

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	Human snRNA-seq data generated by the BRAIN initiative project was accessed at The Neuroscience Multi-omics Archive (NeMO Archive) using gsutil v4.59. Remaining human and mouse RNA-seq data were accessed from the Short Read Archive (SRA) using fastq-dump v2.11.0. Both gsutil and fastq-dump were operated through a modified version of MicroExonator (blencowe branch: https://github.com/hemberg-lab/MicroExonator/tree/blencowe). Genomic variants from SSC, SPARK, and MSSNG were securely stored by the Scherer group and shared with the rest of the co-authors conducting the analyses through an FTP server. RT-PCR gel images were acquired using Bio-Rad Image Lab v6.0.1.
Data analysis	RNA-seq data were analysed using MicroExonator (modified blencowe branch: https://github.com/hemberg-lab/MicroExonator/tree/blencowe), vast-tools v2.5.1, MATT v1.3.0 and Whippet v1.6.1. Variants from SSC, SPARK and MSSNG were processed using BCftools v1.20 within a Snakemake v7.32.4 workflow. Downstream analyses, including the processing of RNA-seq quantification outputs, model predictions and variant annotations, were performed using custom Python and R scripts implemented as Jupyter notebooks with JupyterLab v4.0.8. scikit-learn v1.5.0 was used for kernel-density estimation, and statsmodels v0.14.2 was used for developmental regression analyses. The microexon code classifier was trained using XGBoost v1.5.0. TensorFlow v2.2.0 was used to generate COSSMO-derived input features. Missing PSI values were imputed using probabilistic principal component analysis with the pcaMethods R package v1.94.0, and hierarchical clustering was performed using the R hclust function with method = "ward.D2". Human-mouse conservation analyses used UCSC liftOver and pairwise sequence alignment implemented using Biopython v1.83. Network and enrichment analyses were performed using Cytoscape v3.10.0 with EnrichmentMap v3.3.6, AutoAnnotate v1.4 and clusterMaker2 v2.3.4, as described in the Methods. All code used for model training, variant processing and downstream analyses is available at https://github.com/geparada/microexon-regulatory-code . RT-PCR band intensities were quantified using ImageJ v1.54p or Fiji v2.9.0, as appropriate. GraphPad Prism v10 was used to prepare some figures.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The RNA-seq data generated in this study have been deposited in the NCBI Gene Expression Omnibus (GEO) and are publicly available under accession GSE268269 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE268269>). The complete inventory of reused RNA-seq datasets and their accessions is provided in Supplementary Table 1. Microexon annotations were obtained from VastDB (<https://vastdb.crg.eu>) and GENCODE (<https://www.gencodegenes.org>). Mouse developmental RNA-seq data were obtained from the ENCODE portal (<https://www.encodeproject.org>) and the NCBI Sequence Read Archive under accession SRP055008 (<https://www.ncbi.nlm.nih.gov/sra/?term=SRP055008>). BrainGVEX data were obtained from Synapse under accession syn4590909 (<https://www.synapse.org/Synapse:syn4590909>). MaPSy measurements were obtained from GEO under accession GSE276143 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE276143>). RNA-binding protein motifs were obtained from CisBP-RNA (<http://cisbp-rna.cabr.utoronto.ca>), and population allele frequencies from gnomAD v3 (<https://gnomad.broadinstitute.org>). The human GRCh38/hg38 and mouse GRCh38/mm10 reference genome assemblies were obtained through the UCSC Genome Browser (<https://hgdownload.soe.ucsc.edu>). Whole-genome sequencing and phenotype data from the Simons Simplex Collection and SPARK are available to qualified researchers through SFARI Base (<https://base.sfari.org>) following registration and approval. MSSNG data are available to qualified researchers through the MSSNG Research Portal (<https://research.mssng.org>) following application and approval. These human genomic datasets are subject to their respective controlled-access and data-use conditions.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

We obtained the sex identity information of each donor provided as associated metadata for each of the analysed datasets (SSC, SPARK and MSSNG). Sex was included as a covariate in the sibling-pair-matched conditional logistic regression testing whether cumulative code-score impact predicts ASD status (Methods; Supplementary Table 10).

Reporting on race, ethnicity, or other socially relevant groupings

We did not consider race, ethnicity, or other socially relevant groupings in our analyses. However, as the core of our analyses was based on comparisons between variants present across discordant sibling pairs, we indirectly controlled for variants associated with specific human sub-populations.

Population characteristics

We obtained relevant population characteristics of human donors as metadata of each of the analysed datasets (SSC, SPARK and MSSNG).

Recruitment

We were not involved in the recruitment process of the data collected by SSC, SPARK and MSSNG.

Ethics oversight

We received a Data Use Certification by the National Institute of Mental Health (NIMH) and the associated Human Ethics protocol was approved by University of Toronto.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

No statistical method was used to predetermine sample sizes. For computational analyses, sample sizes were determined by the availability of data meeting the inclusion and quality-control criteria described in the Methods. All available genomes from the SSC, SPARK and MSSNG cohorts that met these criteria were analysed. For cultured-cell experiments, sample sizes were based on established experimental practice, sample and material availability, and independent biological replication where performed. Exact sample sizes and replicate types are reported in the corresponding figure legends. Experiments with fewer than three independent replicates were treated as descriptive, and no inferential statistical analysis was performed on those experiments.

Data exclusions

To control for possible gene annotation artifacts, we only considered microexons present across the main chromosome scaffolds from the hg38 and mm10 assemblies, excluding the mitochondrial chromosome. Additionally, to avoid introducing sex biases into our analyses, we excluded microexons present on chromosome Y.

Replication

Reproducibility of the genetic analyses was assessed across three ASD cohorts: SSC, SPARK and MSSNG. Experimental replication was as

Replication	follows: the QKI minigene reporter assay in Fig. 3f was performed with three biological replicates; the CHX RT-PCR experiment in Extended Data Fig. 2a was performed twice independently; the SMG1i RT-PCR experiment in Extended Data Fig. 2b was performed once, with BCO1 and AFG1L additionally validated in an independent WTC11 iPSC line; QKI knockdown was quantified by qRT-PCR in three biological replicates in Extended Data Fig. 9a; the immunoblot in Extended Data Fig. 9b was performed once and supported by the qRT-PCR experiments; the QKI-regulated microexon RT-PCR assays in Extended Data Fig. 9c were performed with three biological replicates; and the variant minigene assays in Extended Data Fig. 10i were performed with three biological replicates. All planned replication attempts were successful. Exact sample sizes and replicate types are reported in the corresponding figure legends.
Randomization	Samples were not randomly allocated for the computational analyses; analytical groups were defined using the phenotype and family metadata supplied with each cohort. Affected and unaffected relatives from the same pedigrees were compared to account for family structure and population-associated genetic variation. For RV-GDT analyses, empirical P values were obtained by permuting phenotype labels within pedigrees, with 10,000 permutations in each of 10 independent runs per condition. Cultured-cell experiments were not randomized; wells were assigned to control or experimental conditions according to the experimental design and were processed in parallel.
Blinding	Blinding was not performed. For computational analyses, group assignments were defined by pre-existing cohort metadata. For cultured-cell experiments, control and experimental conditions were identifiable from their treatment or transfection assignments. The same data-collection and analysis procedures were applied across the relevant experimental conditions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Mouse monoclonal anti-pan-QKI antibody, clone N147/6 (Sigma-Aldrich, catalogue MABN624; RRID AB_2877387; 1:500 dilution); mouse monoclonal anti-alpha-tubulin antibody, clone B-5-1-2 (Sigma-Aldrich, catalogue T6074; RRID AB_477582; 1:10,000 dilution).
Validation	The anti-pan-QKI antibody was used for immunoblotting of mouse N2A cell lysates. Its specificity was supported in this study by loss of the expected QKI signal following QKI siRNA knockdown (Extended Data Fig. 9b). The manufacturer also validates clone N147/6 for immunoblotting and reactivity with mouse, rat and human QKI. The anti-alpha-tubulin antibody was used as an immunoblot loading control and was not independently validated in this study; manufacturer validation supports its use for immunoblotting across multiple species.

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	Neuro-2a (N2A) mouse neuroblastoma cells were obtained from ATCC (CCL-131). CGR8 mouse embryonic stem cells were obtained from ECACC (07032901) and were used directly or differentiated into neurons. Human iPSC line #37 and its derived neural precursor cells and neurons were previously described and characterized by Cheung et al. (Hum. Mol. Genet. 20, 2103–2115, 2011). WTC11 human iPSCs were obtained from the Coriell Institute for Medical Research (GM25256) and differentiated into neurons.
Authentication	N2A, CGR8 and WTC11 cells were characterized by their respective source repositories, and iPSC line #37 was previously characterized as described by Cheung et al. No cell lines were independently authenticated by the authors for this study.
Mycoplasma contamination	All cells were tested mycoplasma-negative using a commercial PCR kit (e-Myco™ VALiD)
Commonly misidentified lines (See ICLAC register)	Commonly misidentified cell lines were not used in this study.

Seed stocks	Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.
Novel plant genotypes	Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.
Authentication	Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.