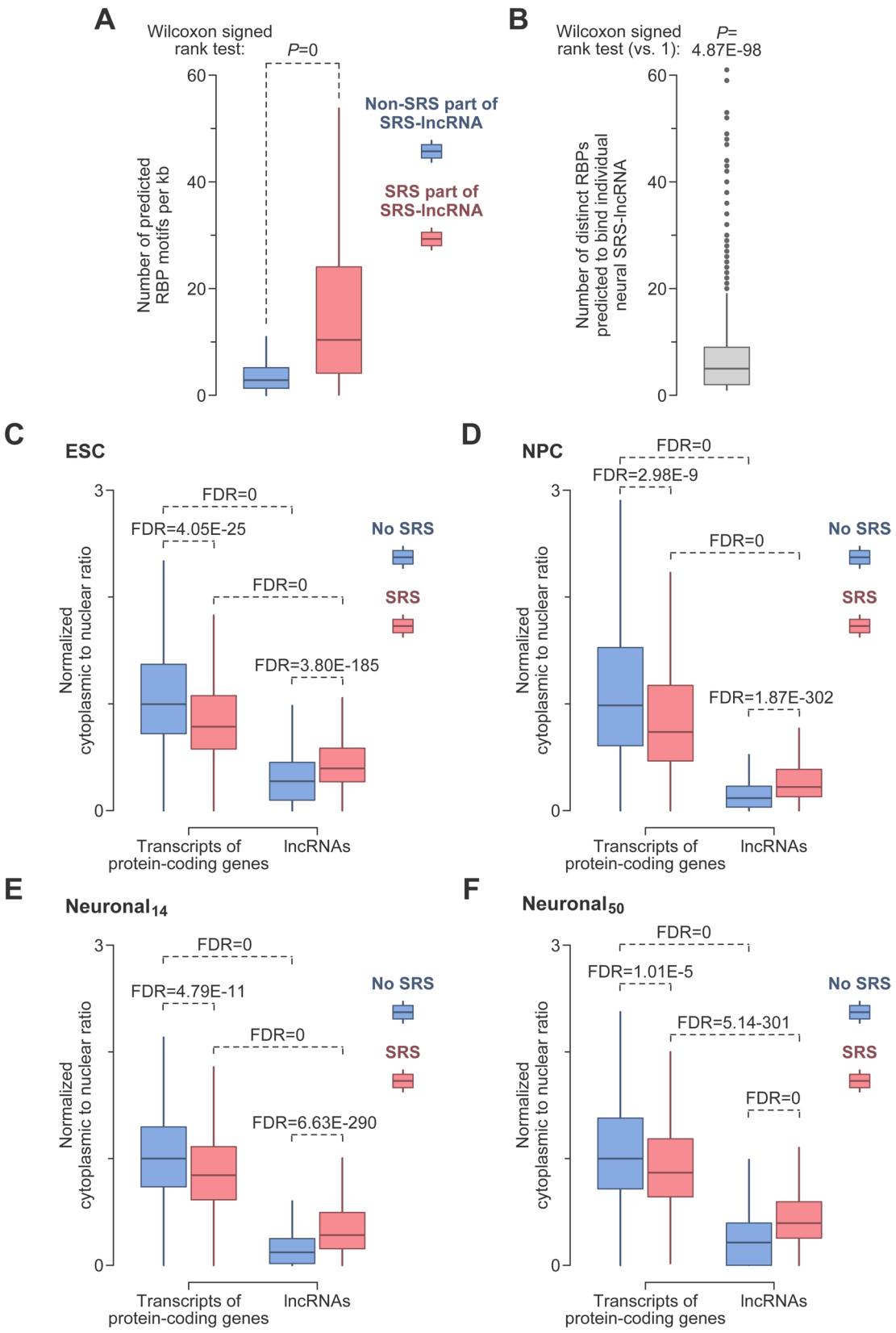
Fig. 2 Genomic distribution of SRS-lncRNAs. Box plots showing the distribution of **A** all genome-encoded SRSs of >200 bp in length and containing ≥3 repeated units, **B** SRSs in all detectably expressed lncRNAs, and **C** SRSs in neurally upregulated lncRNAs along the human chromosome arms separated into 10 equally sized bins, form the middle of the centromere (position 0) to the end of the telomere (position 1). The p-arms of the acrocentric chromosomes 13, 14, 15, 21 and 22 encoding the *47S/45S* rRNA arrays were excluded from these analyses. **D** The Kolmogorov–Smirnov (KS) test shows that the distributions of all SRS-lncRNAs (blue) and neural SRS-lncRNAs (red) along the centromere-to-telomere axis do not significantly differ on a genome-wide scale. The data are presented as empirical cumulative distribution function (ECDF) plots. Karyoplots showing the color-coded density of **E** all detectably expressed SRS-lncRNAs and **F** neurally upregulated SRS-lncRNAs, calculated for individual human chromosomes as a number of loci per 5-Mb window. The centromere-to-telomere distributions of the neural and all detectably expressed SRS-lncRNAs were compared for individual chromosome arms using the KS test. A significant difference was detected only for chromosome 9 (chr9). The panels were generated using the Rideogram package (https://cran.r-project.org/web/packages/RIdeogram/vignettes/RIdeogram.html). **G** A close-up of the centromere-proximal part of chr9 encoding twelve neurally upregulated SRS-lncRNAs organized as six segmentally duplicated bidirectional transcription units (arrows). The panel also displays strand-specific counts-per-million (CPM) normalized coverage plots for the nuclear NPC RNA-seq data. The alpha satellite- and GAAAT repeat-containing SRS-lncRNAs, which constitute the bidirectional unit, are highlighted in magenta and teal, respectively. The coverage plot on the top represents RNAs transcribed in the forward direction, while the bottom part shows RNAs transcribed in the reverse direction (colour figure online)

```
bedtools bamtobed -split -i <input.bam> > <bamtobed_output.bed>
bed2rgb.pl -v -col "<0-255>,<0-255>,<0-255>" <bamtobed_output.bed>
<bed2rgb_output.bed>
cat <all bed2rgb_output.bed files> > <pooled.bed>
```



CLIP peaks were then called as follows:

```
tag2peak.pl -big -ss -v --valley-seeking -p 0.05 --valley-depth 0.9 --gene <gene.bed> \
--multi-test <pooled.bed> \
<peaks.bed> \
--out-boundary <peak_boundaries.bed> \
--out-half-PH <half_peak_height.bed>
```


A
Wilcoxon signed rank test: P=0
Number of predicted RBP motifs per kb
60
40
20
0
Non-SRS part of SRS-lncRNA
SRS part of SRS-lncRNA
B
Wilcoxon signed rank test (vs. 1): P= 4.87E-98
Number of distinct RBPs predicted to bind individual neural SRS-lncRNA
60
40
20
0
C
ESC
Normalized cytoplasmic to nuclear ratio
3
0
FDR=0
FDR=4.05E-25
FDR=0
FDR=3.80E-185
No SRS
SRS
Transcripts of protein-coding genes
lncRNAs
D
NPC
FDR=0
FDR=2.98E-9
FDR=0
FDR=1.87E-302
No SRS
SRS
Normalized cytoplasmic to nuclear ratio
Transcripts of protein-coding genes
lncRNAs
E
Neuronal14
FDR=0
FDR=4.79E-11
FDR=0
FDR=6.63E-290
No SRS
SRS
Normalized cytoplasmic to nuclear ratio
Transcripts of protein-coding genes
lncRNAs
F
Neuronal50
FDR=0
FDR=1.01E-5
FDR=5.14-301
FDR=0
No SRS
SRS
Normalized cytoplasmic to nuclear ratio
Transcripts of protein-coding genes
lncRNAs

◂**Fig. 3** Bioinformatics analyses of SRS-lncRNAs. **A** The density of predicted RBP motifs is significantly higher in the SRS parts of SRS-lncRNAs compared to the non-SRS parts of the same transcript. **B** Individual SRS-lncRNAs expressed in the neural lineage frequently contain ≥10 interaction motifs for more that one type of RBP. Ratios between transcript abundances in the cytoplasm and the nucleus for protein-coding and lncRNA genes containing or lacking SRSs calculated for **C** ESC, **D** NPC, **E** $Neuronal_{14}$ and **F** $Neuronal_{50}$ samples. All ratios are normalized by the differentiation stage-specific median of the non-SRS protein-coding group. The data are compared by the Kruskal–Wallis test ($P=0$ for all panels) with Dunn's post hoc test (P values shown in the panels). The data in **A**–**F** are presented as box plots, with outliers shown in **B** but not the other panels

The resultant half_peak_height.bed file was used to estimate the overlap between nELAVL HITS-CLIP peaks and predicted ELAVL3 binding motifs.

Statistical analyses

Statistical analyses were carried out in R (v. 4.2.2; R Core Team 2021). Data were compared using two-sided Student's *t* test or Wilcoxon signed-rank test, or one-sided Wilcoxon rank sum test. Multiple comparisons were adjusted for multiple testing using the Benjamini–Hochberg (FDR) method. Alternatively, we used one-way ANOVA followed by Tukey's post hoc test or the Kruskal–Wallis test followed by Dunn's post hoc test, as indicated in the figures. SRS-lncRNA distributions along the centromeric-to-telomeric axis were compared using Kolmogorov–Smirnov test. Fisher's exact test was used to compare categorical data. P values <0.05 were considered statistically significant.

Results

Widespread regulation of SRS-lncRNA expression during neuronal differentiation

To investigate the transcriptional status of SRSs encoded in the human genome, we examined a publicly available RNA-sequencing (RNA-seq) dataset for human embryonic stem cells (ESCs) undergoing in vitro differentiation into forebrain-specific neurons (Blair et al. 2017). We selected this study since it contains high-quality sequencing data for four differentiation stages: ESCs, neural progenitor cells (NPCs), and neurons at two stages of maturation, 14 days ($Neuronal_{14}$) and 50 days post-differentiation from NPCs ($Neuronal_{50}$). Furthermore, we reasoned that the nuclear- and cytoplasmic-fraction data provided by this study should increase the likelihood of identifying compartment-enriched transcripts.

Our RNA-seq analysis pipeline was designed to handle multi-mapping reads and incorporated reference-guided transcriptome assembly. This allowed us to include both previously annotated and novel RNAs (Fig. 1A; see "Methods" for further details). We defined SRS-containing lncRNAs as transcripts that lack a known protein-coding sequence and contain > 200 nt-long simple repeat(s) (with at least 3 repeated units) sourced from the UCSC Genome Browser database. The choice of the >200-nt cutoff was based on the traditional definition of lncRNAs (Mattick et al. 2023). To mitigate the artifacts of the ambiguous alignment of multi-mapping reads, we excluded newly predicted transcripts composed entirely of repeated sequences. This uncovered a total of 5430 SRS-lncRNA candidates expressed in the nucleus or/and cytoplasm in at least one of the differentiation stages with the >0.1 TPM cutoff.

To identify differentially expressed SRS-lncRNAs, we compared the NPC, $Neuronal_{14}$ or $Neuronal_{50}$ samples to the corresponding ESC controls (Fig. 1A). This revealed 899 non-redundant candidates that were upregulated either in the nucleus or the cytoplasm (>twofold, FDR < 0.05) in at least one neural sample (i.e. in NPC, $Neuronal_{14}$ or $Neuronal_{50}$; Table S2). Among the upregulated SRS-lncRNAs, 289 SRS-lncRNAs were upregulated in the NPC nuclei and 402 and 465 in the $Neuronal_{14}$ and $Neuronal_{50}$ nuclei, respectively. Similarly, 309 SRS-lncRNAs were upregulated in the NPC cytoplasm and 411 and 522 in the $Neuronal_{14}$ and $Neuronal_{50}$ cytoplasm (Fig. 1B, C). Our pipeline also identified 444 ESC-specific candidates that were consistently downregulated at all the neural stages in at least one compartment (>twofold, FDR < 0.05) and not significantly upregulated in the other compartment (Table S3).

The SRS-lncRNAs shortlisted in this manner included previously characterized transcripts with relevant expression patterns. For instance, LINC00632 (XLOC_334612), the precursor of the brain-enriched circular RNA CDR1as/CiRS-7 containing multiple microRNA miR-7-interacting sequences (Barrett et al. 2017; Hansen et al. 2013; Memczak et al. 2013), was upregulated in NPCs and neurons (Fig. S1A; Table S2). Conversely, the lncRNA CPMER (XLOC_185088) involved in cardiomyocyte differentiation and expressed in embryonic stem cells (Lyu et al. 2022), as well as the p53-regulated pluripotency-specific lncRNA LNCPRESS2 (XLOC_222487; Jain et al. 2016) were downregulated in neural samples (Fig. S1B, C; Table S3). These examples provided internal controls for the performance of our pipeline. Based on the inspection of UCSC Genome Browser data, many other shortlisted transcripts were either novel (e.g., Fig. S2A, B) or matched known lncRNAs not previously associated with the neural lineage (e.g., Fig. S2C).

These data suggest that many SRS-containing transcripts change their expression during normal neuronal differentiation.

SRS-lncRNAs upregulated in neural cells originate from specific genomic regions

To gain initial insights into the 899 SRS-lncRNAs upregulated in NPCs or/and neurons, we compared the positions of their loci with those of all SRSs encoded in the genome and all detectably expressed SRS-lncRNAs. Plotting the overall genomic distribution of SRSs along the centromere-to-telomere axis revealed their expected enrichment in the centromere- and telomere-proximal regions (Fig. 2A). In comparison, all detectable SRS-lncRNAs lost the centromere-proximal peak and showed significant enrichment in a broader telomere-proximal region (Fig. 2B). The subset of neural SRS-lncRNAs exhibited a generally similar profile (Fig. 2C, D). Although we excluded the candidates consisting entirely of repetitious sequences from our main analyses, adding such transcripts back to the shortlist had little effect on the overall genomic distribution of neural SRS-lncRNA (Fig. S3; Table S2).

Our further inspection of individual chromosomes revealed specific enrichment of neural SRS-lncRNAs in centromere-proximal parts of chr9 comprising evolutionarily recent segmental duplications (Fig. 2E, F; Bailey et al. 2002; Crosier et al. 2002; Guy et al. 2000). Interestingly, twelve SRS-lncRNAs encoded in this region are organized into six bidirectional transcription units sharing considerable sequence similarity. Within each unit, one SRS-lncRNA typically contains a 10–36 kb-long alpha-satellite sequence, while the SRS-lncRNA transcribed in the opposite direction has a 3–8 kb-long GAAAT-rich microsatellite repeat (Fig. 2G). Of note, the chromosome-specific patterns of neural SRS-lncRNAs were clearly distinct from the distribution of neurally upregulated protein-coding genes (Fig. S4).

Thus, neural SRS-lncRNA loci are encoded in the genome in a non-random manner.

Bioinformatics characterization of neural SRS-lncRNAs

Several previously characterized SRS-containing transcripts have been shown to recruit multiple copies of specific RBPs (e.g., Yap et al. 2018). With this in mind, we conducted a sequence motif analysis and found that 695 out of the 899 neurally upregulated SRS-lncRNAs contain ≥10 putative interaction sites for individual RBPs within their SRS regions (Table S4; see "Methods" for further details). Importantly, the motif density was significantly higher in the SRS regions compared to the non-SRS parts of the same SRS-lncRNA (Fig. 3A). Furthermore, many SRS-lncRNAs were predicted to engage in multivalent interactions with more than one distinct type of RPBs, with a median of 5 different RBPs (Fig. 3B).

To benchmark our motif predictions, we turned to neuronal ELAV-like RBPs (nELAVL; ELAVL2/HuB, ELAVL3/HuC, and ELAVL4/HuD), whose U/G-rich RNA binding sites have been experimentally identified in the human brain using the HITS-CLIP approach (Scheckel et al. 2016). Based on our predictions, 190 SRS-lncRNAs may form multivalent contacts with members of this protein family (see the ELAVL3 motif in Table S4). Notably, a large fraction of the predicted ELAVL3 motifs in SRS-lncRNAs overlapped with the nELAVL HITS-CLIP peaks, and the extent of such overlap was significantly diminished when we computationally "scrambled" the HITS-CLIP peak positions (Fig. S5).

Since many lncRNAs are retained in the nucleus (Guo et al. 2020; Palazzo and Lee 2018; Tong and Yin 2021), we compared the intracellular distribution of SRS-lncRNAs to that of other types of transcripts. As expected, the cytoplasmic-to-nuclear ratio was noticeably lower for SRS-lncRNAs than for transcripts of protein-coding genes with or without >200 nt-long SRSs. However, SRS-lncRNAs were relatively more abundant in the cytoplasm compared to lncRNAs lacking qualifying SRSs. Interestingly, the presence of such repeats in protein-coding genes led to the opposite effect—a decrease in the cytoplasmic-to-nuclear ratio. These effects were observed for transcripts detectably expressed at all four stages of differentiation (Fig. 3C–F; >0.1 TPM in the nuclear fraction at the corresponding stage).

These data suggest that neural SRS-lncRNAs have a strong potential for multivalent RBP recruitment. Furthermore, our analyses indicate that SRSs may control the relative abundance of RNA in the nucleus and cytoplasm in a transcript-dependent manner.

Doxycycline-inducible differentiation of human iPSCs into neurons

To validate the SRS-lncRNA regulation patterns, we generated a stable human iPSC line with a doxycycline (Dox) inducible *neurogenin 2* (*Ngn2*) transgene (Fig. S6A). Ectopic expression of *Ngn2* is known to trigger efficient neuronal differentiation of proliferating stem cells in vitro (Fernandopulle et al. 2018; Lin et al. 2021; Zhang et al. 2013). Inspired by the high-efficiency RMCE-based knock-in technology available for mouse ESCs (Iacovino et al. 2011), we first knocked in an "acceptor" cassette into the *CLYBL* safe-harbor locus of wild-type human iPSCs using the appropriate TALEN and homology-directed repair constructs (Fig. S6A). The cassette contained a puromycin selection marker, a constitutively expressed reverse tetracycline transactivator (rtTA) and a tetracycline/doxycycline responsive element (*TRE*) promoter driving the expression of a *lox2272* and *loxP* site-flanked *Cre* recombinase gene. To enable the subsequent RMCE-dependent integration step, the cassette also encoded a fragment of the G418 resistance

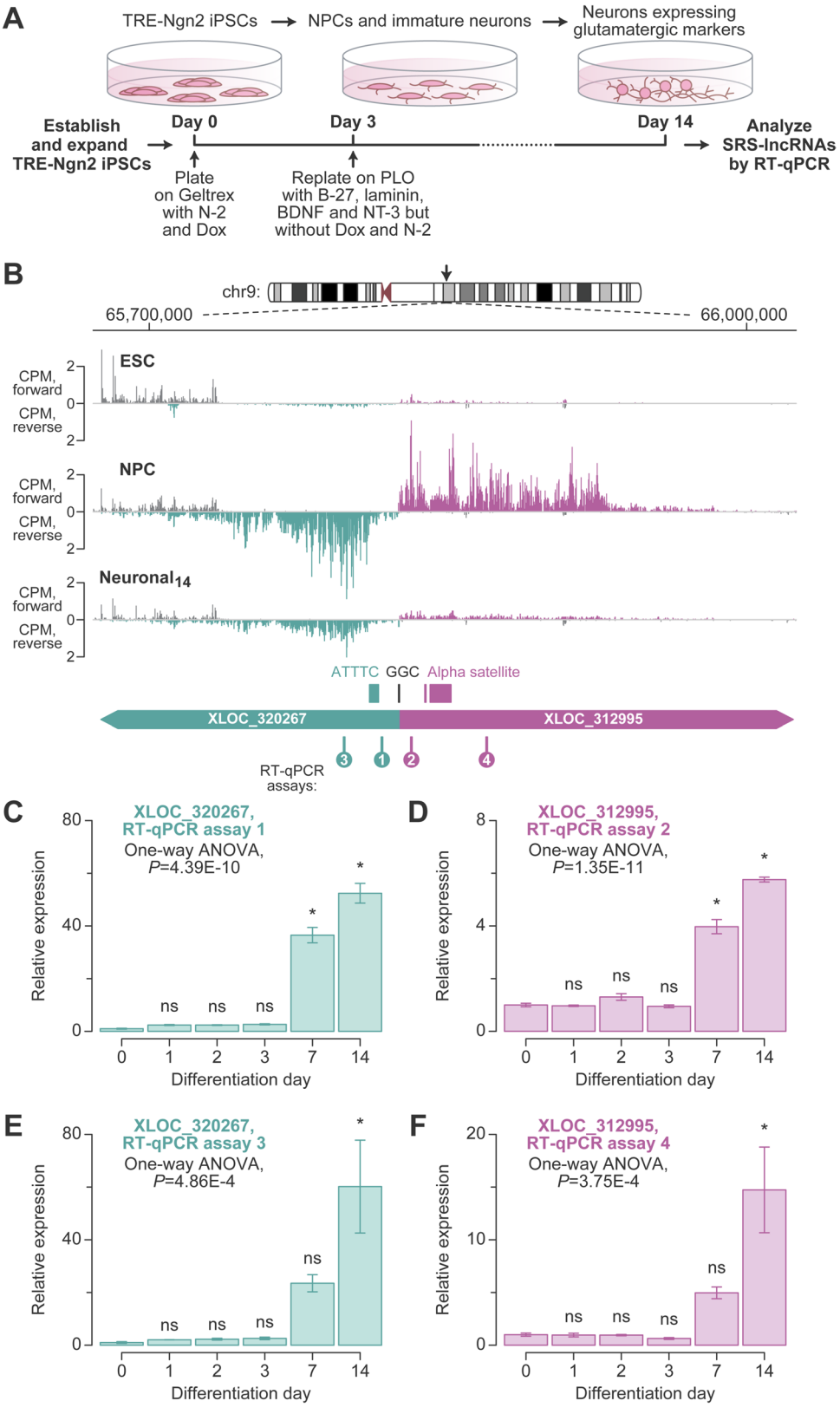


Fig. 4 Neuronal differentiation system for SRS-lncRNA validation. **A** The protocol for Dox-inducible neuronal differentiation of TRE-Ngn2 iPSCs established in this study (see "Methods" for further details). **B** CPM-normalized RNA-seq coverage plots for the nuclear fraction of ESC, NPC, and Neuronal$_{14}$ samples, illustrating one of the six segmentally duplicated chr9 SRS-lncRNA units. The unit encodes the alpha-satellite (XLOC_312995; magenta) and the GAAAT repeat (XLOC_320267; teal) containing SRS-lncRNAs transcribed in the opposite directions from a common GC-rich (GGC) promoter region. The two SRS-lncRNAs and the corresponding SRSs are annotated at the bottom. Note that the microsatellite sequences are shown for the forward strand. The diagram also shows the positions of the four RT-qPCR amplicons (assays 1–4) analyzed in **C**–**F**. RT-qPCR assays validating the regulation of **C** and **E** XLOC_320267 and (D, F) XLOC_312995 expression during neuronal differentiation using primer pairs annealing either upstream (**C**, **D**) or downstream (**E**, **F**) for the corresponding SRSs. Data were averaged for differentiation experiments carried out using three distinct TRE-Ngn2 iPSC clones ±SEM and compared by one-way ANOVA with Tukey's post hoc test. *, Tukey's $P < 0.05$; ns, Tukey's $P \geq 0.05$

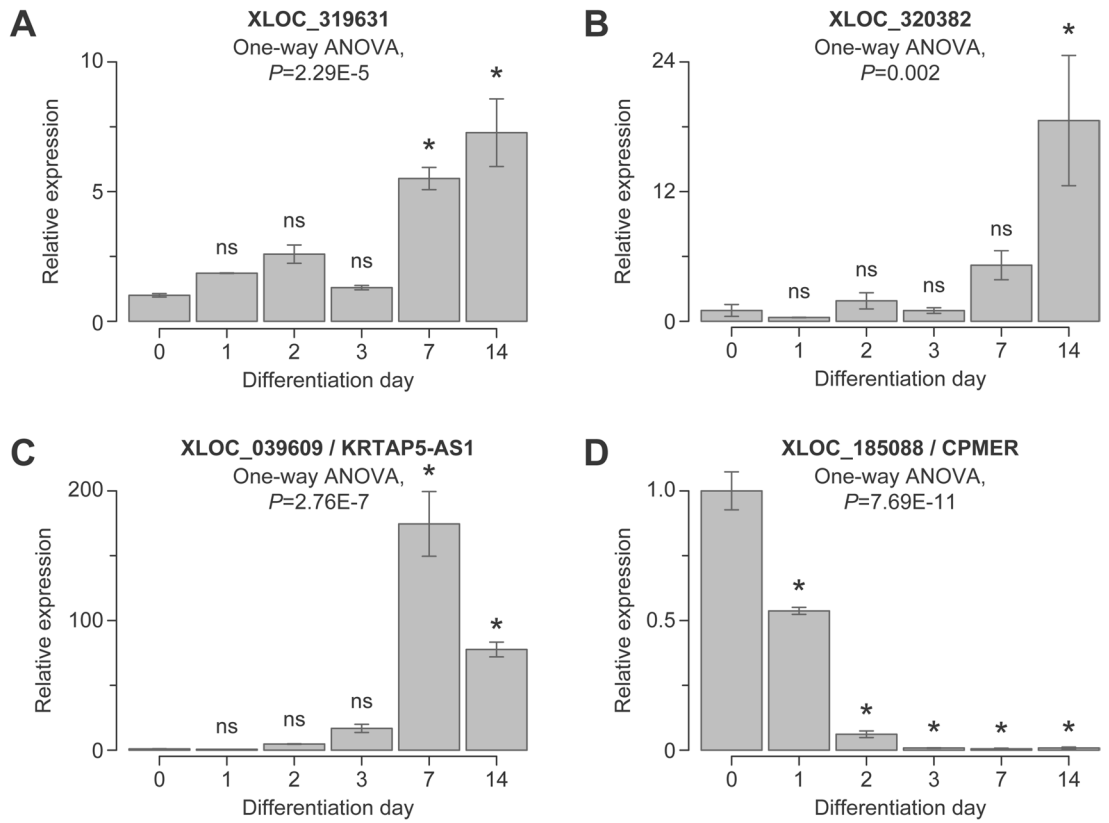


Fig. 5 Experimental validation of SRS-lncRNAs. We used RT-qPCR with SRS-proximal primers to validate the expression dynamics of **A** XLOC_319631, **B** XLOC_320382, and **C** XLOC_039609/KRTAP5-AS1, which are predicted to be upregulated during neuronal differentiation, as well as **D** XLOC_185088/CPMER, which is predicted to be downregulated. Data were averaged for differentiation experiments carried out using three distinct TRE-Ngn2 iPSC clones ±SEM and compared by one-way ANOVA with Tukey's post hoc test. *, Tukey's $P < 0.05$; ns, Tukey's $P \geq 0.05$

gene (*ΔNeoR*) that lacked a promoter and an in-frame translation initiation codon.

After selecting the rtTA-2Lox-Cre line resistant to puromycin (but not G418), we proceeded with the RMCE-dependent knock-in step. The cells were treated with Dox to induce *Cre* expression and transfected with the pML156 "donor" plasmid containing the mouse *Ngn2* along with the *lox2272* and *loxP* sites (Fig. S6A). The pML156 design also precluded transient expression of *Ngn2* that might promote unwanted iPSC differentiation. Cre-mediated RMCE was expected to integrate the *Ngn2* transgene under the *TRE* promoter, replacing the *Cre* gene. Furthermore, the *ΔNeoR* fragment was expected to acquire a start codon and the human PGK promoter (*hPGK*), making the resultant TRE-Ngn2 cells G418 resistant. We confirmed *Ngn2* integration by PCR-genotyping of G418-resistant clones (Fig. S6B). Moreover, TRE-Ngn2 cultures treated with Dox for 48 h expressed readily detectable amounts of the NGN2 protein and the neuronal marker MAP2 (Fig. S6C). Expression of these proteins was not detected in mock-treated cells (Fig. S6C).

To induce neuronal differentiation, we incubated the TRE-Ngn2 cells with Dox for three days and then maintained the cultures until day 14 in a medium supplemented with the neurotrophic factors BDNF and NT-3 (Fig. 4A). RT-qPCR analyses of RNA samples collected on differentiation days 0, 1, 2, 3, 7, and 14 showed that the cells progressively lost the expression of the pluripotency markers POU5F1 and NANOG (Fig. S6D–E) and gained the expression of an early neuronal marker, NCAM1, beginning from differentiation day 1 (Fig. S6F). The mature neuronal marker SYN1 and the glutamatergic marker SLC17A6/VGLUT2 were upregulated on differentiation day 7 and continued to be expressed on day 14 (Fig. S6G, H).

We concluded that Dox-induced TRE-Ngn2 cultures recapitulated the natural dynamics of gene expression in developing neurons.

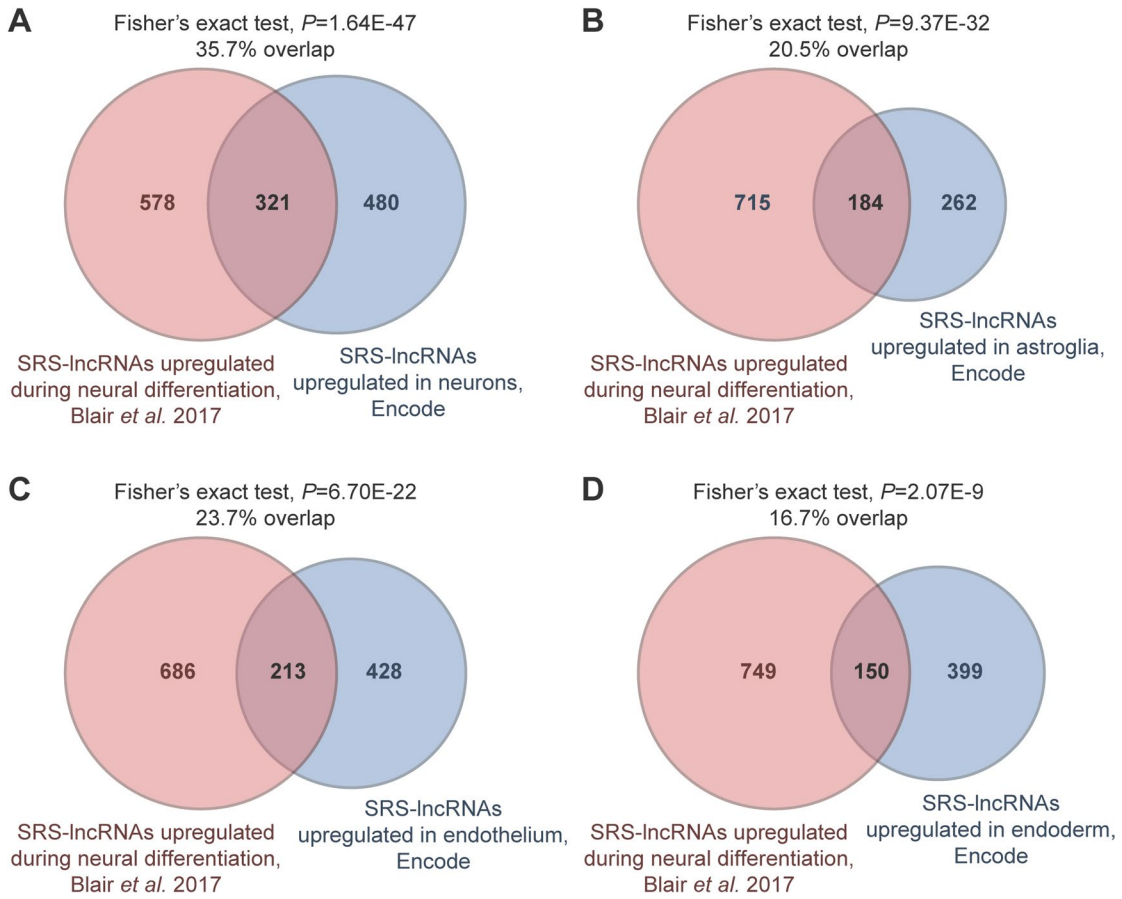


Fig. 6 SRS-lncRNA expression across differentiated cell types. Venn diagrams show overlaps between the neurally upregulated SRS-lncRNAs identified by our pipeline using the data from (Blair et al. 2017) and SRS-lncRNAs upregulated in differentiated **A** neuronal, **B** astroglial, **C** endothelial and **D** endodermal cells from the Encode database. Note that differentiated neurons show the most robust overlap

The expression of neural SRS-lncRNAs is regulated in a developmental stage and cell type-specific manner

We utilized the newly established neuronal differentiation system to validate our bioinformatic predictions. TRE-Ngn2 iPSCs were differentiated as outlined in Fig. 4A and analyzed by RT-qPCR on differentiation days 0, 1, 2, 3, 7, and 14, using appropriate primers. We first investigated the clustered SRS-lncRNAs from the segmentally duplicated part of chr9 (Fig. 2G). According to the RNA-seq data, both members of the bidirectionally transcribed unit were expected to be upregulated in NPCs and retain detectable expression in neurons, albeit at a somewhat decreased level (Fig. 4B).

Our RT-qPCR analyses of the TRE-Ngn2 time series with primers annealing upstream of the GAAAT and alpha-satellite repeats confirmed the upregulation of the corresponding SRS-lncRNAs (e.g., XLOC_320267 and XLOC_312995; Fig. 2G) during neuronal differentiation (assays 1 and 2; Fig. 4B–D). Similar results were obtained when we repeated the analysis using primers designed against downstream sequences (assays 3 and 4; Fig. 4B, E, F). Of note, RT-qPCR showed a higher expression of the two types of SRS-lncRNAs in day-14 neurons compared to the earlier time points. The apparent difference of this dynamics from Fig. 4B might be due to possible variability between the ESC and the TRE-Ngn2 iPSC differentiation protocols.

To confirm the upregulation of XLOC_312995 and XLOC_320267 in neurons, we analyzed differentiation day 0 and 14 samples by RNA-FISH (Fig. S7). Although this approach detected XLOC_312995- and XLOC_320267-specific nuclear foci in both iPSCs and neurons, the foci tended to be noticeably larger in neurons compared to iPSCs (Fig. S7).

Longitudinal RT-qPCR analyses of three additional examples of neural SRS-lncRNAs, XLOC_319631, XLOC_320382, and XLOC_039609/KRTAP5-AS1 (Fig. S2), showed that, their expression significantly increased as a function of development, peaking in day-7–14 neurons (Fig. 5A–C). Conversely, the neurally downregulated SRS-lncRNA candidate XLOC_185088/CPMER (Fig. S1B) decreased its expression, as expected

(Fig. 5D). We validated RT-qPCR assay specificity by Sanger sequencing (Fig. S8). Additionally, transfection of differentiating TRE-Ngn2 cells with an XLOC_312995-specific antisense gapmer (gmXLOC_312995) resulted in a decrease in both assay-2 and assay-4 RT-qPCR signals compared to a negative control gapmer (gmControl; Fig. S9).

To explore the specificity of SRS-lncRNA expression, we mined Encode RNA-seq data and shortlisted SRS-lncRNAs upregulated in diverse types of differentiated human cells. The Encode list for neurons showed a robust overlap with the neurally upregulated SRS-lncRNAs selected by our bioinformatics pipeline (321 out of 899; Fig. 6A; Table S5). Other cell types, including astroglia, endothelial, and endodermal cells, also shared considerable numbers of common SRS-lncRNAs with the neurally upregulated SRS-lncRNAs (Fig. 6B, C; Table S5), but the overlaps in these cases were smaller compared to Fig. 6A. Notably, 129 out of the 321 overlapping neuronal SRS-lncRNAs were not upregulated in the other cell types (Table S5). This subset included, for instance, the XLOC_039609/KRTAP5-AS1 transcript (Fig. S2) and two out of six GAAAT repeat-containing transcripts from the chr9 SRS-lncRNA cluster (XLOC_312965 and XLOC_319986; Fig. 2G).

We also investigated the expression of neural SRS-lncRNAs in developing human cortical organoids (Fiddes et al. 2018). Both the alpha-satellite and the GAAAT repeat-containing chr9 loci were robustly upregulated in this system (Figs. S10, S11), in addition to other neural SRS-lncRNAs (Table S5). Finally, the inspection of RNA-seq data for iPSC-derived day-135 cortical neurons from patients with autism spectrum disorder (ASD) and healthy controls (DeRosa et al. 2018) showed that some SRS-lncRNAs (including the chr9-encoded XLOC_320267 and XLOC_312764) were significantly downregulated, whereas others (e.g. XLOC_320382) were upregulated in the ASD samples (DESeq2 FDR < 0.05; Fig. S12; Table S6).

Thus, many neural SRS-lncRNAs predicted by our pipeline appear to be expressed in a developmental stage and cell type-specific manner, and their expression can be perturbed in disease.

Discussion

Our study argues that numerous long noncoding transcripts containing extensive stretches of simple repeated sequences are expressed in genetically normal human pluripotent stem cells and developing neurons (Fig. 1). Many of these SRS-lncRNAs originate from telomere-proximal regions, despite the abundance of SRSs in both centromere- and telomere-proximal DNA (Fig. 2). While additional work will be required to understand the molecular basis of this genome-wide trend, possible underlying reasons might include more accessible structure or/and more favorable epigenetic modifications of the telomere-proximal chromatin compared to its centromere-proximal counterpart.

A large fraction of the SRS-lncRNAs predicted by our bioinformatics pipeline is significantly upregulated during normal neuronal differentiation, and at least some of these transcripts appear to be specific to neurons (Figs. 1, 4, 5, 6; Figs. S10, S11; Tables S2, S5). Moreover, a subset of SRS-lncRNAs appears to be deregulated in the context of ASD (Fig. S12; Table S6). Similar to other SRS-lncRNAs, many neurally upregulated members of this type of transcripts tend to be encoded in telomere-proximal regions of the genome. An important exception is a broad centromere-proximal region of chr9 that gives rise to a significantly larger number of neural SRS-lncRNAs than expected by chance (Fig. 2E, F). The SRS-lncRNAs encoded in this region are organized as bidirectional pairs of the alpha satellite- and GAAAT repeat-containing transcripts expressed from a common GC-rich promoter (Figs. 2G, 4B). This part of the genome is known to be segmentally duplicated in primates (Bailey et al. 2002; Crosier et al. 2002; Guy et al. 2000), suggesting an intriguing possibility that it encodes primate-specific functions. Additional studies will be needed to test this hypothesis and explore the mechanisms underlying the upregulation of these SRS-lncRNAs in neurons.

Our RBP motif analyses suggest that the chr9-derived and other SRS-lncRNAs upregulated in NPCs and neurons can recruit multiple copies of specific RBPs via SRS-enriched cognate sequence motifs (Fig. 3A, B; Table S4). For example, the medium numbers of multivalent RBP motifs in the SRS parts of the alpha-satellite and GAAAT-repeat SRS-lncRNA families of the chr9 transcripts are 46 and 31.5, respectively (Table S4). The analysis of nELAVL RNA-binding preferences suggests that many computationally predicted motifs serve as bona fide sites for RBP recruitment (Fig. S5; Scheckel et al. 2016). Further validation of such interactions and understanding their possible role in the RNA metabolism of developing neurons will be important directions for future studies.

It will be also interesting to follow up on our finding that SRSs are associated with a decrease in the cytoplasmic-to-nuclear ratio for transcripts of protein-coding genes, while having the opposite effect on lncRNAs (Fig. 3C–F). We hypothesize that this difference relates to the abundance of introns in the former group and their paucity in the latter. Indeed, at least some SRSs are known to interfere with intron excision and mRNA export from the nucleus to the cytoplasm (Monteuuis et al. 2019; Sznajder et al. 2018; Yap et al. 2012).

Although SRS-lnRNAs are still more commonly found in the nucleus compared to protein-coding transcripts (Fig. 3C–F; Fig. S7), it is possible that SRS-lnRNAs

"escaping" the nucleus function in the cytoplasm. A previously characterized example of a repeat-containing lncRNA that acts in both the nucleus and the cytoplasm is NORAD (Elguindy and Mendell 2021; Lee et al. 2016; Munschauer et al. 2018; Tichon et al. 2016). This conserved transcript lacks classically defined SRSs, but contains 18 UGURUAUA motifs that may have originated from ancient duplication events. NORAD utilizes these sequences to sequester Pumilio-family RBPs, thereby altering mRNA stability in the cytoplasm.

Another relevant example is provided by competing endogenous RNAs (ceRNAs) that regulate microRNA activity in the cytoplasm (Ala 2020). Interestingly, our analysis identified two known ceRNA candidates: the circular RNA CDR1as/CiRS-7 and the lncRNA XLOC_039609/KRTAP5-AS1 (Figs. S1A, S2C; Barrett et al. 2017; Hansen et al. 2013; Memczak et al. 2013; Song et al. 2017). We are currently investigating the possibility that other cytoplasmically abundant SRS-lncRNAs might interact with specific microRNAs.

In conclusion, we have identified multiple instances of simple repeated sequences, which are transcribed in a developmentally regulated and cell type-specific manner. We anticipate that comprehensive analyses of the expression dynamics, cellular localization, and interaction partners of neural SRS-lncRNAs using the TRE–Ngn2 system (Fig. S6) and other experimental approaches will illuminate their functional significance in the normal development of the human brain, as well as in neurodevelopmental disorders.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00439-023-02626-1.

Acknowledgements We thank Jizhong Zou and Michael Ward for reagents, and Snezhka Oliferenko for valuable discussions.

Author contributions All authors contributed to the study conception and design. Tek Hong Chung performed the bioinformatics analyses, TRE-Ngn2 differentiation experiments, and RT-qPCR, immunofluorescence and RNA-FISH assays. Anna Zhuravskaya generated and characterized iPSCs encoding the TRE-Ngn2 transgene. Eugene Makeyev supervised the project and obtained the funding. The first draft of the manuscript, including the text, figures and the tables, was prepared by Tek Hong Chung and Eugene Makeyev. All authors read and approved the final manuscript.

Funding This work was supported by the Biotechnology and Biological Sciences Research Council (grant numbers BB/R001049/1 and BB/V006258/1).

Data availability All relevant data are provided in the Supplemental Tables.

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

Ala U (2020) Competing endogenous RNAs, non-coding RNAs and diseases: an intertwined story. Cells 9:1574. https://doi.org/10.3390/cells9071574

Almeida M, Pintacuda G, Masui O, Koseki Y, Gdula M, Cerase A, Brown D, Mould A, Innocent C, Nakayama M, Schermelleh L, Nesterova TB, Koseki H, Brockdorff N (2017) PCGF3/5-PRC1 initiates Polycomb recruitment in X chromosome inactivation. Science 356:1081–1084. https://doi.org/10.1126/science.aal2512

Altemose N (2022) A classical revival: human satellite DNAs enter the genomics era. Semin Cell Dev Biol 128:2–14. https://doi.org/10.1016/j.semcdb.2022.04.012

Aly MK, Ninomiya K, Adachi S, Natsume T, Hirose T (2019) Two distinct nuclear stress bodies containing different sets of RNA-binding proteins are formed with HSATIII architectural noncoding RNAs upon thermal stress exposure. Biochem Biophys Res Commun 516:419–423. https://doi.org/10.1016/j.bbrc.2019.06.061

Arnoult N, Van Beneden A, Decottignies A (2012) Telomere length regulates TERRA levels through increased trimethylation of telomeric H3K9 and HP1alpha. Nat Struct Mol Biol 199:48–56. https://doi.org/10.1038/nsmb.2364

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE (2002) Recent segmental duplications in the human genome. Science 297:1003–1007. https://doi.org/10.1126/science.1072047

Bailey TL, Johnson J, Grant CE, Noble WS (2015) The MEME suite. Nucleic Acids Res 43:W39–W49. https://doi.org/10.1093/nar/gkv416

Balendra R, Isaacs AM (2018) C9orf72-mediated ALS and FTD: multiple pathways to disease. Nat Rev Neurol 14:544–558. https://doi.org/10.1038/s41582-018-0047-2

Barral A, Dejardin J (2020) Telomeric chromatin and TERRA. J Mol Biol 432:4244–4256. https://doi.org/10.1016/j.jmb.2020.03.003

Barrett SP, Parker KR, Horn C, Mata M, Salzman J (2017) ciRS-7 exonic sequence is embedded in a long non-coding RNA locus. PLoS Genet 13:e1007114. https://doi.org/10.1371/journal.pgen.1007114

Baud A, Derbis M, Tutak K, Sobczak K (2022) Partners in crime: proteins implicated in RNA repeat expansion diseases. Wiley Interdiscip Rev RNA 13:e1709. https://doi.org/10.1002/wrna.1709

Biamonti G, Caceres JF (2009) Cellular stress and RNA splicing. Trends Biochem Sci 34:146–153. https://doi.org/10.1016/j.tibs.2008.11.004

Blair JD, Hockemeyer D, Doudna JA, Bateup HS, Floor SN (2017) Widespread translational remodeling during human neuronal differentiation. Cell Rep 21:2005–2016. https://doi.org/10.1016/j.celrep.2017.10.095

Blower MD (2016) Centromeric transcription regulates Aurora-B localization and activation. Cell Rep 15:1624–1633. https://doi.org/10.1016/j.celrep.2016.04.054

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2220. https://doi.org/10.1093/bioinformatics/btu170

Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34:525–527. https://doi.org/10.1038/nbt.3519

Cerbini T, Funahashi R, Luo Y, Liu C, Park K, Rao M, Malik N, Zou J (2015) Transcription activator-like effector nuclease (TALEN)-mediated CLYBL targeting enables enhanced transgene expression and one-step generation of dual reporter human induced pluripotent stem cell (iPSC) and neural stem cell (NSC) lines. PLoS ONE 10:e0116032. https://doi.org/10.1371/journal.pone.0116032

Chu C, Zhang QC, da Rocha ST, Flynn RA, Bharadwaj M, Calabrese JM, Magnuson T, Heard E, Chang HY (2015) Systematic discovery of Xist RNA binding proteins. Cell 161:404–416. https://doi.org/10.1016/j.cell.2015.03.025

Cid-Samper F, Gelabert-Baldrich M, Lang B, Lorenzo-Gotor N, Ponti RD, Severijnen L, Bolognesi B, Gelpi E, Hukema RK, Botta-Orfila T, Tartaglia GG (2018) An integrative study of protein-RNA condensates identifies scaffolding rnas and reveals players in fragile X-associated tremor/ataxia syndrome. Cell Rep 25:3422-3434.e7. https://doi.org/10.1016/j.celrep.2018.11.076

Ciesiolka A, Jazurek M, Drazkowska K, Krzyzosiak WJ (2017) Structural characteristics of simple RNA repeats associated with disease and their deleterious protein interactions. Front Cell Neurosci 11:97. https://doi.org/10.3389/fncel.2017.00097

Crosier M, Viggiano L, Guy J, Misceo D, Stones R, Wei W, Hearn T, Ventura M, Archidiacono N, Rocchi M, Jackson MS (2002) Human paralogs of KIAA0187 were created through independent pericentromeric-directed and chromosome-specific duplication mechanisms. Genome Res 12:67–80. https://doi.org/10.1101/gr.213702

Denegri M, Chiodi I, Corioni M, Cobianchi F, Riva S, Biamonti G (2001) Stress-induced nuclear bodies are sites of accumulation of pre-mRNA processing factors. Mol Biol Cell 12:3502–3514. https://doi.org/10.1091/mbc.12.11.3502

Deng Z, Norseen J, Wiedmer A, Riethman H, Lieberman PM (2009) TERRA RNA binding to TRF2 facilitates heterochromatin formation and ORC recruitment at telomeres. Mol Cell 35:403–413. https://doi.org/10.1016/j.molcel.2009.06.025

DeRosa BA, El Hokayem J, Artimovich E, Garcia-Serje C, Phillips AW, Van Booven D, Nestor JE, Wang L, Cuccaro ML, Vance JM, Pericak-Vance MA, Cukier HN, Nestor MW, Dykxhoorn DM (2018) Convergent pathways in idiopathic autism revealed by time course transcriptomic analysis of patient-derived neurons. Sci Rep 8:8423. https://doi.org/10.1038/s41598-018-26495-1

Elguindy MM, Mendell JT (2021) NORAD-induced Pumilio phase separation is required for genome stability. Nature 595:303–308. https://doi.org/10.1038/s41586-021-03633-w

Fernandopulle MS, Prestil R, Grunseich C, Wang C, Gan L, Ward ME (2018) Transcription factor-mediated differentiation of human iPSCs into neurons. Curr Protoc Cell Biol 79:e51. https://doi.org/10.1002/cpcb.51

Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, Novak AM, van den Bout A, Bishara A, Rosenkrantz JL, Lorig-Roach R, Field AR, Haeussler M, Russo L, Bhaduri A, Nowakowski TJ, Pollen AA, Dougherty ML, Nuttle X, Addor MC, Zwolinski S, Katzman S, Kriegstein A, Eichler EE, Salama SR, Jacobs FMJ, Haussler D (2018) Human-specific NOTCH2NL genes affect notch signaling and cortical neurogenesis. Cell 173:1356-1369.e22. https://doi.org/10.1016/j.cell.2018.03.051

Fujino Y, Nagai Y (2022) The molecular pathogenesis of repeat expansion diseases. Biochem Soc Trans 50:119–134. https://doi.org/10.1042/BST20200143

Goodier JL, Kazazian HH Jr (2008) Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell 135:23–35. https://doi.org/10.1016/j.cell.2008.09.022

Goodwin M, Mohan A, Batra R, Lee KY, Charizanis K, Fernandez Gomez FJ, Eddarkaoui S, Sergeant N, Buee L, Kimura T, Clark HB, Dalton J, Takamura K, Weyn-Vanhentenryck SM, Zhang C, Reid T, Ranum LP, Day JW, Swanson MS (2015) MBNL sequestration by toxic RNAs and RNA misprocessing in the myotonic dystrophy brain. Cell Rep 12:1159–1168. https://doi.org/10.1016/j.celrep.2015.07.029

Gorbunova V, Seluanov A, Mita P, McKerrow W, Fenyo D, Boeke JD, Linker SB, Gage FH, Kreiling JA, Petrashen AP, Woodham TA, Taylor JR, Helfand SL, Sedivy JM (2021) The role of retrotransposable elements in ageing and age-associated diseases. Nature 596:43–53. https://doi.org/10.1038/s41586-021-03542-y

Graf M, Bonetti D, Lockhart A, Serhal K, Kellner V, Maicher A, Jolivet P, Teixeira MT, Luke B (2017) Telomere length determines TERRA and R-Loop regulation through the cell cycle. Cell 170:72-85.e14. https://doi.org/10.1016/j.cell.2017.06.006

Guo CJ, Xu G, Chen LL (2020) Mechanisms of long noncoding RNA nuclear retention. Trends Biochem Sci 45:947–960. https://doi.org/10.1016/j.tibs.2020.07.001

Guy J, Spalluto C, McMurray A, Hearn T, Crosier M, Viggiano L, Miolla V, Archidiacono N, Rocchi M, Scott C, Lee PA, Sulston J, Rogers J, Bentley D, Jackson MS (2000) Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10q. Hum Mol Genet 9:2029–2042. https://doi.org/10.1093/hmg/9.13.2029

Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, Kjems J (2013) Natural RNA circles function as efficient microRNA sponges. Nature 495:384–388. https://doi.org/10.1038/nature11993

Hussong M, Kaehler C, Kerick M, Grimm C, Franz A, Timmermann B, Welzel F, Isensee J, Hucho T, Krobitsch S, Schweiger MR (2017) The bromodomain protein BRD4 regulates splicing during heat shock. Nucleic Acids Res 45:382–394. https://doi.org/10.1093/nar/gkw729

Iacovino M, Bosnakovski D, Fey H, Rux D, Bajwa G, Mahen E, Mitanoska A, Xu Z, Kyba M (2011) Inducible cassette exchange: a rapid and efficient system enabling conditional gene expression in embryonic stem and primary cells. Stem Cells 29:1580–1588. https://doi.org/10.1002/stem.715

Ideue T, Cho Y, Nishimura K, Tani T (2014) Involvement of satellite I noncoding RNA in regulation of chromosome segregation. Genes Cells 19:528–538. https://doi.org/10.1111/gtc.12149

Jain AK, Xi Y, McCarthy R, Allton K, Akdemir KC, Patel LR, Aronow B, Lin C, Li W, Yang L, Barton MC (2016) LncPRESS1 is a p53-regulated LncRNA that safeguards pluripotency by disrupting SIRT6-mediated De-acetylation of Histone H3K56. Mol Cell 64:967–981. https://doi.org/10.1016/j.molcel.2016.10.039

Jiang H, Mankodi A, Swanson MS, Moxley RT, Thornton CA (2004) Myotonic dystrophy type 1 is associated with nuclear foci of mutant RNA, sequestration of muscleblind proteins and deregulated alternative splicing in neurons. Hum Mol Genet 13:3079–3088. https://doi.org/10.1093/hmg/ddh327

Johnson WL, Yewdell WT, Bell JC, McNulty SM, Duda Z, O'Neill RJ, Sullivan BA, Straight AF (2017) RNA-dependent stabilization of SUV39H1 at constitutive heterochromatin. Elife 6:e25299. https://doi.org/10.7554/eLife.25299

Jolly C, Metz A, Govin J, Vigneron M, Turner BM, Khochbin S, Vourc'h C, (2004) Stress-induced transcription of satellite III repeats. J Cell Biol 164:25–33. https://doi.org/10.1083/jcb.200306104

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci USA 100:11484–11489. https://doi.org/10.1073/pnas.1932072100

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann Y, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921. https://doi.org/10.1038/35057062

Leclerc S, Kitagawa K (2021) The role of human centromeric RNA in chromosome stability. Front Mol Biosci 8:642732. https://doi.org/10.3389/fmolb.2021.642732

Lee S, Kopp F, Chang TC, Sataluri A, Chen B, Sivakumar S, Yu H, Xie Y, Mendell JT (2016) Noncoding RNA NORAD regulates genomic stability by sequestering PUMILIO proteins. Cell 164:69–80. https://doi.org/10.1016/j.cell.2015.12.017

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lin HC, He Z, Ebert S, Schornig M, Santel M, Nikolova MT, Weigert A, Hevers W, Kasri NN, Taverna E, Camp JG, Treutlein B (2021) NGN2 induces diverse neuron types from human pluripotency. Stem Cell Reports 16:2118–2127. https://doi.org/10.1016/j.stemcr.2021.07.006

Liquori CL, Ricker K, Moseley ML, Jacobsen JF, Kress W, Naylor SL, Day JW, Ranum LP (2001) Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. Science 293:864–867. https://doi.org/10.1126/science.1062125

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. https://doi.org/10.1186/s13059-014-0550-8

Lu Z, Guo JK, Wei Y, Dou DR, Zarnegar B, Ma Q, Li R, Zhao Y, Liu F, Choudhry H, Khavari PA, Chang HY (2020) Structural modularity of the XIST ribonucleoprotein complex. Nat Commun 11:6163. https://doi.org/10.1038/s41467-020-20040-3

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155. https://doi.org/10.1126/science.290.5494.1151

Lyu Y, Jia W, Wu Y, Zhao X, Xia Y, Guo X, Kang J (2022) Cpmer: a new conserved eEF1A2-binding partner that regulates Eomes translation and cardiomyocyte differentiation. Stem Cell Reports 17:1154–1169. https://doi.org/10.1016/j.stemcr.2022.03.006

Mankodi A, Urbinati CR, Yuan QP, Moxley RT, Sansone V, Krym M, Henderson D, Schalling M, Swanson MS, Thornton CA (2001) Muscleblind localizes to nuclear foci of aberrant RNA in myotonic dystrophy types 1 and 2. Hum Mol Genet 10:2165–2170. https://doi.org/10.1093/hmg/10.19.2165

Masuda A, Andersen HS, Doktor TK, Okamoto T, Ito M, Andresen BS, Ohno K (2012) CUGBP1 and MBNL1 preferentially bind to 3′ UTRs and facilitate mRNA decay. Sci Rep 2:209. https://doi.org/10.1038/srep00209

Mattick JS, Amaral PP, Carninci P, Carpenter S, Chang HY, Chen LL, Chen R, Dean C, Dinger ME, Fitzgerald KA, Gingeras TR, Guttman M, Hirose T, Huarte M, Johnson R, Kanduri C, Kapranov P, Lawrence JB, Lee JT, Mendell JT, Mercer TR, Moore KJ, Nakagawa S, Rinn JL, Spector DL, Ulitsky I, Wan Y, Wilusz JE, Wu M (2023) Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol 24:430–447. https://doi.org/10.1038/s41580-022-00566-8

McHugh CA, Chen CK, Chow A, Surka CF, Tran C, McDonel P, Pandya-Jones A, Blanco M, Burghard C, Moradian A, Sweredoski MJ, Shishkin AA, Su J, Lander ES, Hess S, Plath K, Guttman M (2015) The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. Nature 521:232–236. https://doi.org/10.1038/nature14443

McNulty SM, Sullivan BA (2018) Alpha satellite DNA biology: finding function in the recesses of the genome. Chromosome Res 26:115–138. https://doi.org/10.1007/s10577-018-9582-3

McNulty SM, Sullivan LL, Sullivan BA (2017) Human centromeres produce chromosome-specific and array-specific alpha satellite transcripts that are complexed with CENP-A and CENP-C. Dev Cell 42:226-240.e6. https://doi.org/10.1016/j.devcel.2017.07.001

McStay B (2023) The p-arms of human acrocentric chromosomes play by a different set of rules. Annu Rev Genomics Hum Genet 24:63–83. https://doi.org/10.1146/annurev-genom-101122-081642

Mei Y, Deng Z, Vladimirova O, Gulve N, Johnson FB, Drosopoulos WC, Schildkraut CL, Lieberman PM (2021) TERRA G-quadruplex RNA interaction with TRF2 GAR domain is required for telomere integrity. Sci Rep 11:3509. https://doi.org/10.1038/s41598-021-82406-x

Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen LH, Munschauer M, Loewer A, Ziebold U, Landthaler M, Kocks C, le Noble F, Rajewsky N (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. Nature 495:333–338. https://doi.org/10.1038/nature11928

Meola G, Cardani R (2015) Myotonic dystrophies: an update on clinical aspects, genetic, pathology, and molecular pathomechanisms. Biochim Biophys Acta 1852:594–606. https://doi.org/10.1016/j.bbadis.2014.05.019

Metz A, Soret J, Vourc'h C, Tazi J, Jolly C, (2004) A key role for stress-induced satellite III transcripts in the relocalization of splicing factors into nuclear stress granules. J Cell Sci 117:4551–4558. https://doi.org/10.1242/jcs.01329

Miller JW, Urbinati CR, Teng-Umnuay P, Stenberg MG, Byrne BJ, Thornton CA, Swanson MS (2000) Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy. EMBO J 19:4439–4448. https://doi.org/10.1093/emboj/19.17.4439

Monteuuis G, Wong JJL, Bailey CG, Schmitz U, Rasko JEJ (2019) The changing paradigm of intron retention: regulation, ramifications and recipes. Nucleic Acids Res 47:11497–11513. https://doi.org/10.1093/nar/gkz1068

Munschauer M, Nguyen CT, Sirokman K, Hartigan CR, Hogstrom L, Engreitz JM, Ulirsch JC, Fulco CP, Subramanian V, Chen J, Schenone M, Guttman M, Carr SA, Lander ES (2018) The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. Nature 561:132–136. https://doi.org/10.1038/s41586-018-0453-z

Nalavade R, Griesche N, Ryan DP, Hildebrand S, Krauss S (2013) Mechanisms of RNA-induced toxicity in CAG repeat disorders. Cell Death Dis 4:e752. https://doi.org/10.1038/cddis.2013.276

Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39:121–152. https://doi.org/10.1146/annurev.genet.39.073003.112240

Nemeth A, Grummt I (2018) Dynamic regulation of nucleolar architecture. Curr Opin Cell Biol 52:105–111. https://doi.org/10.1016/j.ceb.2018.02.013

Ninomiya K, Hirose T (2020) Short tandem repeat-enriched architectural RNAs in nuclear bodies: functions and associated diseases. Noncoding RNA 6:6. https://doi.org/10.3390/ncrna6010006

Ninomiya K, Adachi S, Natsume T, Iwakiri J, Terai G, Asai K, Hirose T (2020) LncRNA-dependent nuclear stress bodies promote intron retention through SR protein phosphorylation. EMBO J 39:e102729. https://doi.org/10.15252/embj.2019102729

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, Aganezov S, Hoyt SJ, Diekhans M, Logsdon GA, Alonge M, Antonarakis SE, Borchers M, Bouffard GG, Brooks SY, Caldas GV, Chen NC, Cheng H, Chin CS, Chow W, de Lima LG, Dishuck PC, Durbin R, Dvorkina T, Fiddes IT, Formenti G, Fulton RS, Fungtammasan A, Garrison E, Grady PGS, Graves-Lindsay TA, Hall IM, Hansen NF, Hartley GA, Haukness M, Howe K, Hunkapiller MW, Jain C, Jain M, Jarvis ED, Kerpedjiev P, Kirsche M, Kolmogorov M, Korlach J, Kremitzki M, Li H, Maduro VV, Marschall T, McCartney AM, McDaniel J, Miller DE, Mullikin JC, Myers EW, Olson ND, Paten B, Peluso P, Pevzner PA, Porubsky D, Potapova T, Rogaev EI, Rosenfeld JA, Salzberg SL, Schneider VA, Sedlazeck FJ, Shafin K, Shew CJ, Shumate A, Sims Y, Smit AFA, Soto DC, Sovic I, Storer JM, Streets A, Sullivan BA, Thibaud-Nissen F, Torrance J, Wagner J, Walenz BP, Wenger A, Wood JMD, Xiao C, Yan SM, Young AC, Zarate S, Surti U, McCoy RC, Dennis MY, Alexandrov IA, Gerton JL, O'Neill RJ, Timp W, Zook JM, Schatz MC, Eichler EE, Miga KH, Phillippy AM (2022) The complete sequence of a human genome. Science 376:44–53. https://doi.org/10.1126/science.abj6987

Palazzo AF, Lee ES (2018) Sequence determinants for nuclear retention and cytoplasmic export of mRNAs and lncRNAs. Front Genet 9:440. https://doi.org/10.3389/fgene.2018.00440

Patrat C, Ouimette JF, Rougeulle C (2020) X chromosome inactivation in human development. Development 147:dev183095. https://doi.org/10.1242/dev.183095

Perea-Resa C, Blower MD (2018) Centromere biology: transcription goes on stage. Mol Cell Biol 38:e00263-e318. https://doi.org/10.1128/MCB.00263-18

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc 11:1650–1667. https://doi.org/10.1038/nprot.2016.095

Pintacuda G, Wei G, Roustan C, Kirmizitas BA, Solcan N, Cerase A, Castello A, Mohammed S, Moindrot B, Nesterova TB, Brockdorff N (2017) hnRNPK recruits PCGF3/5-PRC1 to the Xist RNA B-repeat to establish Polycomb-mediated chromosomal silencing. Mol Cell 68:955-969.e10. https://doi.org/10.1016/j.molcel.2017.11.013

Porro A, Feuerhahn S, Delafontaine J, Riethman H, Rougemont J, Lingner J (2014) Functional characterization of the TERRA transcriptome at damaged telomeres. Nat Commun 5:5379. https://doi.org/10.1038/ncomms6379

Quenet D, Dalal Y (2014) A long non-coding RNA is required for targeting centromeric protein A to the human centromere. Elife 3:e03254. https://doi.org/10.7554/eLife.03254

R Core Team (2021) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, Na H, Irimia M, Matzat LH, Dale RK, Smith SA, Yarosh CA, Kelly SM, Nabet B, Mecenas D, Li W, Laishram RS, Qiao M, Lipshitz HD, Piano F, Corbett AH, Carstens RP, Frey BJ, Anderson RA, Lynch KW, Penalva LO, Lei EP, Fraser AG, Blencowe BJ, Morris QD, Hughes TR (2013) A compendium of RNA-binding motifs for decoding gene regulation. Nature 499:172–177. https://doi.org/10.1038/nature12311

Sakaguchi T, Hasegawa Y, Brockdorff N, Tsutsui K, Tsutsui KM, Sado T, Nakagawa S (2016) Control of chromosomal localization of Xist by hnRNP U family molecules. Dev Cell 39:11–12. https://doi.org/10.1016/j.devcel.2016.09.022

Scheckel C, Drapeau E, Frias MA, Park CY, Fak J, Zucker-Scharff I, Kou Y, Haroutunian V, Ma'ayan A, Buxbaum JD, Darnell RB, (2016) Regulatory consequences of neuronal ELAV-like protein binding to coding and non-coding RNAs in human brain. Elife 5:e10421. https://doi.org/10.7554/eLife.10421

Schwartz JL, Jones KL, Yeo GW (2021) Repeat RNA expansion disorders of the nervous system: post-transcriptional mechanisms and therapeutic strategies. Crit Rev Biochem Mol Biol 56:31–53. https://doi.org/10.1080/10409238.2020.1841726

Sellier C, Rau F, Liu Y, Tassone F, Hukema RK, Gattoni R, Schneider A, Richard S, Willemsen R, Elliott DJ, Hagerman PJ, Charlet-Berguerand N (2010) Sam68 sequestration and partial loss of function are associated with splicing alterations in FXTAS patients. EMBO J 29:1248–1261. https://doi.org/10.1038/emboj.2010.21

Shah A, Qian Y, Weyn-Vanhentenryck SM, Zhang C (2017) CLIP Tool Kit (CTK): a flexible and robust pipeline to analyze CLIP sequencing data. Bioinformatics 33:566–567. https://doi.org/10.1093/bioinformatics/btw653

Silva B, Arora R, Bione S, Azzalin CM (2021) TERRA transcription destabilizes telomere integrity to initiate break-induced replication in human ALT cells. Nat Commun 12:3760. https://doi.org/10.1038/s41467-021-24097-6

Song YX, Sun JX, Zhao JH, Yang YC, Shi JX, Wu ZH, Chen XW, Gao P, Miao ZF, Wang ZN (2017) Non-coding RNAs participate in the regulatory network of CLDN4 via ceRNA mediated miRNA evasion. Nat Commun 8:289. https://doi.org/10.1038/s41467-017-00304-1

Sznajder LJ, Swanson MS (2019) Short tandem repeat expansions and RNA-mediated pathogenesis in myotonic dystrophy. Int J Mol Sci 20:3365. https://doi.org/10.3390/ijms20133365

Sznajder LJ, Thomas JD, Carrell EM, Reid T, McFarland KN, Cleary JD, Oliveira R, Nutter CA, Bhatt K, Sobczak K, Ashizawa T, Thornton CA, Ranum LPW, Swanson MS (2018) Intron retention induced by microsatellite expansions as a disease biomarker. Proc Natl Acad Sci USA 115:4234–4239. https://doi.org/10.1073/pnas.1716617115

Tichon A, Gil N, Lubelsky Y, Havkin Solomon T, Lemze D, Itzkovitz S, Stern-Ginossar N, Ulitsky I (2016) A conserved abundant cytoplasmic long noncoding RNA modulates repression by Pumilio proteins in human cells. Nat Commun 7:12209. https://doi.org/10.1038/ncomms12209

Tong C, Yin Y (2021) Localization of RNAs in the nucleus: cis- and trans-regulation. RNA Biol 18:2073–2086. https://doi.org/10.1080/15476286.2021.1894025

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515. https://doi.org/10.1038/nbt.1621

Trigiante G, Blanes Ruiz N, Cerase A (2021) Emerging roles of repetitive and repeat-containing RNA in nuclear and chromatin organization and gene expression. Front Cell Dev Biol 9:735527. https://doi.org/10.3389/fcell.2021.735527

Ugarkovic D, Sermek A, Ljubic S, Feliciello I (2022) Satellite DNAs in health and disease. Genes (basel) 13:1154. https://doi.org/10.3390/genes13071154

Weighardt F, Cobianchi F, Cartegni L, Chiodi I, Villa A, Riva S, Biamonti G (1999) A novel hnRNP protein (HAP/SAF-B) enters a subset of hnRNP complexes and relocates in nuclear granules in response to heat shock. J Cell Sci 112(Pt 10):1465–1476. https://doi.org/10.1242/jcs.112.10.1465

Wong LH, Brettingham-Moore KH, Chan L, Quach JM, Anderson MA, Northrop EL, Hannan R, Saffery R, Shaw ML, Williams E, Choo KH (2007) Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere. Genome Res 17:1146–1160. https://doi.org/10.1101/gr.6022807

Wright SE, Todd PK (2023) Native functions of short tandem repeats. Elife 12:e84043. https://doi.org/10.7554/eLife.84043

Yap K, Lim ZQ, Khandelia P, Friedman B, Makeyev EV (2012) Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. Genes Dev 26:1209–1223. https://doi.org/10.1101/gad.188037.112

Yap K, Mukhina S, Zhang G, Tan JSC, Ong HS, Makeyev EV (2018) A short tandem repeat-enriched RNA assembles a nuclear compartment to control alternative splicing and promote cell survival. Mol Cell 72:525-540.e13. https://doi.org/10.1016/j.molcel.2018.08.041

Yap K, Chung TH, Makeyev EV (2022a) Analysis of RNA-containing compartments by hybridization and proximity labeling in cultured human cells. STAR Protoc 3:101139. https://doi.org/10.1016/j.xpro.2022.101139

Yap K, Chung TH, Makeyev EV (2022b) Hybridization-proximity labeling reveals spatially ordered interactions of nuclear RNA compartments. Mol Cell 82:463-478.e11. https://doi.org/10.1016/j.molcel.2021.10.009

Yum K, Wang ET, Kalsotra A (2017) Myotonic dystrophy: disease repeat range, penetrance, age of onset, and relationship between repeat size and phenotypes. Curr Opin Genet Dev 44:30–37. https://doi.org/10.1016/j.gde.2017.01.007

Zhang Y, Pak C, Han Y, Ahlenius H, Zhang Z, Chanda S, Marro S, Patzke C, Acuna C, Covy J, Xu W, Yang N, Danko T, Chen L, Wernig M, Sudhof TC (2013) Rapid single-step induction of functional neurons from human pluripotent stem cells. Neuron 78:785–798. https://doi.org/10.1016/j.neuron.2013.05.029

Zhuravskaya A, Yap K, Hamid F, Makeyev EV (2023) Alternative splicing coupled to nonsense-mediated decay coordinates downregulation of non-neuronal genes in developing neurons. bioRxiv2:023.09.04.556212. doi:https://doi.org/10.1101/2023.09.04.556212