Cell Genomics, Volume *3*

Supplemental information

Somatic mutations alter the differentiation outcomes of iPSC-derived neurons

Pau Puigdevall, Julie Jerber, Petr Danecek, Sergi Castellano, and Helena Kilpinen

Supplemental Figures

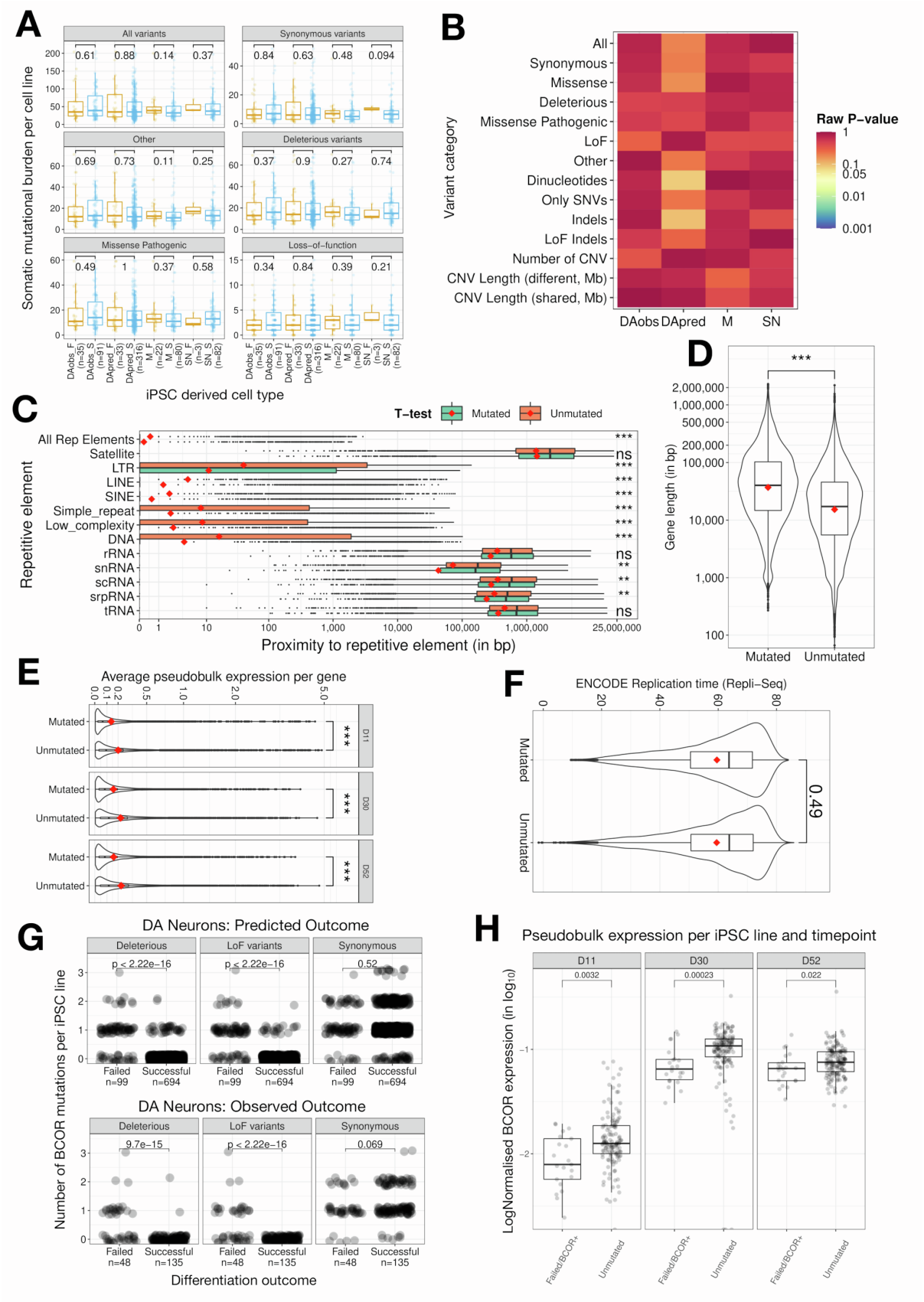

Figure S1. Total somatic burden did not alter the differentiation outcome, but LoF mutations in *BCOR* compromised neuron production and gene expression. Related to Figures 2A-2B.
(A) No significant mean differences on total number of mutations acquired *in vitro* (Wilcox.test, FDR<5%) were observed between the two differentiation outcomes (failed vs successful) of the three different iPSC-derived cell types (dopaminergic neurons - observed and predicted-, macrophages and sensory neurons). We also tested such differences of burden on six variant classes: total variants excluding CNVs; synonymous variants and others (coding, non-coding and unannotated), deleterious variants (union of LoF and missense pathogenic), as well missense pathogenic and loss-of-function variants alone. In none of the 24 tests, the difference reached the statistical significance threshold, even before multiple test-correction.
(B) No association was observed between the binary differentiation outcome (failed or successful groups) and the total number of mutations acquired *in vitro* for any of the tested variant classes. For each mutation category, we fitted a logistic regression and showed the corresponding raw p-values shown in the heatmap. None of the tests reached the statistical significance threshold even before multiple-test correction. We considered all mutations (without CNVs), synonymous, missense (pathogenic and non-pathogenic), deleterious and LoF only, other mutations (coding, non-coding and unannotated), dinucleotides, single nucleotide variants only, indels, indels predicted to be LoF, the number of CNVs, and the region length (in Mb) of shared and different CNVs.
(C) Genes with acquired somatic mutations in our dataset (n=9,235) showed a significantly closer proximity to repetitive elements (UCSC RepeatMasker [S2]) than unmutated genes (n=11,118), except for satellite repeats, ribosomal RNA and transfer RNA (T-test, pAdj<0.05).
(D) Mutated genes were longer on average than unmutated genes (T-test, $p<1.02 \cdot 10^{-149}$).
(E) Mutated genes showed reduced expression when compared with unmutated genes (T-test, p<0.05). Gene expression values were calculated from the average pseudobulk log-normalised values per line and time point.
(F) Genes with somatic acquired mutations did not show differences on replication timing when compared to unmutated genes (T-test, p=0.49). Previously, each gene was annotated with an average replication timing (Wavelet-smoothed signal) from the ENCODE Repli-seq BG02ES ESC line [S3]. The average was computed across all the signal values from the 1-Kb windows overlapping each corresponding gene.
(G) Number of *BCOR* mutations per iPSC line (either LoF, deleterious or synonymous variants) linked to the DA differentiation outcome (observed and predicted). A significant higher burden of damaging variants were observed across failed lines (Wilcox.test, p<0.05). None of the successful lines (N=135) in the DA observed outcome carried a *BCOR* LoF mutation, while 22 out of 48 failed lines have at least one mutation.
(H) Failed lines in DA differentiation with LoF mutations in the *BCOR* gene show significantly reduced *BCOR* expression when compared with unmutated lines, either failed or successful (Wilcox.test, p<0.05). Expression units correspond to pseudobulk log-normalised values per iPSC line and time point.

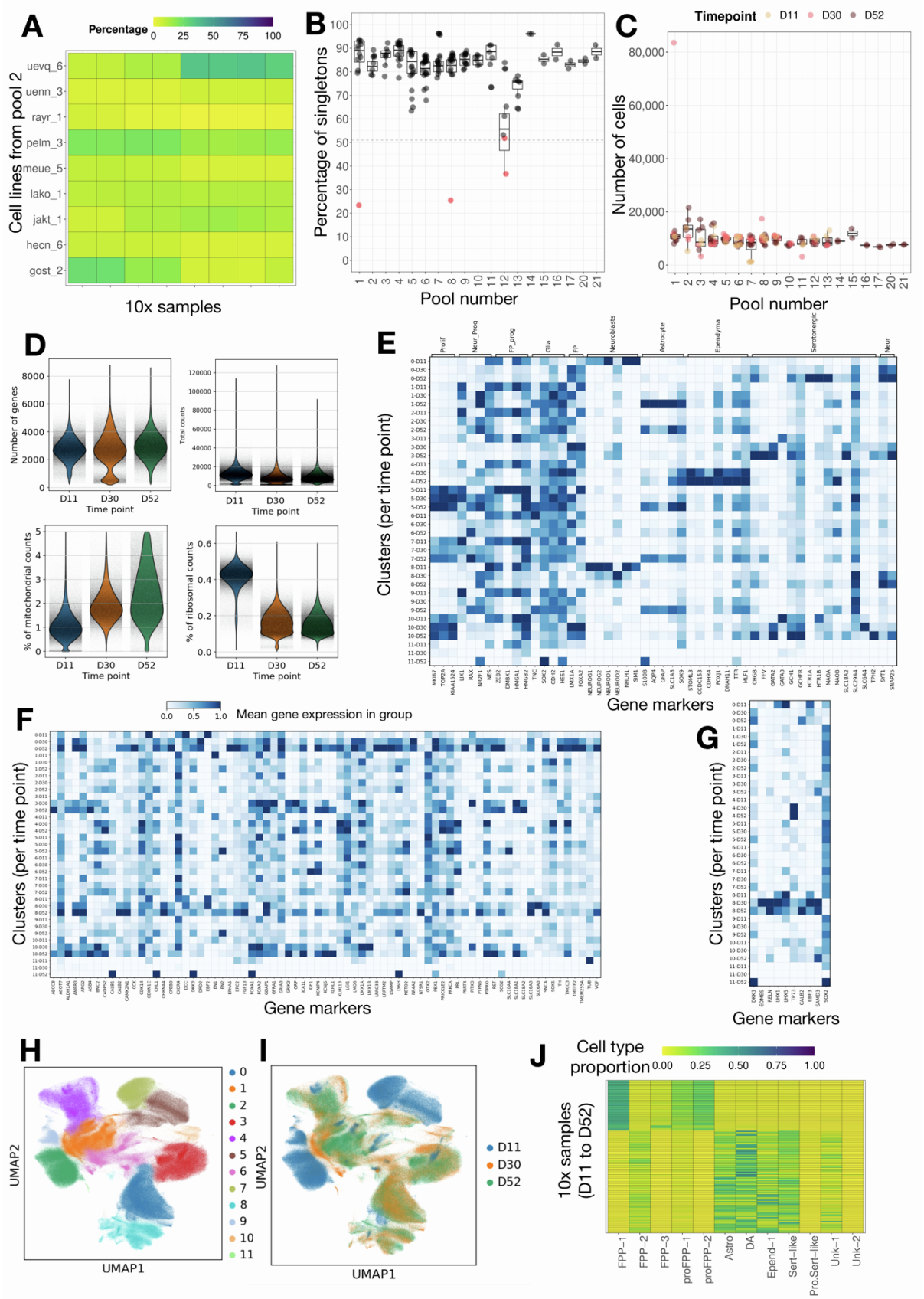

Figure S2. Quality control and annotation of the pooled DA differentiation dataset. Related to Figures 3A and 4A.

(A) Cell line proportion for the different 10x samples run for pool 2. Each heatmap row is a cell line and each column corresponds to a 10x sample. In certain samples, one-to-two cell lines consistently account for up to 50% of the cells.

(B) Percentage of singletons confidently assigned to a 10x sample (each dot) of day 52 grouped by pool. As a quality control step, those 10x samples with less than a 50% of singletons were removed from further analysis due to concerns on data integrity.
(C) Number of cell droplets estimated per 10x sample after CellRanger processing.
(D) Quality control metrics after the merging and the cell deconvolution of the 115 10x samples from the dopaminergic differentiation. The median number of genes per cell was below 3,000, while the total count of reads was around 10,000, as expected. Also, cells at day 52 showed a higher average of mitochondrial counts, while at day 11 the higher percentage corresponded to ribosomal counts.
(E-G), Time-specific cell markers expression from Julie et al [S1] used to annotate our Leiden clustering. General makers are shown in (E), specific dopaminergic markers in (F) and glial cells in (E). The intensity of blue shows the mean expression of the gene for each cluster-time point group.
(H) UMAP based on gene expression with cells coloured by the annotated cell types.
(I) UMAP on gene expression with cells coloured by sampling time point in differentiation (days 11, 30 and 52).
(J) Cell type proportion in each 10x sample highlights the observed variability on cell type composition throughout differentiation. Each row is a 10x sample ordered by time point, without considering the contribution of individual lines.

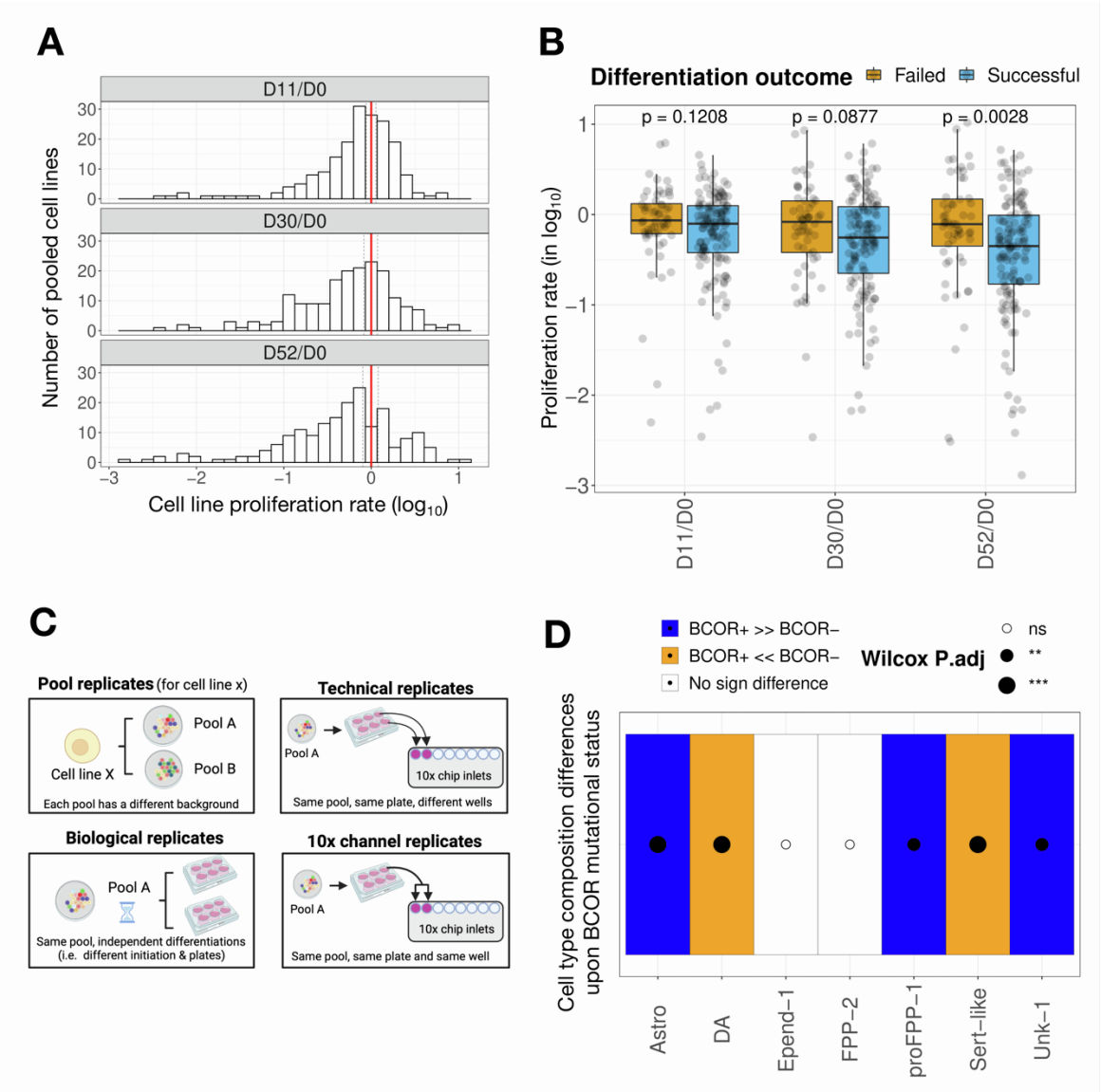

Figure S3. Differentiation failure is driven by increased proliferation rate and *BCOR* LoF mutations. Related to Figures 3B-3C and 4B.

(A) Distribution of the *in vitro* proliferation rate for pooled cell lines (expressed in $\log_{10}$) between each differentiation time point and day 0. The vertical red line indicates the mean cell line proportion ratio and the dashed lines the 95% confidence intervals.

(B) Failed cell lines showed on average a larger proliferation rate at day 52 than successful lines (Wilcoxon Rank Sum Test, $p=2.8 \cdot 10^{-3}$). Proliferation rates are expressed in $\log_{10}$ values.

(C) Illustrative diagram for the four types of replicates considered in DA dataset: pool replicates, referred to cell lines that are included in different pools; biological replicates, referred to lines from the same pool that are differentiated independently (in time and space); technical replicates, referred to lines from the same pool differentiated in the same plate but different wells; and 10x replicates, referred to lines from the same pool, plate and well.

(D) Cell lines carrying at least one deleterious *BCOR* mutation were associated with cell type composition changes at day 52 with a significant depletion of dopaminergic and serotonergic-like neurons (DA and Sert-like) and an excess of astrocytes (Astro), proliferative floor-plate progenitors type 1 (proFPP-1) and glial cells (Unk-1). The test for association was run for all protein-coding genes (Wilcoxon Rank Sum Test) and corrected for multiple-testing (Benjamini & Hochberg). The increasing size of the black-filled circle indicates higher significance levels for gene set enrichment: $p<0.01$ (**), $p<0.001$ (***); while the white-filled circles indicate non-significant association. Cell types with a significantly higher proportion in mutated lines compared to the unmutated ones are labelled in blue. In the opposite scenario, they are labelled in orange.

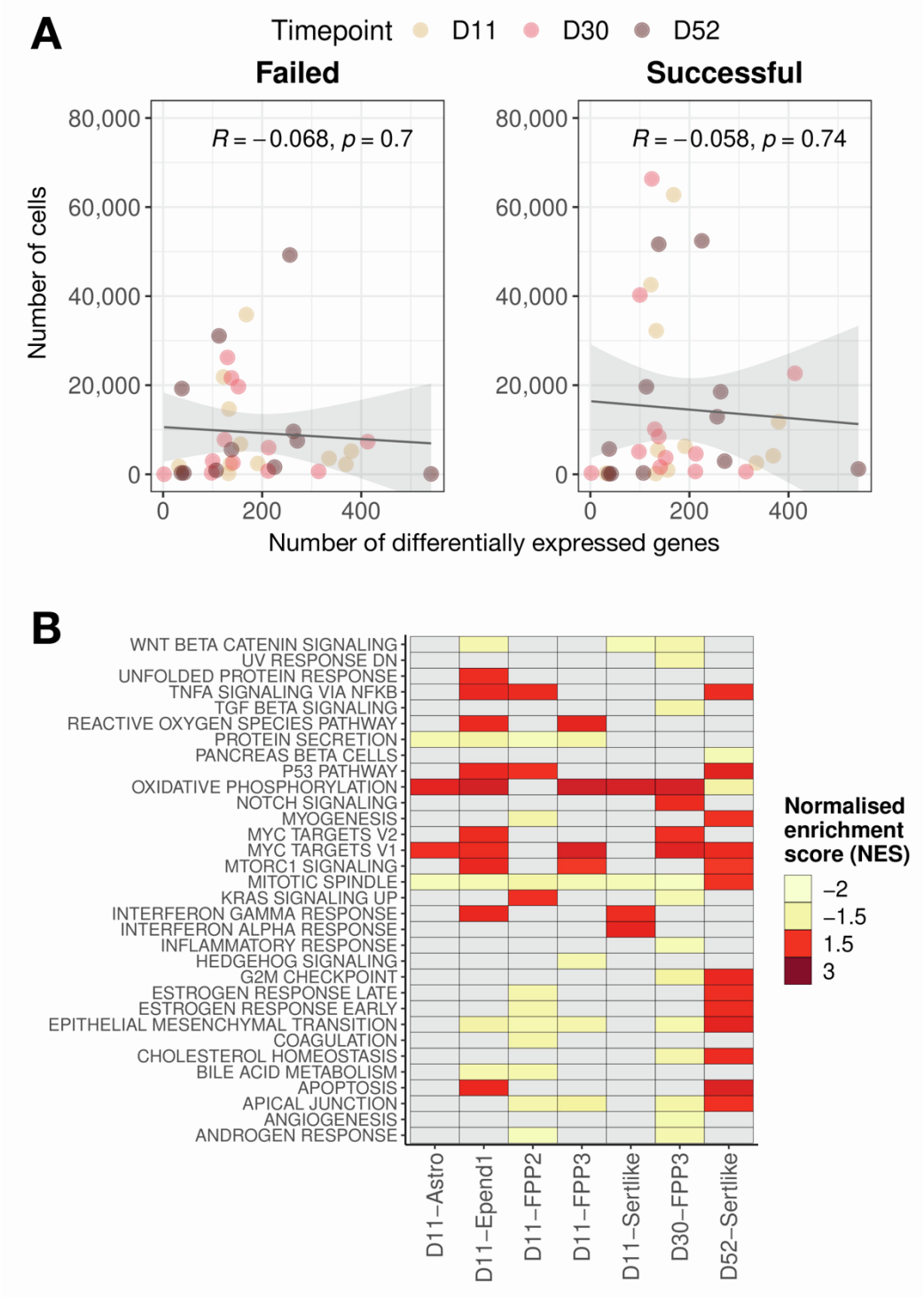

Figure S4. GSEA hallmark signatures for cell types with cancer-associated differential gene expression. Related to Figure 4C.

(A) Quality control: the number of cells per outcome and cell type did not correlate with the detected number of differentially expressed (DE) genes (p>0.05), suggesting that differential gene expression was not biased by the sample sizes. Each dot represents a cell type and is coloured according to the corresponding time point.

(B) Gene set enrichment analysis on MSigDB hallmark signatures [S4] (|NES|>1.5, adj.P<0.05) among those cell types with cancer-associated differential gene expression. Pathways with significant enrichment are coloured in red for upregulation, in yellow for downregulation and in grey when no significant enrichment is observed. Astrocytes at day 52 downregulated the oxidative phosphorylation and upregulated the mitotic spindle assembly, contrary to other cell types.

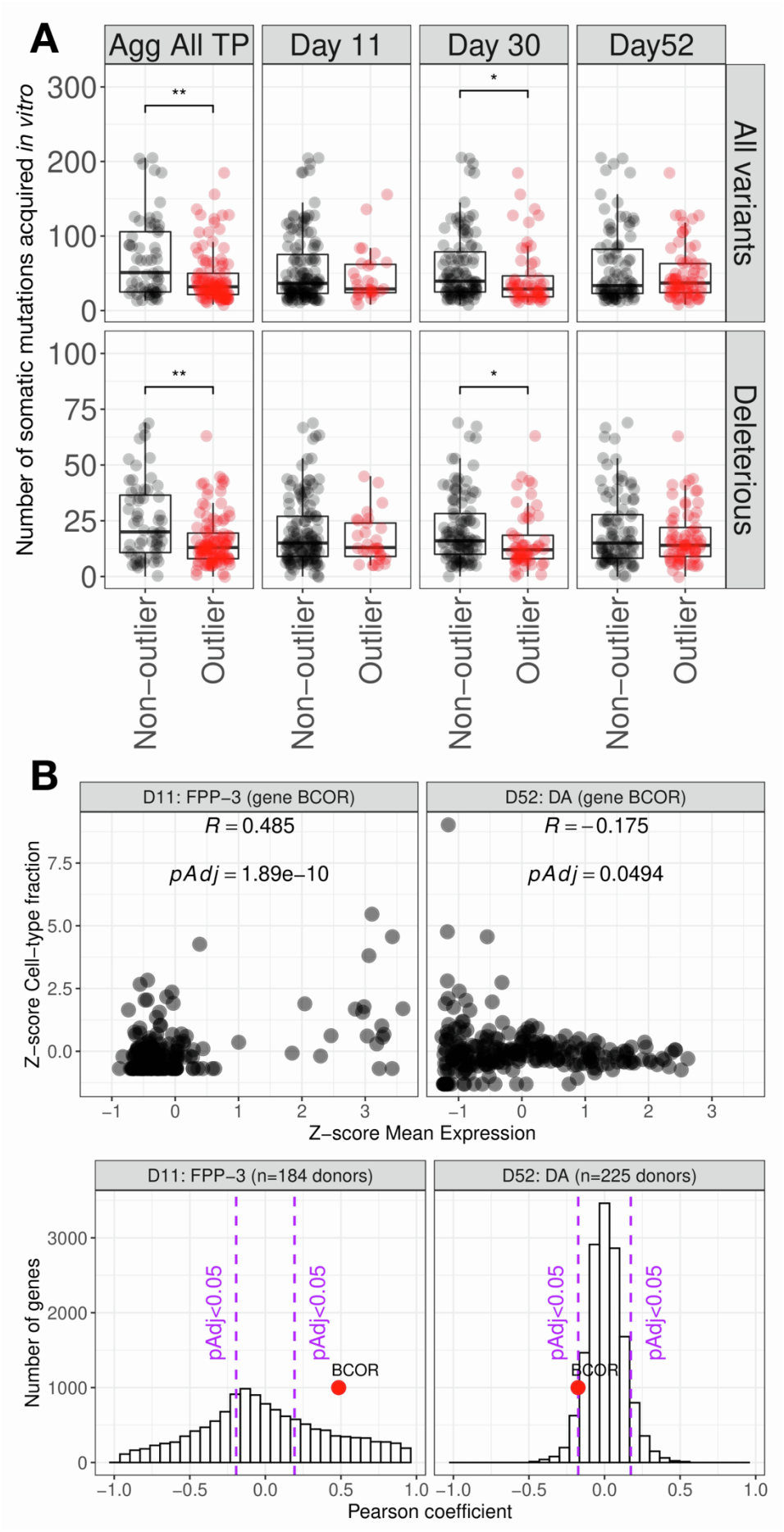
A
Agg All TP
Day 11
Day 30
Day52
All variants
Deleterious
Number of somatic mutations acquired in vitro
Non−outlier
Outlier
B
D11: FPP−3 (gene BCOR)
D52: DA (gene BCOR)
R = 0.485
pAdj = 1.89e−10
R = −0.175
pAdj = 0.0494
Z−score Cell−type fraction
Z−score Mean Expression
D11: FPP−3 (n=184 donors)
D52: DA (n=225 donors)
Number of genes
pAdj<0.05
BCOR
Pearson coefficient

Figure S5. iPSC lines with outlier cell type composition showed reduced somatic burden only at day 30. Related to Figures 5B and 5D.
(A) The differences in the burden of somatic mutations acquired *in vitro* between outliers and non-outliers of cell type composition were not consistent throughout the different stages of the differentiation (Wilcox.test). The definition of outlier lines was based on the observation of an abnormal cell type proportion during the differentiation or at a given time point. Only at day 30, we observed a reduced burden in the outlier group, which was also observed with the aggregated definition, both when accounting for total mutations or for deleterious mutations only.
(B) *upper*: z-score correlation of gene expression (i.e *BCOR*) and cell type proportions of floor-plate progenitors type 3 (FPP-3) at day 11 and dopaminergic neurons (DA) at day 52. Each dot is an iPSC line; *lower*: The distribution of Pearson correlation coefficients for all genes expressed in previous cell types and time points. Point highlighted in red indicates where *BCOR* is in that distribution.

Supplemental References

S1. 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.

S2. Jurka, J. (2000). Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. *16*, 418–420.

S3. Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M., Dorschner, M.O., Gartler, S.M., and Stamatoyannopoulos, J.A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc. Natl. Acad. Sci. U. S. A. *107*, 139–144.

S4. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst *1*, 417–425.