# A reference induced pluripotent stem cell line for large-scale collaborative studies

## Authors and affiliations:

Caroline B. Pantazis[1^], Andrian Yang[2-5^], Erika Lara[1^], Justin A. McDonough[6^], Cornelis Blauwendraat[1,7^], Lirong Peng[1,8,9^], Hideyuki Oguro[6,10], Jizhong Zou[11], David Sebesta[12], Gretchen Pratt[12], Erin Cross[12], Jeffrey Blockwick[12], Philip Buxton[12], Lauren Kinner-Bibeau[12], Constance Medura[12], Christopher Tompkins[12], Stephen Hughes[12], Marianita Santiana[1], Faraz Faghri[1,7,8], Mike A. Nalls[1,7,8], Daniel Vitale[1,7,8], Yue A. Qi[1], Daniel M. Ramos[1], Kailyn M. Anderson[1], Julia Stadler[1], Priyanka Narayan[1,13], Jason Papademetriou[1], Luke Reilly[1], Matthew P. Nelson[1], Sanya Aggarwal[4,5], Leah U. Rosen[2], Peter Kirwan[4,5], Venkat Pisupati[5,14], Steven L. Coon[15], Sonja W. Scholz[16,17], Elena Coccia[18], Lily Sarrafha[18], Tim Ahfeldt[18], Salome Funes[19], Daryl A. Bosco[19], Melinda S. Beccari[20], Don W. Cleveland[20], Maria Clara Zanellati[21], Richa Basundra[21], Mohanish Deshmukh[21], Sarah Cohen[21], Zachary S. Nevin[22], Madeline Matia[22], Jonas Van Lent[23], Vincent Timmerman[23], Bruce R. Conklin[22], Dan Dou[24], Erika L.F. Holzbaur[24], Emmy Li[25], Indigo V.L. Rose[25], Martin Kampmann[25], Theresa Priebe[26], Miriam Öttl[26], Jian Dong[26], Rik van der Kant[26,27], Lena Erlebach[28], Marc Welzer[28], Deborah Kronenberg-Versteeg[28], Dad Abu-Bonsrah[29], Clare L. Parish[29], Malavika Raman[30], Laurin Heinrich[31], Birgitt Schüle[31], Carles Calatayud Aristoy[32], Patrik Verstreken[32], Aaron Held[33], Brian J. Wainger[34], Guochang Lyu[35], Ernest Arenas[35], Ana-Caroline Raulin[36], Guojun Bu[36], Dennis Crusius[37], Dominik Paquet[37,38], Rebecca M.C. Gabriele[39], Selina Wray[39], Katherine Johnson Chase[36], Ke Zhang[36], John C. Marioni[2,3,40], William C. Skarnes[6*], Mark R. Cookson[1,7*], Michael E. Ward[1*], Florian T. Merkle[4,5*]

[^]denotes co-first authorship; [*]denotes co-corresponding authorship

[1] Center for Alzheimer's and Related Dementias, National Institutes of Health, Bethesda, MD, USA

[2] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

[3] Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK

[4] Metabolic Research Laboratories and Medical Research Council Metabolic Diseases Unit, Wellcome Trust - Medical Research Council Institute of Metabolic Science, University of Cambridge, Cambridge CB2 0QQ, UK

[5] Wellcome Trust - Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge CB2 0AW, UK

[6] The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA

[7] Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA

[8] Data Tecnica International LLC, Glen Echo, MD, USA

[9] Integrated Research Facility, National Institute of Allergy and Infectious Diseases, National Institute of Health, Frederick, MD, USA

[10] Department of Cell Biology, University of Connecticut Health Center, Farmington, CT, USA

[11] iPS Cell Core Facility, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA

[12] KromaTiD Inc., Longmont, CO, USA.

[13] Genetics and Biochemistry Branch, NIDDK, National Institutes of Health, Bethesda, MD 20814, USA.

[14] John van Geest Centre for Brain Repair, University of Cambridge, Cambridge CB2 0PY, UK

[15] Molecular Genomics Core, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD, USA

[16] Neurodegenerative Diseases Research Unit, National Institute of Neurological Disorders and Stroke, Bethesda, MD, USA

[17] Department of Neurology, Johns Hopkins University, Baltimore, MD 21287, USA

[18] Nash Family Department of Neuroscience at Mount Sinai, New York, NY, USA; Departments of Neurology and Cell, Developmental and Regenerative Biology at Mount Sinai, New York, NY, USA; Ronald M. Loeb Center for Alzheimer's Disease at Mount Sinai, New York, NY, USA; Friedman Brain Institute at Mount Sinai, New York, NY, USA; Black Family Stem Cell Institute at Mount Sinai, New York, NY, USA.

[19] Department of Neurology, UMass Chan Medical School, Worcester, MA, USA.

[20] Department of Cellular and Molecular Medicine, University of California at San Diego, La Jolla, CA, USA; Ludwig Institute for Cancer Research, University of California at San Diego, La Jolla, CA, USA.

[21] Department of Cell Biology and Physiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[22] Gladstone Institutes, San Francisco, CA, USA

[23] Peripheral Neuropathy Research Group, Department of Biomedical Sciences, University of Antwerp, Antwerp, 2610, Belgium

[24] Department of Physiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

[25] Institute for Neurodegenerative Diseases, University of California, San Francisco, San Francisco, CA, USA; Chan Zuckerberg Biohub, San Francisco, CA, USA; Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA, USA.

[26] Department of Functional Genomics, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, VU University Amsterdam de Boelelaan 1087, 1081 HV Amsterdam, the Netherlands.

[27] Alzheimer Center Amsterdam, Department of Neurology, Amsterdam Neuroscience, Amsterdam UMC, Amsterdam, Netherlands

[28] Department of Cellular Neurology, Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany; German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany.
[29] Florey Institute of Neuroscience and Mental Health, Parkville, VIC 3052, Australia.
[30] Department of Developmental Molecular and Chemical Biology, Tufts University School of Medicine, Boston, MA, USA.
[31] Department of Pathology, Stanford University School of Medicine, Stanford, California, USA
[32] VIB-KU Leuven Center for Brain & Disease Research, 3000 Leuven, Belgium; KU Leuven, Department of Neurosciences, Leuven Brain Institute, Mission Lucidity, Leuven, Belgium
[33] Department of Neurology, Sean M. Healey & AMG Center for ALS, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA.
[34] Department of Neurology, Sean M. Healey & AMG Center for ALS, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA; Department of Anesthesiology, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, MA, USA; Harvard Stem Cell Institute, Cambridge, MA, USA; Broad Institute of Harvard University and MIT, Cambridge, MA, USA.
[35] Division of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden.
[36] Department of Neuroscience, Mayo Clinic, Jacksonville, FL, USA.
[37] Institute for Stroke and Dementia Research, University Hospital Munich, Ludwig-Maximilians-Universität München, 81377 Munich, Germany
[38] Munich Cluster for Systems Neurology (SyNergy), 81377 Munich, Germany
[39] Department of Neurodegenerative Disease, UCL Queen Square Institute of Neurology, London, WC1N 3BG, UK.
[40] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK

*Correspondence: fm436@medschl.cam.ac.uk, bill.skarnes@jax.org, cookson@mail.nih.gov, michael.ward4@nih.gov

## Abstract

Human induced pluripotent stem cells (iPSCs) are a powerful tool for studying development and disease. However, different iPSC lines show considerable phenotypic variation. The lack of common well-characterised cell lines that are used widely frustrates efforts to integrate data across research groups or replicate key findings. Inspired by model organism communities who addressed this issue by establishing a limited number of widely accepted strains, we characterised candidate iPSC lines in unprecedented detail to select a well-performing line to underpin collaborative studies. Specifically, we characterised the morphology, growth rates, and single-cell transcriptomes of iPSC lines in the pluripotent state and assessed their genomic integrity using karyotyping, DNA microarrays, whole genome sequencing, and functional assays for p53 activity. We further tested their ability to be edited by CRISPR/Cas9 and used single-cell RNA sequencing to compare the efficiency with which they could be differentiated into multiple lineages. We found that there was significant variability in the performance of lines across the tested assays that enabled the rational selection of a lead line, KOLF2.1J, which is a gene-corrected derivative of a publicly available line from the Human Induced Pluripotent Stem Cells Initiative (HipSci) resource. We are now using this line in an initiative from the NIH Center for Alzheimer’s and Related Dementias to derive hundreds of gene-edited and functionalized sub-clones to be distributed widely throughout the research community along with associated datasets, with the aim of promoting the standardisation required for large-scale collaborative science in the stem cell field.

## Summary

The authors of this collaborative science study describe a deep characterisation of widely available human induced pluripotent stem cell (hiPSC) lines to rationally select a line that performs well in multiple experimental approaches. Analysis of transcriptional patterns in the pluripotent state, whole genome sequencing, genomic stability after highly efficient CRISPR-mediated gene editing, integrity of the p53 pathway, and differentiation efficiency towards multiple lineages identified KOLF2.1J as an all-around well-performing cell line. The widespread distribution and use of this line makes it an attractive cell line for comparative and collaborative efforts in the stem cell field.

## Highlights

- Deep genotyping and phenotyping reveal KOLF2.1J as an all-around well-performing cell line that is readily distributed and could serve as common reference line
- Despite rare copy-neutral loss of heterozygosity (CN-LOH) events, iPSC lines retain genomic fidelity after CRISPR/Cas9-based gene editing
- Our multifactorial pipeline serves as a blueprint for future efforts to identify other lead iPSC lines

## Introduction

Human iPSCs are an increasingly widespread model for studying human disease. They can be derived from individuals affected by different diseases, capturing genetic contributors to disease risk in pluripotent cells that can be differentiated into many disease-relevant cell populations. Additionally, iPSCs can be edited to introduce or correct specific disease-associated genetic variants, generating isogenic disease models. Indeed, several recent studies have utilized isogenic lines to elucidate the biology contributing to disease states (Guttikonda et al., 2021; Konttinen et al., 2019; Kwart et al., 2019). For example, knock-in of Mendelian variants associated with Alzheimer's disease (AD) into an isogenic genetic background identified convergent transcriptomic events after differentiation into neurons (Kwart et al., 2019). Isogenic cell lines can equally be used to examine mutational effects in non-neuronal cells (Konttinen et al., 2019).

In theory, the results of isogenic experiments could be compared across genetic variants, cell types, and analysis modalities by different groups to reveal convergent molecular mechanisms of disease. However, a plethora of different cell lines are used by different groups, creating a major obstacle to data integration, since genetic background strongly influences cellular phenotypes (Bonyadi et al., 1997; Doetschman, 2009; Threadgill et al., 1995). This issue has been recognised by communities working with model organisms, who appreciated that the benefits of a common reference far outweigh whether a given particular line has idealized characteristics, since key results obtained on one genetic background can always be tested on another (Mackay and Huang, 2018; Sittig et al., 2016; Sterken et al., 2015). Consequently, there is an unmet need in the iPSC field for common, well-characterised cell lines that can be used as the basis of large-scale collaborative studies.

Several recent efforts have sought to address these challenges by developing edited iPSC lines from a common parental cell line. For example, the Allen Cell Collection has generated a series of publicly available and gene-edited iPSC lines on a common, genetically characterised parental line (Roberts et al., 2017, 2019). Here, we aim to identify a common cell line to underpin the recently announced iPSC Neurodegenerative Disease Initiative (iNDI) that will generate hundreds of single nucleotide variant (SNV) knock-in, revertant, gene knockout, and endogenously-tagged CRISPR/Cas9-edited lines relevant to Alzheimer's Disease and Related Dementias (ADRD) on a single deeply characterised genetic background (Ramos et al., 2021). When selecting candidate cell lines, we accounted for practical considerations including the freedom to modify and distribute the line and its derivatives. To this end, we deeply characterised the genomic status, functional characteristics, and differentiation potential of multiple candidate iPSC lines, and identified KOLF2.1J as a standard cell line for future investment.

## Results

### *Rationale and establishment of candidate cell line sub-clones*

We set out to identify one or more deeply characterised iPSC lines to serve as a common reference for the field (Ramos et al., 2021; Reilly et al.). Although it will be important in the long term to have a series of iPSCs from male and female donors of diverse genetic backgrounds, we initially prioritized male lines due to the possibility that random X-chromosome inactivation may contribute to variance in gene expression (Bar et al., 2019; Mekhoubad et al., 2012). Since human embryonic stem cell (hESC) lines face usage restrictions in many countries, we chose to prioritise hiPSC lines to enable broad sharing around the world. We therefore searched public repositories and curated a series of male iPSC lines, many of which already have whole genome sequencing available publicly (see graphical abstract). We then focused on a subset of lines with broad consents for data sharing and further dissemination of the line and its derivatives. These selected lines were: KOLF2_C1, KUCG3, LNGPI1, MS19-ES-H, NCRM1, NCRM5, NN0003932, NN0004297, and PGP1.

After obtaining and thawing these lines, we noticed that the line MS19-ES-H was prone to spontaneous differentiation, and therefore excluded it from further study. We then established clonal sub-lines from each of the eight remaining parental cell lines to reduce heterogeneity from genetic and epigenetic drift in culture. Additionally, we used CRISPR/Cas9 editing (see Methods) to correct a mutation present in one copy of *ARID2* in the KOLF2_C1 line (Hildebrandt et al., 2019). In parallel with our efforts, the same genetic correction was carried out at the Wellcome Sanger Institute, resulting in two distinct KOLF2_C1-corrected cell lines with diverging culture histories. To avoid confusion, we have named our sub-line KOLF2.1J to indicate its derivation at Jackson Laboratories and distinguish it from the KOLF2.1S sub-line derived at the Wellcome Sanger Institute.

We then selected one to four of these sub-lines with typical stem cell morphology per parental cell line for further expansion and Giemsa-band karyotype analysis of metaphase chromosome spreads (Table S1A). The karyotype analysis indicated that almost all tested sub-lines were euploid (46; XY; Table S1B), but some lines harbored a subset of aberrant cells. For example, 2 of the 20 analysed spreads for the selected KUCG3 sub-line showed a gain of chromosome 12. Based on this data, a single clonal sub-line was selected from each parental line for further expansion into replicate stock vials (Table S1A). To complement this analysis, we subjected the selected sub-lines to directional genomic hybridisation (Robinson et al., 2019) and scored 181 to 200 high-quality metaphase spreads per sub-line at chromosomes 1, 2 and 3. We found that the rate of scored events was not significantly different between the eight sub-lines, and that the rate of translocations and inversions was low (Table S1C). Overall, our chromosome-scale analyses revealed that most sub-lines were karyotypically normal, except for KUCG3. Cells were thawed and passaged once prior to several simultaneously executed analyses, including morphological and transcriptional characterisation, karyotyping, P53 activity assay, and a CRISPR-based gene editing experiment.

### *Morphology and proliferation rates*

Each selected sub-line had the morphology expected for hiPSCs including a high nuclear to cytoplasmic ratio, prominent nucleoli, growth in colonies with well-defined borders, and an absence of differentiated cells (Figure 1A). To compare the survival and growth rates of these cell lines, we dissociated each line to a single-cell suspension, plated them in parallel, imaged them on an Incucyte live cell imager at 24 and 48 hours after plating to calculate their confluence, then dissociated and manually counted cells 48 hours after plating. We observed a significant difference between the sub-lines in their total numbers of cells after 48 hours (one-way ANOVA, $F_{7,\ 24}$= 185.1, *p*<0.0001; Figure 1B). Additionally, we observed a significant main effect of time (two repeated measures ANOVA; $F_{2,\ 80}$ = 1836, *p*<0.0001) and cell line ($F_{7,\ 80}$ = 97.85, *p*<0.0001), on confluency as well as a time x cell line interaction ($F_{14,80}$=45.76, *p*<0.0001; Figure 1C, Table S2). Together, these results show that all cell lines had similar morphology but varied in their survival and proliferation rates.

*Pluripotency*

The eight selected sub-lines were immunostained with antibodies against pluripotency markers TRA-1-60 and NANOG, and the percentage of immunopositive cells was quantified by flow cytometry analysis. We found that all analysed sub-lines are >90% positive for both markers, except PGP1, which was 84.2% positive for NANOG (Figure S1).

Next, we performed single-cell RNA sequencing (scRNA-seq) on all eight iPSC sub-lines (Figure 1D), pooling lines together for joint single-cell library preparation and sequencing to minimize technical sources of variation and using genetic diversity to assign each single-cell to a particular cell line (Huang et al., 2019). Since sub-lines NN0003932 and NN0004297 were derived from the same donor, these are represented in our dataset as NN_combined (Figure S2A). UMAP projection of the data from 2,270 single cells revealed two distinct groups of cells, the largest corresponding to six genetically distinct lines and one small outlier group composed primarily of cells from one sub-line, LNGPI1. Louvain clustering identified 5 clusters within the larger group, which appeared to arise from variable cell cycle states of these proliferative cells (Figure S2B-C). Comparison of pluripotency markers across the 7 genetically distinct sub-lines showed consistent expression of core pluripotency-associated genes including *SOX2*, *POU5F1* (OCT4) and *NANOG* (Figure 1E, S2D). Interestingly, we found that the LNGPI1 line expressed higher levels of *UTF1* and other genes associated with inefficient neuronal differentiation (Jerber et al., 2021). Together, these findings indicate that all analysed iPSC cell lines were pluripotent, and 6 of the 7 genetically distinct lines showed similar transcriptional profiles.

*Integrity of the p53 response to DNA damage*

Established iPSC lines are known to acquire genetic changes that impart a growth advantage in culture (Halliwell et al., 2020). Of these, mutations in the *TP53* gene are of particular concern because they are recurrent (Avior et al., 2021; Merkle et al., 2017) and loss of a functional p53 pathway may be selected for during CRISPR editing and clonal expansion (Ihry et al., 2018). To measure p53 pathway function in our hiPSC sub-lines, we transfected cells with a reporter plasmid that contains 13 copies of a p53 DNA binding site linked to mCherry (see Methods), challenged them with either a vehicle control or the DNA-damaging agent doxorubicin, and quantified induction of the mCherry reporter expression by flow cytometry. This analysis showed doxorubicin activated reporter expression in all eight selected sub-lines compared to *TP53* knockout cells (Figures 1F, S3), confirming integrity of the p53 pathway in all eight sub-lines (Hildebrandt et al., 2019), with particularly robust responses in lines KOLF2.1J, LNGPI1, NCRM1, and PGP1.

*Whole Genome Sequencing (WGS)*

To more deeply characterise the genetic diversity in the eight parental cell lines, we sequenced their genomes at >30x coverage with paired-end reads (Illumina). The overall distribution of insertion-deletion (indel), loss-of-function (LOF), and missense single-nucleotide variants (SNVs) was similar across all sub-lines (Figure 2A) and similar to the human population database, gnomAD (Karczewski et al., 2020). When we restricted our analysis to those variants likely to be deleterious (allele frequency <0.001 and CADD score >30, (Huang et al., 2019)), we observed a modest number of potentially deleterious variants per cell line (Figure 2B, Table S3A). Most of these variants were of unknown significance, and none were annotated to be pathogenic in ClinVar (Landrum et al., 2020) or also annotated as haploinsufficient in ClinGen (https://www.clinicalgenome.org/) or predicted to be highly constrained based on LoFtool (Fadista et al., 2017).

Since the goal of iNDI is to model neurodegenerative diseases, we examined WGS for known pathogenic mutations in ADRD-associated genes. We identified potentially damaging LOF variants in *DRD4* (rs587776842 in lines NN0003932 and NN0004297) and *MPDZ* (rs376078512 in line KUCG3) that are pathogenic in a homozygous or compound heterozygous state, but only as heterozygous mutations.

To examine the role of common genetic variants associated with ADRD, we screened all iPSC lines for the AD risk gene *APOE* (rs429358 and rs7412), the frontotemporal dementia variant *TMEM106B* rs3173615, and *MAPT* haplotype rs1800547, associated with risk of Parkinson’s disease (PD). We found, for example, that NCRM1 and NCRM5 carry the *APOE* E4 allele that increases risk for AD (Table S4A). We also identified variants in other genes known to be risk factors for ADRD, such as rs113809142 in the PGP1 line and an *ABCA7* splicing variant associated with increased risk of AD (Steinberg et al., 2015). Finally, we calculated polygenic risk scores for all eight iPSC lines based on their cumulative burden of common genetic variants associated with AD or PD in genome-wide association studies (Kunkle et al., 2019; Nalls et al., 2019). (Figure 2C). Together, these findings show that the analysed cell lines do not have causal ADRD mutations and have a genetic background within the expected range of the population for overall cumulative risk of AD or PD (Figure 2C).

*CRISPR-based editing potential*

Since the selected hiPSC line will be used to generate CRISPR-edited derivatives, we next characterised the efficiency with which a SNV could be edited in each cell sub-line. Using improved conditions for homology-directed repair (Skarnes et al., 2019, 2021), we introduced a G to C SNV in exon 1 of the *TIMP3* gene, resulting in an S38C missense mutation. Twenty-four edited clones from each line were genotyped by Sanger sequencing of PCR amplicons spanning the targeted site of the *TIMP3* locus to quantify frequency of six possible genotypes (WT/WT, WT/SNV, WT/indel, SNV/SNV, SNV/indel, and indel/indel; Figure 3A). We observed that the overall editing efficiency was very high in most cell lines (Figure 3B) with a mean frequency of homozygous (SNV/SNV) edits of over 40%. We also found that ratio of SNV/WT and SNV/SNV edits generated by homology directed repair (HDR) to indels generated from non-homologous end joining (NHEJ) varied across the sub-lines, with a higher frequency of WT/indel and SNV/indel clones observed in NN0003932 and NCRM1, respectively.

Since CRISPR-Cas9-induced DNA double strand breaks can lead to undesired editing (Merkle et al., 2015a; Weisheit et al., 2020) and because cells in culture may be selected for chromosomal abnormalities that confer growth advantage, we evaluated genomic fidelity of edited clones to the parental lines using high-density SNP arrays. We analysed genomic DNA from 20 to 24 gene-edited clones from each parental line, using the NeuroChip DNA microarray

(Blauwendraat et al., 2017) and compared these results with the whole genome sequencing for each parental line. First, we confirmed using pi-hat as a point estimate of genomic similarity between parental lines that each line was distinct (pihat <0.12) except for the two clones NN0003932 and NN004297 from the same donor (pihat ~0.999, Figure S4A). Examining subclones from each line after editing, the majority of lines showed very high concordance with the parental genotypes across the genome (pihat>0.986), suggesting that the number of acquired SNVs during CRISPR editing was lower than that detectable with our achieved call rates for this array (>0.968; Table S4B, Figure S4B).

We then used the same DNA microarray genotyping data to evaluate the frequency of large chromosomal abnormalities in the edited subclones (Table S4B). Among the 185 analysed subclones, we identified 10 clones with chromosomal abnormalities involving chr12 (2 clones), chr20 (2 clones) or chr22 (6 clones). The parental line with the greatest number of abnormalities was KUCG3, whose subclones included two with duplications of Chr12 (Figure S4D), consistent with the Giemsa-band karyotyping above, and two subclones with duplications of the long arm of Chr20 (Figure S4E). These results are indicative of two distinct subclonal aneuploidies in the KUCG3 sub-line. We did not find either abnormality in the edited clones from any other parental line.

We observed recurrent chr22 abnormalities in two homozygous edited clones of KOLF2.1J and one clone each of KUCG3, NCRM1, NN0003932 and PGP1. This abnormality represents copy-neutral loss of heterozygosity (CN-LOH) from chr22q12.3 to the terminus (Figures 4, S4C). The distal end of this event corresponds to the location of the edited gene *TIMP3 (Apte et al., 1994)* suggesting that it was induced by CRISPR/Cas9 editing (Weisheit et al., 2020). However, it is important to note that at an estimated frequency of 6/185 (~3%) clones with this abnormality is lower than that seen with other methods of up to 40% (Weisheit et al., 2020).

### *Differentiation Potential*

Since the generation of disease-relevant cell types is an important use of hiPSCs, we tested the ability of our candidate sublines to support cell differentiation. We tested four distinct established protocols, two of which use small molecules to recapitulate developmental differentiation trajectories, and two of which were based on the overexpression of transcription factors to force iPSCs to adopt a transcriptional program similar to the target cell type (Figure 5A). Specifically, we directed the differentiation of iPSCs into either cortical or hypothalamic neurons and expressed either the transcription factor *NGN2* to induce the formation of excitatory forebrain neurons (iNeuron), or the transcription factors *NGN2, ISL1*, and *LHX3* to induce the formation of lower motor neurons (iLowerMotorneurons). To assess differentiation efficiency, we profiled the differentiated cells using single-cell RNA-sequencing.

For hypothalamic directed differentiation, cell lines were differentiated individually up to day 37 and then combined for scRNA-seq library preparation, with each line having 2 technical replicates (Figure S6A). For the cortical directed differentiation, cell lines were pooled together in equal proportion, and pools were differentiated up to day 34 by distinct investigators in two different tissue culture facilities, with each investigator having two technical replicates (Figure S6B). This pooling approach reduces technical variability and has been shown to yield minimal non-cell-autonomous effects on differentiation efficiency (Jerber et al., 2021). The technical replicate pools from each investigator were then labelled with unique lipid-oligo MULTIseq barcodes before being combined for scRNA-seq library preparation. For iNeuron and iLowerMotorneuron differentiations, cell lines were individually differentiated in parallel up to day 17 before being processed for scRNAseq in two different conditions, with each condition having two technical replicates. In one condition, cell lines were pooled in equal amounts three days after differentiations were started, and in the second condition, cell lines were differentiated individually and combined immediately before scRNAseq library preparation (Figure S6C and D).

After quality control, demultiplexing and doublet removal, we obtained 12,818 cells from the hypothalamic protocol, 9,656 cells from the cortical protocol, 27,708 cells from the iNeuron protocol, and 18,008 cells from the iLowerMotorneurons protocol. As described above for the demultiplexing of scRNAseq data from iPSCs in the pluripotent state, cell line identity was assigned to each cell using genotype information from the scRNA-seq reads. Louvain clustering revealed 17 clusters for hypothalamic differentiation and 13 clusters for the other 3 differentiations, which were then annotated using literature-curated genes indicative of cell identity (Figure S5). Clusters were then grouped into four categories: 1) target neuron, defined as clusters that clearly expressed genes indicative of the target cell population 2) neurons, defined as clusters other than the target neurons that expressed high levels of genes indicative of neurons 3) immature neurons, defined as clusters that expressed genes indicative of neuroblast identity (e.g. neurogenins) or progenitor identity (e.g. vimentin), and 4) progenitor/other, defined as any other cell types (Figure 5C). We then defined differentiation efficiency as the percentage of cells from each cell line that gave rise to the target cell type.

We found that across all four differentiation protocols, cell lines consistently generated the target cell type but with line-to-line differences in differentiation efficiency (Figure S6). We observed that the KOLF2.1, NCRM5, and NN_combined lines consistently generated a substantial fraction of target cell types, whereas LNGPI1 did not perform well in directed differentiation protocols. Indeed, UMAP projections show that this cell line does not cluster well with the other 6 genetically distinct lines in most differentiations consistent with the high expression of *UTF1* and other genes that predict poor neuronal differentiation identified in the pluripotent state (Figure 1D). Overall, these data provide another important line of evidence to guide the rational selection of a cell line for future development.

### *Selection of KOLF2.1J as a lead cell line*

Since the overall aim of this study was to identify a candidate cell line to underpin large-scale collaborative projects and improve reproducibility in the field, we asked if any of the eight lines we tested showed favourable properties across all of the measures we tested (Table S5). We removed LNGPI1 from consideration due to its unusual gene expression in the pluripotent state and poor differentiation properties. We also eliminated PGP1 due to possible residual expression of reprogramming factors suggested by GFP expression during FACS analysis, though other integration-free versions of this cell line exist (Lee et al., 2009). Though all lines were amenable to CRISPR/Cas9-mediated gene editing of individual DNA bases, lines NN0003932 and NCRM1 showed relatively low gene editing efficiencies at the tested locus. Lines KUCG3, NCRM5, and NN0004297 were fairly amenable to gene editing and differentiated well, but appeared to have slow growth kinetics relative to the other cell lines, including KOLF2.1J. Consequently, since KOLF2.1J performed well across all tested assays, we selected it as a candidate lead line.

### *KOLF2.1J lacks obvious disease-causing genetic variants*

We reasoned that any cell line selected for large-scale studies should be extensively tested for the presence of genetic variants that might hinder the interpretation of molecular and cellular phenotypes. Furthermore, we reasoned that insights into the long-range haplotype structure of KOLF2.1J would facilitate the use of this line in disease modelling and might enable the identification of complex structural variants. We therefore submitted the genomes of both KOLF2.1J and the parental KOLF2-C1 line for 10x Genomics linked-read sequencing to generate phased genotyping data to complement our earlier Illumina short-range (150 bp) 30x whole genome sequencing of KOLF2.1J. We then called high confidence and high quality SNV and insertion/deletion variants present across all three datasets. Only 25 coding variants were unique to KOLF2.1J and the remainder were shared between KOLF2.1J and the parent KOLF2-C1 (Figure 6). None of the variants unique to KOLF2.1J were predicted to be rare (GnomAD allele frequency < 0.001) or deleterious (CADD score >30) (Table S3B), suggesting the clonal bottlenecks and cell expansion accompanying the genetic correction of *ARID2* did not select for these types of deleterious variants. Among the variants shared between KOLF2.1J and its parental line, we found four flagged as potentially deleterious using the criteria described above, including loss-of-function (LOF) variants in genes *FMO2*, *ACYP2*, and *FBF1*, and a missense variant in *FGD6*. These variants were annotated to be either benign or of unknown clinical significance, and do not affect genes predicted to be loss-intolerant by LoFtool (Fadista et al., 2017). Expanding our search beyond these criteria, we confirmed the presence of a previously described (Hildebrandt et al., 2019) splice site disruption in *COL3A1*, a loss-intolerant gene (LoFtool score of 0.02) associated with the autosomal dominant vascular disease Ehlers-Danlos Syndrome (OMIM 130050). Given the role of this gene in extracellular matrix (ECM) production, we speculate that the variant will have little effect on most neural cell types, but urge caution if using KOLF2.1J to study cell lineages or co-culture systems that interact with the ECM. Together with G-band karyotyping and dGH data showing the absence of large structural variants, these findings suggest that the genome of KOLF2.1J does not harbour genetic variants that would substantially compromise the utility of this line for modelling neurological disease.

### *Distribution and community-based characterisation of KOLF2.1J*

In addition to KOLF2.1J being used as the lead line for the iNDI study, we hope it will be widely adopted by the community to facilitate data sharing, integration, and reproducibility. To facilitate adoption of this deeply-characterised KOLF2.1J cell line and its gene-edited derivatives, we distributed cells to multiple requesters, who then returned information on cell growth and differentiation potential using independent approaches. As of August 2021, 56 investigators across 3 continents and 10 countries have received the KOLF2.1J sub-line and have successfully differentiated iPSCs into numerous cell types, including three-germ layer cells (Figure S7A), NGN2-induced cortical neurons (Figure 7A-F; Figure S7B-C), skeletal myocytes (Figure 7G), forebrain neurons (Figure 7H), motor neurons (Figure 7I-J; Figure S7D), microglia and macrophages (Figure 7K-M), and dopaminergic neurons (Figure 7N; Figure S7E) and organoids (Figure 7O), using established differentiation protocols. Thus, KOLF2.1J can be robustly and reproducibly differentiated into many cellular phenotypes in laboratories across the world.

## Discussion

The discovery of hiPSCs (Takahashi et al., 2007) has generated new ways to study developmental biology, model disease, and inform cell-based therapies. While banks of genetically distinct iPSCs enable studies to probe the functional effects of common genetic variants (Jerber et al., 2021; Mitchell et al., 2020) or serve as HLA-matched donors for cell transplantation (Umekage et al., 2019), the diversity of iPSCs in use today also presents challenges to the field. Since genetic background contributes to molecular and cellular phenotypes (Kilpinen et al., 2017), results obtained on one genetic background may not always be replicated (Sittig et al., 2016). To overcome this limitation, other research communities have adopted common reference lines to facilitate replication and data integration, which we have attempted to adopt for hiPSC lines here. To the best of our knowledge, this study represents the first deep multimodal genetic and phenotypic comparison to facilitate rational hiPSC cell line selection.

We believe that KOLF2.1J is an excellent choice to become a commonly used cell line. First, the parental KOLF2-C1 line was reprogrammed using non-integrating methods under feeder-free conditions in chemically defined media and substrates (Kilpinen et al., 2017). The provenance and ethical derivation of the parental KOLF2-C1 line is well documented. Second, both the parental line (Bruntraeger et al., 2019; Skarnes et al., 2021) and KOLF2.1J (Bruntraeger et al., 2019) retain genomic fidelity during efficient CRISPR/Cas9-based gene editing. Third, the previous genetic characterisation of the line (Hildebrandt et al., 2019) and subsequent genetic correction of *ARID2*, followed by our detailed genomic characterisation here suggest that KOLF2.1J is free of obviously deleterious genetic variants. Fourth, we have functionally compared KOLF2.1J head-to-head with 7 other cell lines and found that it performed as well as, if not better than, other lines in all tested assays (see below). Fifth, to complement the international collaborative effort that went into the selection of KOLF2.1J, our community science approach indicates that this cell line behaves well across many independent groups and differentiation protocols (Figure 7, Figure S7).

By comparison, several of the other lines evaluated here were less robust in all assays. For example, the PGP1 iPSC line (Lee et al., 2009) is widely used but the clone we analysed showed evidence of residual retroviral activity and was not prioritized. We also found clones that differentiated poorly (LNGPI1) or carried recurrent chromosomal abnormalities (KUGC3). We also found lines that were less efficiently edited than KOLF2.1J, at least for the test edit we performed here (NCRM1 and NN0004297). We also initially considered the WTC-11 iPSC line that is readily obtainable (Coriell GM25256), has been used to generate reporter lines (Roberts et al., 2017, https://www.allencell.org/genomics.html), and formed the basis of CRISPR screens (Tian et al., 2019, 2021). However, we did not consider this line further due to the presence of a potentially neuroprotective genetic variant (rs1990621) in *TMEM106B* (Li et al., 2020). Thus, it should be noted that we make no statement that KOLF2.1J is an idealized line for all purposes. Rather, it is fit for purpose for the development of a community-facing resource of isogenic edited lines relevant to ADRD (Ramos et al., 2021), an initiative of the Center for Alzheimer's and Related Dementias (CARD) that aims to generate transformative foundational data and resources for AD/ADRD.

KOLF2.1J is derived from a male Northern European donor, which is a result of our choice to initially work with male lines to avoid complications that might arise from the erosion of X-chromosome inactivation during culture (Bar et al., 2019; Mekhoubad et al., 2012). Future studies should be directed towards the use of female lines from more varied genetic backgrounds, and we believe that the workflow described in this study can serve as a blueprint for such evaluations. All code used in these studies can be found at https://github.com/NIHCARD/INDI-README/tree/main/INDI-genetics to facilitate such efforts. While the characterisation performed in this study was aimed to be thorough, it was not comprehensive. We examined eight hiPSC lines and cannot exclude the possibility that other lines might be equally suitable for our purposes or for other studies. We did not assess structural variants smaller than the detection limit of G-band karyotyping and directional genomic hybridization (approximately 5 Mbp) or larger than approximately 50 bp as assessed by short-read whole genome sequencing. However, multiple whole genome sequencing data have been generated for KOLF2.1J and are available to the entire community on the Alzheimer's Disease Workbench (https://fair.addi.addatainitiative.org/#/data/datasets/a_reference_induced_pluripotent_stem_cell_line_for_large_scale_collaborative_studies). We did not fully investigate the possibility of "on target" effects of CRISPR in homozygous clones, such as deletions/insertions at the target site, which would require an NGS-based approach or qgPCR (Weisheit et al., 2020) to verify that both alleles contain the edit of interest without additional local alterations. However, we were pleased to observe that the copy-neutral LOH events that initiate at the CRISPR target site and extend out to the telomere and are reported to affect up to 40% of edited clones using some editing approaches (Weisheit et al., 2020) only affected 6/185 clones (~3%) in this study, suggesting that the high gene editing efficiencies we observed were not achieved at the expense of genomic stability. Furthermore, we found that the p53 pathway was robustly inducible in all tested cell lines, suggesting that any advantages in growth rate for lines are not due to acquisition of oncogenic potential. For KOLF2.1J, we have not found recurrent abnormalities that support cell survival such as chr20 duplication (Assou et al., 2020), although it will be important to survey lines after additional editing and passages in culture.

Finally, we have made the data in this study freely available to the community on the Alzheimer's Disease Workbench (https://fair.addi.addatainitiative.org/#/data/datasets/a_reference_induced_pluripotent_stem_cell_line_for_large_scale_collaborative_studies) and have established a pipeline to distribute these cell lines to minimise the regulatory and logistical hurdles that can frustrate the sharing of other cell lines. Groups can obtain the KOLF2.1J cell line at a similar passage as that analysed in this study to minimise the likelihood of genetic drift. Our vision is that the deep genotypic and phenotypic characterisation of this cell line, its proven performance in many laboratories, and its relative ease of distribution will lead to its widespread adoption by groups seeking to work with a trusted iPSC line.

## Acknowledgements

This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging (NIA), National Institutes of Health, Department of Health and Human Services; project number ZO1 AG000535, as well as the National Institute of Neurological Disorders and Stroke. A.Y. is supported by a fellowship from the EMBL-EBI/Cambridge Computational Biomedical Postdoctoral Fellowship (EBPOD) programme. V.P. is supported by the UK Regenerative Medicine Platform (MR/R015724/1). The costs of directed differentiation and single-cell sequencing were supported by the Chan Zuckerberg Initiative [191942]. F.T.M. is a New York Stem Cell Foundation - Robertson Investigator [NYSCF-R-156] and is supported by the Wellcome Trust and Royal Society [211221/Z/18/Z]. J.C.M. acknowledges core funding from the European Molecular Biology Laboratory and from Cancer Research UK (C9545/A29580). We thank Greg Strachnan and the MRC Metabolic Diseases Unit Imaging Core Facility for assistance with imaging, and Katarzyna Kania and the staff at the Genomics Core Facility of the Cancer Research UK Cambridge Institute in Cambridge for assistance with single-cell library preparation and sequencing. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD, USA. (http://biowulf.nih.gov). This research has been conducted using the UK Biobank Resource under Application Number 33601. We gratefully acknowledge the contribution of the Scientific Services at The Jackson Laboratory for work described in this publication, including the Cellular Engineering service, the Genome Technologies service for expert assistance with whole genome sequencing and DNA and RNA preparation, the Flow Cytometry service for expertise in the p53 assay, and Scientific Instrument Services. We further thank the Cytogenetics laboratory at JAX for G-band karyotyping and analysis of several iPSC sublines

Funding from the Dan and Diane Riccio Fund for Neuroscience was provided to D.A.B. M.S.B. is supported by T32AG066596-01 training grant. D.W.C. is funded by R01-NS27036. M.D. and S.C. are supported by a CZI Neurodegeneration Challenge Network Collaborative Pairs Award.

J.V.L. is supported by a DOC-PR04 PhD fellowship from the University of Antwerp and an FWO-Flanders travel grant to perform the research at Gladstone Institute. E.L is funded by the National Defense Science and Engineering Graduate Fellowship. D.D. was supported by T32-AG-000255. E.L.F.H. is funded by NINDS R01 NS060698. I.V.L.R. is funded by the California Institute for Regenerative Medicine Scholars Research Training Program (EDUC4-12812) and the Ruth L. Kirschstein National Research Service Award (T32 NS115706). M.K. is funded by a Chan Zuckerberg Initiative Ben Barres Early Career Acceleration award. R.v.d.K is supported by a Chan Zuckerberg Initiative Neurodegeneration Challenge Network Pilot Grant (2020-222244(5022)). L.E. is supported by the Studienstiftung des deutschen Volkes. M.R is supported by R01GM127557 and the Tufts Springboard Award. L.H is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; project number: 471227244). D.K-V. is supported by grants from the Chan Zuckerberg Initiative (2020-221779(5022) & 2021-235147). B.S. and P.V are supported by the Chan-Zuckerberg (CZI) Neurodegeneration Challenge Network and Amici Lovanienses. C.C. is funded by the Marie Skłodowska-Curie Actions - Seal of Excellence FWO and postdoctoral fellowship. P.V. is supported by Mission Lucidity, an ERC consolidator grant, and Research Foundation Flanders (FWO). B.J.W is a New York Stem Cell Foundation - Robertson Investigator and is supported by The New York Stem Cell Foundation [NYSCF-I-R44]. A.H. is supported by NRSA postdoctoral fellowship F32NS114319. G.L. and E.A. are supported by: EU H2020-MSCA-ITN-2018 (813851), CZI Foundation Neurodegeneration Challenge Network (2018-191929), Wallenberg Scholar (KAW2018.0232), ERC Advanced (PreciseCellPD, 884608), SSF (SB16-0065), VR (2020-01426), Hjarnfonden (FO2019-0068), Cancerfonden (CAN2016/572), Karolinska Institutet (StratRegen 2018). D.P. is supported by grants from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy within the framework of the Munich Cluster for Systems Neurology (EXC 2145 SyNergy – ID 390857198). S.W and R.G are supported by Alzheimer's Research UK (ARUK-SRF2016B-2), the UCL Neurogenetic Therapies Programme, generously funded by The Sigrid Rausing Trust and the National Institute for Health Research University College London Hospitals Biomedical Research Centre. K.J.C. and K.Z is supported by grants from NIH-NINDS/NIA (R01NS117461), DoD (W81XWH-21-1-0082), Target ALS, and the Frick Foundation for ALS Research.

## Author contributions

J.C.M., F.T.M., M.E.W., M.R.C., and W.C.S. conceived of and oversaw the study and wrote the manuscript with C.B.P., A.Y., E.L., J.A.M., C.B., and L.P. All authors read and provided feedback on the manuscript prior to submission. All co-first authors (C.B.P., A.Y., E.L., J.A.M., C.B., and L.P.) contributed equally to this manuscript and all authors agree that they can both indicate their equal contribution and re-order the list of co-first authors in their own publication records. E.L., A.Y., J.A.M., C.B., L.P., H.O., J.Z., D.S., G.P., E.C., J.B., P.B., L.K-B., C.M., C.T., S.H., S.A., P.K., V.P. S.C., S.S. performed experiments for the manuscript, with assistance from M.S., M.A.N., F.F., D.V., Y.A.Q., D.R., J.S., P.N., J.P., L.P., and M.P.N., under the guidance of M.R.C., J.C.M., W.C.S., M.E.W., and F.T.M.

C.B.P., E.L., A.Y., J.A.M., C.B., L.P., H.O., J.Z., D.S., G.P., E.C., J.B., P.B., L.K-B., C.M., C.T., S.H., L.U.R, S.C., analysed and interpreted results for these studies.

C.B.P., E.L., A.Y., J.A.M., C.B., L.P., H.O., J.Z., and K.A. prepared figures for the publication.

E.C., L.S., T.A, S.F, D.A.B., M.S.B., D.W.C., M.C.Z., R.B., M.D., S.C., Z.S.N., M.M., J.V.L, V.T., B.R.C., D.D., E.L.F.H, E.L., I.V.L.R., M.K., T.P., M.O., J.D., R.v.d.K., L.E., M.W., D.K-V., D.A-B., C.L.P., M.R, L.H., B.S., C.C.A, P.V., A.H., B.J.W, G.L., E.A., A-C.R., G.B., D.C., D.P., R.M.C.G., S.W., K.J.C, and K.Z. conceptualized use case studies to differentiate KOLF2.1J into different cellular phenotypes, developed the methods, wrote the methods and results, and prepared figures for these studies.

C.B.P. managed and coordinated manuscript organization and resource/data sharing, with E.L., A.Y., J.A.M., C.B., L.P., under the guidance of J.C.M., W.C.S., M.R.C., M.E.W., and F.T.M.

## Declaration of interests

None to declare

## Methods

*Plasmids*
The plasmids PB-TO-hNGN2 (Addgene plasmid #172115), PB-TO-hNIL (NGN2, ISL1, and LHX3, Addgene plasmid #172113), EFa1-Transposase (Addgene plasmid #172116) were used for transcription factor-based differentiation experiments. The PG13-mCherry p53 reporter plasmid was generated by replacing the luciferase sequence of PG13-Luc reporter plasmid (el-Deiry et al., 1993) (Addgene plasmid # 16442), which contains 13 copies of a p53-binding site followed by the polyoma promoter, with the mCherry sequence by the In-Fusion HD Cloning Plus kit (Takara, 638910).

*Sub-cloning*
Eight candidate iPSC lines were put through a uniform workflow for subline generation, clone expansion, and archiving of clonal lines, as described in detail in the Supplementary Information.

*Proliferation rates*
Lines were maintained in Essential 8 medium (Thermo Fisher Scientific) on Matrigel (1:100, Corning)-coated plates and passaged at 70-80% confluence with Accutase (Thermo Scientific) to a single-cell suspension. Dissociated iPSC were plated onto a Matrigel coated 48 well plates at 30,000 cells/well in Essential 8 medium (Thermo Fisher Scientific) and 10µM Rock inhibitor Y-27632 (Selleck Chem, n=6 wells per line). After 24 hr, the media was changed to Essential 8 medium. Plates were scanned in an Incucyte® S3 Live-Cell Analysis System every 24 h and confluence was analysed with Incucyte software. After 48 hr, iPSCs were dissociated with Accutase and total cell numbers counted (n=4 wells per line).

*Flow cytometry analysis of pluripotency markers*
iPSCs were dissociated into single cells using TrypLE (Thermo) for 5-10 minutes, pelleted by centrifugation at 200x*g* for 5 minutes, and then fixed in 500 µL 4% paraformaldehyde for 10-15 minutes at room temperature. After fixation, cell pellets were washed with 1mL PBS, and incubated with 50 µl of permeabilization buffer (PBS plus 2% FBS and 0.2% Tween-20) for 10 minutes. During the permeabilization, 1 µl antibody was diluted in a 5 µl permeabilization buffer and added to 96-well for each staining reaction, then 50 µl permeabilized cells were added to each well and incubated for 1 hour at 4°C with mixing occasionally. After staining, cell pellets were washed and resuspended with 200 µl per-well PBS for flow cytometry analysis. The antibodies and isotype controls used were: TRA-1-60 Monoclonal Antibody (TRA-1-60), DyLight 488 (Life Technology MA1-023-D488X); Mouse IgM Isotype Control, DyLight 488 conjugate (Life Technology MA1-194-D488); CABS352A4 Milli-Mark™ Anti-Nanog-Alexa Fluor 488 Antibody, NT (EMD Millipore FCABS352A4); Rabbit IgG isotype control, AlexaFluor 488 conjugate, (Cell Signalling 4340S).

*p53 reporter assay*

iPSCs were maintained in StemFlex medium (Thermo Fisher, A3349401) on Synthemax II-SC substrate (Corning). At 70% confluence, the culture medium was replaced with fresh medium supplemented with RevitaCell (Thermo Fisher) five hours before nucleofection. Cells were dissociated with pre-warmed Accutase (STEMCELL Technologies) at 37°C for 7 minutes and $4x10^5$ cells were transferred to a well of a 96-well V-bottom plate (Corning) then centrifuged at 100x*g* for 3 minutes. Cell pellets were resuspended with 20 µL of P3 Primary Cell Buffer (Lonza) containing 5 µg of the PG13-mCherry reporter plasmid and transferred to a well of a 16-well Nucleocuvette strip, followed by nucleofection with the 4D-Nucleofector Unit (Lonza) using the CA-137 pulse code. After nucleofection, $1.5x10^5$ cells (for the no treatment group) or $2.5x10^5$ cells (for the doxorubicin treatment group) were seeded in the StemFlex medium supplemented with RevitaCell on a well of a 48-well plate coated with Matrigel hESC-Qualified Matrix (Corning). One day after nucleofection, the medium was changed to StemFlex without RevitaCell. Two days after nucleofection, the medium was changed to StemFlex with or without 20 nM doxorubicin (Bio-Techne, 2252). Three days after nucleofection, cells were dissociated with Accutase and mCherry expression in the singlet cell population was analysed using a FACSymphony A5 flow cytometer (BD Biosciences). *TP53*-deficient KOLF2 cells (W Skarnes, unpublished) were used as a negative control. Non-viable cells were excluded by staining with 4',6-diamidino-2-phenylindole (Thermo Fisher, D1306). Flow-cytometric data were analysed using FlowJo software (BD Biosciences).

*DNA and RNA preparation*

DNA and RNA extraction was performed by the JAX Genome Technologies service, quantified by TapeStation (Agilent), and assigned a DIN or RIN value. The extracted DNA and RNA from each of the 8 sublines was submitted to Psomagen for Illumina short read whole genome sequencing. From an additional well, high molecular weight genomic DNA extraction was performed by the JAX Genome Technologies service, quantified by TapeStation (Agilent), and assigned a DIN value. The DNA from each of the 8 sublines was submitted to Psomagen for 10x Genomics long read whole genome sequencing. From an additional well, genomic DNA for each line was also prepared using the DNeasy Blood and Tissue kit (Qiagen) and submitted to the JAX Genome Technologies service for Illumina short read whole genome sequencing.

*Whole genome sequencing and annotation of variants*

The parental iPSC lines and a National Institute of Standards and Technology (NIST) reference (HG-002) were sequenced with 30x coverage and paired-end through the Illumina short read sequencing and the 10x Genomics linked-read sequencing by Jax and/or Psomagen, Inc. For Illumina short read data, SNVs and indels were called using the HaplotypeCaller (link) following the Genome Analysis Toolkit (GATK) best practices and executed through the Google genomics alpha pipeline. FASTQ files were processed into unmapped BAM files using the paired-fastq-to-unmapped-bam workflow on the human GRCh38 build. Initial variant calling was performed using the PairedSingleSampleWf. The joint discovery was then executed using the JointGenotypingWf. Variants were filtered using the variant quality score recalibration (VQSR) with default filtering parameters. The structural variant calling was performed using the Manta algorithm (Version 1.6.0) and then standardised using the structural variant tool kit (SVTK). For the 10x Genomics linked-read data, the SNP and indel variants and the structural variants were called using the 10x Genomics LongRanger wgs (version 2.2) pipeline. Sequencing reads were aligned to the human GRCh38 build containing decoy contigs and subjected to variant calling and phasing. The GATK's HaplotypeCaller mode was applied to call SNPs and Indels.

Variants were also annotated using ANNOVAR (Wang et al., 2010) including the ClinVar database (version clinvar_20200316) to identify potential known pathogenic variants.

Additionally, all data were screened for loss of function variants (stop, frame-shift and splicing) in INDI project genes and specific variants of interest including *APOE* haplotype (rs429358 and rs7412), *MAPT* haplotype (rs1800547) and *TMEM106B* (rs3173615) genotype. Polygenic risk scores for AD and PD were calculated using PLINK (v1.9) with the weights of recent GWAS (Kunkle et al., 2019; Nalls et al., 2019). As a reference population for the polygenic risk score we used AD (Data-Field 131037), PD (Data-Field 131023) and controls (no known neurodegenerative disease, no parent with a known neurodegenerative disease and >=60 years old at recruitment) from the UK Biobank (application ID: 33601)(Bycroft et al., 2018).

*CRISPR/Cas9 genome editing*

Editing was performed on each iPSC subline by high-throughput engineering of a missense mutation (S38C) in exon 1 of the *TIMP3* gene, using optimized conditions for homology-directed repair (HDR) (Skarnes et al., 2019).

Cas9 sgRNA to *TIMP3* (CCAGGAGCGCTTACCGATGT/CGG) was chemically synthesized with 2'-O-methyl and 3'-phosphorothioate end modifications (Synthego CRISPRevolution sgRNA) and resuspended in TE buffer at a concentration of 4 µg/µl. RNP was formed by combining SpCas9 nuclease (HiFi V3, IDT) with sgRNA at a molar ratio of 1:4. A 100-nt single stranded oligo donor (ssODN) containing a G to C SNV was synthesized with HDR-optimized end modifications (Alt-R™ HDR Donor Oligo, IDT) and resuspended in DPBS-/- at a concentration of 200 pmol/µl. For high-throughput introduction of Cas9 RNP and ssODN into human iPS cells, 8 wells were transfected using Amaxa nucleofection with P3 Primary Cell 4D-Nucleofector 16-well Strips (Lonza). Each well contained a single-cell suspension of 1.6 x $10^5$ cells in 20 µl of Primary Cell P3 buffer with supplement (Lonza) containing 2 µg Cas9, 1.6µg sgRNA, and 40 pmol ssODN. Nucleofection was performed using Amaxa program CA-137. Immediately following electroporation, cells were distributed to wells of a Matrigel-coated 24-well plate containing StemFlex, RevitaCell, and 30 µM final Alt-R® HDR enhancer (IDT). Cells were incubated at 32°C for 3 days before transfer to 37°C. At 24h post-nucleofection, and every other day thereafter, the media was replaced with only StemFlex. Upon reaching near confluency, cells were single-cell-plated into Synthemax-coated 10cm dishes as described above. At Day 10, 24 colonies per cell line were manually picked as described (Skarnes et al., 2019) and incubated in Matrigel-coated 96-well plates for 4-5 days before being frozen down.

Crude cell lysates for each clone were prepared as described (Skarnes et al., 2019) and used to amplify a 896 bp genomic region containing the CRISPR target site. Sanger sequencing of purified PCR products was carried out by the Genome Technologies service at The Jackson Laboratory in Bar Harbor, Maine. Sequence traces were aligned and analysed using SeqManPro (https://www.dnastar.com/software/lasergene/molecular-biology/) and Synthego ICE (https://ice.synthego.com/).

Two additional wells of each clone were lysed, pooled, and genomic DNA was purified using the 96-well high-throughput DNeasy Blood and Tissue kit (Qiagen). Array-based genotyping was performed on the resulting genomic DNA.

*Genotyping array genotyping of subclones*

To assess the genomic fidelity of IPSC lines and subclones after editing, DNA was isolated and Illumina genotyping array was performed using the NeuroChip array and standard Illumina genotyping protocols (Blauwendraat et al., 2017). In total 185 subclones were successfully genotyped including at least 20 clones per included IPSC line. Genomic fidelity was assessed using two strategies: 1) comparison between genotyping array data and WGS data and 2) Assessment of genome wide B-allele frequency and Log R ratio values. For the comparison between genotyping array and WGS data, all data was merged using PLINK (v1.9) and only overlapping variants were kept. Potential genetic differences were identified using the --merge-mode 7 option in plink which reports mismatching non-missing calls between two datasets. Variants discordant in more than 33% of the genotyped clones were excluded due to high likelihood of being genotyping errors. Mismatching non-missing calls were plotted using R (v3.6.1) per chromosome and visually inspected for large clusters of discordant array genotypes and WGS. Genotyping array data was also assessed for large events based on the B-allele frequency and Log R ratio values. The B-allele frequency and Log R ratio values were downloaded from Illumina GenomeStudio and processed and plotted using the GWASTools package in R (v3.6.1) (Gogarten et al., 2012). GenCall score variant filtering thresholds of >0.4 and >0.7 were used to filter out calls likely arising from genotyping errors.

*Comparative whole genome sequence analysis of KOLF2-C1 and KOLF2.1*

To retrieve the highly confident variants for KOLF2.1, the three variant sets originally discovered from its default variant calling pipeline was subjected to filtering to exclude those the did not pass the thresholds of PASS, QUAL >= 30, DP >= 10, QP >= 2.0, and MQ >= 40. After filtering, the variant sets were intersected to generate the 3,278,414 common variants (SNPs/Indels) of high confidence, illustrated in Figure 5B. The common variants were subjected to annotation and effect prediction using VEP. About 88.9% variants were SNVs (Figure 5). Of the coding variants, 54.4% were found to be synonymous, 43.4% were missense, and the remainder were LOF, splice site, or other types of variant (Figure 5). KOLF2.1 highly confident SNPs/Indels were compared to KOLF-C1 SNPs/Indels to exclude the possibility that it gained deleterious variants after editing (Figure 5B). The rare and deleterious coding genes that were unique to KOLF2.1 were predicted using VEP (gnomAD_NFE_AF < 0.001 and CADD_PHRED > 30). Clinical relevance and dosage sensitivity of the predicted deleterious variants was annotated through ClinGen (Rehm et al., 2015). 25 protein-coding SNPs/Indels were found in KOLF2.1 but not in KOLF2-C1 (Table S3B). However, none had minor allele frequencies less than 0.001 or a CADD score greater than 30, suggesting that KOLF2.1 didn't gain deleterious mutations after genomic editing and passaging. Four rare and deleterious variants were found in both KOLF2.1 and KOLF-C1, but were not classified as pathogenic (Table S3C).

*Transcription Factor-NGN2 differentiation into cortical neurons*

We expressed human NGN2 under a tetracycline-inducible promoter as previously described (Fernandopulle et al., 2018) using a PiggyBac system for delivery. iPSCs were transfected with PB-TO-hNGN2 vector in a 1:2 ratio (transposase:vector) using Lipofectamine Stem (Invitrogen), then selected after 72 hrs for 2 weeks with 8 µg/mL of puromycin (Sigma-Aldrich).

iPSCs with a stably integrated human NGN2 were single-cell dissociated using Accutase (Thermo Scientific), plated 1.5 million onto a Matrigel (1:100, Corning) coated-6 well for 3 days with neuronal induction media (NIM: Knockout DMEM/F12, 1X N2 Supplement, 1X Non-Essential Amino Acids, 1X Glutamax (all from Thermo Fisher Scientific), 10 µM Rock inhibitor Y-27632 (SelleckChem) and 2 µg/mL Doxycycline (Sigma-Aldrich)). On day 3, 1.5x106 cells were replated onto a poly-L-ornithine (PLO, Sigma-Aldrich) coated-6 well for 14 days using Brainphys (Stem Cell Technologies) 1X B27 Supplement (Thermo Fisher Scientific), 10 ng/mL BDNF (PeproTech), 10 ng/mL NT3 (PeproTech), 1 µg/mL Laminin (R&D) and 2 µg/mL Doxycycline (Sigma-Aldrich). For neuronal maintenance, half of the media was changed every 2-3 days.

*Transcription Factor-based differentiation into hNIL-expressing Lower Motor Neurons*

Over-expression of NGN2-ISL1-LHX3 (hNIL) (Fernandopulle et al., 2018) was performed as described with a PiggyBac system for delivery. iPSCs were transfected with PB-TO-hNIL vector in a 1:2 ratio (transposase:vector) using Lipofectamine Stem (Invitrogen), then selected after 72 hours for 2 weeks with 8 µg/mL of puromycin (Sigma-Aldrich). The iPSCs with a stably integrated human NIL under a tetracycline-inducible promoter were exposed to doxycycline in neuronal induction medium (NIM) The iPSCs with a stably integrated human NIL under a tetracycline-inducible promoter were exposed to doxycycline in neuronal induction medium (NIM). Briefly, on day 0 the iPSCs were single-cell dissociated using Accutase (Thermo Scientific), plated 1.5 million onto a Matrigel (1:100, Corning) coated-10cm dish in Essential 8 medium (Thermo Fisher Scientific) and 10µM Rock inhibitor Y-27632 (Selleck Chem), on day 1 the media was changed with NIM: DMEM/F12, 1X N2, 1X Non-Essential Aino Acids, 1X Glutamax (all reagents were from Thermo Fisher Scientific), added 10µM Rock inhibitor Y-27632, 2µg/ml Doxycycline (Selleck Chem) and 0.2 µM Compound E (Stem Cell Technologies). On day 3, 1 million cells were re-plated onto a poly-L-ornithine (PLO, Sigma-Aldrich)-15 µg/mL Laminin (R&D) coated-6 well in NIM with 10µM Rock inhibitor Y-27632 (SelleckChem), 2 µg/mL Doxycycline (Sigma-Aldrich), 0.2 µM Compound E (Stem Cell Technologies), 1 µg/mL Laminin (R&D) and 40 µM BrdU (Thermo Fisher Scientific). On day 4, the media was changed to NIM with 1X B27 Supplement (Thermo Fisher Scientific), 1X Culture One Supplement (Thermo Fisher Scientific), 1 µg/mL Laminin (R&D), 20 ng/mL BDNF (PeproTech), 20 ng/mL GDNF (PeproTech) and 10 ng/mL NT3 (PeproTech), on day 7, ½ of the media was changed. On day 10, half of the media was changed with Neurobasal, 1X B27, 1X N2, 1X Culture One (all from Thermo Fisher Scientific), 40 ng/mL BDNF (PeproTech), 40 ng/mL GDNF (PeproTech), 20 ng/mL NT3 (PeproTech), 1 µg/mL Laminin (R&D) and 2 µg/mL Doxycycline (Sigma-Aldrich). For neuronal maintenance, half media was changed every 2-3 days.

*Directed differentiation to cortical and hypothalamic lineages*

Prior to differentiation, hiPSC cell lines were maintained in mTeSR1 on geltrex (1:100, Thermo Fisher Scientific) and passaged with EDTA (0.5 µM, pH 8.0, Thermo Fisher, 15575-020) at 60-80% confluence. For each line, we confirmed that colonies had clearly defined borders and cultures lacked differentiated cells when viewed under a phase contrast microscope. Lines were passaged at least twice under these conditions before differentiation experiments were initiated, and lines were synchronised by adjusting split ratios so that the last passage was 3-4 days before plating for differentiation. For differentiation, hiPSC lines were dissociated to a single-cell suspension with TrypLE, counted, and plated onto coated plates. Cell lines were pooled for cortical differentiation, and grown separately for hypothalamic differentiation. For cortical and hypothalamic differentiation we followed published methods (Kirwan et al., 2017; Merkle et al., 2015b).

Briefly, dissociated hiPSCs were plated at 10,000 cells/cm$^2$ into 6-well plates in the presence of 10 µM ROCK inhibitor (Y-27632, Tocris Bioscience, Bristol, UK). Cortical and hypothalamic differentiation took place on a substrate of geltrex (1:100) in N2B27 media containing 500 ml Neurobasal-A (Thermo Fisher Scientific, cat. no 10888022), 500 ml DMEM/F12 with GlutaMAX (Thermo Fisher Scientific, cat. no 31331093), 10 ml Glutamax (Thermo Fisher Scientific, cat. no 35050038), 10 ml sodium bicarbonate (Thermo Fisher Scientific, cat. no 25080-094), 5 ml MEM Nonessential Amino Acids (Thermo Fisher Scientific, cat. no 11140035), 1 ml 200 mM ascorbic acid (Sigma-Aldrich, cat. no A4403) 10 ml 100x penicillin-streptomycin (Thermo Fisher Scientific, cat. no. 15140122), 20 ml 50x B27 supplement (Thermo Fisher Scientific, cat. no 17504044), 10 ml 100x N2 supplement (Thermo Fisher Scientific, cat. no 17502048). Patterning to forebrain progenitors in this medium was directed using 100 nM LDN-193189 (Stemgent, cat. no. 04-0074), 10 µM SB431542 (Sigma-Aldrich, cat. no. S4317), and 2 µM XAV939 (Stemgent, cat. no. 04-0046). The concentrations of these small molecules were adjusted over time as previously described (Kirwan et al., 2017), and for hypothalamic differentiation the small molecules SAG (Fisher Scientific, cat. no. 56-666-01MG) and purmorphamine (Calbiochem, cat. no. 540220) were each added to a final concentration of 1 µM from day 2-8 of differentiation. To assess the efficiency of differentiation and the identity of progenitors, cells were plated in a separate 48-well plate at 400,000 cells/well on day 11 and fixed for immunofluorescence assay on day 16.

*Cell dissociation for single-cell RNA sequencing*

iNeurons and iLowerMotorneurons were washed once with PBS after 17 days of differentiation. The lines were then either differentiated in separate wells and pooled in a single tube at the end of differentiation, or the lines were pooled at the beginning of differentiation and were resuspended in 1x PBS-0.04% BSA (Jackson Immunoresearch) and washed 3 additional times with this solution after single-cell dissociation. Single-cell pellets were resuspended in 1x PBS-0.04% BSA, counted using an automated cell counter (Countess II) and the concentration was adjusted to 1x10$^6$ cells/mL.

Cortical and hypothalamic neurons were washed once with 1X DPBS (Thermo Fisher, 14190-144) before adding TrypLE containing 20 units/ml of papain after 5 weeks of differentiation (34 days for cortical and 37 days for hypothalamic). Cultures were incubated at 37°C until the cells physically detached from each other when viewed under a phase contrast microscope and could be readily dissociated with a P1000 pipette. Enzyme mix was aspirated and cells were dissociated with a P1000 in DMEM:F12 (Thermo Fisher Scientific, 10565-018) supplemented with 10 µM Rock inhibitor, 33 µg/ml DNase I (Worthington, LK003170), and 45 uM Actinomycin D to block dissociation-induced transcription. The resulting cell suspension for each separate well was passed through a 40 µm cell strainer and brought to 10 ml in the dissociation solution centrifuged at 160x g for 5 minutes. For cortical differentiations, each well was uniquely barcoded using cholesterol-modified MULTIseq oligonucleotides to facilitate cell pooling during droplet-based single-cell cDNA library preparation based on a published protocol (McGinnis et al., 2019). After two additional washes, cells were resuspended in 1x DPBS containing 0.04% BSA (Sigma, A0281) and washed 3 additional times in 1x DPBS containing 0.04% BSA. Single-cell suspensions were counted using an automated cell counter (Countess II).

*Chromium 10x Genomics library and sequencing*

For iPSC, iNeurons, and iLowerMotorneurons, single-cell RNA sequencing was performed using Chromium Single Cell 3' Reagent kit V3.1 (PN-1000128) and 25,000 cells per condition were loaded into the 10x Genomics chip G. For cortical and hypothalamic neurons, Single-cell suspensions were processed by the Chromium Controller (10x Genomics) using Chromium Single Cell 3' Reagent Kit v3 (PN-1000075) according to the manufacturer's specifications. On average, 15,000 cells from each 10x reaction were directly loaded into one inlet of the 10x Genomics chip. Barcoded libraries were sequenced using the Illumina Novaseq 6000 (one lane per 10x chip position) with 75 bp paired-end reads to an average depth of approximately 50,000 reads per cell.

*scRNA-seq data processing and quality control*

Raw sequencing libraries were processed using 10x Genomics' Cell Ranger platform (version 3.1). Reads were aligned and quantified to the 10x Genomics provided human reference genome (GRCh38, Ensembl 93). The samples were then grouped based on differentiation protocols and each group were processed independently in subsequent downstream analysis.

Droplets containing captured cells were called using the emptyDrops function from the DropletUtils R package (Lun et al., 2019), using varying UMI threshold per differentiation protocol groups and an FDR of 0.001. Low quality cells and outlier cells were then filtered based on the total unique molecular identifier (UMI) content, number of detected features/genes and fraction of mitochondrial content. Cells were discarded if their UMI content is more than ± 3 median absolute deviation (MAD) away from the median, or the detected features is more than ± 3 MAD away from the median, or the fraction of mitochondrial content is higher than 3 MAD from median. Gene expression levels were normalized using the logNormCounts function from the scran R package (Lun et al., 2016), with size factors estimated using the computeSizeFactors function.

*Doublet detection*

Doublet identification was done in two stages – at the individual sample level and across samples within the same differentiation protocol. First, doublets were detected at the individual sample level using the hybrid method (cxds_bcds_hybrid function with estNdbl parameter set to true) from the scds R package (Bais and Kostka, 2020). This is followed by identification of 'guilt-by-association' doublets, where doublets were further identified if there is enrichment of scds' hybrid-based doublets in the neighboring cells (number of neighbor = 3 for iPSC and 5 for other differentiations). Clustering was then performed on cells in each sample (see below), and this was repeated for each identified cluster to form smaller sub-clusters. Finally, cells were also classified as doublets if cells belong to sub-clusters containing more than 50% of Vireo-identified or MULTI-seq-identified doublets.

For doublet identification across samples within the same differentiation protocol, samples were first batch corrected (see below) into a single dataset per differentiation protocol, followed by two rounds of clustering to identify cell sub-clusters. Cells were classified as cross-sample doublets if they belonged to sub-clusters with an enriched fraction of per-sample doublets (>3 MAD away from the median). Cells which were classified as either per-sample doublets, cross-sample doublets, Vireo doublets or MULTIseq doublets were excluded from further downstream analysis.

*Batch correction and dimensionality reduction*

For each differentiation protocol, samples were combined into a single dataset and corrected for batch effect using the fastMNN function from Batchelor R package (Haghverdi et al., 2018) on the first 50 principal components computed from highly variable genes (HVG). HVG were selected by fitting the mean-variance curve on the normalized gene expression across all samples within a differentiation group with modelGeneVar from scran R package and filtering for genes which have higher variance than the fitted trend. Mitochondrial genes and ribosomal genes for large and small ribosomal subunits were excluded from mean-variance curve fitting as these genes have both high variance and expression. For visualization, Uniform Manifold Approximation and Projection (UMAP) two-dimensional embedding (McInnes et al., 2018) were calculated from the corrected principal component with the following settings: spread = 1 and minimum distance = 0.4.

*Clustering and annotation*

Cells were grouped into clusters for each differentiation protocol using the community detection-based Louvain clustering method. Briefly, shared nearest-neighbor graphs were constructed from the 50 corrected principal components, followed by clustering using the Louvain method (cluster_louvain function) from the igraph R package (Nepusz, 2006). Each cluster was then manually annotated with cell type based on a list of curated markers (Table S6) and further assigned into one of four cell type groups for evaluating differentiation efficiency of each cell line.

*Cell line and replicate demultiplexing*

Cell line identity was inferred based on genotype information using 10x Genomics VarTrix and Vireo tools (Huang et al., 2019). Variant count matrices for captured cells were produced by VarTrix using aligned reads from Cell Ranger output and variants called from whole genome sequencing data (see above). Cell line identity for captured cells were then determined with Vireo using the variant count matrix and variant information. Only variants from 7 cell lines were

utilized for cell line demultiplexing as 2 cell lines (NN0003932 and NN0004297 – denoted as NN_combined) were derived from the same parental lines.
Replicate demultiplexing of MULTI-seq labelled samples was performed using the deMULTIplex R package. Briefly, MULTI-seq barcode reads from captured cells were aligned to the MULTI-seq barcodes used for labelling each sample, followed by read deduplication based on UMI and generation of MULTIseq barcode count matrix. Replicate classification was then performed on the barcode count matrix iteratively until there are no negative classified cells, followed by a negative-cell reclassification to recover incorrectly classified negative cells. Cells were assigned to cell cycle phases based on the expression of the G2/M and S phase markers (Tirosh et al., 2016) using the CellCycleScoring function from Seurat R package (Hao et al., 2021).

*Immunofluorescence and imaging*

Cells were fixed in 4% paraformaldehyde for 10 min at room temperature and washed 3x in PBS. Afterwards, the cells were incubated in blocking solution composed of PBS, 0.2% Triton-X and 4% donkey serum for 1 hr at 4°C. Primary antibody, anti-FOXA2 (R&D Systems AF2400) and anti-Lmx1 (Merck Millipore, AB10533) were added 1:100 into blocking solution (PBS, 5 % donkey serum, 0.1 % Triton X10) and the cells were incubated for 2 hr at room temperature. Samples were washed in 3x blocking solution before the addition of 1:500 donkey anti-goat IgG secondary antibody, Alexa Fluor® 561 (Thermo Fisher Scientific, Waltham, USA) and 1:500 donkey anti-goat IgG (H+L) secondary antibody, Alexa Fluor® 647 (Thermo Fisher Scientific) to the cells for 1 hr at room temperature. Samples were washed 3x in blocking solution and 1x in PBS. 1:1000 Hoechst 33342 (Invitrogen, Thermo Fisher Scientific) was added to the cells in PBS before imaging them.

## References

Apte, S.S., Mattei, M.G., and Olsen, B.R. (1994). Cloning of the cDNA encoding human tissue inhibitor of metalloproteinases-3 (TIMP-3) and mapping of the TIMP3 gene to chromosome 22. Genomics *19*, 86–90.

Assou, S., Girault, N., Plinet, M., Bouckenheimer, J., Sansac, C., Combe, M., Mianné, J., Bourguignon, C., Fieldes, M., Ahmed, E., et al. (2020). Recurrent Genetic Abnormalities in Human Pluripotent Stem Cells: Definition and Routine Detection in Culture Supernatant by Targeted Droplet Digital PCR. Stem Cell Reports *14*, 1–8.

Avior, Y., Lezmi, E., Eggan, K., and Benvenisty, N. (2021). Cancer-Related Mutations Identified in Primed Human Pluripotent Stem Cells. Cell Stem Cell 28, 10–11.

Bais, A.S., and Kostka, D. (2020). scds: computational annotation of doublets in single-cell RNA sequencing data. Bioinformatics *36*, 1150–1158.

Bar, S., Seaton, L.R., Weissbein, U., Eldar-Geva, T., and Benvenisty, N. (2019). Global Characterization of X Chromosome Inactivation in Human Pluripotent Stem Cells. Cell Rep. *27*, 20–29.e3.

Blauwendraat, C., Faghri, F., Pihlstrom, L., Geiger, J.T., Elbaz, A., Lesage, S., Corvol, J.-C., May, P., Nicolas, A., Abramzon, Y., et al. (2017). NeuroChip, an updated version of the NeuroX genotyping platform to rapidly screen for variants associated with neurological diseases. Neurobiol. Aging *57*, 247.e9–e247.e13.

Bonyadi, M., Rusholme, S.A.B., Cousins, F.M., Su, H.C., Biron, C.A., Farrall, M., and Akhurst, R.J. (1997). Mapping of a major genetic modifier of embryonic lethality in TGFβ1 knockout mice. Nature Genetics *15*, 207–211.

Bruntraeger, M., Byrne, M., Long, K., and Bassett, A.R. (2019). Editing the Genome of Human Induced Pluripotent Stem Cells Using CRISPR/Cas9 Ribonucleoprotein Complexes. Methods Mol. Biol. *1961*, 153–183.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.

Chal, J., Al Tanoury, Z., Hestin, M., Gobert, B., Aivio, S., Hick, A., Cherrier, T., Nesmith, A.P., Parker, K.K., and Pourquié, O. (2016). Generation of human muscle fibers and satellite-like cells from human pluripotent stem cells in vitro. Nat. Protoc. *11*, 1833–1850.

el-Deiry, W.S., Tokino, T., Velculescu, V.E., Levy, D.B., Parsons, R., Trent, J.M., Lin, D., Mercer, W.E., Kinzler, K.W., and Vogelstein, B. (1993). WAF1, a potential mediator of p53 tumor suppression. Cell *75*, 817–825.

Doetschman, T. (2009). Influence of Genetic Background on Genetically Engineered Mouse Phenotypes. Methods in Molecular Biology 423–433.

Fadista, J., Oskolkov, N., Hansson, O., and Groop, L. (2017). LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. Bioinformatics *33*, 471–474.

Fernandopulle, M.S., Prestil, R., Grunseich, C., Wang, C., Gan, L., and Ward, M.E. (2018). Transcription Factor-Mediated Differentiation of Human iPSCs into Neurons. Curr. Protoc. Cell Biol. *79*, e51.

Gantner, C.W., Hunt, C.P.J., Niclis, J.C., Penna, V., McDougall, S.J., Thompson, L.H., and Parish, C.L. (2021). FGF-MAPK signaling regulates human deep-layer corticogenesis. Stem Cell Reports *16*, 1262–1275.

Gogarten, S.M., Bhangale, T., Conomos, M.P., Laurie, C.A., McHugh, C.P., Painter, I., Zheng, X., Crosslin, D.R., Levine, D., Lumley, T., et al. (2012). GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. Bioinformatics *28*, 3329–3331.

Guttikonda, S.R., Sikkema, L., Tchieu, J., Saurat, N., Walsh, R.M., Harschnitz, O., Ciceri, G., Sneeboer, M., Mazutis, L., Setty, M., et al. (2021). Fully defined human pluripotent stem cell-derived microglia and tri-culture system model C3 production in Alzheimer's disease. Nat. Neurosci. *24*, 343–354.

Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. *36*, 421–427.

Halliwell, J., Barbaric, I., and Andrews, P.W. (2020). Acquired genetic changes in human pluripotent stem cells: origins and consequences. Nat. Rev. Mol. Cell Biol. *21*, 715–728.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29.

Hildebrandt, M.R., Reuter, M.S., Wei, W., Tayebi, N., Liu, J., Sharmin, S., Mulder, J., Lesperance, L.S., Brauer, P.M., Mok, R.S.F., et al. (2019). Precision Health Resource of Control iPSC Lines for Versatile Multilineage Differentiation. Stem Cell Reports *13*, 1126–1141.

Huang, Y., McCarthy, D.J., and Stegle, O. (2019). Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. Genome Biol. *20*, 273.

Ihry, R.J., Worringer, K.A., Salick, M.R., Frias, E., Ho, D., Theriault, K., Kommineni, S., Chen, J., Sondey, M., Ye, C., et al. (2018). p53 inhibits CRISPR-Cas9 engineering in human pluripotent stem cells. Nat. Med. *24*, 939–946.

Jerber, J., Seaton, D.D., Cuomo, A.S.E., Kumasaka, N., Haldane, J., Steer, J., Patel, M., Pearce, D., Andersson, M., Bonder, M.J., et al. (2021). Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. Nat. Genet. *53*, 304–312.

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

Kilpinen, H., Goncalves, A., Leha, A., Afzal, V., Alasoo, K., Ashford, S., Bala, S., Bensaddek, D., Casale, F.P., Culley, O.J., et al. (2017). Common genetic variation drives molecular heterogeneity in human iPSCs. Nature *546*, 370–375.

Kirwan, P., Jura, M., and Merkle, F.T. (2017). Generation and Characterization of Functional Human Hypothalamic Neurons. Curr. Protoc. Neurosci. *81*, 3.33.1–3.33.24.

Konttinen, H., Cabral-da-Silva, M.E.C., Ohtonen, S., Wojciechowski, S., Shakirzyanova, A., Caligola, S., Giugno, R., Ishchenko, Y., Hernández, D., Fazaludeen, M.F., et al. (2019). PSEN1ΔE9, APPswe, and APOE4 Confer Disparate Phenotypes in Human iPSC-Derived Microglia. Stem Cell Reports *13*, 669–683.

Kunkle, B.W., Grenier-Boley, B., Sims, R., Bis, J.C., Damotte, V., Naj, A.C., Boland, A., Vronskaya, M., van der Lee, S.J., Amlie-Wolf, A., et al. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nat. Genet. *51*, 414–430.

Kwart, D., Gregg, A., Scheckel, C., Murphy, E.A., Paquet, D., Duffield, M., Fak, J., Olsen, O., Darnell, R.B., and Tessier-Lavigne, M. (2019). A Large Panel of Isogenic APP and PSEN1 Mutant Human iPSC Neurons Reveals Shared Endosomal Abnormalities Mediated by APP β-CTFs, Not Aβ. Neuron *104*, 1022.

Landrum, M.J., Chitipiralla, S., Brown, G.R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., et al. (2020). ClinVar: improvements to accessing data. Nucleic Acids Res. *48*, D835–D844.

Lee, J.-H., Park, I.-H., Gao, Y., Li, J.B., Li, Z., Daley, G.Q., Zhang, K., and Church, G.M. (2009). A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. PLoS Genet. *5*, e1000718.

Li, Z., Farias, F.H.G., Dube, U., Del-Aguila, J.L., Mihindukulasuriya, K.A., Fernandez, M.V., Ibanez, L., Budde, J.P., Wang, F., Lake, A.M., et al. (2020). The TMEM106B FTLD-protective variant, rs1990621, is also associated with increased neuronal proportion. Acta Neuropathol. *139*, 45–61.

Lun, A.T.L., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res. *5*, 2122.

Lun, A.T.L., Riesenfeld, S., Andrews, T., Dao, T.P., Gomes, T., participants in the 1st Human Cell Atlas Jamboree, and Marioni, J.C. (2019). EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. *20*, 63.

Mackay, T.F.C., and Huang, W. (2018). Charting the genotype-phenotype map: lessons from the Drosophila melanogaster Genetic Reference Panel. Wiley Interdiscip. Rev. Dev. Biol. *7*.

McGinnis, C.S., Patterson, D.M., Winkler, J., Conrad, D.N., Hein, M.Y., Srivastava, V., Hu, J.L., Murrow, L.M., Weissman, J.S., Werb, Z., et al. (2019). MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nat. Methods *16*, 619–626.

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. J. Open Source Softw. *3*, 861.

Mekhoubad, S., Bock, C., de Boer, A.S., Kiskinis, E., Meissner, A., and Eggan, K. (2012). Erosion of dosage compensation impacts human iPSC disease modeling. Cell Stem Cell *10*, 595–609.

Merkle, F.T., Neuhausser, W.M., Santos, D., Valen, E., Gagnon, J.A., Maas, K., Sandoe, J., Schier, A.F., and Eggan, K. (2015a). Efficient CRISPR-Cas9-mediated generation of knockin human pluripotent stem cells lacking undesired mutations at the targeted locus. Cell Rep. *11*, 875–883.

Merkle, F.T., Maroof, A., Wataya, T., Sasai, Y., Studer, L., Eggan, K., and Schier, A.F. (2015b). Generation of neuropeptidergic hypothalamic neurons from human pluripotent stem cells. Development *142*, 633–643.

Merkle, F.T., Ghosh, S., Kamitaki, N., Mitchell, J., Avior, Y., Mello, C., Kashin, S., Mekhoubad, S., Ilic, D., Charlton, M., et al. (2017). Human pluripotent stem cells recurrently acquire and expand dominant negative P53 mutations. Nature *545*, 229–233.

Mitchell, J.M., Nemesh, J., Ghosh, S., Handsaker, R.E., Mello, C.J., Meyer, D., Raghunathan, K., de Rivera, H., Tegtmeyer, M., Hawes, D., et al. (2020). Mapping genetic effects on cellular phenotypes with “cell villages.”

Nalls, M.A., Blauwendraat, C., Vallerga, C.L., Heilbron, K., Bandres-Ciga, S., Chang, D., Tan, M., Kia, D.A., Noyce, A.J., Xue, A., et al. (2019). Identification of novel risk loci, causal insights, and heritable risk for Parkinson’s disease: a meta-analysis of genome-wide association studies. Lancet Neurol. *18*, 1091–1102.

Nepusz, G.C.T. (2006). The igraph software package for complex network research. InterJournal *Complex Systems*, 1695.

Ramos, D.M., Skarnes, W.C., Singleton, A.B., Cookson, M.R., and Ward, M.E. (2021). Tackling neurodegenerative diseases with genomic engineering: A new stem cell initiative from the NIH. Neuron *109*, 1080–1083.

Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., et al. (2015). ClinGen--the Clinical Genome Resource. N. Engl. J. Med. *372*, 2235–2242.

Reilly, L., Peng, L., Lara, E., Ramos, D., Fernandopulle, M., Pantazis, C., Stadler, J., Santiana, M., Dadu, A., Iben, J.R., et al. A fully automated FAIMS-DIA proteomic pipeline for high-throughput characterization of iPSC-derived neurons.

Roberts, B., Haupt, A., Tucker, A., Grancharova, T., Arakaki, J., Fuqua, M.A., Nelson, A., Hookway, C., Ludmann, S.A., Mueller, I.A., et al. (2017). Systematic gene tagging using CRISPR/Cas9 in human stem cells to illuminate cell organization. Mol. Biol. Cell *28*, 2854–2874.

Roberts, B., Hendershott, M.C., Arakaki, J., Gerbin, K.A., Malik, H., Nelson, A., Gehring, J., Hookway, C., Ludmann, S.A., Yang, R., et al. (2019). Fluorescent Gene Tagging of Transcriptionally Silent Genes in hiPSCs. Stem Cell Reports *12*, 1145–1158.

Robinson, E., McKenna, M.J., Bedford, J.S., Goodwin, E.H., Cornforth, M.N., Bailey, S.M., and Ray, F.A. (2019). Directional Genomic Hybridization (dGH) for Detection of Intrachromosomal Rearrangements. Methods Mol. Biol. *1984*, 107–116.

Sittig, L.J., Carbonetto, P., Engel, K.A., Krauss, K.S., Barrios-Camacho, C.M., and Palmer, A.A. (2016). Genetic Background Limits Generalizability of Genotype-Phenotype Relationships. Neuron *91*, 1253–1259.

Skarnes, W.C., Pellegrino, E., and McDonough, J.A. (2019). Improving homology-directed repair efficiency in human stem cells. Methods *164-165*, 18–28.

Skarnes, W.C., Ning, G., Giansiracusa, S., Cruz, A.S., Blauwendraat, C., Saavedra, B., Holden, K., Cookson, M.R., Ward, M.E., and McDonough, J.A. (2021). Controlling homology-directed repair outcomes in human stem cells with dCas9.

Steinberg, S., Stefansson, H., Jonsson, T., Johannsdottir, H., Ingason, A., Helgason, H., Sulem, P., Magnusson, O.T., Gudjonsson, S.A., Unnsteinsdottir, U., et al. (2015). Loss-of-function variants in ABCA7 confer risk of Alzheimer’s disease. Nat. Genet. *47*, 445–447.

Sterken, M.G., Snoek, L.B., Kammenga, J.E., and Andersen, E.C. (2015). The laboratory domestication of Caenorhabditis elegans. Trends Genet. *31*, 224–231.

Takahashi, K. *et al.* (2007) ‘Induction of pluripotent stem cells from adult human fibroblasts by defined factors’, *Cell*, 131(5), pp.861-72.

Threadgill, D.W., Dlugosz, A.A., Hansen, L.A., Tennenbaum, T., Lichti, U., Yee, D., LaMantia, C., Mourton, T., Herrup, K., and Harris, R.C. (1995). Targeted disruption of mouse EGF receptor: effect of genetic background on mutant phenotype. Science *269*, 230–234.

Tian, R., Gachechiladze, M.A., Ludwig, C.H., Laurie, M.T., Hong, J.Y., Nathaniel, D., Prabhu, A.V., Fernandopulle, M.S., Patel, R., Abshari, M., et al. (2019). CRISPR Interference-Based

Platform for Multimodal Genetic Screens in Human iPSC-Derived Neurons. Neuron *104*, 239–255.e12.

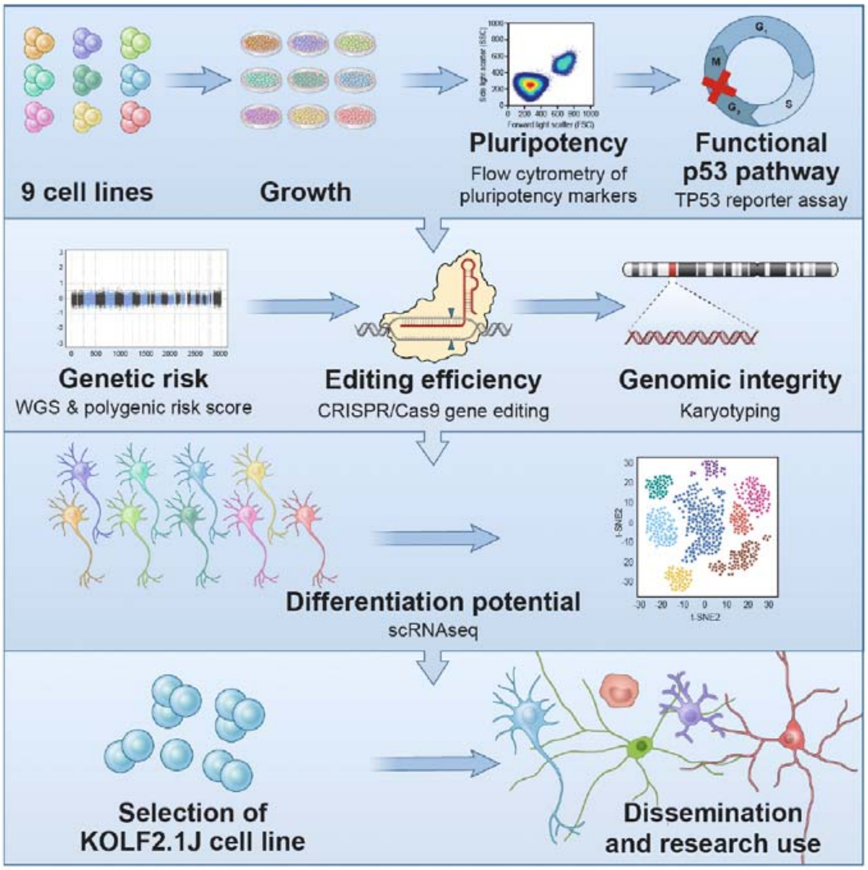
Tian, R., Abarientos, A., Hong, J., Hashemi, S.H., Yan, R., Dräger, N., Leng, K., Nalls, M.A., Singleton, A.B., Xu, K., et al. (2021). Genome-wide CRISPRi/a screens in human neurons link lysosomal failure to ferroptosis. Nat. Neurosci. *24*, 1020–1034.

Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., 2nd, Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189–196.

Umekage, M., Sato, Y., and Takasu, N. (2019). Overview: an iPS cell stock at CiRA. Inflamm. Regen. *39*, 17.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.

Weisheit, I., Kroeger, J.A., Malik, R., Klimmt, J., Crusius, D., Dannert, A., Dichgans, M., and Paquet, D. (2020). Detection of Deleterious On-Target Effects after HDR-Mediated CRISPR Editing. Cell Rep. *31*, 107689.



Graphical abstract

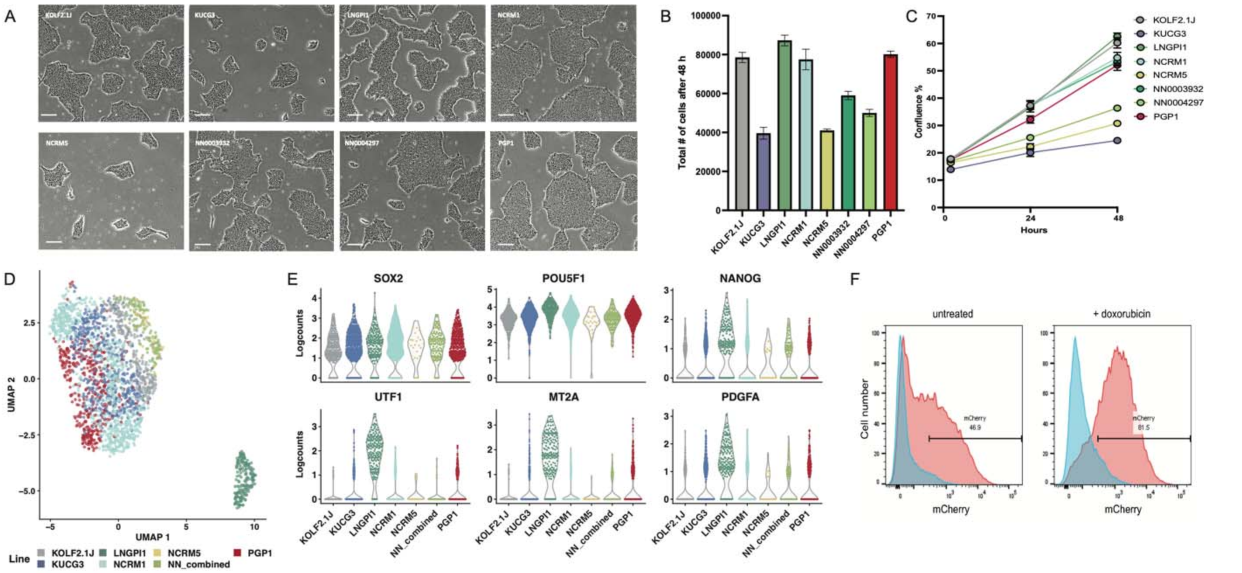


**Figure 1. Functional characterization of candidate cell lines. (A)** Representative images of colony morphology of the eight iPSC sub-lines on day 3 after replating. Scale bar indicates 100 µm. (**B**) Mean and SEM (n=4) of the total number of cells 48h after plating 30,000 cells/well. (**C**) Mean and SEM (n=6) of percent confluence at 0, 24 and 48 h after plating 30,000 cells/well. (**D**) UMAP projection of 2,270 iPSC cells color-coded by cell sub-line. (**E**) Beeswarm plots showing expression of selected genes associated with pluripotency (top row) or poor neuronal differentiation potential (bottom row). (**F**) A representative image of the TP53 reporter assay assessing p53 function in KOLF2.1J (red) or *TP53* knockout cells (blue). Basal expression of the p53-responsive fluorescent reporter increases in the presence of the DNA damaging agent doxorubicin. See also Figures S1-S3.

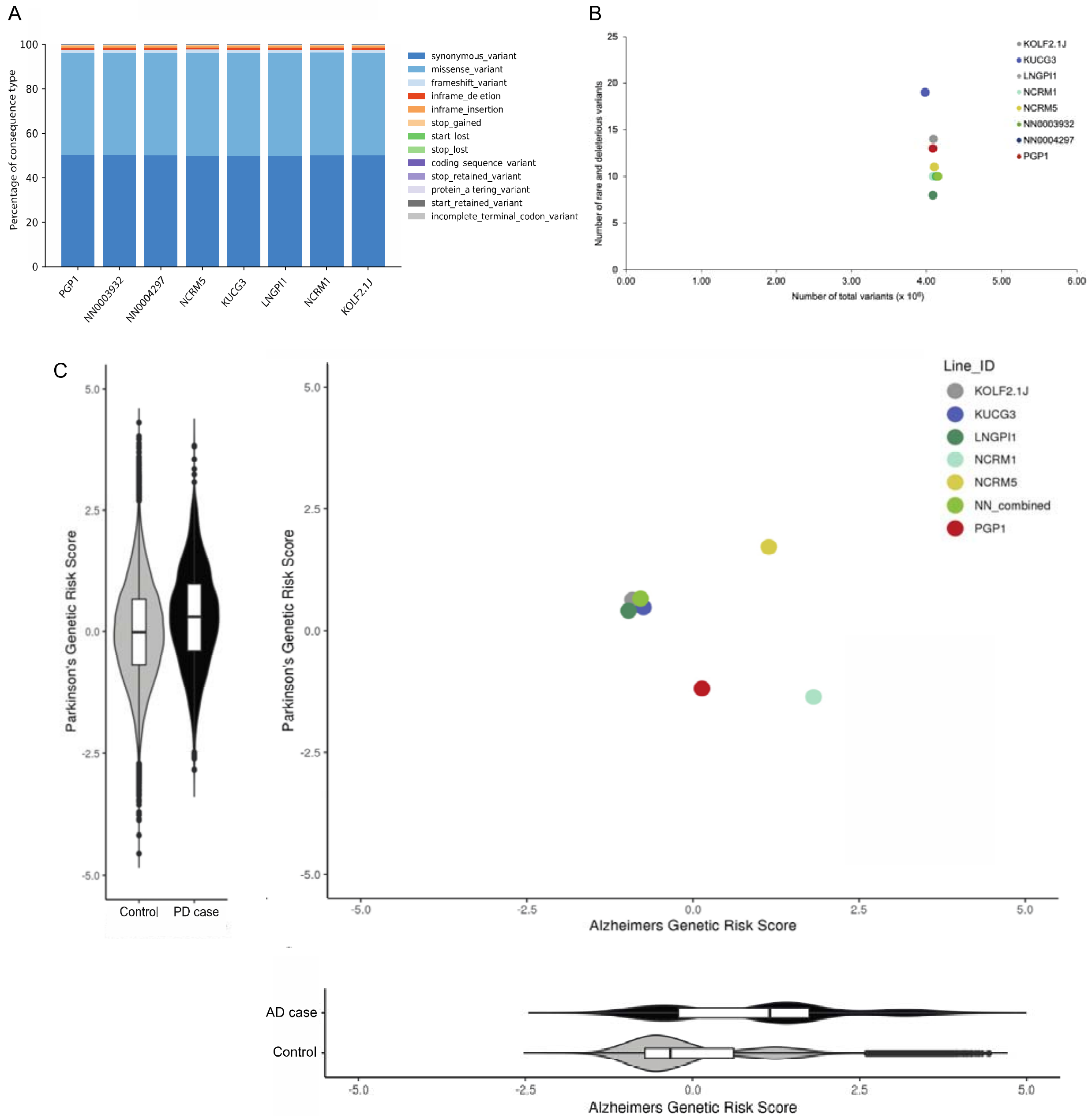


**Figure 2. Genetic analyses of eight candidate iPSC lines.** (**A**) The percentage of variant classes present in the eight candidate iPSC lines, grouped by their consequences on coding sequences. (**B**) The number of total variants and predicted rare deleterious variants identified in the 8 iPSC lines. (**C**) Polygenic risk scores for Alzheimer disease and Parkinson disease for the eight iPSC lines. The y-axis is the population-centered Z score distribution for the Parkinson's disease genetic risk score in 2995 Parkinson's disease cases (orange) and 96,215 controls (blue) from the UK Biobank. The x-axis is the population centered Z score distribution for the Alzheimer's disease genetic risk score in 2337 Alzheimer's disease cases (also orange) and the same controls.

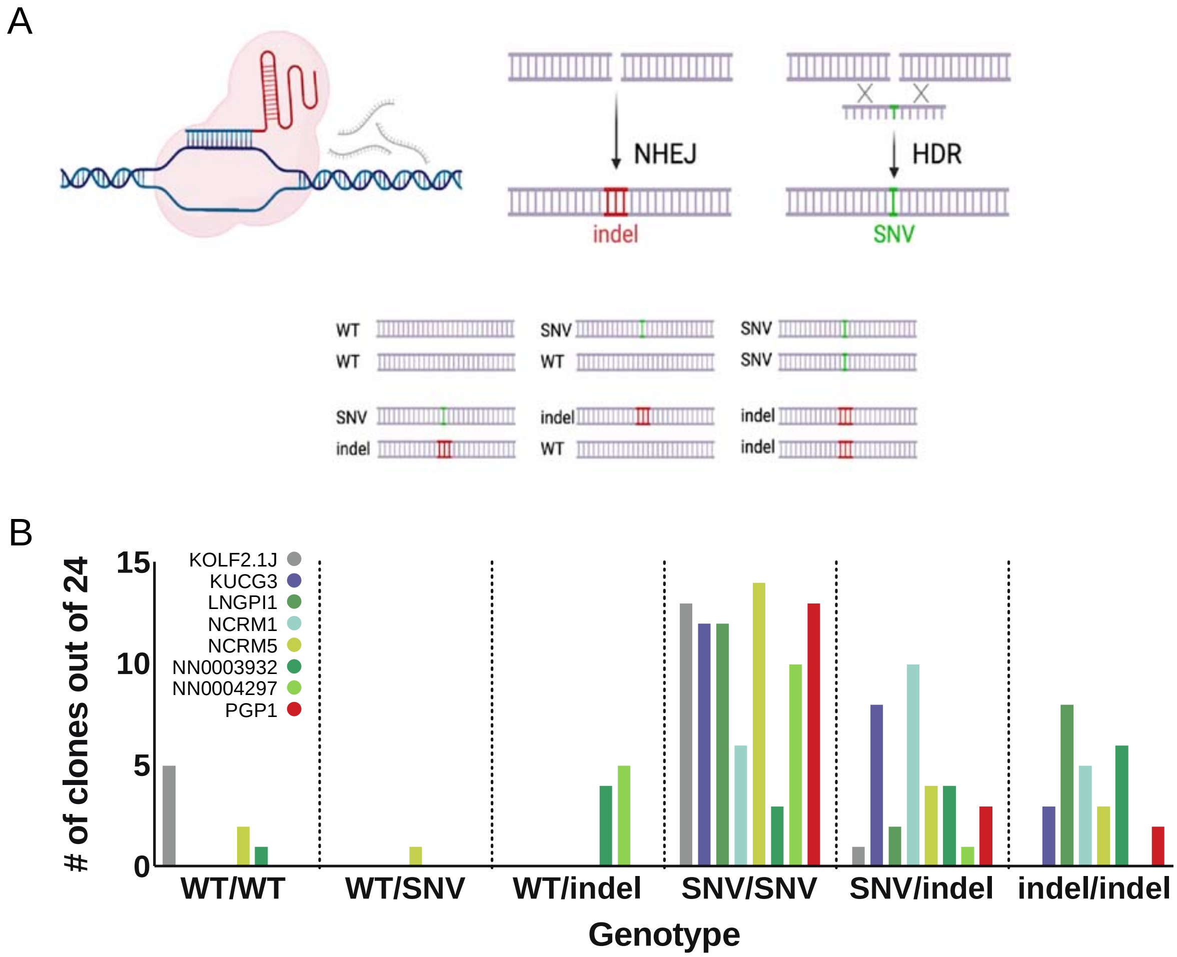


**Figure 3. Mock-editing experiment on the 8 candidate lines.** (**A**) Schematic of experiment, created with BioRender.com (**B**) The number of clones out of 24 expressing each possible genotype at the targeted SNV.

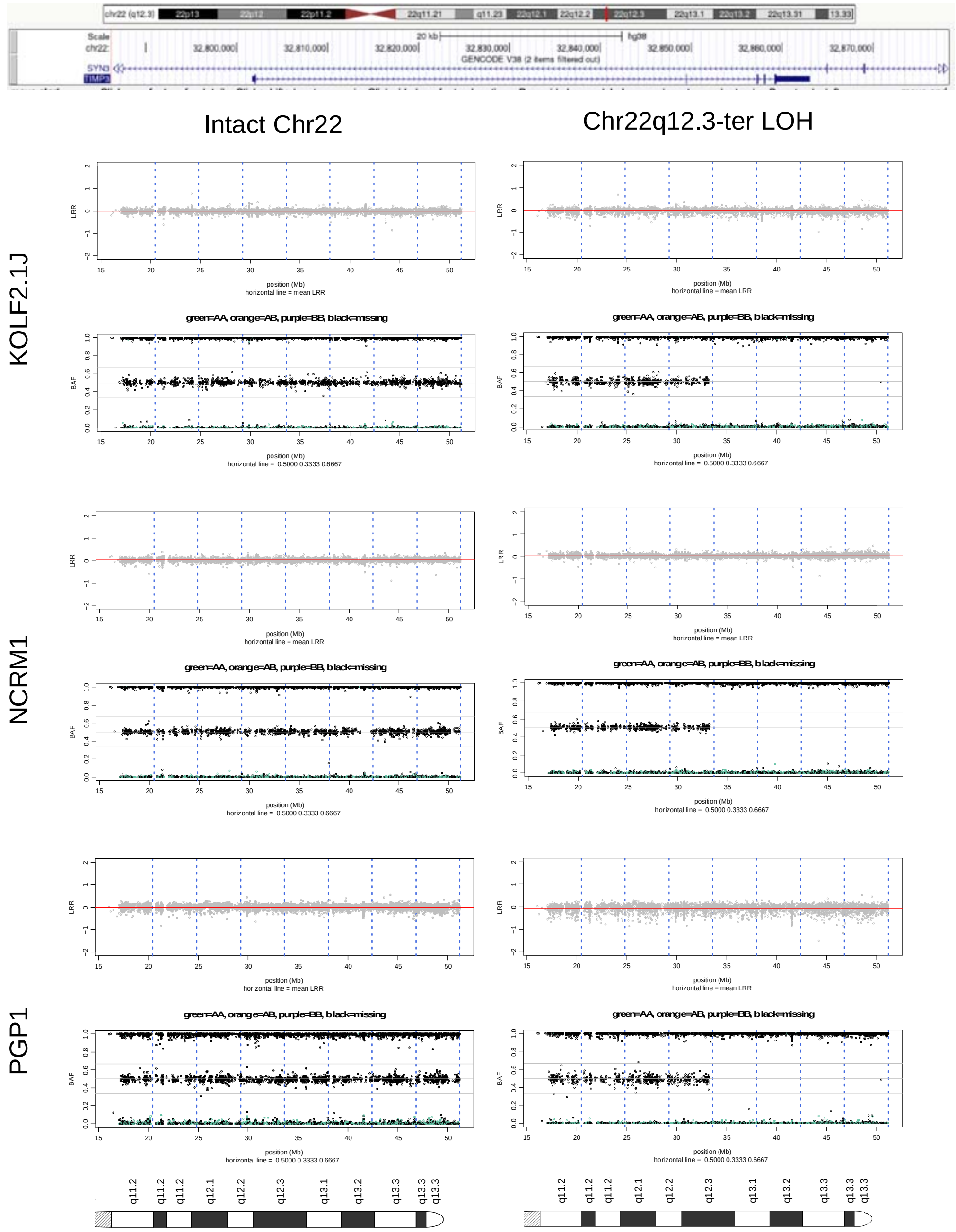


**Figure 4. (A)** Ideogram of chromosome 22 with the *TIMP3* gene indicated by a red bar, using the UCSC Genome Browser. **(B)** Chr22 loss of heterozygosity (LOH) in TIMP3-edited subclones of KOLF2.1J. Subclones of KOLF2.1 after TIMP3 editing were genotyped using NeuroArray. TIMP3 is located on 22q12.3. One homozygous subclone was found to have LOH from chr22q12.3-ter; a normal homozygous clone (KOLF2.1J-02) is shown for comparison. Upper plots show Log R ratio (LRR) for bead arrays where mean LRR=0 (red line) across the length of Chr22. Middle panels show B allele frequency for bi-allelic probes along the arrays showing loss of heterozygosity (LOH). Ideograms of chr22 are shown below each image for scale. **(C) and (D)** Chromosome 22 loss of heterozygosity (LOH) in *TIMP3* edited subclones from the **(C)** NCRM1 and **(D)** PGP1 lines. Three subclones from three different parental lines after TIMP3 editing were genotyped using NeuroArray and found to exhibit LOH from chr22q12.3-ter. Upper plots show Log R ratio (LRR) for bead arrays, where mean LRR=0 (red line) across the length of Chr22. Middle panels show B allele frequency for bi-allelic probes along the arrays showing loss of heterozygosity (LOH) from chr22.q12.3-ter. Ideograms of chr22 are shown below each image for scale. See also Figure S4.

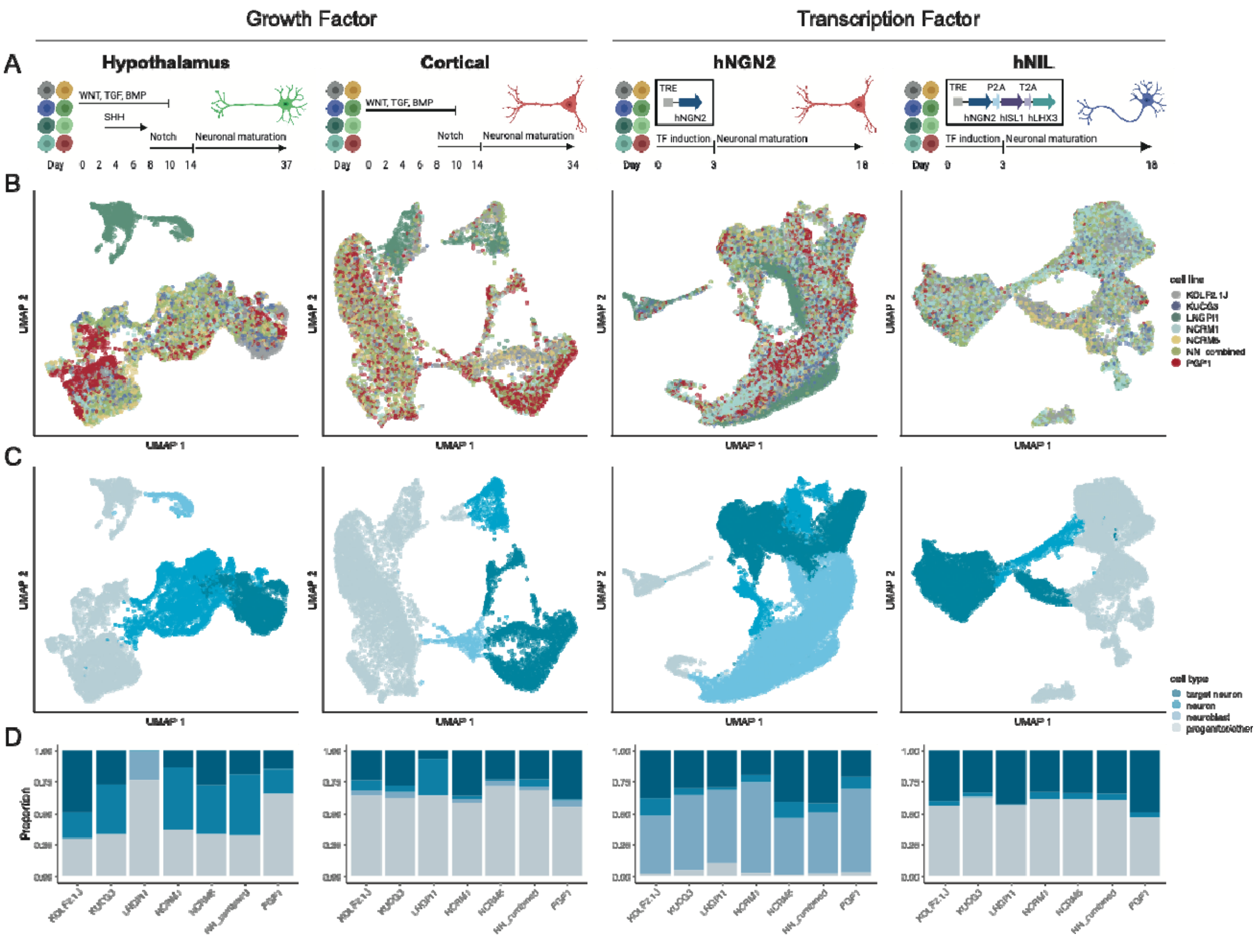


**Figure 5. Differentiation potential of candidate cell lines. A.** Schematic of experimental design for four differentiation protocols evaluated in this study: hypothalamus and cortical differentiation (growth factor-based protocols), and hNGN2 and hNIL differentiation (transcription factor-based protocols). **B and C.** UMAP plot for each differentiation colored by cell line (**B**) and cell type (**C**). Cell type classification is derived from grouping each cluster into target neuron, neuron, neuroblast and progenitor/other categories based on the cluster annotation (Figure S5). (**D)** Bar plot showing proportion of cells assigned to each cell type per cell line. See also Figures S5 and S6.

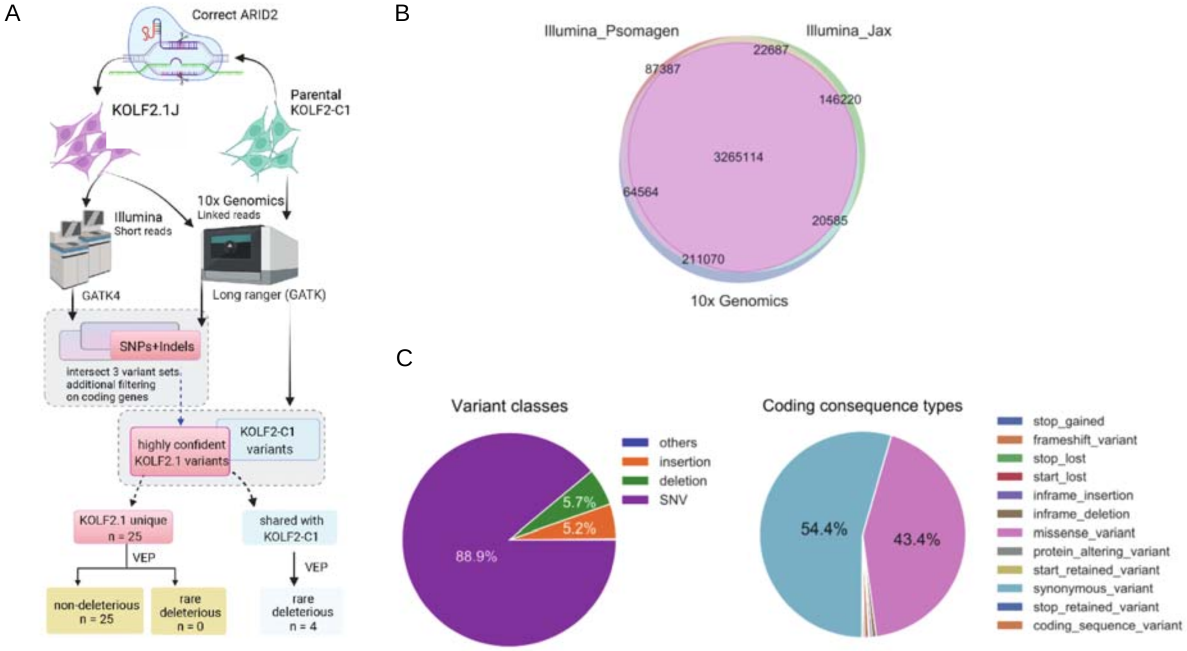


**Figure 6.** Whole genome sequencing of KOLF2.1J and KOLF2-C1 parental line. (**A**) Flowchart of the discovery, filtering, annotations, and comparisons of SNP/indel variants in the *ARID2*-corrected KOLF2.1J and its parental line KOLF2-C1. Schematic created with Biorender.com (**B**) The intersection of the three SNP/indel variant call datasets from two Illumina short sequencing services (Psomagen and Jax) and one long-read linked sequencing platform (10x Genomics). (**C**) The genetic compositions of the variant classes (pie plot on left) and their effect on coding genes (right) for the 3.28 million SNP/indels identified as high confidence.

-

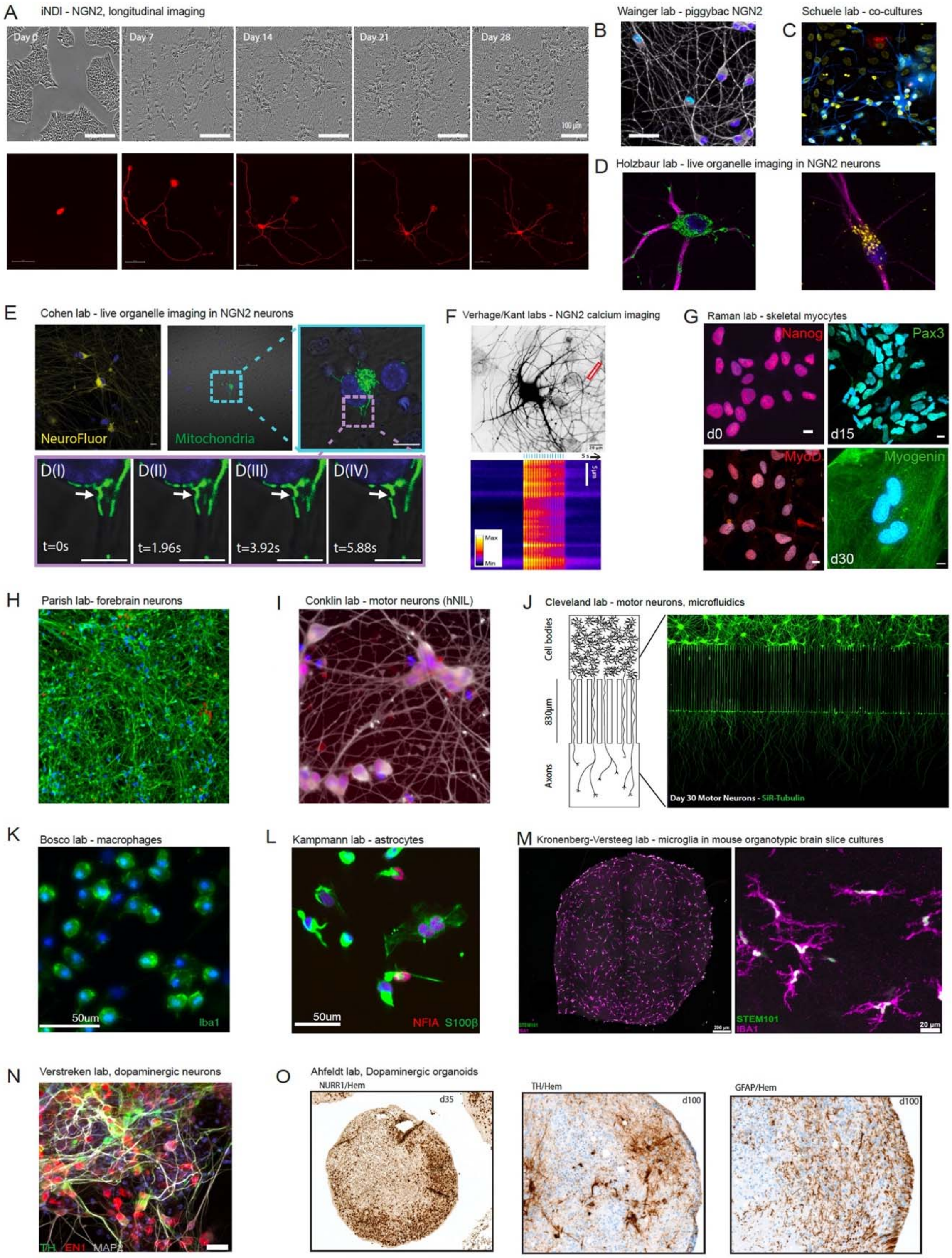

A
iNDI - NGN2, longitudinal imaging
Day 0
Day 7
Day 14
Day 21
Day 28
100 µm
B
Wainger lab - piggybac NGN2
C
Schuele lab - co-cultures
D
Holzbaur lab - live organelle imaging in NGN2 neurons
E
Cohen lab - live organelle imaging in NGN2 neurons
NeuroFluor
Mitochondria
D(I)
D(II)
D(III)
D(IV)
t=0s
t=1.96s
t=3.92s
t=5.88s
F
Verhage/Kant labs - NGN2 calcium imaging
5 s
5 µm
Max
Min
G
Raman lab - skeletal myocytes
Nanog
Pax3
d0
d15
MyoD
Myogenin
d30
H
Parish lab- forebrain neurons
I
Conklin lab - motor neurons (hNIL)
J
Cleveland lab - motor neurons, microfluidics
Cell bodies
830µm
Axons
Day 30 Motor Neurons - SiR-Tubulin
K
Bosco lab - macrophages
50um
Iba1
L
Kampmann lab - astrocytes
50um
NFIA
S100β
M
Kronenberg-Versteeg lab - microglia in mouse organotypic brain slice cultures
200 µm
STEM101
IBA1
20 µm
N
Verstreken lab, dopaminergic neurons
TH
EN1
MAP2
O
Ahfeldt lab, Dopaminergic organoids
NURR1/Hem
d35
TH/Hem
d100
GFAP/Hem
d100

**Figure 7. (A)** Differentiation of KOLF2.1J into NGN2-expressing cortical neurons. Bright field images of KOLF2.1J-hNGN2 throughout differentiation. Scale bar indicates 100 µm (top). Time-course of a neuron transduced with cytosolic mScarlet to identify neurites. Scale bar is 50 µm (bottom). **(B)** KOLF2.1 differentiated into NGN2-expressing cortical neurons and stained for Tuj1 (grey) and the cortical layer 2/3 markers Brn2 (green) or FoxG1 (red). Scale bar denotes 50 µm. **(C)** Mixed neuronal culture of cortical, striatal, and dopaminergic KOLF2.1J-derived neurons, co-cultured with primary human astrocytes. Cells immuno-positive for cTip2 (magenta), TBR2 (cyan), MAP2 (blue), and tyrosine hydroxylase (TH; red) were observed. DAPI stain is yellow. **(D)** Live organelle imaging of d21 KOLF2.1J differentiated into NGN2-expressing cortical neurons. Cells were stained for MAP2 (purple) and either expressed a mEmerald-mito plasmid (green, left) or were stained for LAMP1 (yellow, right) at d21. **(E)** Time-lapse images of KOLF2.1J cells differentiated into NGN2-expressing cortical neurons. Images were taken every 1.96 seconds. Scale Bar 5µm **(F)** KOLF2.1J neurons show robust calcium influx upon repetitive stimulation. Representative example of a KOLF2.1J neuron during calcium imaging (top). Kymograph of traces of intracellular calcium (Fluo5-AM) levels upon repetitive electrical stimulation (blue bars) in KOLF2.1J. (**G**) KOLF2.1J cells stained for the pluripotency marker Nanog (top left), then cells were differentiated into skeletal myocytes as in (Chal *et al.*, 2016). Pax3 staining at d15 indicates pre-myogenic progenitors (top right), myoD at d20 labels myoblasts (bottom left), and myogenin at d30 indicates myocytes (bottom right). Scale bar denotes 10 µm. **(H)** KOLF2.1J differentiated into cortical neurons using a dual SMAD protocol described in (Gantner *et al.*, 2021). By d55 of differentiation, cultures displayed a dense network of postmitotic TUJ1+ (green) neurons that were predominantly of deep cortical layers, as indicated by the expression of CTIP2 (blue) and TBR1 (red). Scale bar denotes 150 µm. **(I)** KOLF2.1J iPSCs were differentiated into motor neurons using an inducible transgenic system. Beta-III-tubulin (cell body/axons) and HB9 (nuclear) expression indicates cells that successfully differentiated by d10. Scale bar denotes 50 µm. **(J)** KOLF2.1J-derived motor neurons (d30) in microfluidic chambers live-imaged via SiR-tubulin (green), differentiated to motor neurons. **(K)** iPSC-derived macrophages at d7 of differentiation. Differentiated cells express the myeloid marker ionized calcium-binding adapter molecule 1 (IBA1) in KOLF2.1J-derived macrophages. Scale bar indicates 50 µm. **(L)** KOLF2.1J cells were differentiated into astrocytes and fixed at day 12 in astrocytic media. **(M)** KOLF2.1J iPSC-derived microglial precursors at d21 post-engraftment onto mouse organotypic brain slice cultures at low (left) or high (right) magnification. Scale bars denote 200 µm (left) or 20 µm (right). Cultures were immunostained for IBA1 (magenta) and STEM101 (green, against human nuclear protein Ku80). **(N)** KOLF2.1J differentiation towards ventral midbrain dopamine neurons **(O)** A d35 midbrain organoid section indicating NURR1-positive nuclei (brown). Scale bar 100µm (left). d100 midbrain organoid sections are depicted in center and right images. Dopaminergic (DA) neurons are indicated by tyrosine hydroxylase-positive staining (TH; center), and astrocytes are indicated by GFAP-positive staining (right) in brown. Scale bar 50 µm. Unless indicated in the subpanel, nuclei are labeled in blue. See also Figure S7.

## Supplemental Figures

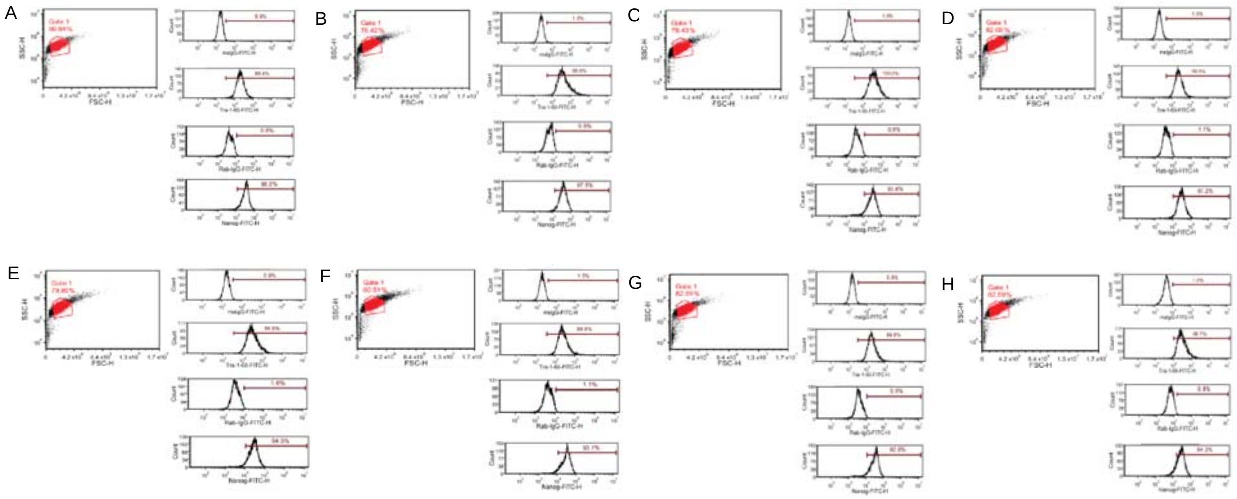


**Figure S1. Flow cytometry sorting of pluripotency markers in each of the candidate cell lines. A.** KOLF2.1J, **B.** KUCG3, **C.** LNGPI1, **D.** NCRM1, **E.** NCRM5, **F.** NN0003932, **G.** NN0004297, and **H.** PGP1. Related to Figure 1.

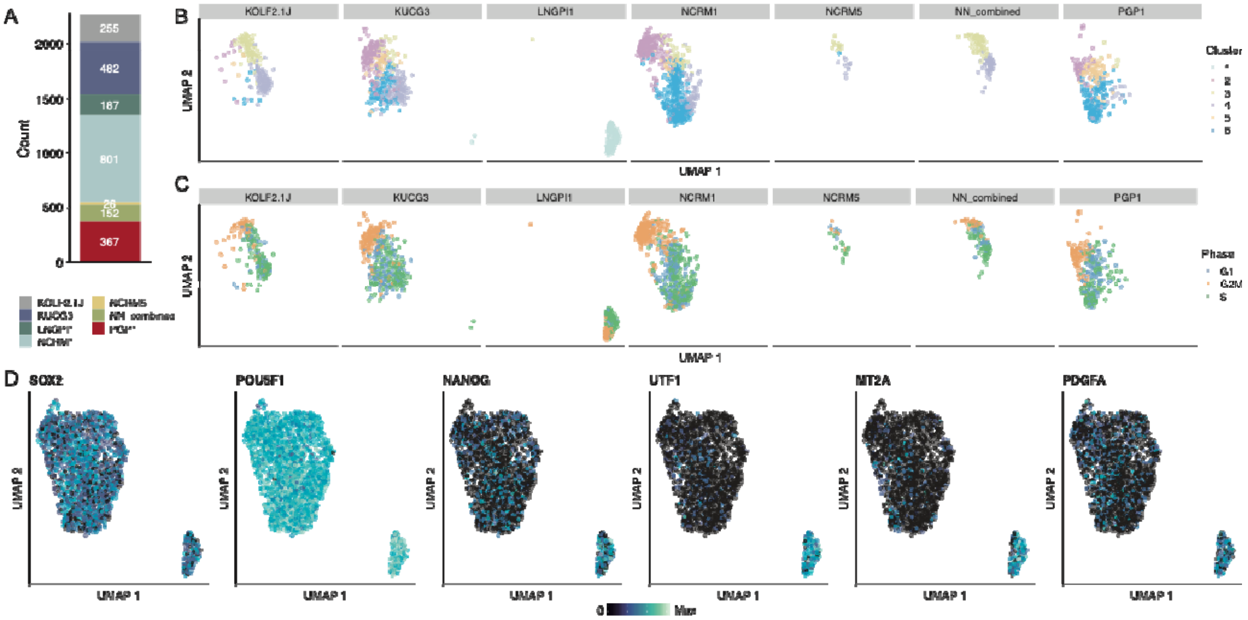


**Figure S2. Transcriptomic characterization of candidate lines. A.** Bar plot displaying the proportion of the cell line assayed. **B, C.** UMAP plots of all iPSC cells faceted by cell line and colored by clusters identified (**B**) and predicted cell cycle phase (**C**). **D.** UMAP plots of all iPSC cells coloured by expression of selected pluripotency markers. Related to Figure 1.

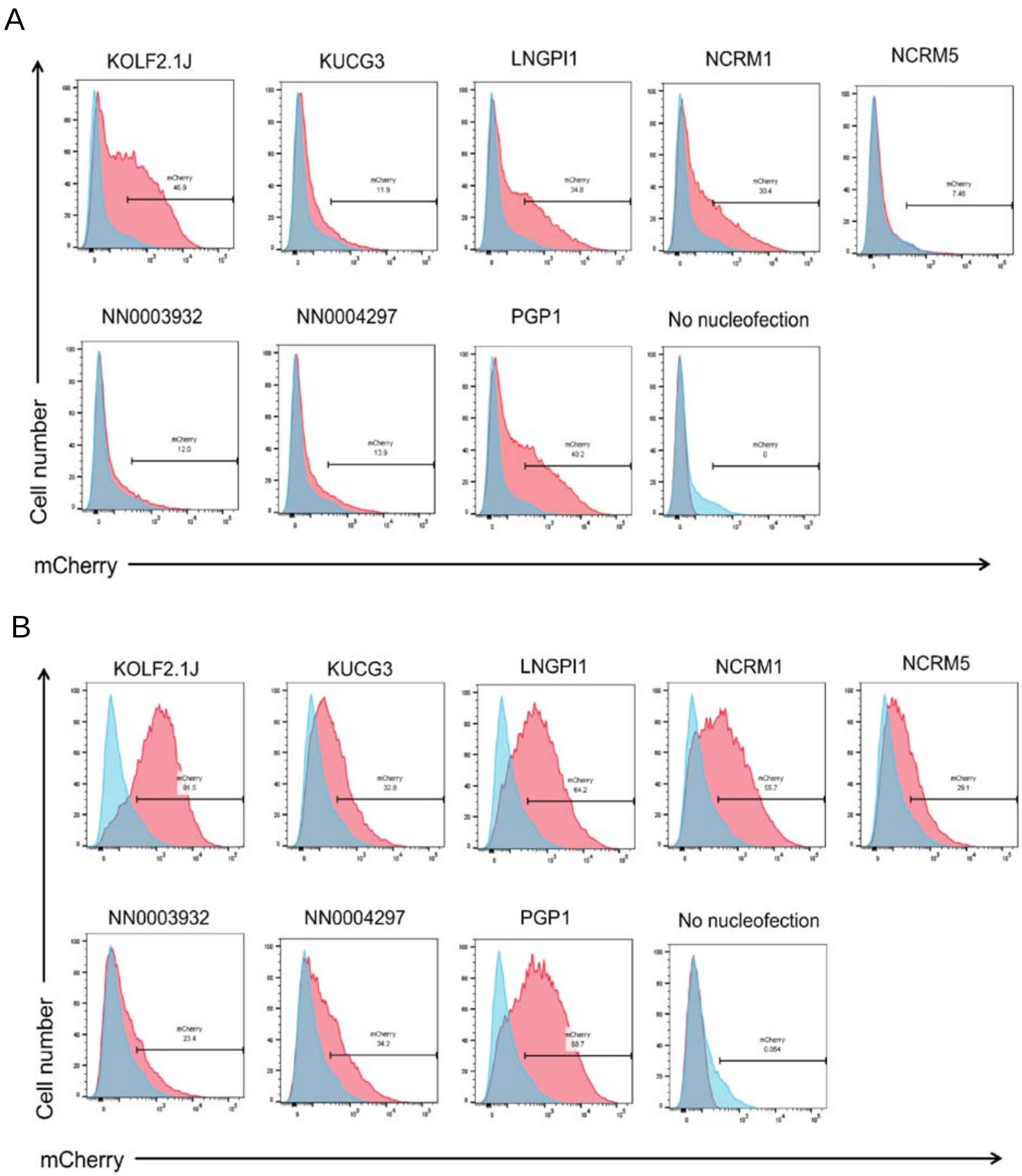


**Figure S3.** p53 reporter assay assessing p53 function in each candidate cell line (red) or *TP53* knockout cells (blue) in the presence of vehicle control (**A**) or the DNA damaging agent doxorubicin (**B**). Related to Figure 1.

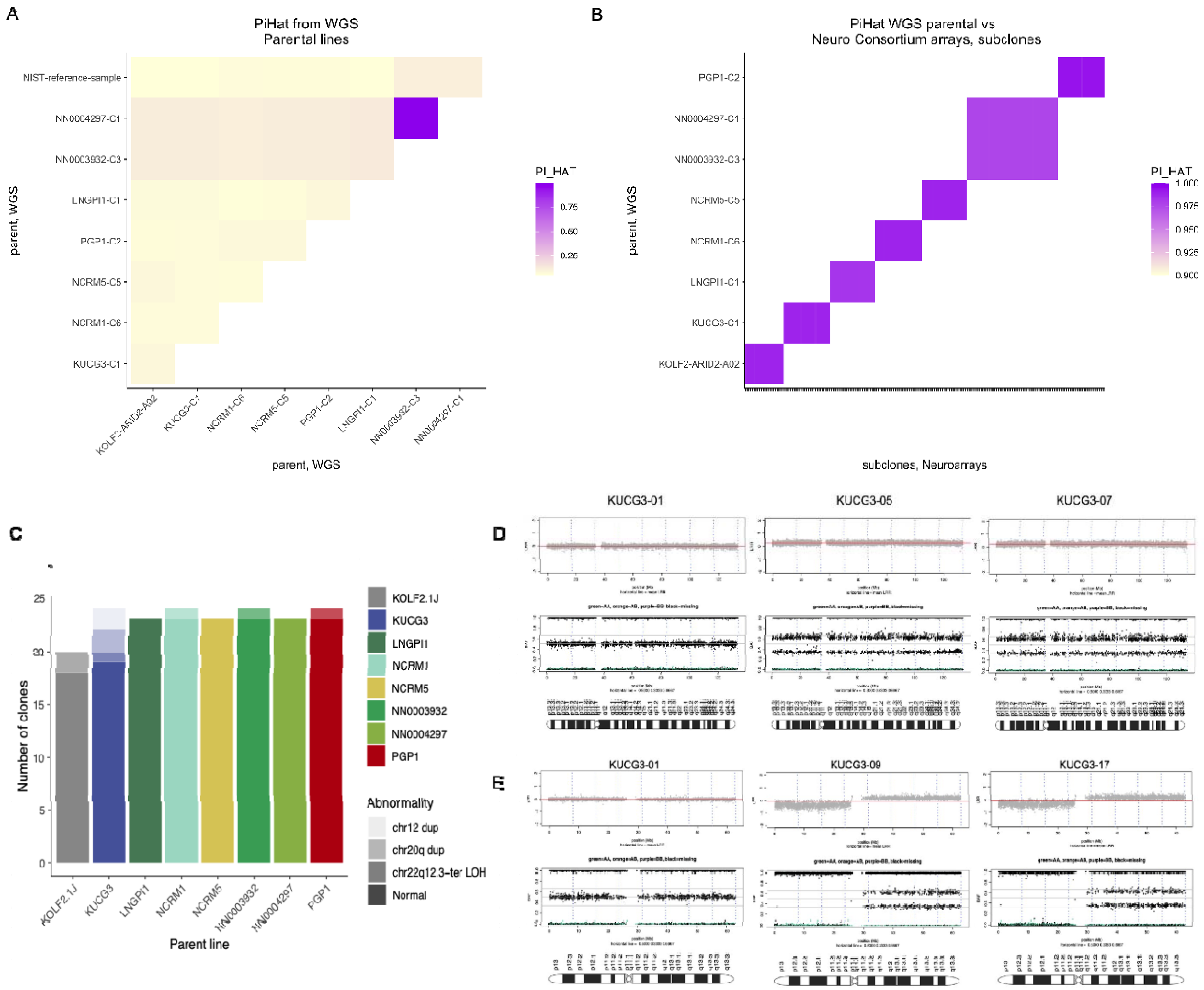


**Figure S4 Genomic fidelity of 8 candidate cell lines. (A)** Pairwise comparisons of pi-hat (color scale) between all parental lines from whole genome sequencing. Only the two lines from the same donor, effectively clonal to each other, show high pi-hat values while the remaining lines show pi-hat values similar to expected for unrelated samples of a given population. **(B)** Pairwise comparison of pi-hat (color scale) from WGS for all parental lines on the vertical axis against subclones from array genotyping on the y-axis. Note that all edited clones show high genomic fidelity to the parental line from which they were derived. **(C)** Bar graph of the number of chr22 abnormalities observed in the edited candidate lines. **(D)** Three subclones of KUCG3 after *TIMP3* editing were genotyped using NeuroArray. Two subclones (KUCG3-05 and KUCG3-07) were found to have duplication of chr12; a normal clone (KUCG3-01) is shown for comparison. **(E)** Chromosome 20q duplication in two subclones of KUCG3 (KUCG3-09 and KUCG3-17) after *TIMP3* editing; a normal clone (KUCG3-01) is shown for comparison. Upper plots show Log R ratio (LRR) for bead arrays where mean LRR=0 for the normal clone and LRR>1 for the abnormal clones (red line). Middle panels show B allele frequency for bi-allelic probes along the arrays with evidence of duplicated alleles across the chromosome. Ideograms of chr12 (D) or chr20 (E) are shown below each image for scale. Related to Figure 4.

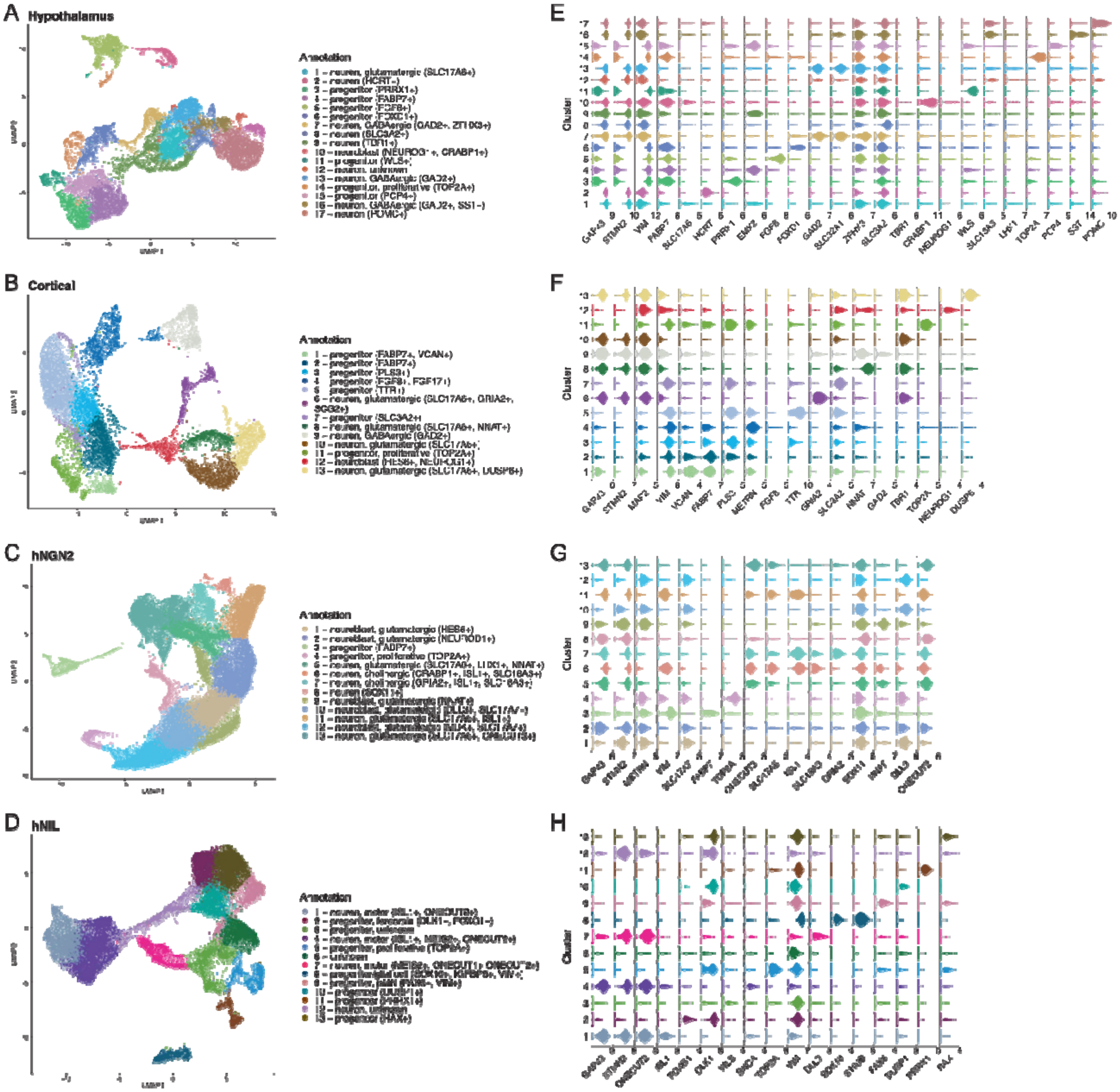


**Figure S5. Cluster identity and marker expression across differentiation protocols. A-D.** UMAP plots for each of the four differentiation protocols evaluated in this study – hypothalamus (**A**), cortical (**B**), hNGN2 (**C**) and hNIL (**D**) – colored by cluster identity. **E-H.** Beeswarm plots showing expression of informative differentiation-specific curated markers for each of the four differentiation protocols; hypothalamus (**E**), cortical (**F**), hNGN2 (**G**) and hNIL (**H**). Related to Figure 5.

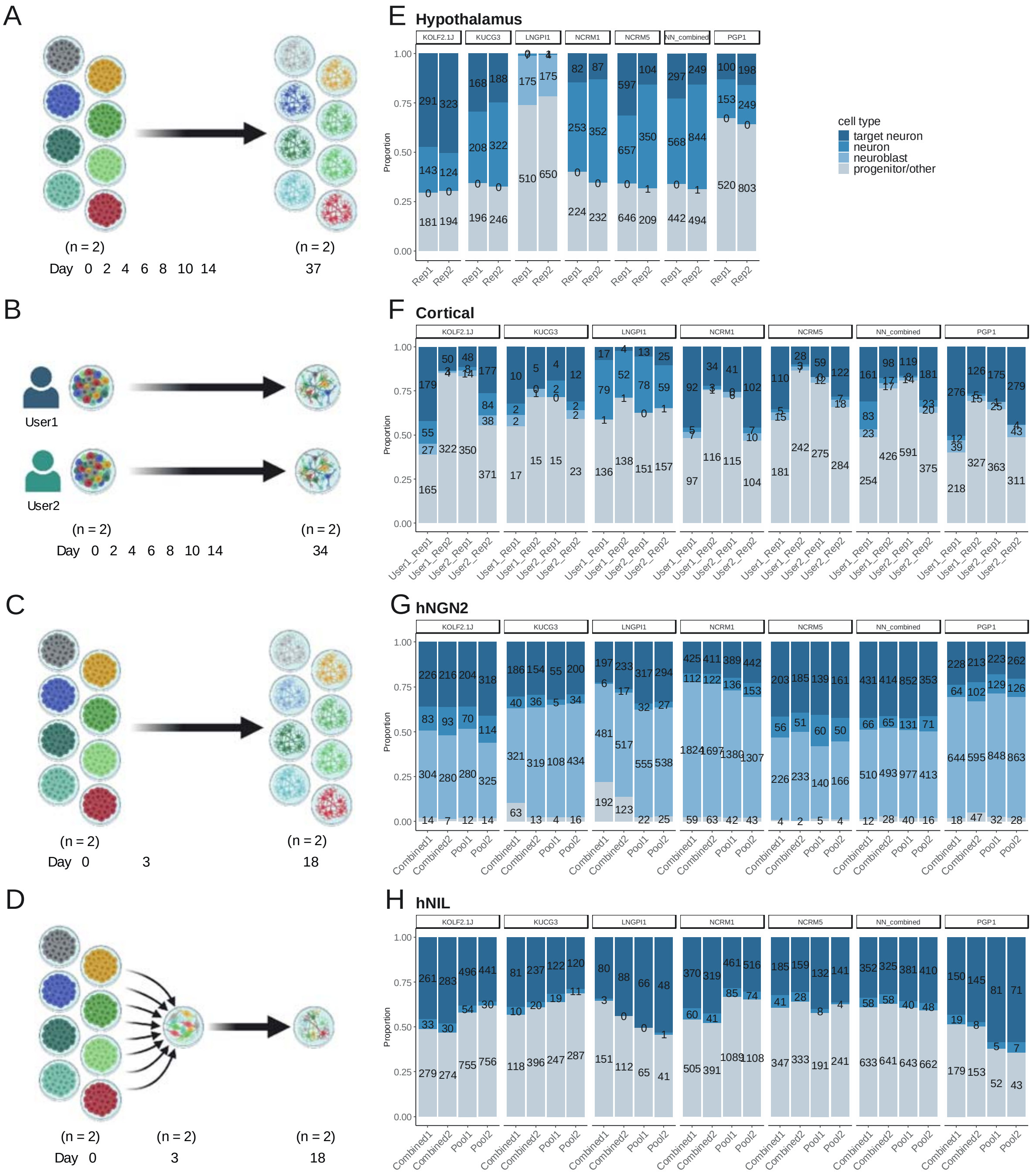


**Figure S6. Differentiation performance of candidate cell lines across replicates. A-D.** Schematic of experimental replicate structure for the four differentiation protocols evaluated in this study. For hypothalamic differentiation, the differentiations were performed individually for each cell line with each line having two replicates (**A**); while for cortical differentiation, the differentiations were performed with all the lines pooled together by two different users, with each user having two replicates (**B**). In the hNGN2 and hNIL differentiations, cells lines were differentiated both individually for each cell line (**C**, indicated as Combined in **G and H**) and with all lines pooled together after initial transcription factor induction at day 3 (**D**, indicated as Pool in **G and H**), with each differentiation method having two replicates. **E-H.** Bar plots showing proportion of cells assigned to each cell type per replicate, faceted by cell line, for each of the four differentiation protocols – hypothalamus (**E**), cortical (**F**), hNGN2 (**G**) and hNIL (**H**). Lefthand schematics were created with Biorender.com. Related to Figure 5.

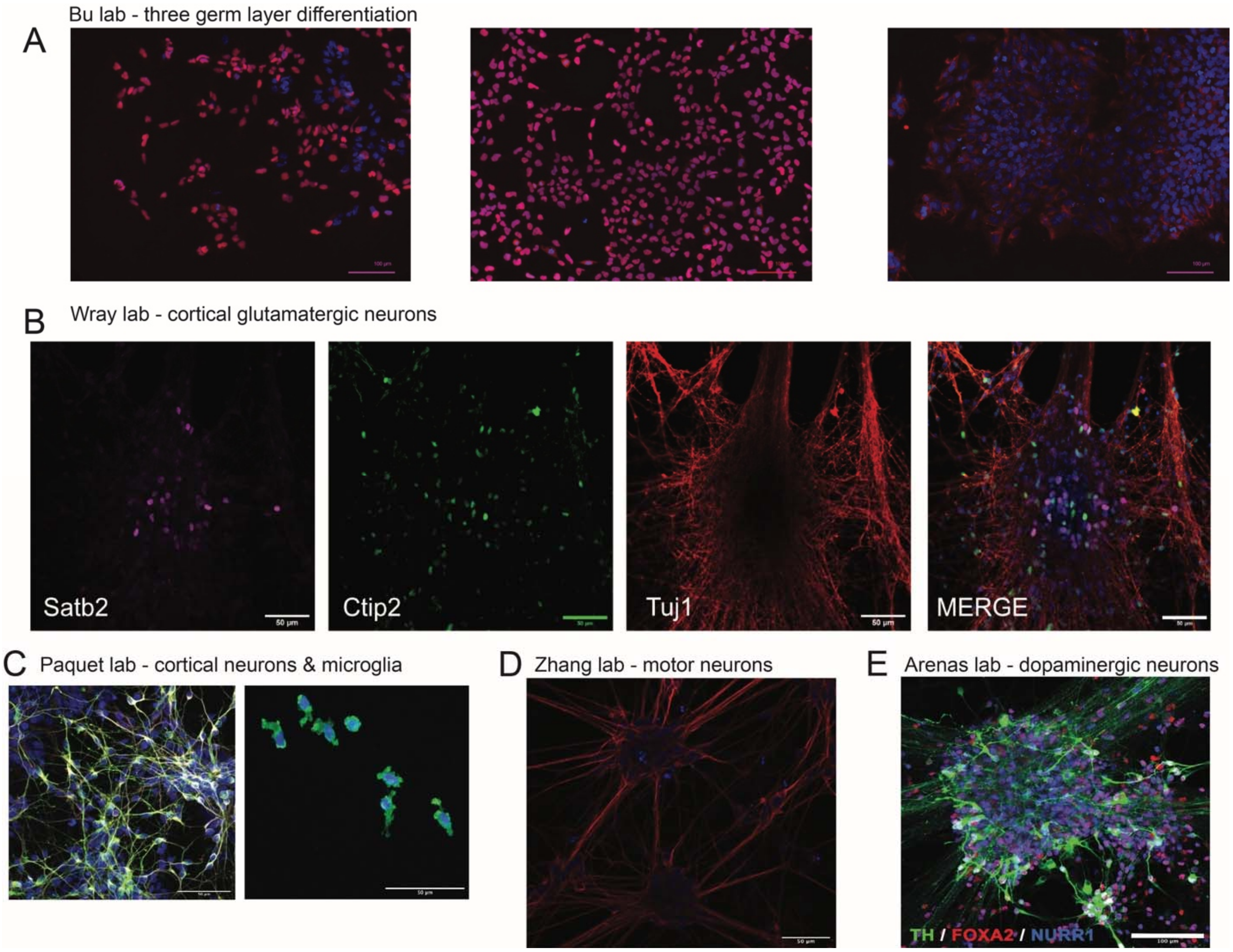

A
Bu lab - three germ layer differentiation
B
Wray lab - cortical glutamatergic neurons
Satb2
Ctip2
Tuj1
MERGE
50 µm
C
Paquet lab - cortical neurons & microglia
D
Zhang lab - motor neurons
E
Arenas lab - dopaminergic neurons
TH / FOXA2 / NURR1
100 µm

**Figure S7. Additional KOLF2.1J use cases.** (**A**) KOLF2.1J iPSCs differentiate into endoderm (SOX17-positive, left), mesoderm (brachyury-positive, center), and ectoderm (nestin-positive, right) layer cells. Scale bar indicates 100 µm. (**B**) KOLF2.1J cells differentiated into cortical glutamatergic neurons. At d100, KOLF2.1J cells were positive for the upper-layer cortical marker Satb2 (magenta) and deep-layer cortical marker Ctip2 (green). Neuronal identity was confirmed by immunostaining for Tuj1 (red). Scale bar is 50 µm. (**C**) KOLF2.1J neurons differentiated to cortical neurons and immunostained for Tau DA9 (green), synapsin (red), and MAP2 (gray; left), or to microglia and immunostained for Iba1 (green). Scale bar indicates 50 µm. (**D**) KOLF2.1J neurons differentiated to motor neurons and stained positively for SMI32 (red). Scale bar indicates 50 µm. (**E**) KOLF2.1J-derived midbrain dopaminergic neurons were triple-positive for TH (green), FOXA2 (red), and NURR1 (blue) at d28. Scale bar indicates 100 µm. Nuclear stains are in blue, unless denoted in the figure panel. Related to Figure 7.