science.sciencemag.org/content/371/6530/eaax2537/suppl/DC1

Supplementary Materials for

Reintroduction of the archaic variant of *NOVA1* in cortical organoids alters neurodevelopment

Cleber A. Trujillo, Edward S. Rice, Nathan K. Schaefer, Isaac A. Chaim, Emily C. Wheeler, Assael A. Madrigal, Justin Buchanan, Sebastian Preissl, Allen Wang, Priscilla D. Negraes, Ryan A. Szeto, Roberto H. Herai, Alik Huseynov, Mariana S. A. Ferraz, Fernando S. Borges, Alexandre H. Kihara, Ashley Byrne, Maximillian Marin, Christopher Vollmers, Angela N. Brooks, Jonathan D. Lautz, Katerina Semendeferi, Beth Shapiro, Gene W. Yeo, Stephen E. P. Smith, Richard E. Green, Alysson R. Muotri*

*Corresponding author. Email: muotri@ucsd.edu

Published 12 February 2021, *Science* **371**, eaax2537 (2021)
DOI: 10.1126/science.aax2537

This PDF file includes:

Other Supporting Online Material for this manuscript includes the following:
(available at science.sciencemag.org/content/371/6530/eaax2537/suppl/DC1)

Tables S3 to S7 (.xlsx)
MDAR Reproducibility Checklist (.pdf)

Materials and Methods

Haplotype analyses

We downloaded phased variant calls from the 1000 Genomes Project (*11*) and the Simons Genome Diversity Project (*12*), along with high-coverage BAM files for the Vindija (*5*) and Altai (*4*) Neanderthals, and the high-coverage Denisovan (*3*). Considering only the autosomes, we used snpAD (*38*) to call variants in the ancient data, with the parameter –Q 25 and separating UDG-treated and non-UDG treated libraries (Vindija33.19) and single-stranded and double-stranded libraries (Altai and Denisovan) during parameter estimation. We merged all archaic hominin VCFs with the 1000 Genomes data (*11*) using bcftools (*39*), with the –missing-to-ref option and –m all to merge all types of variants. We filtered these files using bcftools norm with the –d all option to delete all sites with multiple, conflicting variant calls and the –c x option to ensure reference allele calls agreed with the human reference genome version hg19. Next, we filtered to only biallelic sites with no missing genotype calls, using bcftools view –m 2 –M 2 –c 2. We ran SnpEff (*40*) on these files with reference genome version GRCh37.75 to obtain functional annotations, and we considered the following annotations to denote nonsynonymous variants:

5_prime_UTR_premature_start_codon_gain_variant, conservative_inframe_deletion, conservative_inframe_insertion, disruptive_inframe_deletion, disruptive_inframe_insertion, exon_loss_variant, frameshift_variant, initiator_codon_variant, missense_variant, start_lost, stop_gained, stop_lost, structural_interaction_variant.

We then scanned through our VCFs, using minimum genotype quality 25 and requiring the Altai Neanderthal to be between 23-70X coverage, the Denisovan to be between 10-43X coverage, and the Vindija33.19 Neanderthal to be between 10-47X coverage at each variant. We also required that one of the two alleles of each variant match the ancestral allele from the 1000 Genomes Project data (and we discarded variants for which the ancestral allele was unknown). Of the sites passing these filters, we compiled a list of sites for which all 1000 Genomes Project haplotypes and no archaic hominin haplotypes carried the derived allele.

If there were sites containing variation in modern humans that were filtered or missing from the 1000 Genomes Project data, but for which we observed non-reference alleles in the Neanderthals or Denisovan, they would at this point appear as false-positive fixed derived alleles in modern humans, because we merged our variant call sets using the –missing-to-ref option. To mitigate this problem, we required that all derived alleles that we found to be fixed in modern humans were also fixed in, or missing from, the Simons Genome Diversity Project panel. This produced a conservative set of fixed mutations for further analysis. We note that the Simons Genome Diversity Project data set, as downloaded, was processed in a way that caused all heterozygous genotypes to be called homozygous reference wherever there is no known chimpanzee allele; this should have only affected a minority of sites where the 1000 Genomes Project reference sequence differs from the chimpanzee genome; of this set of sites, it would only interfere with our filtering if all non-reference alleles were in the heterozygous state. This could have produced a small number of sites where we did not observe human-specific

polymorphism that exists, and this in turn would have made our filtering less stringent. As a result, our set of human-specific fixed derived alleles might be slightly inflated. The same problem could arise from sites that were filtered from the Simons Genome Diversity Project panel but not the 1000 Genomes Project panel.

After compiling our list of human-specific fixed derived alleles, we used the 1000 Genomes Project data to scan for unrecombined haplotypes around each. We defined an unrecombined haplotype as the span of sites upstream and downstream of the allele for which no modern human in the 1000 Genomes Project data set shares a derived allele with an archaic hominin. Any haplotype that fell within a centromeric or telomeric region was discarded. We calculated pairwise nucleotide diversity and Watterson's estimator of theta (*41*) within each human-specific haplotype using only biallelic SNPs for which the ancestral allele is known. We then calculated Tajima's D from these two values and then normalized Tajima's D within each haplotype by dividing it by its minimum possible value (*42*).

Cell source

The neurotypical iPSC lines (WT83 and ASC) and related clones were previously characterized and validated (*20*, *43*). These iPSCs have been tested for the expression of the pluripotency markers, including OCT4, SOX2, SSEA4, and TRA-1-60, and functional pluripotency via teratoma formation analysis. Human iPSC colonies were expanded on feeder-free conditions on Matrigel-coated dishes (BD Biosciences, San Jose, CA, USA) with mTeSR1 medium (StemCell Technologies, Vancouver, Canada) changed daily. The cells were routinely checked by karyotype and CNV arrays to avoid genomic alterations in the culture. The study was approved by the University of California San Diego IRB/ESCRO committee (protocol 141223ZF).

NOVA1 cell line generation

Human edited NOVA1$^{Ar/Ar}$, NOVA1$^{Ko/Ar,}$ and NOVA1$^{Ko/Ko}$ iPSC lines were generated by CRISPR-Cas9 genome-editing system to induce a point mutation insertion (V200I) by substituting codon 200 GTA/V with ATC/I (rs906-08) at both alleles. Four guide RNA (gRNA) candidates targeting the point mutation region were designed, generated, and evaluated for their abilities to mediate cleavage and indel formation. After gRNA validation, gRNA-AAGGCTGTAATGGAGCAGTCAGG was identified to be highly active and the donor sequence was synthesized based on the desired alteration. For the V200I point mutation, a 169bp donor sequence was designated to optimize the likelihood of being used as a repair template at the gRNA cut site. The donor sequence used is
5’-TTTAATTACAGGTAAAGATTATAGTTCCCAACAGCACAGCAGGTCTGATAA
TAGGGAAGGGAGGTGCTACTGTGAAGGCTATCATGGAACAGTCAGGGGcttggg
tgcagctttcccagaaacctgatgggatcaacttgcaagagagggttgtcactgtgagtggaga-3’.

The gRNA cut site maps to the approximate middle of the donor sequence. The incorporation of ATC at rs906-08 not only generated the desired point mutation, but also generated a Fat I restriction site at rs408-11. Additionally, a silent mutation was incorporated into the donor (GAG/E→GAA/E at rs914) to prevent the gRNA from re-cutting sequences that were incorporated into this site.

The hiPSCs were cultured and transfected with gRNA, Cas9, and donor using the Neon Transfection System (Life Technologies). After transfection, the iPSCs were evenly distributed into three wells of a 6 well plate coated with Matrigel. Since the Cas9/Puro construct provided transient puromycin resistance, the transfected iPSCs were subjected to puromycin selection two days after transfection. The selection conditions were 0.3-5 µg/mL puromycin for 48 hours. The iPSCs were cultured in feeder-free conditions with daily media changes until the appearance of iPSC colonies large enough to passage onto a 48-well plate coated with Matrigel (~10 days). Only isolated individual colonies were picked (one colony per well). For logistical purposes, each well/colony was given a clone number. For genotyping analysis, a portion of each colony was placed in a duplicate 48-well plate containing 50 µL of gDNA extraction solution (QuickExtract, EpiCenter, QE09050). Several colonies were picked for genotyping analysis. Once the mixed culture showed point mutation, it was subjected to a single-cell cloning process. In brief, the mixed culture was diluted to less than one cell per 200 µL culture media and dispersed into each well of a 96-well plate. The cells were allowed to grow for 4 to 6 weeks. Genomic DNA from single-cell colonies was extracted, and PCR was performed to amplify the targeted region. Primer sequences for amplifying the point mutation region are F-GTTCCCAACAGCACAGCAGGTC and R-GGACCTGTCACATTGGCATAACTG; or F-CAAAATTCACAGTGACTCAAGG and R-ATAAGGAGATCCGGTTGGATTGG. Amplification of the PCR product of 251 bp and 555 bp, respectively.

The genotyping strategy was designed to use PCR amplification and Fat I enzyme digestion to distinguish between clones containing the wildtype NOVA1$^{Hu/Hu}$ allele (GTA/V at rs906-08) and clones containing the NOVA1$^{Ar/Ar}$, NOVA1$^{Ko/Ar}$ and NOVA1$^{Ko/Ko}$ alleles. PCR products were further purified and sequenced to identify desired mutated clones. For one neurotypical iPSC line (WT83), three homozygous point mutation clones (c1, c15 and c28), three isogenic wildtype NOVA1$^{Hu/Hu}$ clones that passed through the same genome editing process (c6, c12 and c27), one homozygous NOVA1$^{Ko/Ko}$ clone (c9), and one NOVA1$^{Ko/Ar}$ clone (c13). For the second neurotypical iPSC (ASC), two homozygous point mutation clones NOVA1$^{Ar/Ar}$ (asc1 and asc2), two isogenic wildtype human variant clones NOVA1$^{Hu/Hu}$ (asc1 and asc2), one homozygous NOVA1$^{Ko/Ko}$ clone (2F9), and one NOVA1$^{Ko/Ar}$ (3G1). Clones that passed through all the genome editing processes and did not show any alteration in the target region were used as isogenic controls (NOVA1$^{Hu/Hu}$) (fig. S1A). Therefore, the target NOVA1$^{Hu/Hu}$ region translates to ATVKAVMEQS, NOVA1$^{Ar/Ar}$ translates to ATVKAVMEQS, and NOVA1$^{Ko/Ko}$ translate to ATVKA*, where an early stop codon will not produce a functional NOVA1. Then the selected clones were subjected to expansion. We track the expression of p53 and iPSC viability and proliferation rates due to a possible toxic role in the genome editing process (*44*, *45*).

Edit confirmation and off-target effects

We confirmed that the cells had the intended edits or lack thereof to *NOVA1* using whole-exome sequencing. We aligned these reads to hg19 using bwa mem (*46*) with default parameters and then filtered out alignments with MAPQ<20 and those that did not intersect with the genomic coordinates targeted by the capture, as provided by the manufacturer. We used the output of samtools mpileup (*47*) with default parameters to

calculate for each site with coverage of at least 20 the percentage of alignments supporting a SNP compared to the reference hg19 allele, and made a list of sites where these ratios were significantly different between the NOVA1$^{Hu/Hu}$, NOVA1$^{Ar/Ar}$, NOVA1$^{Ko/Ko}$ and NOVA1$^{Ko/Ar}$ cell lines (tables S3 and S4).

Beadchip array computational analysis

The analysis of beadchip arrays was performed using PennCNV software (version 1.05) (*48*) for the detection of Copy Number Variation (CNV). The analysis is based on a hidden Markov model (HMM) that integrates multiple sources of information to infer CNV calls for individually genotyped samples. The entire pipeline started by splitting the entire beadchip into single files each corresponding to an individual sample (kcolumn script). Next, individual samples were converted (compile_pfb script) to PFB (Population frequency of B allele). These files were then subjected to PennCNV analysis for quality measure (QC control, with waviness factor (WF value) set to be between -0.04 and 0.04) and CNV detection (detect_cnv script). CNV analysis investigates variation in allele-level copy number variation, indicating whether an allele is duplicated or deleted. CNV calls were set to regions with minimum 100 nucleotide length. To annotate the variations found, it was used the human reference genome (build hg19, annotation from UCSC genes) (*49*).

Whole exome sequencing analysis

The whole-exome sequencing (WES) was outsourced and was performed by the UCSD IGM Genomics Center, in 101 bp paired-end and an average of 100x coverage per sample. Raw exome data were filtered to recover high-quality (Phred-scaled quality score higher than 25) sequencing reads using NGS QC software package (*50*). Next, these WES pre-filtered libraries, in FASTQ file format, were aligned to the human reference genome (build Hg38) using BWA-MEM software (*51*) to report all best alignments. Unrelated alignments to the human exome were filtered out of the analysis. Valid alignments in SAM text file format were directly converted to sorted BAM binary file using SAMTOOLS software (*47*). Duplicated reads, as well as redundant alignments, were removed for all sequenced WES libraries. Alignments were then analyzed using FreeBayes to analyze and detect genetic variants. Next, we extracted the genomic loci of *NOVA1* gene to identify the CRISPR-induced mutations and also the related control isogenic cell lines. Genetic variants were considered homozygous when a minimum of 95% of covered reads showed the same mutation or heterozygous when a minimum of 30% of covered reads showed the same mutation. To verify whether these WES samples do have off-target variants that could be introduced by CRISPR system (*52*), the output BED files of FreeBayes software were annotated using several independent databanks, including Clinvar (*53*), dbSNP (https://www.ncbi.nlm.nih.gov/snp/), 1k genomic data consortium (*54*), GNomad (https://gnomad.broadinstitute.org/). Annotation was performed using bedTools software package in combination with GATK (*55*) and SpeedSeq software (*56*). Off-target variants were only considered positive when a genetic variant has the following attributes: rare in population databanks (minimum allele frequency (MAF) <0.001), occurs in exons of the protein-coding region, is annotated as pathogenic according to Clinvar annotation. To analyze all protein coding exons of *NOVA1* gene, we used the alignment files of exome analysis and used IGV (Integrate

Genomics Viewer) software (version 2.8.2). All the figures were extracted using IGV without any alterations using image processing applications to keep the fidelity of exome sequencing locus visualization.

Generation of functional cortical organoids

We used the protocol described elsewhere to generate functional cortical organoids (*20*). Briefly, iPSC colonies were dissociated and resuspended in mTeSR1 supplemented with 10 µM SB431542 (SB; Stemgent, Cambridge, MA, USA) and 1 µM Dorsomorphin (Dorso; R&D Systems, Minneapolis, MN, USA), and kept in suspension under rotation (95 rpm). After 3 days, mTeSR1 was substituted for Neurobasal (Life Technologies) supplemented with 1% Glutamax, 2% Gem21 NeuroPlex (Gemini Bio-Products, West Sacramento, CA, USA), 1% N2 NeuroPlex (Gemini Bio-Products), 1% MEM nonessential amino acids (NEAA; Life Technologies), 1% penicillin/streptomycin (PS; Life Technologies), 10 µM SB and 1 µM Dorso for 7 days. Then, the cells were maintained in Proliferative media supplemented with 20 ηg/mL of FGF2 and 20 ηg/mL EGF (PeproTech, Rocky Hill, NJ, USA). Next, cells were transferred to media supplemented with 10 µg/mL of BDNF, 10 µg/mL of GDNF, 10 µg/mL of NT-3 (all from PeproTech), 200 µM L-ascorbic acid and 1 mM dibutyryl-cAMP (Sigma-Aldrich). The cortical organoids were maintained using Neurobasal (Life Technologies) supplemented with 1% Glutamax, 2% Gem21 NeuroPlex, 1% N2 NeuroPlex, 1% NEAA, 1% PS, and analyzed after 1- or 2-months post-induction.

Mycoplasma testing

All cellular cultures were routinely tested for mycoplasma by PCR. Only negative samples were used in the study. Ten microliters of each sample were used for a PCR with the following primers:

Forward: GGCGAATGGGTGAGTAAC;
Reverse: CGGATAACGCTTGCGACCT.

Immunofluorescence staining

Cortical organoids were fixed with 4% paraformaldehyde, cryopreserved in 30% sucrose, and sliced in a cryostat (20 µm slices). The sliced samples were permeabilized/blocked with 0.1% triton X-100 and 3% FBS in PBS, and incubated with primary antibodies overnight at 4°C. Primary antibodies used in this study were: mouse anti-Nestin, Abcam (Cambridge, UK) ab22035, 1:250; rat anti-CTIP2, Abcam ab18465, 1:500; rabbit anti-SATB2, Abcam ab34735, 1:200; chicken anti-MAP2, Abcam ab5392, 1:2000; rabbit anti-Synapsin1, EMD-Millipore AB1543P, 1:500; mouse anti-NeuN, EMD-Millipore MAB377, 1:500; rabbit anti-Ki67, Abcam ab15580, 1:1000; rabbit anti-SOX2, Cell Signaling Technology 2748, 1:500; mouse anti-NOVA1, Santa Cruz sc-100334, 1:1000. After wash, the slices were incubated with secondary antibodies (Alexa Fluor 488-, 555- and 647-conjugated antibodies, Life Technologies, 1:1000) for 2 hours at room temperature. The nuclei were stained using DAPI solution (1 µg/mL). The slides were mounted using ProLong Gold antifade reagent and analyzed under a fluorescence microscope (Axio Observer Apotome, Zeiss).

Cell cycle and apoptosis assays

The NucleoCounter NC-3000 system (Chemometec, Denmark) was used to access the cell cycle and apoptosis. For the cell cycle, the system allows to automatically measure DNA content in stained cells using either a 365 ηm LED and DAPI, or a 530 ηm LED and propidium iodide. For cell death, the early apoptotic cells were quantified based on Annexin V binding and PI exclusion. Cells were stained with an Annexin V-CF488A conjugate and Hoechst 33342. After staining, cells were loaded into a slide and single cell fluorescence was quantified.

Reconstruction of 3D organoid surface models from their 2D outlines.

We utilized Fiji platform (*57*) for automatic image processing and watershed segmentation, varying processing parameters to achieve the best possible 2D organoid outline extraction (for details see ijm custom scripts). Only well-segmented outlines were selected that resulted in a consistent sample size of $n = 12$ 2D organoid outline per neurodevelopmental stage (Induction, Proliferation and Maturation), which resulted in a total of $n = 36$ organoids for each genotype.

We took *homology-free* approaches as cortical organoids have no specific biologically homologous structures. 3D organoid surface models/meshes were reconstructed using the following approach (see https://github.com/alikhuseynov/make3dOrganoid and/or Rmd scripts for data analysis). We performed Procrustes analysis i.e., linear transformations (on all surface mesh vertices) to remove the size and positional differences between 3D organoid models. We utilized mesh volume of 3D models as a measure of an organoid size (alternatively - centroid size of mesh vertices), and Dirichlet normal energy (DNE) of the surface as a shape metric or surface curvature/complexity measure (ie. the bending of the surface, high DNE values correspond to a ridged surface). All the data analysis and visualization were performed in R, mainly using the packages *Morpho*, *Rvcg*, *mesheR, Arothron, molaR, rgl, ggplot2* and *mgcv*.

Construction of RNA-seq Libraries

All RNA libraries were prepared using a modified SmartSeq2 method (*58*). A total of 2 µL of RNA (50 ηg) of each sample was reverse transcribed using Smartscribe Reverse Transcriptase (Clontech) in a 10 µL reaction containing a Smart-seq2 TSO according to manufacturer's instructions for 60 min at 42°C (*59*). The resulting cDNA was treated with 1 µL of 1:10 dilutions of RNAse A (Thermo) and Lambda Exonuclease (NEB) for 30 min at 37ºC. The cDNA was then amplified using KAPA Hifi Readymix 2x (KAPA) and incubated at 95ºC for 3 min, followed by 15 cycles of (98ºC for 20 s, 67ºC for 15 s, 72ºC for 4 min), with a final extension of 72ºC for 5 min. PCR product was then treated with our Tn5 enzyme (*60*) custom loaded with Tn5ME-A/R and Tn5ME-B/R. The Tn5 reaction was performed using 5 µL of the amplified product, 1 µL of the loaded Tn5 enzyme, 10 µL of water and 4 µL of 5x TAPS-PEG buffer. The sample was incubated at 55ºC for 5 min. The Tn5 reaction was then inactivated using 5 µL of 0.2% SDS. Five microliters of the Tn5 product were nicktranslated at 72°C for 6 min and further amplified using KAPA Hifi Polymerase (KAPA) using both Nextera Primer B and Nextera Primer A primers. The sample was incubated at 98ºC for 30 s, followed by 10 cycles of (98ºC for 10 s, 63ºC for 30 s, 72ºC for 2 min) with a final extension at 72ºC for

5 min. The Tn5 treated PCR product was then size selected using a 2% EX E-gel (Thermo) to a size range of 300-850 bp. All libraries were quantified using qPCR and Qubit prior to sequencing and pooled equally based on concentration. Furthermore, all libraries were prepared in parallel on the same day and pooled once prior to sequencing so they should be at similar ratios on each sequencing lane, which should mitigate any batch effects. The libraries were sequenced on an Illumina HiSeq 2x100 run on 2 lanes.

Expression verification

To ensure that the expected version of NOVA1 was being expressed in each sample, we amplified a 204 bp sequence of NOVA1 in the cDNA library with the primers 5' GGTAAAGATTATAGTTCCCAACAGC 3' and 5' CTTCTGGATGATAAGTTCAACAGC 3' using KAPA Taq polymerase with the provided kit protocol, an annealing temperature of 61^{o}C, 40 cycles, and 20 µL reaction volumes. We then ran the product on a gel and purified the band at 204 bp with the Zymoclean Gel DNA Recovery kit. Finally, we Sanger-sequenced the purified DNA with the same primers on an Applied Biosystems 310 Genetic Analyzer and compared it with the reference cDNA.

Expression quantification

To measure gene expression across all samples, we first aligned reads to hg19 using TopHat2 v2.0.8 (*61*), a spliced aligner, with default parameters. We calculated raw counts of fragments mapping to gene features in each library using featureCounts v1.5.1 (*62*) with the parameters -t exon -p -g gene id and GENCODE v19 as the annotation set (*63*). For comparisons of expression levels across both genes and samples, we normalized the raw counts using the transcripts per million (TPM) normalization method (*64*). For differential expression analysis, we used DESeq2 with unnormalized raw counts and parameters 'lfcThreshold = 1, altHypothesis ="greaterAbs", alpha = 0.01' (*65*).

Single-nucleus RNA-sequencing data generation

Nuclei preparation was adapted from elsewhere (*66*). Frozen whole cortical organoids were thawed at 37°C for 2 minutes before centrifugation for 3 minutes at 100 g at 4°C. Supernatant was discarded, and organoids were resuspended in 1 mL of douncing buffer consisting of 0.25 M sucrose (S1888, Sigma), 25 mM KCl (AM9610G, Invitrogen), 5 mM $MgCl_2$ (194698, Mp Biomedicals Inc.), 10 mM Tris-HCl pH 7.5 (15567027, Thermo Fischer Scientific), 1 mM DTT (D9779, Sigma), protease inhibitor (05056489001, Roche), 0.1% Triton X-100 (T8787-100ML, Sigma), and 0.2 U/µL RNasin RNase inhibitor (PAN21110, Promega) in molecular biology grade water (46000-CM, Corning). Organoids were transferred to a dounced homogenizer on ice and dounce 10 times with a loose plunger and 25 times with a tight plunger. The suspension was passed through a Celltrix 30 µm filter (04-004-2326, Sysmex) and washed with 300 µL of douncing buffer before being centrifuged for 10 minutes at 1000 g using the previously described settings. Pellet was then washed with an additional 1 mL of douncing buffer without Triton X-100 and centrifuged again for 10 minutes at 1000 g. Supernatant was then discarded, and the pellet was resuspended in 600 µL of sort buffer consisting of 1 mM EDTA (15575020, Invitrogen), 0.2 U/ µL RNasin, and 2% BSA in PBS before staining with 3 µM DRAQ7 (#7406S, Cell Signaling Technology). Nuclei

were incubated on ice for 10 minutes and ~75,000 nuclei were sorted using a 100 µm chip in an SH800 sorter (Sony) into 50 µL of collection buffer consisting of 1 U/µL RNasin and 5% BSA in PBS. Samples were centrifuged for 15 minutes at 1000 g and resuspended in 35 µL of reaction buffer consisting of 0.2 U/µL RNasin and 1% BSA in PBS, and nuclei were visually inspected and manually counted using a hemocytometer before loading 12,000 onto a Chromium Controller for 10x GEM generation in the Single Cell 3' v3 kit (1000075, 10x Genomics).

Libraries were generated using the Chromium Single Cell 3′ Library Construction Kit v3 (1000078, 10x Genomics) according to manufacturer specifications. CDNA was amplified for 12 PCR cycles. SPRISelect reagent (Beckman Coulter) was used for size selection and clean-up steps. Final library concentration was assessed by Qubit dsDNA HS Assay Kit (Thermo-Fischer Scientific) and fragment size was checked using Tapestation High Sensitivity D1000 (Agilent) to ensure that fragment sizes were distributed normally about 500 bp. Libraries were sequenced using a NextSeq500 or HiSeq4000 (Illumina) using these read lengths: Read 1: 28 cycles, Read 2: 91 cycles, Index 1: 8 cycles.

Single nucleus RNA-seq analysis

Raw sequencing data was demultiplexed and preprocessed using the Cell Ranger software package v3.0.2 (10x Genomics). Raw sequencing files were first converted from Illumina BCL files to FASTQ files using cellranger mkfastq. Demultiplexed FASTQs were aligned to the GRCh38 reference genome (10x Genomics), and reads for exonic and intronic reads mapping to protein coding genes, long non-coding RNA, antisense RNA, and pseudogenes were used to generate a counts matrix using cellranger count; expect-cells parameter was set to 5,000. Next, count matrices for individual datasets were processed using the Seurat v3.0.2 R package (*67*) to assess dataset quality. Features represented in at least 3 cells and barcodes with between 500 and 5,000 genes were used for downstream processing. Counts were log-normalized and scaled by a factor of 10,000 using NormalizeData. To identify variable genes, FindVariableFeatures was run with default parameters except for nfeatures = 3000 to return the top 3,000 variable genes. All genes were then scaled using ScaleData, which transforms the expression values for downstream analysis. Next, principal component analysis was performed using RunPCA with default parameters and the top 3,000 variable features as input. The first 20 principal components were used to run clustering using FindNeighbors and FindClusters (parameter res = 0.5). To generate UMAP coordinates RunUMAP was run using the first 20 principal components; the default UMAP dependency for this version ran with umap-learn and was used here. Doublet scores (pANN) were generated for cell barcodes using DoubletFinder (*68*) using the parameters pN =0.15 and pK = 0.001; the anticipated collision rate was set by specifying 1% collisions per thousand nuclei for individual datasets.

To understand the significance of *NOVA1* variants, datasets for the 1-month time-point corresponding to $NOVA1^{Hu/Hu}$, $NOVA1^{Ar/Ar}$, $NOVA1^{Ko/Ar}$ and $NOVA1^{Ko/Ko}$ were combined; for the time-course analysis, the 1 month and 2 months $NOVA1^{Hu/Hu}$ and $NOVA1^{Ar/Ar}$ individual datasets were combined. For both merged datasets, all downstream processing was identical. First, individual datasets were merged using the merge function in Seurat to combine the count matrices and designate unique barcodes.

Cell barcodes classified as doublets by DoubletFinder were removed from downstream analysis. Metadata was also encoded for each barcode, and used to generate the percentage contribution tables and bar plots of *NOVA1* variants and time points to the final annotated clustering. The merged datasets were processed as described above; clusters were identified using FindNeighbors and FindClusters (res = 0.4), and UMAP coordinates were generated using the first 30 principal components as input for RunUMAP; the UMAP dependency used was umap-learn. To regress out dataset-specific effects, the Harmony R package (*69*) was used, and the recomputed principal components were used to re-cluster the cells and rerun UMAP using the above parameters. Finally, an initial differential gene expression test was run using FindAllMarkers with parameters logFC = 0.25, min.pct = 0.25, and only.pos = FALSE to identify cluster-specific genes. These genes, along with canonical markers previously used to identify cortical organoid populations (*20*), were used to manually aggregate the unsupervised clustering by assigning multiple clusters the same identity on the basis that such cluster were biologically alike. Cluster-specific genes were then identified for the annotated clustering using FindAllMarkers with the mentioned parameters.

eCLIP library preparation

The assay was performed as previously described (*28, 70*). Briefly, cortical organoids were mechanically and enzymatically dissociated into cell suspension. Fifteen to forty million cells were UV-crosslinked (400 mJ/cm^2, 254 ηm) and snap-frozen. Lysed pellets were sonicated and treated with RNaseI for RNA fragmentation. Two percent of lysate was retained for preparation of a size-matched input library, and the remaining 98% was subject to immunoprecipitation (IP) using 50 µL of anti-NOVA1 antibody (Santa Cruz; 512Y sc-100334), coupled to magnetic dynabeads (Invitrogen 11203D). Bound RNA fragments were dephosphorylated and 3'-end ligated with an RNA adapter. Protein-RNA complexes from both input and IP samples were run on SDS polyacrylamide gel and transferred to nitrocellulose membrane for extraction of bound RNA fragments. Membrane regions from the size of the protein to 75 kDa above the protein size were cut and RNA was released with proteinase K. Input samples were then dephosphorylated and 3'-end ligated with an RNA adapter. Reverse transcription was performed with AffinityScript (Agilent) and cDNAs were 5'-end ligated with a DNA adaptor. cDNA products were amplified with Q5 PCR mix (NEB) to yield a sequencing library. Libraries were sequenced on the Illumina HiSeq4000 in SE75 mode to a depth of ~40 million reads per library.

Computational analysis of eCLIP sequencing data

Reads were processed as described previously (*28*). Briefly, reads were adapter-trimmed and mapped to human-specific repetitive elements from RepBase (version 18.05) by STAR (*71*). Repeat-mapping reads were removed, and remaining reads mapped to the human genome assembly hg19 with STAR. PCR duplicate reads were removed using the unique molecular identifier (UMI) sequences in the 5' adapter to generate 'usable reads' used in peak calling. Peaks were called on the usable reads by CLIPper (*72*) and assigned to gene regions annotated in GENCODE (v19) with the following descending priority order: 3' splice site, 5' splice site, CDS, 3'UTR, 5'UTR, proximal intron, and distal intron, noncoding sequences. Proximal intron regions are defined as

extending up to 500 bp from an exon-intron junction. Each peak was normalized to the size-matched input (SMInput) by calculating the fraction of the number of usable reads from immunoprecipitation to that of the usable reads from the SMInput. Peaks were deemed significant at ≥4-fold enrichment and $p \leq 10^{-5}$ (χ^2 test, or Fisher's exact test if the observed or expected read number in eCLIP or SMInput was below 5). Peaks passing significance thresholds in either replicate were kept for downstream analyses.

HOMER sequence analyses were performed as following: Fasta files were generated from peak regions with bedtools getfasta run in stranded mode. The homer function 'findMotifs' was used with the -rna flag in stranded mode to determine enriched motifs. A list of consisting of the same number of peak regions matched for genic distribution was used as the background. Top two significant motifs and associated p-values are shown.

Kmer analyses were performed as following: 6-mer sequences were counted in peak regions using kvector (*https://github.com/olgabot/kvector*). The resulting counts were summed and normalized to total count to generate a normalized enrichment score for each 6-mer within peaks.

Splicing quantification

To quantify splicing, we used juncBASE (*36*) to calculate a PSI (percent spliced in) value for each alternative splicing event. We ran juncBASE on the read alignments, which we created as described above, using GENCODE v19 (*63*) as the annotation set and the parameters "--by chrin steps 1B, 2, 4, and 5"; "--majority_rules" in step 1B; and "--jcn_seq_len 188" in steps 5 and 6. To call differentially spliced events, we used the pairwise Fisher's test script with parameters "--jcn_seq_len 188 --method BH". We considered a splicing event to be differentially spliced if the replicates of the human control were not significantly different from each other, but the replicates of the sample were all significantly different from the control. To visualize the differences in splicing between the different cell lines and time points, we performed principal components analysis (PCA) on the PSI values for each sample and cassette exon splicing event. The replicates from each sample clustered together in the first two principal components. We compared each of the principal components to *NOVA1* expression and found that the second component is negatively correlated with *NOVA1* expression with $R^2 = 0.65$.

Gene Ontology analysis

We found Gene Ontology (GO) terms that were significantly enriched among genes with differential splicing using func v0.4.8 (*73*). For a given time-point, we used as a background set all genes with expression level higher than 5 TPM in at least one replicate at that time point. We then used the func hyper script and the refinement script it creates to calculate FDR-corrected p-values for overrepresentation of terms.

Western blot

Cortical organoids were lysed in RIPA buffer with protease inhibitors. NeuN, GFAP, CTIP2, TBR1, FOXG1, Homer1, Syn1, VGlut1, PSD95 and NOVA1 were used as primary antibodies, as previously performed (*20*). IRDye 800CW goat anti-rabbit and IRDye 680RD goat anti-mouse (1:6000) were used as secondary antibodies. Signal

intensities were measured using the Odyssey Image Studio and normalized by actin relative quantification.

Synaptic puncta quantification

Co-localized Vglut and Homer1 puncta were quantified after three-dimensional reconstruction of z-stack images. Images were taken randomly from three different experiments. Only puncta overlapping MAP2-positive processes were scored.

Quantitative Multiplex co-immunoprecipitation

QMI analysis was performed as previously described (*29, 30*). Briefly, cortical organoids were homogenized in lysis buffer [150 mM NaCl, 50 mM Tris (pH 7.4), 1% NP-40, 10 mM NaF, 2 mM sodium orthovanadate + Protease/phosphatase inhibitor cocktails (Sigma)] using a glass tissue homogenizer, incubated in lysis buffer for 15 minutes, centrifuged at high speeds to remove nuclei and debris, and protein concentration was determined using a Pierce BCA kit. A master mix containing equal numbers of each antibody-coupled Luminex bead class was prepared and distributed into cell lysate samples in duplicate. Protein complexes were immunoprecipitated from samples containing equal amounts of protein overnight at 4^{o}C, washed twice in ice-cold Fly-P buffer [50 mM tris (pH 7.4), 100 mM NaCl, 1% bovine serum albumin, and 0.02% sodium azide], and distributed into as many wells of a 96-well plate as there were probes, on ice. Biotinylated detection antibodies were added and incubated for 1 hour, with gentle agitation at 500 rpm at 4^{o}C. Following incubation, microbeads and captured complexes were washed three times in Fly-P buffer using a Bio-Plex Pro II magnetic plate washer in the cold room. Microbeads were incubated for 30 min with streptavidin-PE on ice, washed three times, and resuspended in 125 μL of ice-cold Fly-P buffer. Fluorescence data were acquired on a customized, refrigerated Bio-Plex 200 instrument calibrated and routinely validated according to the manufacturer's recommendations.

Data preprocessing and inclusion criteria XML output files were parsed to acquire the raw data for use in MATLAB while XLS files were used for input into R statistical packages. For each well from a data acquisition plate, data were processed to (i) eliminate doublets based on the doublet discriminator intensity (>5000 and <25,000 arbitrary units; Bio-Plex 200), (ii) identify specific bead classes within the bead regions used, and (iii) pair individual bead PE fluorescence measurements with their corresponding bead regions. This processing generated a distribution of PE intensity values for each pairwise protein co-association measurement.

Adaptive nonparametric analysis with empirical alpha cutoff (ANC) (*30*) is used to identify high-confidence, statistically significant differences (corrected for multiple comparisons) in bead distributions on an individual interaction basis. ANC was conducted as described in (*29*). We required that hits are present in 4 of 4 replicates at an adjusted $P < 0.05$. The α-cutoff value required per experiment to determine statistical significance was calculated to maintain an overall type I error of 0.05 (adjusted for multiple hypothesis testing with Bonferroni correction), with further empirical adjustments to account for technical errors.

Bead distributions used in ANC were collapsed into a single median fluorescent intensity (MFI), which was averaged across duplicate samples and input into the WGCNA package for R (*31*). Data were filtered to remove weakly detected interactions

('noise', MFI <100), and batch effects were removed using the COMBAT function for R (*29*), with "experiment number" as the "batch" input. Post-Combat data was log2 transformed prior to CNA analysis. Closely related protein co-associations were assigned to arbitrary color-named modules by the WGCNA program. Modules whose eigenvectors significantly ($P < 0.05$) correlated with a given genotype were considered significant, and protein co-associations belonging to modules of interest were defined as those with a probability of module membership (p.MM) < 0.05.

ANC data and CNA data were merged as described previously (*29*, *30*) to produce a high-confidence set of interactions that were both individually significantly different in comparisons between experimental groups, and that belonged to a larger module of co-regulated interactions that was significantly correlated with the experimental group. Hierarchical Clustering by Principal Components (HCPC) was performed on the Combat-normalized data in R using the PCA function in the FactoMineR package, followed by the HCPC function (*74*).

Multi-electrode array (MEA) recording

MEA electrophysiological recordings were performed as described elsewhere (*12*). Briefly, cortical organoids were plated on 12-well MEA plates (Axion Biosystems, Atlanta, GA, USA). Recordings were performed using a Maestro MEA system and AxIS Software Spontaneous Neural Configuration (Axion Biosystems). Spikes were detected with AxIS software using an adaptive threshold crossing set to 5.5 times the standard deviation of the estimated noise for each electrode. Bright-field images were captured from each well to assess for neural density and electrode coverage over time.

Spike sorting algorithm was used for the data in each electrode. PCA was used in the shapes of the spikes (*75*). The shapes were defined as the LFP sequence of 12 temporal points before, and 12 temporal points after the peak of the spike, including the peak, totalizing 25 temporal points. In addition, a clustering algorithm was applied based on k-means in order to detect the different neurons, using the first and second component and gap criteria (*76*). Following the spike sorting process, both firing rate (FR) and coefficient of variation (CV) were determined from the interspike interval (ISI) series.

Statistical analysis

Data are presented as mean ± s.e.m., unless otherwise indicated, and it was obtained from different samples. No statistical method was used to predetermine the sample size, and no adjustments were made for multiple comparisons. The statistical analyses were performed using Prism software (GraphPad, San Diego, CA, USA). Student's t test, Mann–Whitney-test, or ANOVA with post hoc tests were used as indicated. Significance was defined as $P < 0.05$(*), $P < 0.01$(**), or $P < 0.001$(***). Blinding was used for comparing samples.

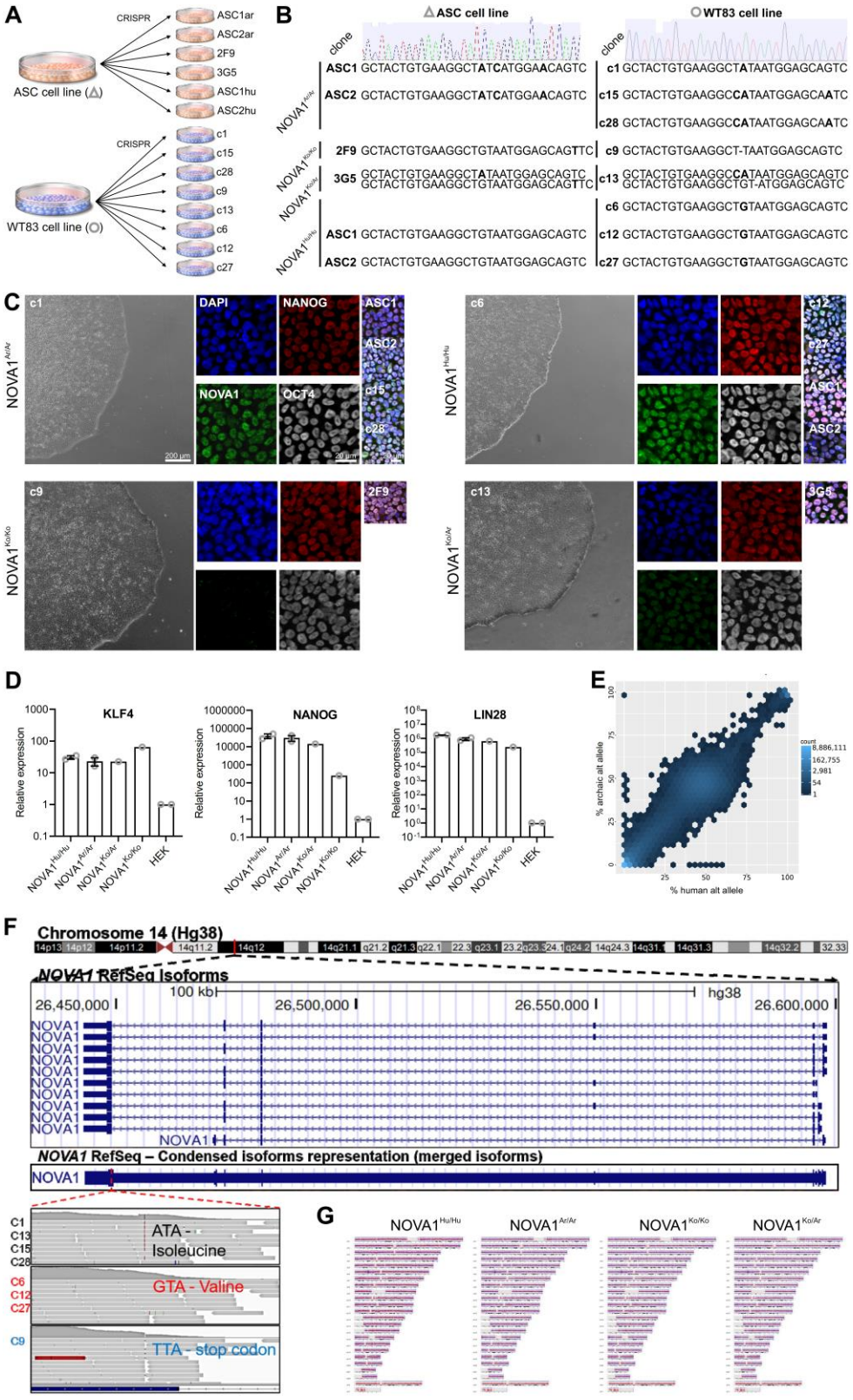


Fig. S1. Characterization of NOVA1$^{Ar/Ar}$ genetic variant introduction in modern human cell lines.

(A) Summary of neurotypical cell lines used and the CRISPR-edited clones representing all genotypes. Note that each cell line is depicted with a symbol. **(B)** Sanger sequencing of RT-PCR products from NOVA1 transcript shows that NOVA1$^{Hu/Hu}$ and NOVA1$^{Ar/Ar}$ cell lines both express the intended version of *NOVA1*. The genotypes were confirmed for all clones and all cell lines by sequencing. **(C)** Brightfield and immunofluorescence images of isogenic pairs of NOVA1 iPSC colonies showing the expression of the pluripotency markers Nanog and OCT4 for all clones. **(D)** Quantitative PCR of isogenic pairs of NOVA1 cell lines showing the expression of the pluripotency markers *KLF4*, *Nanog* and *LIN-28*. **(E, F)** Alignment of exome-sequencing reads from NOVA1$^{Hu/Hu}$ and NOVA1$^{Ar/Ar}$ cell lines to hg19 shows that the intended edit to *NOVA1* is the only homozygous change to the NOVA1$^{Ar/Ar}$ line compared to NOVA1$^{Hu/Hu}$. **(G)** BeadArray analysis was performed to infer CNV calls for individual genotyped samples. No aberration was observed in all the cell line genotypes. The list of CNV alterations and off-targets can be found in Tables S3 and S4.

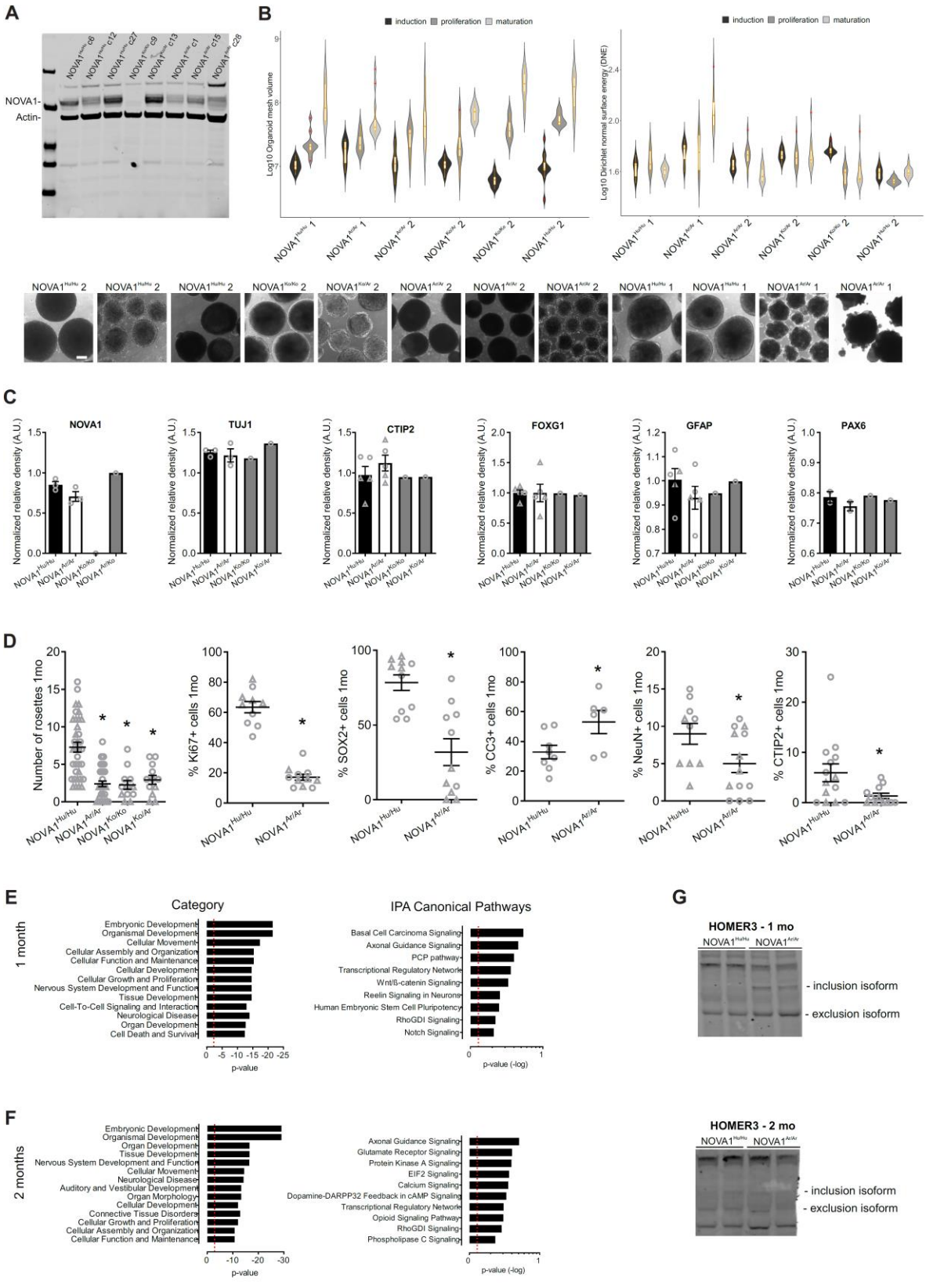
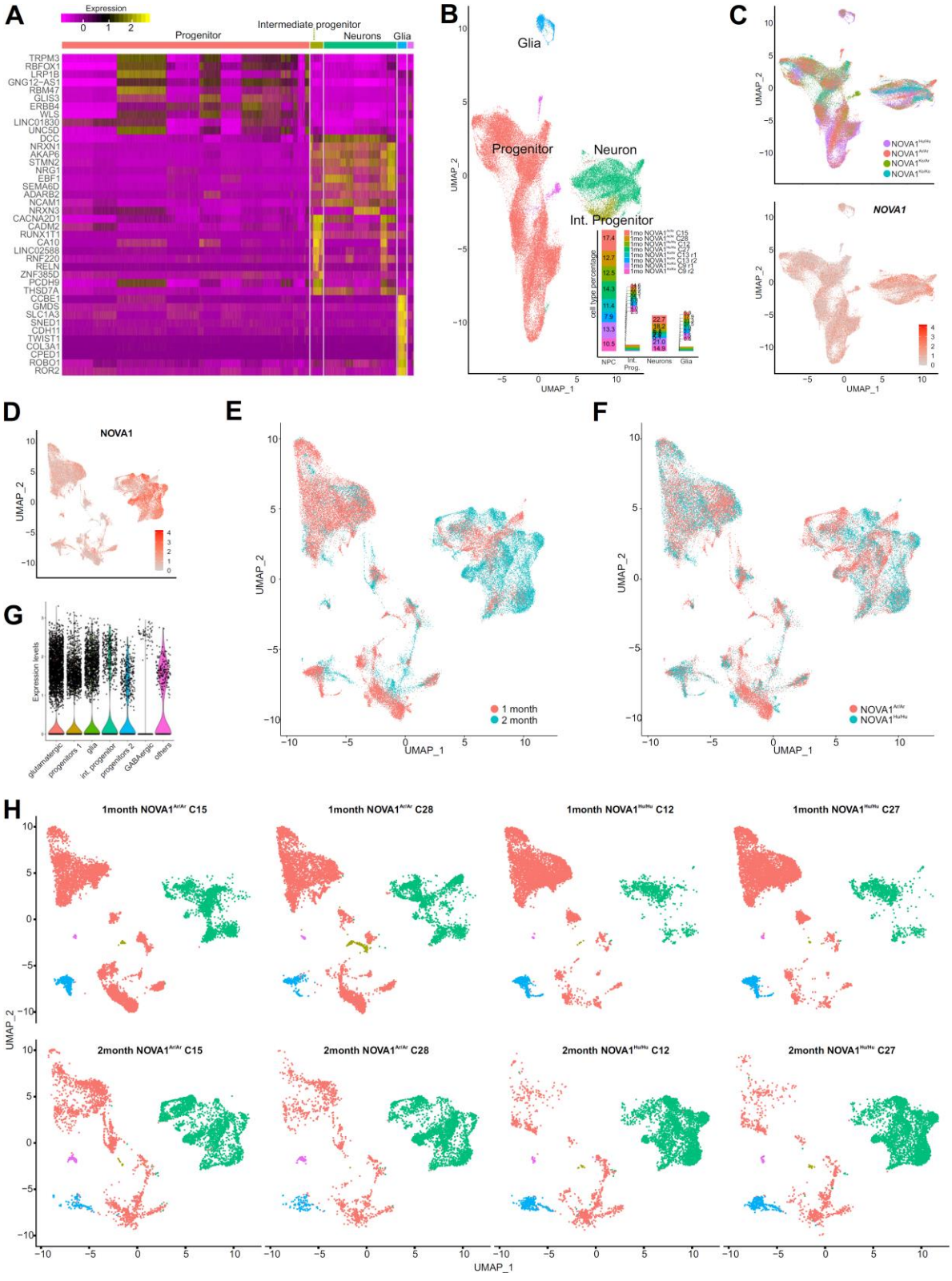
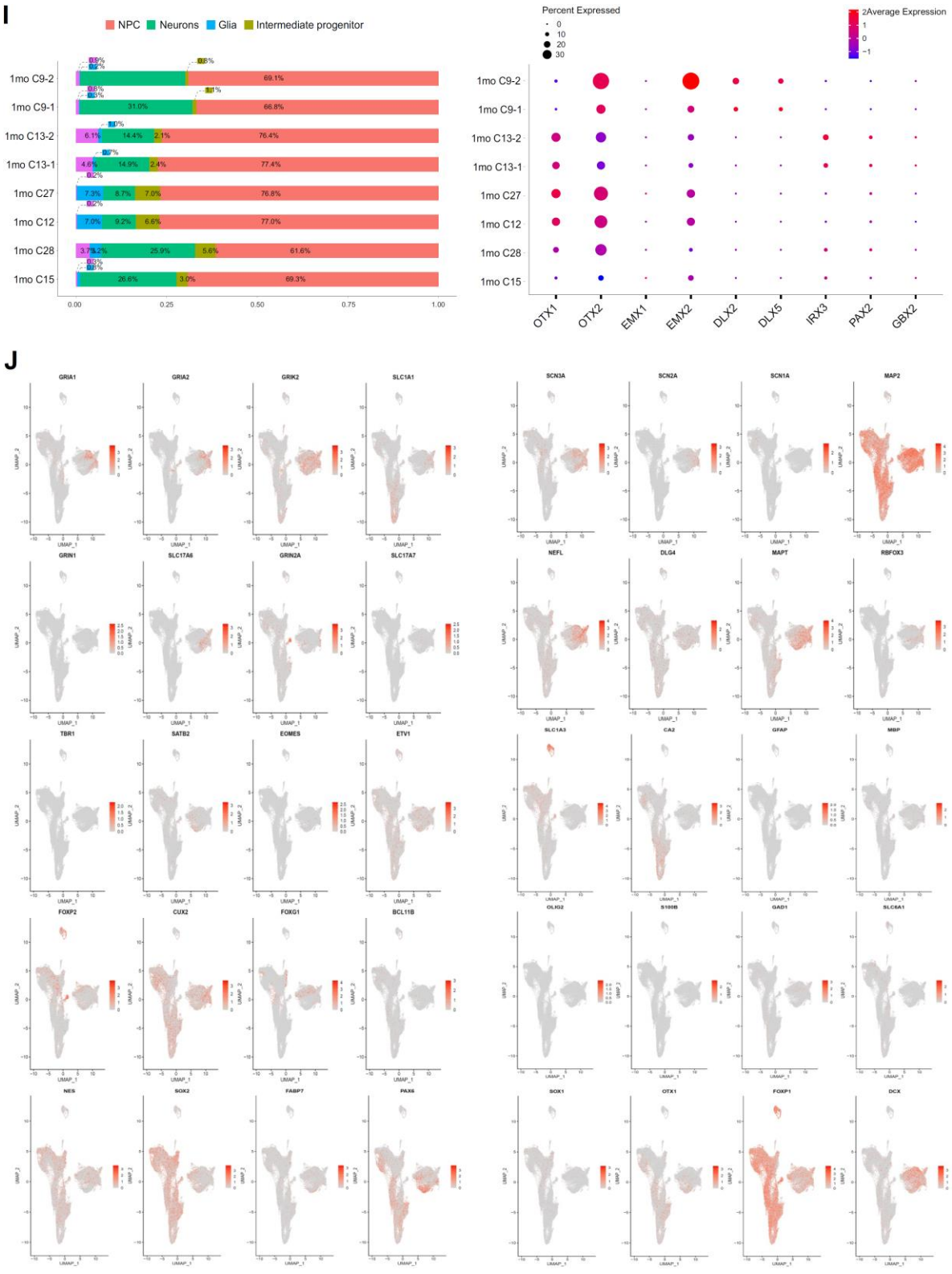


Fig. S2. Independent validation of morphology, gene expression, splicing and pathway analyses on cortical organoids. (A) Western blot of NOVA1 expression of all

genotypes and clones of cortical organoids. **(B)** Correlations between organoid size (mesh volume) and shape (Dirichlet normal surface energy). DNE was used as a shape metric or surface rugosity/curvature measure. All cell lines were used. **(C)** Western blot of specific markers indicating a similar level of neuronal and glial markers in 1-month NOVA1 cortical organoids (each cell line is depicted as a different symbol). **(D)** Number of rosettes per 1-month cortical organoids and proportion of different cell markers quantified by percentage in immunostainings. Number of rosettes: one-way ANOVA Dunnett's multiple comparisons test; other markers: two-sided unpaired Student's t test, $*P < 0.05$. **(E, F)** Enriched GO terms in the set of metabolic categories and canonical pathways differentially regulated in $NOVA1^{Ar/Ar}$ in early and late stages of maturation. **(G)** Western blot analysis validating the presence of different splicing variants of HOMER3 at different maturation stages predicted by the alternative splicing expression analyses.


A
Expression
0 1 2
Progenitor
Intermediate progenitor
Neurons
Glia
TRPM3
RBFOX1
LRP1B
GNG12-AS1
RBM47
GLIS3
ERBB4
WLS
LINC01830
UNC5D
DCC
NRXN1
AKAP6
STMN2
NRG1
EBF1
SEMA6D
ADARB2
NCAM1
NRXN3
CACNA2D1
CADM2
RUNX1T1
CA10
LINC02588
RNF220
RELN
ZNF385D
PCDH9
THSD7A
CCBE1
GMDS
SLC1A3
SNED1
CDH11
TWIST1
COL3A1
CPED1
ROBO1
ROR2
B
Glia
Progenitor
Neuron
Int. Progenitor
cell type percentage
NPC
Int. Prog.
Neurons
Glia
UMAP_1
UMAP_2
C
NOVA1
D
NOVA1
E
1 month
2 month
F
G
Expression levels
glutamatergic
progenitors 1
glia
int. progenitor
progenitors 2
GABAergic
others
H
1month NOVA1Ar/Ar C15
1month NOVA1Ar/Ar C28
1month NOVA1Hu/Hu C12
1month NOVA1Hu/Hu C27
2month NOVA1Ar/Ar C15
2month NOVA1Ar/Ar C28
2month NOVA1Hu/Hu C12
2month NOVA1Hu/Hu C27

Cont.


I
NPC
Neurons
Glia
Intermediate progenitor
1mo C9-2
1mo C9-1
1mo C13-2
1mo C13-1
1mo C27
1mo C12
1mo C28
1mo C15
0.00
0.25
0.50
0.75
1.00
Percent Expressed
0
10
20
30
2Average Expression
1
0
−1
OTX1
OTX2
EMX1
EMX2
DLX2
DLX5
IRX3
PAX2
GBX2
J
GRIA1
GRIA2
GRIK2
SLC1A1
SCN3A
SCN2A
SCN1A
MAP2
GRIN1
SLC17A6
GRIN2A
SLC17A7
NEFL
DLG4
MAPT
RBFOX3
TBR1
SATB2
EOMES
ETV1
SLC1A3
CA2
GFAP
MBP
FOXP2
CUX2
FOXG1
BCL11B
OLIG2
S100B
GAD1
SLC6A1
NES
SOX2
FABP7
PAX6
SOX1
OTX1
FOXP1
DCX

Cont.

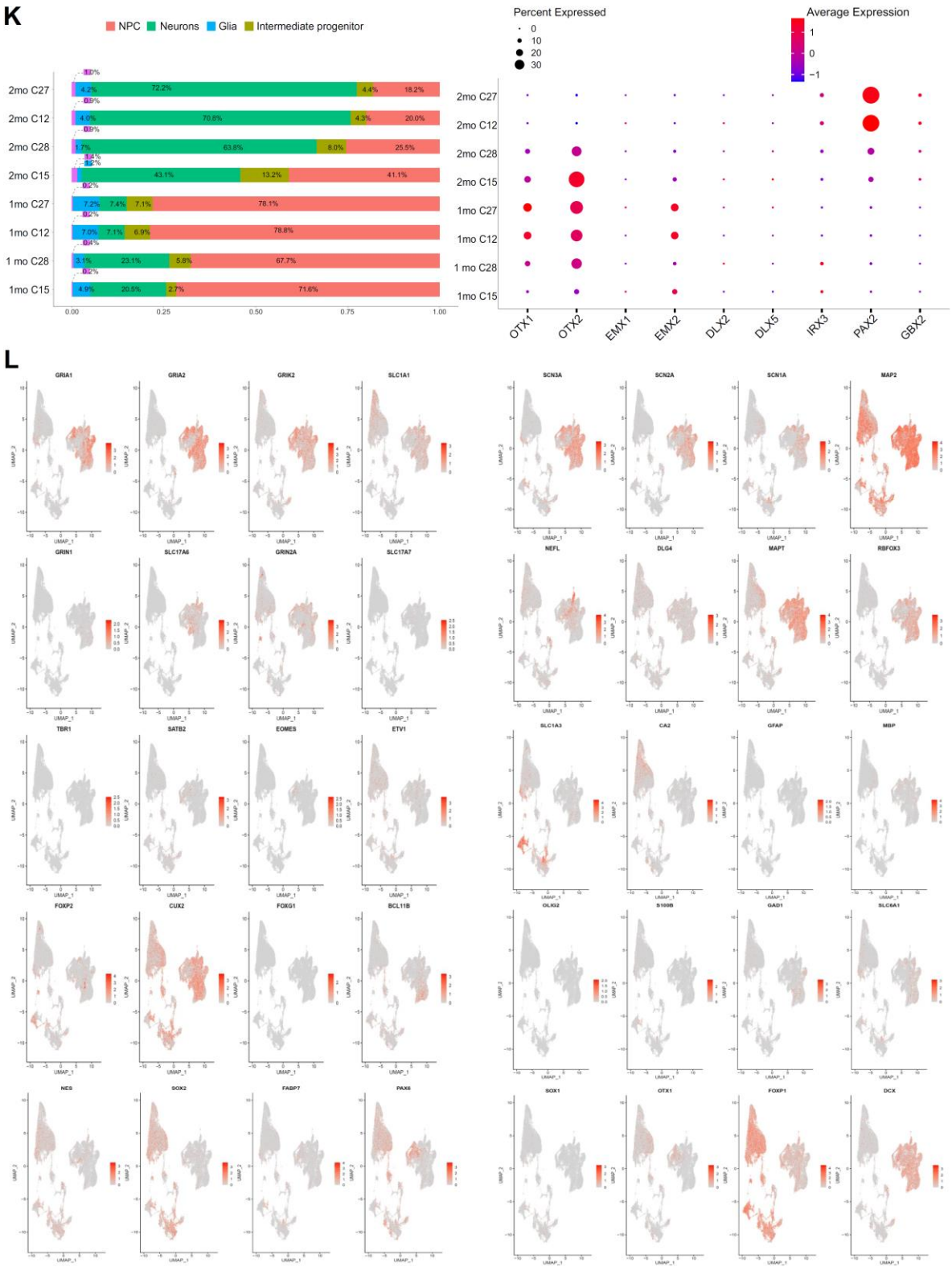


Fig. S3. Single-cell characterization of NOVA1 cortical organoids. (A) Heatmap of the top differentially expressed genes in different cell populations of 1-month old organoids. **(B)** Uniform manifold approximation and projection (UMAP) of the

integrated datasets of 53,734 nuclei for four NOVA1 genotypes of 1 month old organoids colored by four main cell clusters (13,333 nuclei for $NOVA1^{Hu/Hu}$, 17,456 nuclei for $NOVA1^{Ar/Ar}$, 13,376 nuclei for $NOVA1^{Ko/Ko}$, and 9569 nuclei from one clone of $NOVA1^{Ko/Ar}$). The inset represents population abundance of the different genotypes across the annotated cell types. **(C)** UMAP from integrating datasets of different NOVA1 genotypes of 1-month cortical organoids colored by genotypes. *NOVA1* gene expression is highlighted in the bottom plot. **(D)** UMAP plot of *NOVA1* gene expression in 1 and 2-month-old cortical organoids. **(E, F)** UMAP plot of the integrated datasets colored by main cell clusters, time point and genotype. **(G)** Violin plot showing the expression of *NOVA1* in different cell types during 10-month cortical organoid maturation. Resource data obtained from elsewhere (*20*). **(H)** UMAP plot depicting different NOVA1 clones and its cellular variability in different genotypes. **(I, J)** Cell type proportion and UMAP plot overlaid with gene expression of different cortical markers in 1-month old organoids with four different NOVA1 genotypes ($NOVA1^{Hu/Hu}$, $NOVA1^{Ar/Ar}$, $NOVA1^{Ko/Ko}$, $NOVA1^{Ko/Ar}$). **(K, L)** Cell type proportion and UMAP plot overlaid with gene expression of different cortical markers in the integrated dataset of 1- and 2-month old organoids with two different NOVA1 genotypes ($NOVA1^{Hu/Hu}$ or $NOVA1^{Ar/Ar}$).

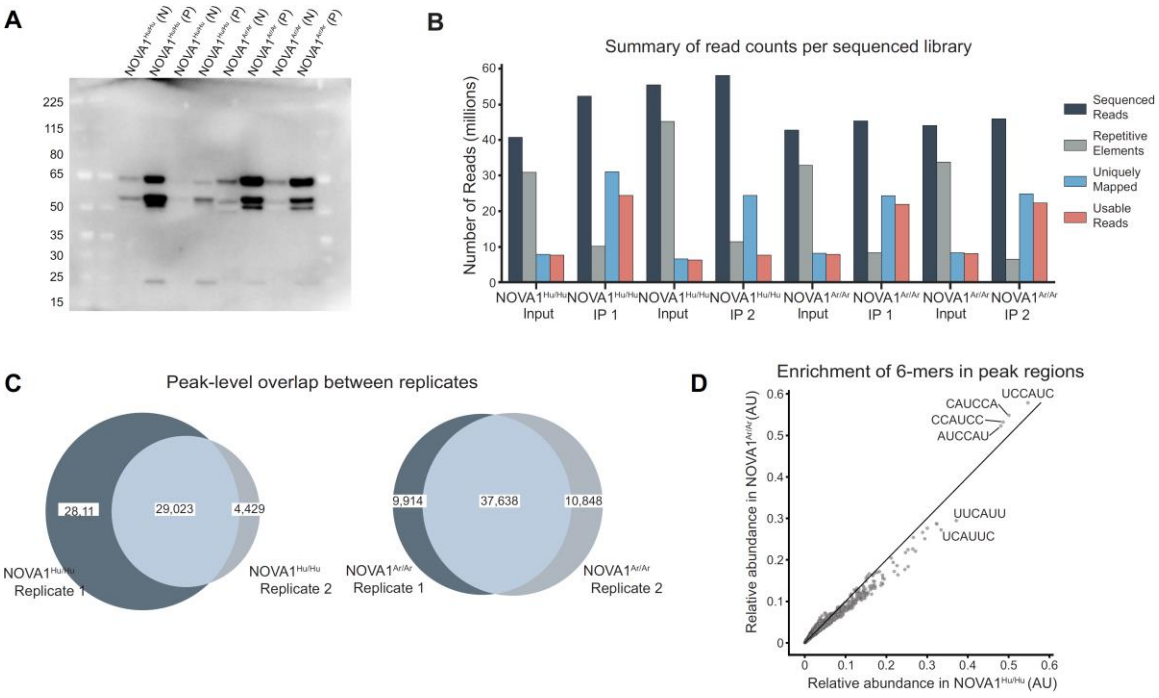


Fig. S4. eCLIP, splicing and motif analyses of *NOVA1* on cortical organoids. (A) IP-western blot for samples used in eCLIP. The experiment was performed on biological duplicates of each genotype. N represents input sample (2% of sample loaded on the gel), P represents IP sample (20% of sample loaded on gel). IP and western were both performed with the same NOVA1 antibody. NOVA1 band is at 50 kDa. **(B)** Read counts from eCLIP libraries. Total sequenced reads (dark grey), reads mapped to a database of repetitive elements (light grey), reads uniquely mapped to the hg19 genome (blue), and uniquely mapping reads that have been filtered for unique molecular barcodes (UMIs) to the final set of 'Usable reads' (red). Usable reads were used as the input for peak calling. **(C)** Venn diagram of peaks overlapping between biological duplicates of each genotype. All peaks have fold change > 4 relative to input and $P < 0.001$ (chi squared test). **(D)** Relative abundance of all 6-mer sequences found within peak regions per genotype. AU=arbitrary units.

Table S1.

List of variants for which no modern humans have the archaic version.

chr (hg19)	pos (hg19)	Ensembl Gene ID	HGNC name
1	6694660	ENSG00000007923	DNAJC11
1	79106805	ENSG00000137959	IFI44L
1	245582905	ENSG00000162849	KIF26B
2	27535311	ENSG00000115204	MPV17
2	71062833	ENSG00000116031	CD207
2	73438011	ENSG00000214513	NOTO
2	95831534	ENSG00000163067	ZNF2
3	47469149	ENSG00000114650	SCAP
3	196674495	ENSG00000119227	PIGZ
4	89410317	ENSG00000138646	HERC5
5	10250094	ENSG00000150753	CCT5
5	77745853	ENSG00000085365	SCAMP1
5	82837946	ENSG00000038427	VCAN
5	139197136	ENSG00000146005	PSD2
5	156721863	ENSG00000055163	CYFIP2
7	17375392	ENSG00000106546	AHR
7	149426305	ENSG00000133619	KRBA1
8	39537618	ENSG00000168619	ADAM18
8	39564352	ENSG00000168619	ADAM18
8	48805816	ENSG00000253729	PRKDC
8	54975904	ENSG00000120992	LYPLA1
9	6606647	ENSG00000178445	GLDC
9	125563200	ENSG00000165204	OR1K1
9	140139881	ENSG00000188163	FAM166A
10	118383463	ENSG00000165862	PNLIPRP2
11	795366	ENSG00000177542	SLC25A22
11	6654769	ENSG00000166341	DCHS1
11	28119295	ENSG00000121621	KIF18A
11	64884957	ENSG00000174276	ZNHIT2
11	64893151	ENSG00000149792	MRPL49
11	65154602	ENSG00000126391	FRMD8
11	111853106	ENSG00000150764	DIXDC1
11	124253181	ENSG00000204293	OR8B2
11	129772293	ENSG00000170325	PRDM10
12	6883790	ENSG00000089692	LAG3
12	7080210	ENSG00000126749	EMG1
14	26918100	ENSG00000139910	NOVA1
15	40912860	ENSG00000137812	CASC5
15	40915640	ENSG00000137812	CASC5
15	42742312	ENSG00000103994	ZNF106
15	81173308	ENSG00000103888	KIAA1199
16	138772	ENSG00000103148	NPRL3
16	66947064	ENSG00000166589	CDH16
16	88804443	ENSG00000103335	PIEZO1
17	27959034	ENSG00000141298	SSH2
17	27959258	ENSG00000141298	SSH2
17	41931199	ENSG00000161649	CD300LG
17	62290457	ENSG00000136478	TEX2
17	73753035	ENSG00000132470	ITGB4
17	79514871	ENSG00000185504	C17orf70
17	80006980	ENSG00000169733	RFNG
19	6685111	ENSG00000125730	C3
19	24116551	ENSG00000213967	ZNF726
19	36214632	ENSG00000272333	KMT2B
19	46265288	ENSG00000237452	AC074212.3
19	51835892	ENSG00000186806	VSIG10L
19	55489189	ENSG00000022556	NLRP2
20	33337529	ENSG00000198646	NCOA6
21	45671052	ENSG00000142182	DNMT3L
22	20779973	ENSG00000244486	SCARF2
22	40760978	ENSG00000239900	ADSL

Table S2.

Description of cell lines and clones used for each experiment.

Experiment \ Cell lines	WT83 C6 Hu/Hu	WT83 C12 Hu/Hu	WT83 C27 Hu/Hu	WT83 C9 Ko/Ko	WT83 C13 Ko/Ar	WT83 C1 Ar/Ar	WT83 C15 Ar/Ar	WT83 C28 Ar/Ar	ASC C1 Hu/Hu	ASC C2 Hu/Hu	ASC 2F9 Ko/Ko	ASC 3G5 Ko/Ar	ASC C1 Ar/Ar	ASC C2 Ar/Ar
NOVA1 gene sequencing	X	X	X	X	X	X	X	X	X	X	X	X	X	X
SNP-based array for CNV	X	X	X	X	X	X	X	X						
Exome sequencing	X	X	X	X	X	X	X	X	X	X			X	X
iPSC staining	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Mycoplasma test	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Pluripotency expression	X	X		X	X	X	X							
Organoid generation	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Morphology analysis	X	X	X	X	X	X	X	X	X	X			X	X
Organoid staining	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Annexin positive cells	X	X	X	X	X	X	X	X	X	X			X	X
Cell cycle	X	X	X	X	X	X	X	X	X	X			X	X
RNA sequencing (gene expression and splicing)	X	X	X	X	X	X	X	X	X	X			X	X
Single-nuclei RNA Seq		X	X	X	X		X	X						
NOVA1 Western Blot	X	X	X	X	X	X	X	X	X		X		X	
Markers Western Blot	X	X	X	X	X	X	X	X	X	X			X	X
Splicing Analysis	X					X			X	X	X		X	X
ECLIP	X					X								
Co-localized Puncta	X	X		X	X	X	X		X		X		X	
Co-Immunoprecip. (QMI)									X	X			X	X
Multi-electrode Array (MEA)									X	X			X	X

Table S3.

Beadchip array analysis for CNV identification.
(Excel file)

Table S4.

Off-target effects of CRISPR-Cas9.
(Excel file)

Table S5.

Differential gene expression between NOVA1$^{Hu/Hu}$ and NOVA1$^{Ar/Ar}$.
(Excel file)

Table S6.

Differential splicing between NOVA1$^{Hu/Hu}$ and NOVA1$^{Ar/Ar}$ (ASC line).
(Excel file)

Table S7.

Differential splicing between NOVA1$^{Hu/Hu}$ and NOVA1$^{Ar/Ar}$ (WT83 line).
(Excel file)

Table S8.

Enriched GO terms in the sets of genes differentially spliced at different timepoints.

Term ID	Term description	FDR α
	Early timepoint (1 month)	
GO:0007156	Homophilic cell adhesion via plasma membrane molecules	5.5×10^{-38}
GO:0005509	Calcium ion bonding	2.3×10^{-25}
GO:0007399	Nervous system development	5.2×10^{-19}
GO:0005887	Integral component of plasma membrane	9.4×10^{-17}
GO:0007267	Cell-cell signaling	2.8×10^{-14}
GO:0016339	Calcium-dependent cell-cell adhesion via plasma membrane	5.3×10^{-3}
GO:0051085	Chaperone mediated protein folding requiring cofactor	5.3×10^{-3}
	Late timepoint (2 months)	
GO:0005509	Calcium ion binding	2.3×10^{-22}
GO:0007156	Homophilic cell adhesion via plasma membrane molecules	3.5×10^{-20}
GO:0005887	Integral component of plasma membrane	7.9×10^{-14}
GO:0007267	Cell-cell signaling	4.9×10^{-11}
GO:0060789	Hair follicle placode formation	3.8×10^{-4}
GO:0070527	Platelet aggregation	1.8×10^{-3}
GO:0097718	Disordered domain specific binding	2.1×10^{-3}
GO:1990023	Mitotic spindle midzone	3.4×10^{-3}
GO:0061077	Chaperone-mediated protein folding	8.5×10^{-3}
GO:0016339	Calcium-dependent cell-cell adhesion via plasma membrane	9.8×10^{-3}
GO:0031116	Positive regulation of microtubule polymerization	9.8×10^{-3}

Reference and Notes

1. B. Wood, M. Collard, The human genus. *Science* **284**, 65–71 (1999). doi:10.1126/science.284.5411.65 Medline

2. R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H. Y. Fritz, N. F. Hansen, E. Y. Durand, A. S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Ž. Kucan, I. Gušic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, S. Pääbo, A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010). doi:10.1126/science.1188021 Medline

3. M. Meyer, M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prüfer, C. de Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, A. P. Derevianko, N. Patterson, A. M. Andrés, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012). doi:10.1126/science.1224344 Medline

4. K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. de Filippo, H. Li, S. Mallick, M. Dannemann, Q. Fu, M. Kircher, M. Kuhlwilm, M. Lachmann, M. Meyer, M. Ongyerth, M. Siebauer, C. Theunert, A. Tandon, P. Moorjani, J. Pickrell, J. C. Mullikin, S. H. Vohr, R. E. Green, I. Hellmann, P. L. F. Johnson, H. Blanche, H. Cann, J. O. Kitzman, J. Shendure, E. E. Eichler, E. S. Lein, T. E. Bakken, L. V. Golovanova, V. B. Doronichev, M. V. Shunkov, A. P. Derevianko, B. Viola, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014). doi:10.1038/nature12886 Medline

5. K. Prüfer, C. de Filippo, S. Grote, F. Mafessoni, P. Korlević, M. Hajdinjak, B. Vernot, L. Skov, P. Hsieh, S. Peyrégne, D. Reher, C. Hopfe, S. Nagel, T. Maricic, Q. Fu, C. Theunert, R. Rogers, P. Skoglund, M. Chintalapati, M. Dannemann, B. J. Nelson, F. M. Key, P. Rudan, Ž. Kućan, I. Gušić, L. V. Golovanova, V. B. Doronichev, N. Patterson, D. Reich, E. E. Eichler, M. Slatkin, M. H. Schierup, A. M. Andrés, J. Kelso, M. Meyer, S. Pääbo, A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658 (2017). doi:10.1126/science.aao1887 Medline

6. L. Abi-Rached, M. J. Jobin, S. Kulkarni, A. McWhinnie, K. Dalva, L. Gragert, F. Babrzadeh, B. Gharizadeh, M. Luo, F. A. Plummer, J. Kimani, M. Carrington, D. Middleton, R. Rajalingam, M. Beksac, S. G. E. Marsh, M. Maiers, L. A. Guethlein, S. Tavoularis, A.-M. Little, R. E. Green, P. J. Norman, P. Parham, The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science* **334**, 89–94 (2011). doi:10.1126/science.1209202 Medline

7. F. Racimo, S. Sankararaman, R. Nielsen, E. Huerta-Sánchez, Evidence for archaic adaptive introgression in humans. *Nat. Rev. Genet.* **16**, 359–371 (2015). doi:10.1038/nrg3936 Medline

8. I. Juric, S. Aeschbacher, G. Coop, The strength of selection against Neanderthal introgression. *PLOS Genet.* **12**, e1006340 (2016). doi:10.1371/journal.pgen.1006340 Medline

9. S. Sankararaman, S. Mallick, M. Dannemann, K. Prüfer, J. Kelso, S. Pääbo, N. Patterson, D. Reich, The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354–357 (2014). doi:10.1038/nature12961 Medline

10. B. Vernot, J. M. Akey, Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343**, 1017–1021 (2014). doi:10.1126/science.1245938 Medline

11. A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). doi:10.1038/nature15393 Medline

12. S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. Di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, D. Reich, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016). doi:10.1038/nature18964 Medline

13. R. J. Buckanovich, R. B. Darnell, The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. *Mol. Cell. Biol.* **17**, 3194–3201 (1997). doi:10.1128/MCB.17.6.3194 Medline

14. K. B. Jensen, B. K. Dredge, G. Stefani, R. Zhong, R. J. Buckanovich, H. J. Okano, Y. Y. L. Yang, R. B. Darnell, Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron* **25**, 359–371 (2000). doi:10.1016/S0896-6273(00)80900-9 Medline

15. J. Ule, A. Ule, J. Spencer, A. Williams, J.-S. Hu, M. Cline, H. Wang, T. Clark, C. Fraser, M. Ruggiu, B. R. Zeeberg, D. Kane, J. N. Weinstein, J. Blume, R. B. Darnell, Nova regulates brain-specific splicing to shape the synapse. *Nat. Genet.* **37**, 844–852 (2005). doi:10.1038/ng1610 Medline

16. Y. Xin, Z. Li, H. Zheng, J. Ho, M. T. V. Chan, W. K. K. Wu, Neuro-oncological ventral antigen 1 (NOVA1): Implications in neurological diseases and cancers. *Cell Prolif.* **50**, e12348 (2017). doi:10.1111/cpr.12348 Medline

17. N. N. Parikshak, V. Swarup, T. G. Belgard, M. Irimia, G. Ramaswami, M. J. Gandal, C. Hartl, V. Leppa, L. T. Ubieta, J. Huang, J. K. Lowe, B. J. Blencowe, S. Horvath, D. H. Geschwind, Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* **540**, 423–427 (2016). doi:10.1038/nature20612 Medline

18. J. Ule, G. Stefani, A. Mele, M. Ruggiu, X. Wang, B. Taneri, T. Gaasterland, B. J. Blencowe, R. B. Darnell, An RNA map predicting Nova-dependent splicing regulation. *Nature* **444**, 580–586 (2006). doi:10.1038/nature05304 Medline

19. M. Teplova, L. Malinina, J. C. Darnell, J. Song, M. Lu, R. Abagyan, K. Musunuru, A. Teplov, S. K. Burley, R. B. Darnell, D. J. Patel, Protein-RNA and protein-protein recognition by dual KH1/2 domains of the neuronal splicing factor Nova-1. *Structure* **19**, 930–944 (2011). doi:10.1016/j.str.2011.05.002 Medline

20. C. A. Trujillo, R. Gao, P. D. Negraes, J. Gu, J. Buchanan, S. Preissl, A. Wang, W. Wu, G. G. Haddad, I. A. Chaim, A. Domissy, M. Vandenberghe, A. Devor, G. W. Yeo, B. Voytek, A. R. Muotri, Complex oscillatory waves emerging from cortical organoids model early human brain network development. *Cell Stem Cell* **25**, 558–569.e7 (2019). doi:10.1016/j.stem.2019.08.002 Medline

21. O. Nygård, H. Nika, Identification by RNA-protein cross-linking of ribosomal proteins located at the interface between the small and the large subunits of mammalian ribosomes. *EMBO J.* **1**, 357–362 (1982). doi:10.1002/j.1460-2075.1982.tb01174.x Medline

22. H. H. Lin, E. Bell, D. Uwanogho, L. W. Perfect, H. Noristani, T. J. D. Bates, V. Snetkov, J. Price, Y.-M. Sun, Neuronatin promotes neural lineage in ESCs via Ca^{2+} signaling. *Stem Cells* **28**, 1950–1960 (2010). doi:10.1002/stem.530 Medline

23. N. C. Boles, S. E. Hirsch, S. Le, B. Corneo, F. Najm, A. P. Minotti, Q. Wang, S. Lotz, P. J. Tesar, C. A. Fasano, NPTX1 regulates neural lineage specification from human pluripotent stem cells. *Cell Rep.* **6**, 724–736 (2014). doi:10.1016/j.celrep.2014.01.026 Medline

24. T. Shimizu, M. Hibi, Formation and patterning of the forebrain and olfactory system by zinc-finger genes *Fezf1* and *Fezf2*. *Dev. Growth Differ.* **51**, 221–231 (2009). doi:10.1111/j.1440-169X.2009.01088.x Medline

25. L. K. Davis, K. J. Meyer, D. S. Rudd, A. L. Librant, E. A. Epping, V. C. Sheffield, T. H. Wassink, Pax6 3′ deletion results in aniridia, autism and mental retardation. *Hum. Genet.* **123**, 371–378 (2008). doi:10.1007/s00439-008-0484-x Medline

26. M. Heide, Y. Zhang, X. Zhou, T. Zhao, A. Miquelajáuregui, A. Varela-Echavarría, G. Alvarez-Bolado, Lhx5 controls mamillary differentiation in the developing hypothalamus of the mouse. *Front. Neuroanat.* **9**, 113 (2015). doi:10.3389/fnana.2015.00113 Medline

27. K. K. Szumlinski, P. W. Kalivas, P. F. Worley, Homer proteins: Implications for neuropsychiatric disorders. *Curr. Opin. Neurobiol.* **16**, 251–257 (2006). doi:10.1016/j.conb.2006.05.002 Medline

28. E. L. Van Nostrand, G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, G. W. Yeo, Robust transcriptome-wide discovery of RNA-binding protein

binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **13**, 508–514 (2016). doi:10.1038/nmeth.3810 Medline

29. J. D. Lautz, E. A. Brown, A. A. Williams VanSchoiack, S. E. P. Smith, Synaptic activity induces input-specific rearrangements in a targeted synaptic protein interaction network. *J. Neurochem.* **146**, 540–559 (2018). doi:10.1111/jnc.14466 Medline

30. S. E. P. Smith, S. C. Neier, B. K. Reed, T. R. Davis, J. P. Sinnwell, J. E. Eckel-Passow, G. F. Sciallis, C. N. Wieland, R. R. Torgerson, D. Gil, C. Neuhauser, A. G. Schrum, Multiplex matrix network analysis of protein complexes in the human TCR signalosome. *Sci. Signal.* **9**, rs7–rs7 (2016). doi:10.1126/scisignal.aad7279 Medline

31. P. Langfelder, S. Horvath, WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008). doi:10.1186/1471-2105-9-559 Medline

32. P. Monteiro, G. Feng, SHANK proteins: Roles at the synapse and in autism spectrum disorder. *Nat. Rev. Neurosci.* **18**, 147–157 (2017). doi:10.1038/nrn.2016.183 Medline

33. J. A. Ronesi, K. A. Collins, S. A. Hays, N.-P. Tsai, W. Guo, S. G. Birnbaum, J.-H. Hu, P. F. Worley, J. R. Gibson, K. M. Huber, Disrupted Homer scaffolds mediate abnormal mGluR5 function in a mouse model of fragile X syndrome. *Nat. Neurosci.* **15**, 431–440, S1 (2012). doi:10.1038/nn.3033 Medline

34. W. G. I. V. Walkup IV, T. L. Mastro, L. T. Schenker, J. Vielmetter, R. Hu, A. Iancu, M. Reghunathan, B. D. Bannon, M. B. Kennedy, A model for regulation by SynGAP-α1 of binding of synaptic proteins to PDZ-domain 'Slots' in the postsynaptic density. *eLife* **5**, e16813 (2016). doi:10.7554/eLife.16813 Medline

35. X. Zhang, M. H. Chen, X. Wu, A. Kodani, J. Fan, R. Doan, M. Ozawa, J. Ma, N. Yoshida, J. F. Reiter, D. L. Black, P. V. Kharchenko, P. A. Sharp, C. A. Walsh, Cell-type-specific alternative splicing governs cell fate in the developing cerebral cortex. *Cell* **166**, 1147–1162.e15 (2016). doi:10.1016/j.cell.2016.07.025 Medline

36. A. N. Brooks, L. Yang, M. O. Duff, K. D. Hansen, J. W. Park, S. Dudoit, S. E. Brenner, B. R. Graveley, Conservation of an RNA regulatory map between *Drosophila* and mammals. *Genome Res.* **21**, 193–202 (2011). doi:10.1101/gr.108662.110 Medline

37. N. Jelen, J. Ule, M. Zivin, R. B. Darnell, Evolution of Nova-dependent splicing regulation in the brain. *PLOS Genet.* **3**, e173 (2007). doi:10.1371/journal.pgen.0030173 Medline

38. G. A. Watterson, On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975). doi:10.1016/0040-5809(75)90020-9 Medline

39. S. W. Schaeffer, Molecular population genetics of sequence length diversity in the Adh region of *Drosophila pseudoobscura*. *Genet. Res.* **80**, 163–175 (2002). doi:10.1017/S0016672302005955 Medline

40. C.-H. Hsieh, A. Shaltouki, A. E. Gonzalez, A. Bettencourt da Cruz, L. F. Burbulla, E. St Lawrence, B. Schüle, D. Krainc, T. D. Palmer, X. Wang, Functional impairment in miro degradation and mitophagy is a shared feature in familial and sporadic Parkinson's disease. *Cell Stem Cell* **19**, 709–724 (2016). doi:10.1016/j.stem.2016.08.002 Medline

41. E. L. Van Nostrand, P. Freese, G. A. Pratt, X. Wang, X. Wei, R. Xiao, S. M. Blue, J.-Y. Chen, N. A. L. Cody, D. Dominguez, S. Olson, B. Sundararaman, L. Zhan, C. Bazile, L. P. B. Bouvrette, J. Bergalet, M. O. Duff, K. E. Garcia, C. Gelboin-Burkhart, M. Hochman, N. J. Lambert, H. Li, M. P. McGurk, T. B. Nguyen, T. Palden, I. Rabano, S. Sathe, R. Stanton, A. Su, R. Wang, B. A. Yee, B. Zhou, A. L. Louie, S. Aigner, X.-D. Fu, E. Lécuyer, C. B. Burge, B. R. Graveley, G. W. Yeo, A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020). doi:10.1038/s41586-020-2077-3 Medline

42. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013). doi:10.1093/bioinformatics/bts635 Medline

43. J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, T. J. Hubbard, GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012). doi:10.1101/gr.135350.111 Medline

44. K. Prüfer, snpAD: An ancient DNA genotype caller. *Bioinformatics* **34**, 4165–4171 (2018). doi:10.1093/bioinformatics/bty507 Medline

45. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011). doi:10.1093/bioinformatics/btr509 Medline

46. P. Cingolani, A. Platts, L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, D. M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012). doi:10.4161/fly.19695 Medline

47. R. J. Ihry, K. A. Worringer, M. R. Salick, E. Frias, D. Ho, K. Theriault, S. Kommineni, J. Chen, M. Sondey, C. Ye, R. Randhawa, T. Kulkarni, Z. Yang, G. McAllister, C. Russ, J. Reece-Hoyes, W. Forrester, G. R. Hoffman, R. Dolmetsch, A. Kaykas, p53 inhibits CRISPR-Cas9 engineering in human pluripotent stem cells. *Nat. Med.* **24**, 939–946 (2018). doi:10.1038/s41591-018-0050-6 Medline

48. E. Haapaniemi, S. Botla, J. Persson, B. Schmierer, J. Taipale, CRISPR-Cas9 genome editing induces a p53-mediated DNA damage response. *Nat. Med.* **24**, 927–930 (2018). doi:10.1038/s41591-018-0049-z Medline

49. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN] (16 March 2013).

50. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). doi:10.1093/bioinformatics/btp352 Medline

51. K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S. F. A. Grant, H. Hakonarson, M. Bucan, PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007). doi:10.1101/gr.6861907 Medline

52. W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, D. Haussler, The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002). doi:10.1101/gr.229102 Medline

53. R. K. Patel, M. Jain, NGS QC Toolkit: A toolkit for quality control of next generation sequencing data. *PLOS ONE* **7**, e30619 (2012). doi:10.1371/journal.pone.0030619 Medline

54. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010). doi:10.1093/bioinformatics/btp698 Medline

55. R. H. Herai, Avoiding the off-target effects of CRISPR/cas9 system is still a challenging accomplishment for genetic transformation. *Gene* **700**, 176–178 (2019). doi:10.1016/j.gene.2019.03.019 Medline

56. M. J. Landrum, J. M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, W. Jang, K. Katz, M. Ovetsky, G. Riley, A. Sethi, R. Tully, R. Villamarin-Salomon, W. Rubinstein, D. R. Maglott, ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016). doi:10.1093/nar/gkv1222 Medline

57. G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, G. A. McVean, 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010). doi:10.1038/nature09534 Medline

58. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010). doi:10.1101/gr.107524.110 Medline

59. C. Chiang, R. M. Layer, G. G. Faust, M. R. Lindberg, D. B. Rose, E. P. Garrison, G. T. Marth, A. R. Quinlan, I. M. Hall, SpeedSeq: Ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12**, 966–968 (2015). doi:10.1038/nmeth.3505 Medline

60. J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, A. Cardona, Fiji: An open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012). doi:10.1038/nmeth.2019 Medline

61. A. Byrne, A. E. Beaudin, H. E. Olsen, M. Jain, C. Cole, T. Palmer, R. M. DuBois, E. C. Forsberg, M. Akeson, C. Vollmers, Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* **8**, 16027 (2017). doi:10.1038/ncomms16027 Medline

62. S. Picelli, O. R. Faridani, Å. K. Björklund, G. Winberg, S. Sagasser, R. Sandberg, Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014). doi:10.1038/nprot.2014.006 Medline

63. S. Picelli, Å. K. Björklund, B. Reinius, S. Sagasser, G. Winberg, R. Sandberg, Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014). doi:10.1101/gr.177881.114 Medline

64. D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S. L. Salzberg, TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013). doi:10.1186/gb-2013-14-4-r36 Medline

65. Y. Liao, G. K. Smyth, W. Shi, featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014). Medline

66. G. P. Wagner, K. Kin, V. J. Lynch, Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012). doi:10.1007/s12064-012-0162-3 Medline

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

68. B. Lacar, S. B. Linker, B. N. Jaeger, S. R. Krishnaswami, J. J. Barron, M. J. E. Kelder, S. L. Parylak, A. C. M. Paquola, P. Venepally, M. Novotny, C. O'Connor, C. Fitzpatrick, J. A. Erwin, J. Y. Hsu, D. Husband, M. J. McConnell, R. Lasken, F. H. Gage, Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nat. Commun.* **7**, 11022 (2016). doi:10.1038/ncomms11022 Medline

69. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019). doi:10.1016/j.cell.2019.05.031 Medline

70. C. S. McGinnis, L. M. Murrow, Z. J. Gartner, DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* **8**, 329–337.e4 (2019). doi:10.1016/j.cels.2019.03.003 Medline

71. I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P. R. Loh, S. Raychaudhuri, Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019). doi:10.1038/s41592-019-0619-0 Medline

72. M. T. Lovci, D. Ghanem, H. Marr, J. Arnold, S. Gee, M. Parra, T. Y. Liang, T. J. Stark, L. T. Gehman, S. Hoon, K. B. Massirer, G. A. Pratt, D. L. Black, J. W. Gray, J. G. Conboy, G. W. Yeo, Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat. Struct. Mol. Biol.* **20**, 1434–1442 (2013). doi:10.1038/nsmb.2699 Medline

73. K. Prüfer, B. Muetzel, H.-H. Do, G. Weiss, P. Khaitovich, E. Rahm, S. Pääbo, M. Lachmann, W. Enard, FUNC: A package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics* **8**, 41 (2007). doi:10.1186/1471-2105-8-41 Medline

74. E. A. Brown, J. D. Lautz, T. R. Davis, E. P. Gniffke, A. A. W. VanSchoiack, S. C. Neier, N. Tashbook, C. Nicolini, M. Fahnestock, A. G. Schrum, S. E. P. Smith, Clustering the autisms using glutamate synapse protein interaction networks from cortical and

hippocampal tissue of seven mouse models. *Mol. Autism* **9**, 48–48 (2018). doi:10.1186/s13229-018-0229-1 Medline

75. M. S. Lewicki, A review of methods for spike sorting: The detection and classification of neural action potentials. *Network* **9**, R53–R78 (1998). doi:10.1088/0954-898X_9_4_001 Medline

76. R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. B* **63**, 411–423 (2001). doi:10.1111/1467-9868.00293