Cell Genomics Transparent Peer Review Record

Summary

Initial submission:	Received : 6/11/2022
	Scientific editor: Laura Zahn
First round of review:	Number of reviewers: 2 Revision invited : 8/4/2022 Revision received : 11/24/2022
Second round of review:	Number of reviewers: 2 Accepted : 2/23/2023
Data freely available:	Yes
Code freely available:	Yes

Referees' reports, first round of review

Reviewer #1: Comments enter in this field will be shared with the author; your identity will remain anonymous.This manuscript primarily reports a second order analysis of previously published data on the differentiation of multiple human induced pluripotent stem cell (hPSC) lines generated as part of the HipSci consortium. Studies that define the genotypic and phenotypic variation in human iPSCs are fundamental to defining many important aspects of human variation in health and disease. The vast scope of this emerging field is leading naturally to a series of very recent studies that extend our understanding of human developmental biology and the origins of disease risk. The focus of this study is to compare the behavior of the same set of multiple iPSC lines in differentiation protocols enriching for Dopamine (DA) neurons, sensory neurons, macrophages and definitive endoderm. This study includes exome sequence data on parental donor fibroblasts and their derived iPSC lines. Mutations were found in every cell line and half of these were predicted to be mis-sense or loss-of-function (LoF). However, the authors report no overall link between differentiation and mutational burden in all four fates examined.

There are many specific questions that could be followed here but the next focus in this report is the correlation between LoF mutation in a specific gene, BCOR, and failure to differentiate into DA neurons. This fate is of great interest as clinical trials are currently underway to test the efficacy of in vitro derived DA neurons in Parkinsonian patients. The authors stress that BCOR mutations associate with both the binary classifier first employed (failed/success in differentiation) and a more temporally restricted measure at day 52 of DA neuron differentiation. This result gains interest because it occurred against a background where there was great variability in clonal expansion in different experiments.

BCOR interacts with BCL6 to modulate polycomb epigenetic regulation and mutation had been reported in cancer and stem cell systems. The current report goes onto make clearly valuable contributions to the cell types and molecular mechanisms at work here. From this work, it seems as if the BCOR/BCL6 system is part of a generalized differentiation mechanism but I understand the main point in this manuscript is to emphasize the value in further higher resolution genotype/phenotype studies using large numbers of human iPSCs.

There is no question that as iPSCs differentiate they follow constrained paths leading to specific fates. This is now well established. The current study stresses that measuring the variability in differentiation outcome is also worthwhile. It is technically important that the field pays detailed attention to this variability and this is the primary justification given for the general interest of this report.

I want to encourage the authors to place more stress on the conceptual importance of their work understanding cell line. Perhaps the easiest way to do this is to point out the increasing use of lineage based developmental analyses in the study of psychiatric disorders. This manuscript refers to these studies in passing but these and other recent studies for example on the origins of Tuberous Sclerosis show that we are gaining a systematic understanding of the clinical implications of the developmental genetic mechanisms that construct the human central and peripheral nervous systems.

It will be very interesting to understand better compare these different developmental risks. For example, the clinical impact of de novo CNVs is central to psychiatric genetic research. Here CNVs are mentioned as a minor source of genetic variation but it would be helpful to know specifically how the CNV search algorithms used here compare with others. More generally, I wonder if analytical tools that are already available might reveal mutational burden reported by others in ASD cell lines in vitro.

It would be helpful to have the authors views on these topics but I should stress in closing that this work is presented in a thoughtful style. The authors address important issues in a straightforward style. I find the story a compelling addition to our knowledge of human iPSC biology.

Reviewer #2: In their manuscript entitled "Somatic mutations alter the differentiation outcomes of iPSC-derived neurons" by Costa et al, the authors meta-analyze existing single-cell RNA and whole-exome sequencing for a pooled dopaminergic neuron differentiation experiment of 238 iPSC lines. They observed

deleterious somatic mutations in key developmental genes, notably the BCOR gene, associated with perturbed gene expression, increased proliferation ,and reduced dopaminergic neuron differentiation. Overall, this is an interesting finding that cautions about interpreting differentiation phenotypes in iPSC disease models. Nonetheless, more needs to be done to confirm the causality of these associations and the mechanisms driving somatic mosaicism during neural differentiation. With major revisions, this report might be suitable for Cell Genomics.

Major concerns:

1. Exome-wide burden of acquired mutations does not explain the differentiation outcome: The authors analysed "differentiation outcomes from four independent experiments that produced different target cell types from iPSC lines of the HipSci project (Kilpinen et al., 2017): dopaminergic neurons, DN (Jerber et al., 2021), macrophages (Alasoo et al., 2018), sensory neurons (Schwartzentruber et al., 2018) and definitive endoderm tissue (Cuomo et al., 2020) (Fig. 1A). [They] compared the exome-wide burden of acquired mutations and rare germline variants to the differentiation outcomes of each line (Fig. 1B), and used single-cell transcriptomes of the dopaminergic neurons (Jerber et al., 2021) to delineate in detail how deleterious variants influence the differentiation process of iPSC-derived neurons."
a. Inter- and intra-donor variation: The authors should clarify how many donors, iPSC clones, and independent experiments were included in each analysis. Were some donors particularly prone to acquired deleterious mutations, or did they occur at random in all individuals?

2. Causality and Phenotypic effects: The authors "found that only one gene, BCOR, was significantly more mutated in failed lines compared to the successful lines… [Amongst failed line that] do not carry any deleterious BCOR mutation… The most mutated genes (log2FC>1.5) contributing to the enrichment of those biological processes in failed lines are well-known disease associations: NSUN5 in William Beuren Syndrome) or ASCL1 in neuropsychiatric disorders… or key regulators of neuron differentiation like NRBP2 or SOX14…" Moreover, the authors "correlated the cell type-specific expression [of mutated genes in the abnormal cell type fraction] with the changes in cell type composition, but did not empirically test the impact of mutated genes on gene expression.
a. Dopaminergic differentiation: were individual mutations (e.g. in BCOR) confirmed by sequence changes and/or expression levels reductions by scRNAseq?
b. Given the pooled design used for the DA neurons, the authors should consider the possible impact of non-cell-autonomous effects on their analysis.
c. If would be extremely informative if the authors knocked out/down BCOR and/or NSUN5, ASCL1, NRBP2, or SOX14 in one or more donor backgrounds and confirmed changes in DA patterning, maturation, function.
d. Given that many of these genes, notably BCOR, are robustly expressed in astrocytes, it would be interesting to access what cell types the failed DA neurons are becoming.
e. Because exome data was only available for source fibroblasts and donor hiPSCs, the experimental design fails to capture the impact of somatic mosaicism that occurs during neuronal differentiation. Could exome sequencing of dopaminergic neurons from a subset of iPSC lines be added?

3. Mechanisms and biological relevance: it would be interesting if the authors uncovered the genetic or epigenetic factors driving somatic mosaicism in vitro. Previous reports of somatic mosaicism have shown that it can be drive by L1 (or other) repetitive elements (Coufal et al Nature 2009), high transcription levels (Lodato et al, Science 2015), large gene size and late-replication(Wei Schwer Cell 2016), and/or DNA repair (Reid et al Science 2021)
a. Are the genes with somatic mutations correlate to features such as proximity to repetive elements, expression levels, length, or cell replication, transcribed, and late-replicating genes? Moreover, it would be useful to compare conduct enrichments of mutated genes not just to gene ontology, but to published lists of somatic mutations during neural differentiation (e.g. almost 90% of RDC-containing genes identified are involved in synapse function and/or neural cell adhesion, with a substantial fraction also implicated in tumor suppression and/or mental disorders. (Wei Schwer Cell 2016))
b. The authors state that "Among the significant associations, including positively and negatively correlated genes, we observed a strong enrichment of DDD genes in most of the cell types at the progenitor stage, at day 11 …" They should test association across other psychiatric, neurodegenerative, and non-brain disorders.

Authors' response to the first round of review

Comments,
Reviewer #1
This manuscript primarily reports a second order analysis of previously published data on the differentiation of multiple human induced pluripotent stem cell (hPSC) lines generated as part of the HipSci consortium. Studies that define the genotypic and phenotypic variation in human iPSCs are fundamental to defining many important aspects of human variation in health and disease. The vast scope of this emerging field is leading naturally to a series of very recent studies that extend our understanding of human developmental biology and the origins of disease risk. The focus of this study is to compare the behavior of the same set of multiple iPSC lines in differentiation protocols enriching for Dopamine (DA) neurons, sensory neurons, macrophages and definitive endoderm. This study includes exome sequence data on parental donor fibroblasts and their derived iPSC lines. Mutations were found in every cell line and half of these were predicted to be mis-sense or loss-of-function (LoF). However, the authors report no overall link between differentiation and mutational burden in all four fates examined.

There are many specific questions that could be followed here but the next focus in this report is the correlation between LoF mutation in a specific gene, BCOR, and failure to differentiate into DA neurons. This fate is of great interest as clinical trials are currently underway to test the efficacy of in vitro derived DA neurons in Parkinsonian patients. The authors stress that BCOR mutations associate with both the binary classifier first employed (failed/success in differentiation) and a more temporally restricted measure at day 52 of DA neuron differentiation. This result gains interest because it occurred against a background where there was great variability in clonal expansion in different experiments.

BCOR interacts with BCL6 to modulate polycomb epigenetic regulation and mutation had been reported in cancer and stem cell systems. The current report goes onto make clearly valuable contributions to the cell types and molecular mechanisms at work here. From this work, it seems as if the BCOR/BCL6 system is part of a generalized differentiation mechanism but I understand the main point in this manuscript is to emphasize the value in further higher resolution genotype/phenotype studies using large numbers of human iPSCs.

There is no question that as iPSCs differentiate they follow constrained paths leading to specific fates. This is now well established. The current study stresses that measuring the variability in differentiation outcome is also worthwhile. It is technically important that the field pays detailed attention to this variability and this is the primary justification given for the general interest of this report.

We thank the reviewer for the comprehensive review of the text and for acknowledging our effort to better understand how somatic mutations influence iPSC models of differentiation. We agree that the primary justification for carrying out this work is to raise awareness of variation in differentiation, its association with acquired deleterious mutations, and the implications to iPSCbased disease modelling.

I want to encourage the authors to place more stress on the conceptual importance of their work understanding cell line [?]. Perhaps the easiest way to do this is to point out the increasing use of lineage based developmental analyses in the study of psychiatric disorders. This manuscript refers to these studies in passing but these and other recent studies for example on the origins of Tuberous Sclerosis show that we are gaining a systematic understanding of the clinical implications of the developmental genetic mechanisms that construct the human central and peripheral nervous systems.

We thank the reviewer for bringing this point to our attention. We have reinforced this message by extending one introductory paragraph:
["... Controlling differentiation variability of iPSCs is an essential step in achieving reliable disease models, in particular in the field of developmental biology where substantial efforts are underway to model the cell-level consequences of genetic findings in developmental and neuropsychiatric disorders."]

We also included additional text in the Discussion:

["... Further, abnormal proliferation has been reported as a phenotype in multiple iPSC-based disease models, such as Kabuki Syndrome (Carosso et al., 2019) or in tuberous sclerosis, where an overproliferating population of interneuron progenitors in brain organoids was identified as the underlying cause for the distinctive phenotype of brain tumours and cortical malformations (Eichmüller et al., 2022).]

It will be very interesting to understand better compare these different developmental risks. For example, the clinical impact of de novo CNVs is central to psychiatric genetic research. Here CNVs are mentioned as a minor source of genetic variation but it would be helpful to know specifically how the CNV search algorithms used here compare with others. More generally, I wonder if analytical tools that are already available might reveal mutational burden reported by others in ASD cell lines in vitro.

The reviewer raises an interesting point regarding the important role of de novo CNVs in neurodevelopmental (NDDs) and neuropsychiatric disorders (Drakulic et al. 2020; Rees and Kirov 2021), and suggests to contrast known genetic risk factors of NDDs, such as autism, to the somatic variation observed in our dataset.

In the manuscript, we included CNVs that were acquired or positively selected in iPSCs relative to parental fibroblasts, profiled with genotyping arrays and previously reported in (Kilpinen et al. 2017). The CNV method used here was conceived to validate the genomic integrity of cultured iPSCs using SNP genotyping, analogous to the tumour and matched normal tissue comparison in cancer. Our approach combined two commands: "BCFtools/polysomy", to screen for contamination and chromosome aberrations, with "BCFtools/cnv", to detect CNV variation in local chromosomal regions. This outperforms in robustness other existing methods using genotyping data and t-tests, LRR values, segmentation methods or hidden Markov Models, and does not assume diploidy in controls as illustrated in (Danecek et al. 2016).

The dataset used here contained 460 CNVs across 367 iPSC lines, with only 202 lines presenting at least one CNV region different to their parental cells (CNVs spanning ~1.5 Mb on average). To compare this burden with existing CNV variation, we leveraged dbVar, the NCBI database (September 1, 2021 version) for large human genomic structural variation that includes 3,849,954 CNV regions, (Lappalainen et al. 2013)). Most of the CNVs annotated in dbVar have been

detected using microarrays as part of routine clinical testing, allowing the comparison of CNVs with similar allele frequency spectrum. Indeed, all somatic CNVs overlapped with CNV regions annotated in dbVar, but only 137 overlapped with de novo CNVs annotated as pathogenic by ClinVar dbVar. Among them, seven somatic CNVs (four duplications and three deletions) overlapped regions associated with autism spectrum disorder (MedGen annotation: C1510586, C4014435, C4025832, C0856975), and many other CNVs overlapped regions associated with developmental disorders (top3: Global developmental delay, n=20; Intellectual disability; n=18; abnormal facial shape, n=12). Still, the genes overlapping de novo pathogenic CNVs from dbVAR belonged disproportionately to developmental disorder genes (Wright et al. 2015) when compared to the genes overlapping the somatic CNVs (Fisher test, $p=6.4.10_{-3}$). This suggests that the impact of somatic CNVs on development risk is reduced when strictly compared to pathogenic structural variation, but it is not neutral.

In future experiments, and as an alternative to arrays, profiling lines with WGS may provide better detection of private CNVs and better comparison to disease-associated CNVs. As for the existing WES in this study, we did not call CNVs from WES, as several authors report limitations in accuracy compared to WGS CNV calls, mainly due to the noncontiguous nature of the targeted capture, and because WES was not available for all the parental fibroblasts, making a paired analysis impractical.

It would be helpful to have the authors views on these topics but I should stress in closing that this work is presented in a thoughtful style. The authors address important issues in a straightforward style. I find the story a compelling addition to our knowledge of human iPSC biology.

Again, we thank the reviewer for the positive comments and analytical suggestions helping us to better contextualise the main messages of the manuscript.

Comments, Reviewer #2

In their manuscript entitled "Somatic mutations alter the differentiation outcomes of iPSC-derived neurons" by Costa et al, the authors meta-analyze existing single-cell RNA and whole-exome sequencing for a pooled dopaminergic neuron differentiation experiment of 238 iPSC lines. They observed deleterious somatic mutations in key developmental genes, notably the BCOR gene, associated with perturbed gene expression, increased proliferation ,and reduced dopaminergic neuron differentiation. Overall, this is an interesting finding that cautions about interpreting differentiation phenotypes in iPSC disease models. Nonetheless, more needs to be done to confirm the causality of these associations and the mechanisms driving somatic mosaicism during neural differentiation. With major revisions, this report might be suitable for Cell Genomics.

We would like to thank the reviewer for the thorough review and careful consideration of our results. His/her comments have been very helpful in improving the manuscript and planning future work. We have addressed all of the major concerns when technically and analytically feasible.

Major concerns:
1. Exome-wide burden of acquired mutations does not explain the differentiation outcome: The authors analysed "differentiation outcomes from four independent experiments that produced different target cell types from iPSC lines of the HipSci project (Kilpinen et al., 2017): dopaminergic neurons, DN (Jerber et al., 2021), macrophages (Alasoo et al., 2018), sensory neurons (Schwartzentruber et al., 2018) and definitive endoderm tissue (Cuomo et al., 2020) (Fig. 1A). [They] compared the exome-wide burden of acquired mutations and rare germline variants to the

differentiation outcomes of each line (Fig. 1B), and used single-cell transcriptomes of the dopaminergic neurons (Jerber et al., 2021) to delineate in detail how deleterious variants influence the differentiation process of iPSC-derived neurons."

a. Inter- and intra-donor variation: The authors should clarify how many donors, iPSC clones, and independent experiments were included in each analysis.

Thank you for bringing this to our attention. For the different research questions, different numbers of iPSC lines (and donors) have been analysed, depending on the available information. For example, when looking for cell type composition outliers, we considered all lines with confident estimates of cell type frequency (n=227 lines from 219 donors), while for the impact of iPSC mutational burden on observed DN differentiation outcomes, we required iPSC lines profiled with WES and sampled at day 52, which reduces the number of lines considered (n=183 lines, one line per donor). We have now updated and extended the relevant numbers in the main text to provide a more comprehensive view of such variation, while specifying when multiple lines from the same donor were used. We also made available a full description of these numbers for each analysis in their corresponding Methods section so as to facilitate reproducibility. We have also updated Methods Section 7, where we specify the number of cell lines that were employed for each replicate group and the replicate comparisons performed. For instance, to assess the replicate correlation for pool replicates (one line placed in different pools), 41 different lines were employed, but 90 comparisons were done given the different time points and the scRNA-seq pool design.

Were some donors particularly prone to acquired deleterious mutations, or did they occur at random in all individuals?

This is a highly relevant question. When we initially looked into the acquired mutational burden of the 384 lines, we identified a subset of 17 hypermutated lines, as described in Methods (>240 acquired mutations, corresponding to a Z-score>2, Fig. R1A). Because of this significantly increased burden, we questioned the genetic integrity of these lines and decided not to include them in further analyses. This high mutational burden clearly indicates that certain lines are more prone to acquire mutations. This observation applies also to deleterious mutations, as the total number of acquired mutations per iPSC line correlates with the number of deleterious mutations (Fig. R1B), with hypermutated lines (highlighted in red) also having the highest deleterious burden.

We were also interested in exploring whether the acquired mutational burden was line or donor specific, so we compared the deleterious burden correlation for those donors for whom two iPSC lines were derived (Fig. R1C). Although those replicate lines still preserve some correlation (R=0.774), observing one hypermutated line did not necessarily implicate the same mutational status for the other replicate pair (3 out of the 6 replicate pairs had just one line hypermutated (in blue), while the other 3 pairs were both hypermutated (in red)). In summary, the acquisition of mutations is donor-specific, but among hypermutated lines, the variability on burden can be driven by line-specific effects.

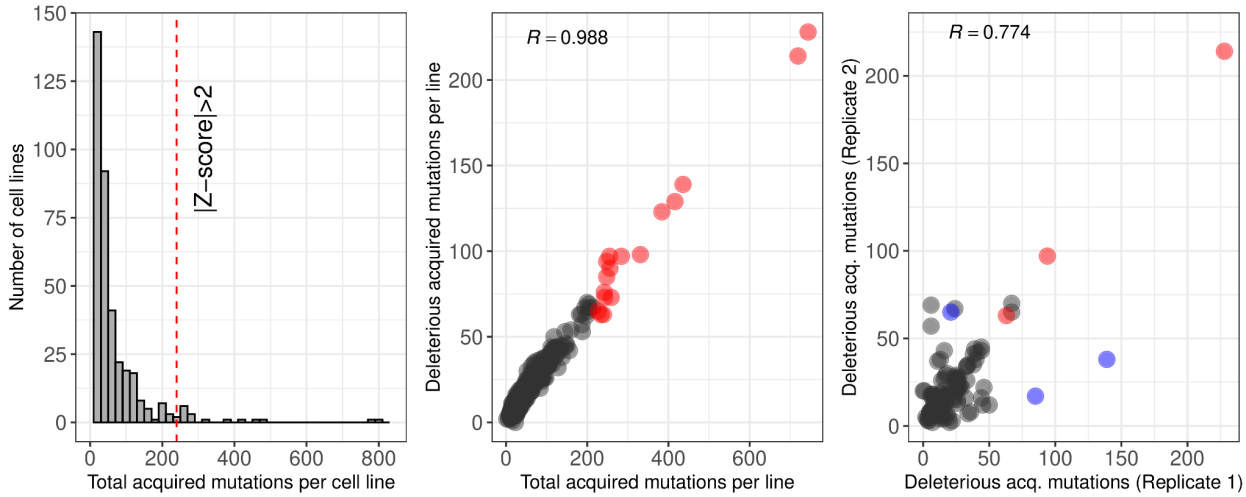

Fig. R1. a, Distribution of the number of acquired somatic mutations per cell line. Those lines with more than 240 mutations (Z-score>2) were defined as hypermutated lines and discarded from further analysis. b, Correlation between the total burden of acquired mutations and the subset of deleterious mutations (Pearson's correlation, R=0.988). Hypermutated lines are highlighted in red. c, Correlation between the deleterious burden of cell line pairs derived from the same donor (Pearson's correlation, R=0.774). In red, replicate pairs that are both hypermutated and in blue, when just one of the pairs is hypermutated.
Following the reviewer's comment, we have made this point explicit in the main text by mentioning the exclusion of such hypermutated lines (page 3).

2. Causality and Phenotypic effects: The authors "found that only one gene, BCOR, was significantly more mutated in failed lines compared to the successful lines… [Amongst failed line that] do not carry any deleterious BCOR mutation… The most mutated genes (log2FC>1.5) contributing to the enrichment of those biological processes in failed lines are well-known disease associations: NSUN5 in William Beuren Syndrome) or ASCL1 in neuropsychiatric disorders… or key regulators of neuron differentiation like NRBP2 or SOX14…" Moreover, the authors "correlated the cell type-specific expression [of mutated genes in the abnormal cell type fraction] with the changes in cell type composition, but did not empirically test the impact of mutated genes on gene expression.

We thank the reviewer for this useful comment and have now tested the impact of mutations on gene expression. We confirmed a reduction of BCOR gene expression in those failed lines with BCOR LoF mutations. While carrying out this analysis, we discovered an error in Figure 2D affecting the selection of genes used as input for the enrichment analysis. This error masked the impact of deleterious mutations in successful lines and provided wrong log_2FC values for the reported gene candidates (NSUN5, ASCL1, NRBP2, SOX14). We have now corrected this error and used a less conservative threshold for the selection of mutated genes: top-10% ($log_2$2CFC$\geq$ 0.95) in failed lines, and bottom-10% ($log_2$2CFC$\leq$ -1.37) in successful lines. This provided better resolution to identify brain-related processes disrupted by deleterious mutations, as the newly identified processes all ranged among the top-30 enriched items, as ordered by odds ratio. Previously, only one in six of the brain-related processes were among the top-30.

Updated results still indicate that the deleterious mutations observed specifically in failed lines disrupt key neurodevelopmental processes, such as neuron fate commitment (GO:0048663), response to axon injury (GO:0048678), and midbrain development (GO:0030901). In addition, deleterious mutations specific to successful lines disrupt the positive regulation of neuron apoptotic process (GO:0043525), which might prevent neuronal death, and the disruption of neuron fate specification (GO:0048665), which might lead to stronger commitment to neuronal

fate (Fig. R2).

Finally, we updated the list of genes contributing to these biological processes and found again relevant disease associations for those genes mutated in failed lines, for e.g. CDC42 in intellectual disability (Martinelli et al. 2018), BMP4 in syndromic microphthalmia (Reis et al. 2011) and PMP22 in hereditary neuropathy (van Paassen et al. 2014), but also in successful lines, as expected by the actual enriched brain-related processes (e.g. DLL1 in genetic epilepsy syndromes). All of these genes have diagnostic evidence for disease (Stark et al. 2021), contrary to the candidates mistakenly reported before. Main text, Figure 2D, and Supplemental Table 2 have been modified accordingly.

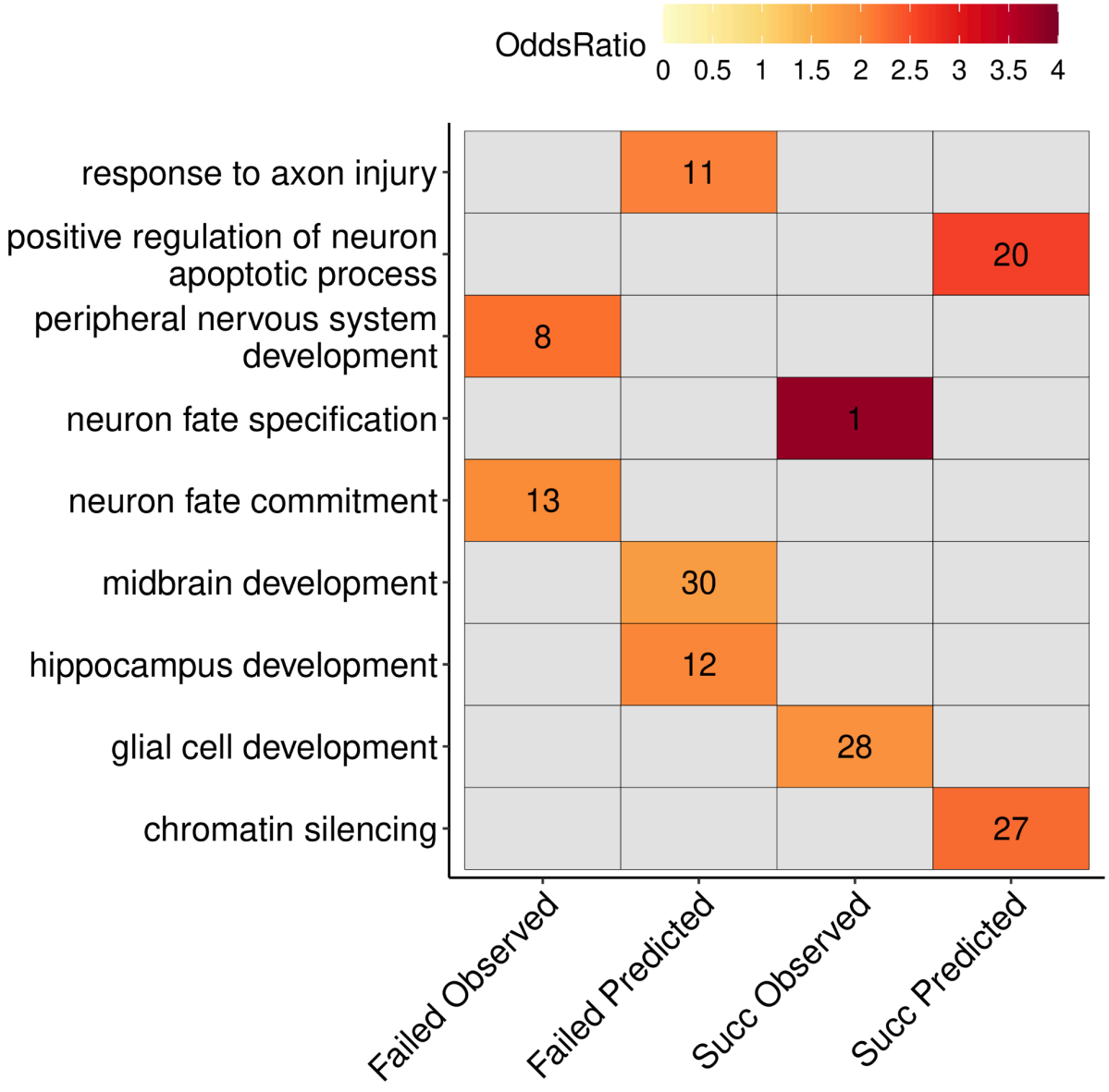

Figure R2 (Fig. 2D in main text). Gene ontology enrichment analysis revealed the impact of deleterious mutations on brain-related biological processes (pAdj<0.05, red-coloured tiles) contributing to either failed or successful differentiation, in lines whose outcome was either observed or predicted. The number in each tile corresponds to the rank of the significant GO term, ordered by decreasing odds ratio within each

analysis (only top-30 positions are illustrated).

a. Dopaminergic differentiation: were individual mutations (e.g. in BCOR) confirmed by sequence changes and/or expression levels reductions by scRNAseq?

We thank the reviewer for this valuable comment. This information was indeed missing and we have now added it. While it was not possible to experimentally validate individual BCOR mutations identified from WES (no access to the cell lines), we have now included a new supplementary table (Supp. Table 7), where we detail LoF BCOR mutations identified in 22 failed iPSC lines, out of the 183 cell lines for which WES and dopaminergic differentiation were available. This table includes several fields: the chromosome (X chromosome), the position, the reference and alternative alleles, the genotypes (21 heterozygous and 5 ALT homozygous) and the iPSC line where it was identified, as well as the predicted VEP variant consequences (8 SNPs and 18 indels causing stop-gained codons and frameshift mutations, respectively).

However, we confirmed a reduction in the expression level of the BCOR gene in lines carrying LoF mutations (Failed/BCOR+) compared to unmutated lines (Failed/BCOR- and Successful/BCOR-), a difference that was significant across all time points (p-value<0.05, Wilcoxon Rank Sum Test). As a result, we now provide evidence for LoF mutations in the BCOR gene reducing its gene expression, in addition to being associated with failed dopaminergic differentiation. Given its importance, we have included this result in the main text and in the corresponding supplementary figure (Supp. Fig 1D), which is shown below. The main text (page 4) now reads as:

["We confirmed a significant reduction of BCOR expression between mutated (presence of LoF in BCOR) and unmutated lines across all time points (p<0.05, Wilcox.test) (Fig. R3)."]

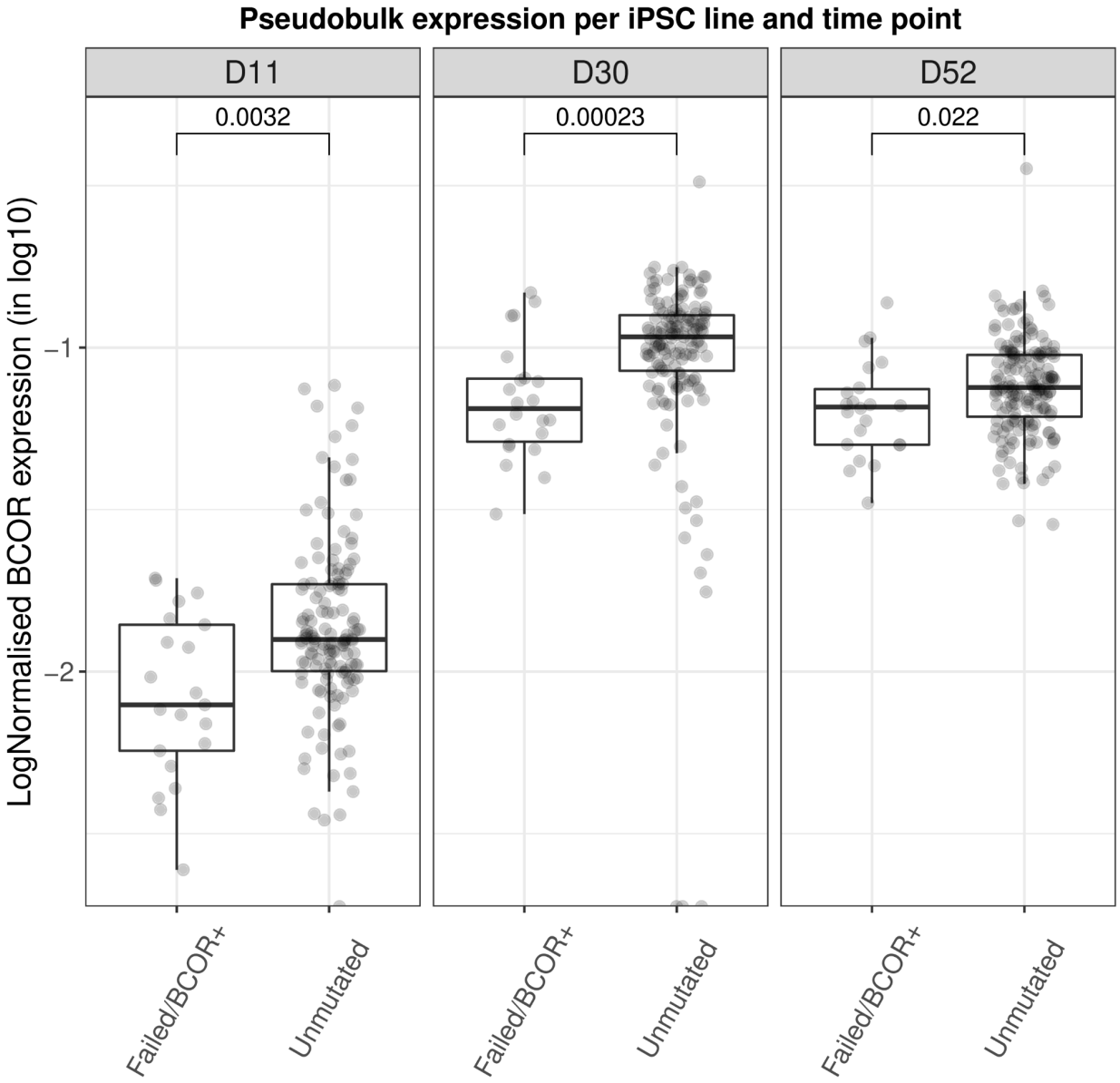

Figure R3 (Supp. Fig 1D in main text). Failed lines with LoF mutations in the BCOR gene show significantly reduced BCOR expression when compared with unmutated lines, either failed or successful (Wilcox.test, $p<0.05$). Expression units correspond to pseudobulk log-normalised values per iPSC line and time point.

b. Given the pooled design used for the DA neurons, the authors should consider the possible impact of non-cell-autonomous effects on their analysis.

We agree with the reviewer that the pooled design used for dopaminergic differentiation has both strengths and weaknesses. While it allows the profiling of a larger number of cell lines with reduced costs and batch effects, it might also expose lines to non-cell-autonomous (NCA) effects. These can alter cell behaviour, either through cell-to-cell interactions from different donors or via induced changes of the microenvironment. Although we did not directly model such effects on gene expression, different replicate types were introduced in the experimental design to address

the extent of NCA effects in pools.

First, we analysed how a single line (pool replicate) grew in different pool backgrounds, i.e. mixed with different cell lines. Second, we looked at independent differentiations of the same pool, so as to track line abundance in the same background (biological replicates). Finally, we focused on technical replicates (cell lines from the same differentiation and pool, but different wells of the same plate) and 10X replicates (cell lines from the same pool, same differentiation and same well). This information is available in Methods - Section 7. As illustrated in the main text (Fig. 3C), we observed that cell lines in a pool can influence the growth dynamics of individual lines, as pool replicates have weaker correlation in cell line abundance (R_2=0.136) than biological replicates (R_2=0.693), or technical and 10X replicates (R_2=0.996). Despite these results suggesting that NCA effects are not negligible, the differentiation outcome of lines in different pool backgrounds was unaltered, as only 3 out of 36 iPSC lines were discordant, that is, the same line failing in one pool and succeeding in another. Therefore, to be more explicit about the extent of NCA effects, we have modified the main text (page 5) and discussed the limitations of the pooled design (page 10).

c. It would be extremely informative if the authors knocked out/down BCOR and/or NSUN5, ASCL1, NRBP2, or SOX14 in one or more donor backgrounds and confirmed changes in DA patterning, maturation, function.

We agree with the reviewer that it would be extremely interesting to validate the phenotypic consequences of deleterious variation in some of the candidate genes, but such effort goes beyond the scope of this manuscript. Here, we assessed the impact of acquired somatic mutations on the capacity to produce neurons in a large iPSC cohort, rather than engineering individual gene knock-out/knock-in models to study potential disease mechanisms. Still, this line of enquiry would be very informative and we will include it in our future research. We also agree with the reviewer that the effect of such mutations should be assessed on several lines to better understand their interaction with the donor background.

d. Given that many of these genes, notably BCOR, are robustly expressed in astrocytes, it would be interesting to access what cell types the failed DA neurons are becoming.

We cannot speculate what failed DA neurons are becoming based on BCOR expression alone, as it consistently showed low expression across astrocytes, dopaminergic neurons and serotonergic neurons at day 52 (Fig. R4). Instead, we addressed this specific question by analysing cell type composition differences between failed and successful lines using a negative binomial regression model. In the Results section of the main manuscript (page 6), we say that "At day 52, the overall lower fraction of neurons in failed lines was accompanied by a significantly larger proportion of astrocytes (pAdj=$1.5.10^{-36}$), ependymal-like cells (pAdj=$7.7.10^{-14}$) and of the unknown cell type 1 (pAdj=$1.1.10^{-19}$)". In addition, we found an association of cell line proliferation rate with the abundance of astrocytes (pAdj=0.004). Therefore, our results indicate that those lines with a low production of DA neurons give rise to a higher proportion of astrocytes. Based on the gene ontology enrichment, we also hypothesise that the disruption of glial cell development by mutations mostly seen in successful lines (Fig. R2) could favour a higher production of neurons at the expense of astrocytes.

Finally, we have also looked at the expression of the other candidate genes driving differentiation failure (PMP22, CDC42, BMP4) in astrocytes and neurons. As mentioned above, these genes are mostly mutated in failed lines, they contributed to enriched brain related processes (Fig. R2) and are well-known disease-causing genes. At day 52, we observed that PMP22 showed

expression in astrocytes and Sert-like neurons, but not in DA neurons, as opposed to CDC42 which showed the highest expression in this cell type (Fig. R4). Some of the expression patterns are shared with genes that are mostly mutated in successful lines (PMP22 with GRN, GSN and GSTP1; and CDC42 with SOX4 and DMRTA2). Only BMP4 showed robust expression exclusively in astrocytes, with reduced expression across failed lines ($p=5.96.10^{-5}$).

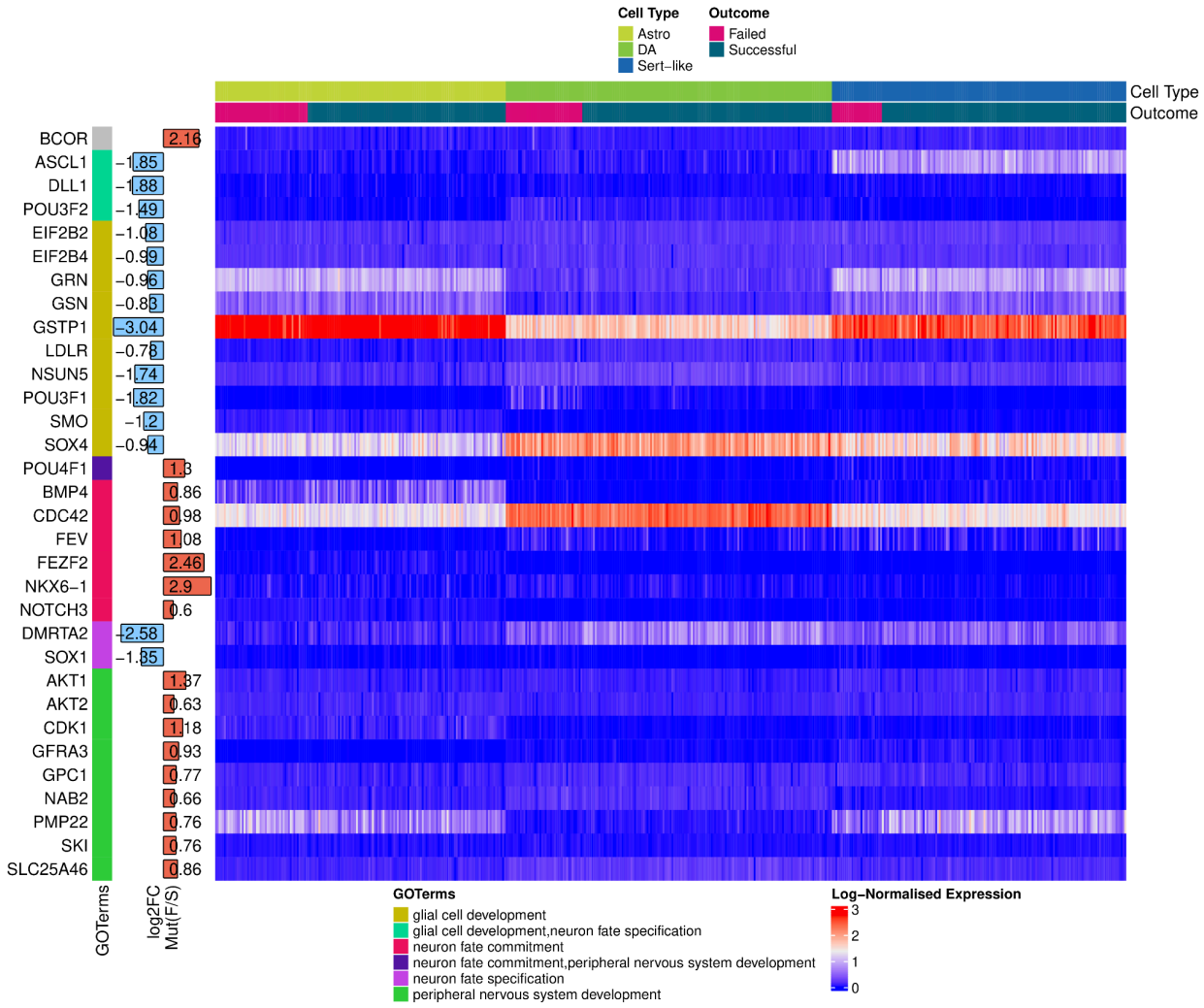

Fig. R4. Expression of the genes mostly mutated in failed (red bars, top-10% log_2FC) and successful lines (blue bars, bottom-10% log_2FC) that contribute to the enrichment of brain-related processes. Lognormalised expression values per line (corresponding to a heatmap tile) are displayed for astrocytes, dopaminergic neurons (DA) and serotonergic-like neurons (Sert-like) at day 52. Within each cell type, cell lines are grouped as those with a failed (left) or successful outcome (right) (observed outcome as in Fig. contributions.

e. Because exome data was only available for source fibroblasts and donor hiPSCs, the experimental design fails to capture the impact of somatic mosaicism that occurs during neuronal differentiation. Could exome sequencing of dopaminergic neurons from a subset of iPSC lines be added?

We thank the reviewer for bringing up this question. We acknowledge that somatic mosaicism could further contribute to differentiation failure, especially in failed lines without a LoF BCOR mutation. However, since the original differentiations (Jerber et al.) were assayed only with scRNA-seq, it was unfortunately not possible to perform mutation analysis of the individual cell lines at the neuronal stage. As an alternative, we considered repeating a subset of the differentiations, but to perform mutation analysis of individual lines we would have needed to either perform individual differentiations or barcoded the cell lines in a given pool, neither of which would have fully replicated the original experiment and its mutagenic processes.
We have now brought up these limitations and the potential effect of somatic mosaicism in neurons in the Discussion (page 9). It reads:

["While our approach is limited to identifying acquired mutations only during iPSC reprogramming, it is possible that somatic mosaicism in neurons also plays a role in determining the final differentiation outcome. Future efforts in profiling exomes of neurons will complement our results on the role of somatic mutations in differentiation systems."]

3. Mechanisms and biological relevance: it would be interesting if the authors uncovered the genetic or epigenetic factors driving somatic mosaicism in vitro. Previous reports of somatic mosaicism have shown that it can be drive by L1 (or other) repetitive elements (Coufal et al Nature 2009), high transcription levels (Lodato et al, Science 2015), large gene size and late-replication (Wei Schwer Cell 2016), and/or DNA repair (Reid et al Science 2021)

We agree with the reviewer that this is a very interesting line of future research and we thank him/her for the suggestions and the proposed bibliography. However, as our manuscript does not focus on somatic mosaicism in vitro, we consider this beyond the scope of the paper. We believe that the analysis presented here covers the most important aspects of somatic mutations that affect iPSC lines and their subsequent differentiation.

a. Are the genes with somatic mutations correlate to features such as proximity to repetitive elements, expression levels, length, or cell replication, transcribed, and late-replicating genes?

This is an interesting question and the reviewer cites relevant publications on the mechanisms of DNA damage in post-mitotic neurons, emphasising the role of endogenous transposable elements, replication timing, strand-bias and active transcription in driving somatic mosaicism in neurons. Our identified somatic mutations, however, originated either in skin fibroblasts or during iPSC reprogramming. For this reason, we expected them to best correlate with features from cancer and early developmental stage rather than from postmitotic neurons, as suggested by the high prevalence of UV-associated DNA damage (Jerber et al., 2021).

Nevertheless, we looked for differences in the proximity to repetitive elements between genes with somatic mutations (n=9,235) and unmutated genes (n=11,118). We used repetitive elements from RepeatMasker at UCSC (Smit et al. 2004). On average, mutated genes were significantly closer to repetitive elements than unmutated ones, except for satellite repeats, ribosomal RNAs and transfer RNAs (T-test, pAdj<0.05, Fig. R5A). This difference remained when aggregating categories together (All Rep Elements, pAdj=1.16.10^{-23}). This correlation agrees with the contribution of L1 retrotransposition in driving somatic mutations (Coufal et al. 2009), not only in human brain, but also in iPSC acquired mutations, as they are closer to long terminal repeat elements (LTR), a group that includes retrotransposons.

We also correlated gene mutational status with their pseudobulk expression averaged across

iPSC lines and each differentiation time point (Fig. R5B). Mutated genes have consistently reduced expression compared to unmutated ones across days 11, 30 and 52 (T-test, $p<0.05$). In contrast, genes mutated in single neurons have higher expression than unmutated ones in the prefrontal cortex of the post-mortem human brain (Lodato et al. 2015).

Neural genes are overrepresented in fragile genomic regions with recurrent double-strand DNA breaks (Wei et al. 2016). These genes are long, actively transcribed and late replicating, accumulating more mutations in response to replication stress. In our dataset, we confirmed that mutated genes were significantly longer than unmutated ones (T-test: $p<1.02.10^{-149}$, Fig. R5C), but iPSC lines were not directly assayed for chromatin state marks or replication timing limiting further comparisons. As an alternative, we used available genome-wide annotations from the Roadmap Epigenomics Projects and ENCODE.

First, we assessed whether the chromatin state of promoters from mutated genes was more active than from unmutated genes, but found no significant composition differences between them ($p=0.34$, Chi-squared test). For this comparison, we used 13 iPSC and ESC lines from the Roadmap Epigenomics Project (Roadmap Epigenomics Consortium et al....), binning the 25 chromatin states into active, bivalent and inactive categories.

Second, we compared the average replicating times between mutated and unmutated genes using annotations from the ENCODE BG02ES ESC line. No significant differences were observed in Wavelet-smoothed signal (T-test, $p=0.49$, Sup. Fig. R5D). Still, a subset of our somatic variants were associated with UV-substitution signatures having strong transcriptional biases, with an excess of C>T and CC>TT on the non transcribed strand (Jerber et al., 2021). Both excesses were enriched in early replication domains, in contrast to late replicating enrichment of genes recurrently mutated in neurons, potentially indicating different mutational processes.

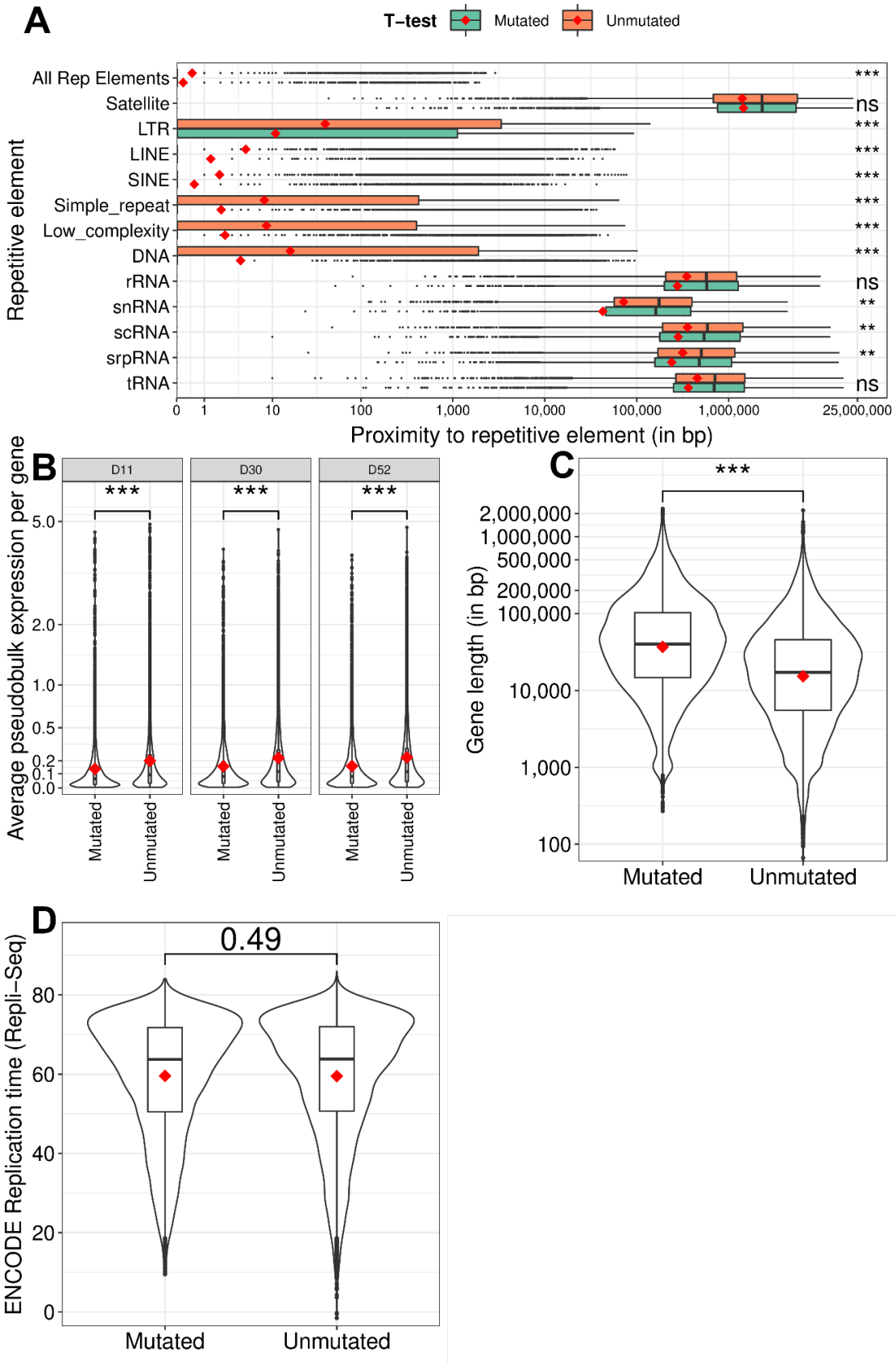
A
T-test
Mutated
Unmutated
All Rep Elements
Satellite
LTR
LINE
SINE
Simple_repeat
Low_complexity
DNA
rRNA
snRNA
scRNA
srpRNA
tRNA
Repetitive element
***
ns
***
***
***
***
***
ns
**
**
**
ns
0
1
10
100
1,000
10,000
100,000
1,000,000
25,000,000
Proximity to repetitive element (in bp)
B
D11
D30
D52
***
Average pseudobulk expression per gene
5.0
2.0
1.0
0.5
0.2
0.1
0.0
Mutated
Unmutated
C
***
Gene length (in bp)
2,000,000
1,000,000
500,000
200,000
100,000
10,000
1,000
100
Mutated
Unmutated
D
0.49
ENCODE Replication time (Repli-Seq)
80
60
40
20
0
Mutated
Unmutated

Fig R5. (Supp. Fig. 2 in main text). a, Genes with somatic acquired mutations in our dataset (n=9,235) showed significantly closer proximity to repetitive elements (UCSC RepeatMasker) than unmutated genes (n=11,118), except for satellite repeats, ribosomal RNA and transfer RNA (T-test, pAdj<0.05). b, The same set of 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. c, Also, mutated genes were on average longer than unmutated genes (T-test, $p<1.02.10^{-149}$). d, 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 computed across all the signal values from the 1-Kb windows overlapping each corresponding gene.

Moreover, it would be useful to compare conduct enrichments of mutated genes not just to gene ontology, but to published lists of somatic mutations during neural differentiation (e.g. almost 90% of RDC-containing genes identified are involved in synapse function and/or neural cell adhesion, with a substantial fraction also implicated in tumor suppression and/or mental disorders. (Wei Schwer Cell 2016))

We have now considered three additional published sets of somatic mutations before and during neuronal differentiation:

1. The landscape of somatic mutations in cerebral cortex of autistic and neurotypical individuals revealed by ultra-deep genome sequencing (Rodin et al. 2021): It includes 51 genes carrying exonic mutations in both controls (n=15) and subjects with autism spectrum disorder (n=59). This approach allows to identify somatic mutations starting from the postzygotic stage.
2. Long neural genes harbor recurrent DNA breaks clusters in Neural Stem/Progenitor Cells (Wei et al. 2016): It includes 27 genes that are susceptible to genomic instability given their long size, active transcription and late replication at stem cell/progenitor neural stage.
3. Comprehensive identification of somatic nucleotide variants in human brain tissue (Wang et al. 2021): This Brain Somatic Mosaicism Network study identified 43 bona-fide somatic SNVs following best calling practices in postmortem brain tissue, from which 21 fall unequivocally in one gene (Ensembl gene annotation: GRCh37, version 87). This list contains a mixture of SNVs originated in post-zygotic and later stages of development, given the wide range of variant allele frequencies (0.005 to 0.28).

We conducted enrichment tests of mutated genes in our study (n=8,608) on the three published sets of somatic mutations, and found significant overlap for set 1 ($pAdj=9.46.10^{-4}$, hypergeometric test) and set 2 ($pAdj=9.02.10^{-5}$), but not for set 3 (pAdj=0.29). Note however that set 3 has the smallest gene size and is likely to be compounded by a higher proportion of somatic mutations originating in later stages of neurodevelopment. This lack of overlap might reflect different mutational pressures than those that trigger somatic mutations in iPSC populations.

We have modified the main text (page 3) to include the results of this analysis on somatic acquired mutations:

[“In our study, mutated genes are enriched in published sets of somatic mutations originating early in development, mainly in the postzygotic stage ($pAdj=9.46.10^{-4}$, hypergeometric test) (Rodin et al. 2021) or in the stem cell and neural progenitor stage ($pAdj=9.02.10^{-5}$) (Wei et al. 2016).”]

b. The authors state that "Among the significant associations, including positively and negatively correlated genes, we observed a strong enrichment of DDD genes in most of the cell types at the

neurodegenerative, and non-brain disorders.

We thank the reviewer for his constructive and valuable comment. We have now expanded our enrichment tests for those genes whose expression correlates (or anti-correlates) with cell type abundance, by including four additional disease-relevant gene panels:

1. Psychiatric traits: Schizophrenia (n=1468, Open Targets Platform) (Ochoa et al. 2020)
2. Neurodegeneration: Adult onset neurodegenerative disorder (n=423, Genomics England, Panel App v2.178) (Martin et al. 2019)
3. Non-brain disorders (1): Bleeding and platelet disorders (n=112, Genomics England, Panel App v1.2) (Martin et al. 2019)
4. Non-brain disorders (2): Family pulmonary fibrosis (n=75, Genomics England, Panel App v1.29) (Martin et al. 2019)

These additional analyses confirmed that the panels of developmental disorder genes (DDD) and cancer-associated genes (Cosmic-Tier 1) are mostly enriched in progenitor populations at days 11 and 30 (Fig. R6, shown below). A similar but more limited enrichment was simultaneously observed in adult onset neurodegenerative disorder genes for floor plate progenitors (FPP). Yet, the enrichment diverged in schizophrenia, a trait for which weaker but significant associations occurred in dopaminergic neurons (day 30) and serotonergic neurons (day 52). These results are in agreement with an iPSC-based neuronal model, as iPSC-derived neurons are known to correlate better with fetal than adult brain tissue (Handel et al. 2016). On the other hand, none of the two non-brain disorder panels were enriched in any cell type across differentiation timepoints.

For consistency with previous enrichment tests, we have substituted Chi-squared tests with hypergeometric tests of overrepresentation (updated in Methods). As before, raw p-values have been corrected for multiple testing (combinations of 12 cell types, 3 time points and 6 panels employed) using the Benjamini-Hochberg correction.

We have updated the main text (pages 7-8) to incorporate these additional results, as well as corresponding Figure 5D.

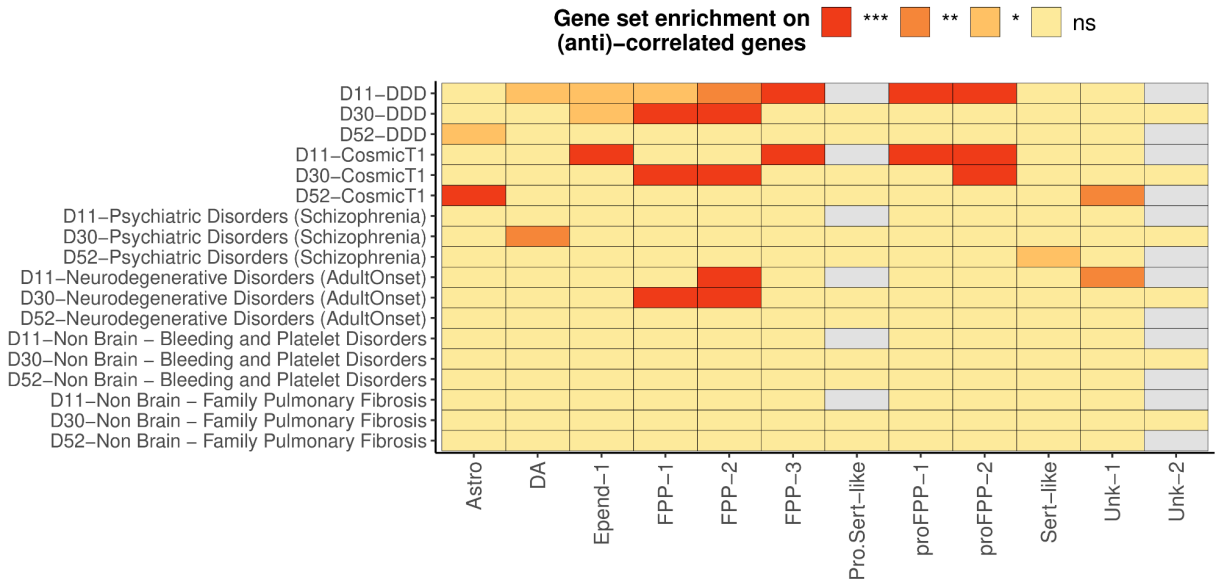

Fig. R6 (Fig. 5D in main text). The abundance of progenitor populations was associated with the

expression of genes in developmental disorder (DD) and cancer-associated (Cosmic-Tier1) genes, and to a lesser extent in adult-onset neurodegenerative disorder genes. In schizophrenia, the association is only observed with genes expressed in neurons and no enrichment is observed for non-brain disorders (hypergeometric test for overrepresentation, Benjamini-Hochberg multiple-test correction). Significance levels: pAdj<0.05 (*, golden yellow), pAdj<0.01 (**, orange), pAdj<0.001 (**, red).

Referees' report, second round of review

Reviewer #1: Comments enter in this field will be shared with the author; your identity will remain anonymous.

In this revised manuscript, the authors have adequately addressed.the majority of the issues raised in the first cycle of review. Here I will restrict myself to emphasizing the general interest in results reported in this work. The work reported here was initiated by analyzing four distinct differentiation experiments using human induced pluripotent stem cells (iPSC) but quickly focusses on one of these where iPSCs were directed to dopaminergic fates.

In the two decades that human embryonic and pluripotent stem cells have been extensively studied there are several reports that these cells acquire genetic change in cell culture. These genetic alterations include gain or loss of specific chromosomal sites and mutation in genes known to be involved in human cancer including TP53. Theses genetic alterations raise important concerns when considering the use of iPSC derived cells in cell therapeutic strategies. Parkinson's disease is one of the most extensively explored indications for iPSC derived cell therapy making the current study particularly relevant as it is based on the ability of hIPSCs to succesfully generate dopaminergic neurons in vitro.. This manuscript presents convincing evidence that mutations in the BCL6 corepressor BCOR occur in many human iPSC lines impairing their differentiation to dopaminergic fates and conferring a selective advantage on these differentiation-impaired lines.

These results are of specific interest to the development of protocols for efficient and safe differentiation of dopaminergic neurons. They will also be of interest to the wider community developing in vitro iPSC protocols to model and develop therapeutic strategies for neurological and psychiatric disorders. Here the focus on neuronal differentiation as the primary data set is unusual. In particular the analysis of the cell lines at 3 differentiation time points along the neural trajectory draws attention to the importance of defining genetic variation at different stages of fate acquisition aligning with ongoing work analyzing cell line competition in chimeric organoids.

Reviewer #2: For the authors:
The authors robustly demonstrate that deleterious somatic mutations in key developmental genes are strongly associated with failure in dopaminergic neuron differentiation. Their revised manuscript now includes confirmation of reduced BCOR expression in failed lines with LoF mutations, correlation of genes affected by somatic mosaicism in the brain with genes with acquired somatic mutations in hiPSCs, and expanded enrichment analysis to additional disease-related gene sets.

Can the authors speculate as to why they only detect the impact of somatic mutation on dopaminergic differentiation and not sensory neurons, endoderm, or macrophages? Likewise, why do they not detect p53 mutations, which have been repeatedly identified as somatic mutations in pluripotent stem cells and are expected to favor replication and impair differentiation across many lineages (PMID: 28445466).

As written, it is too easy for the reader to miss the value of this analysis. It has been well established for more than ten years that somatic mosaicism occurs in stem cell lines and that this can impact replication and differentiation [1-6], which is why stem call labs perform extensive quality control of hiPSC lines to rule out lines where novel mutations have been selected for. Repeatedly, in the abstract, introduction, and discussion, the authors highlight "somatic mutations as an important source of variation in iPSC-based models and call for caution when interpreting differentiation-related phenotypes to understand disease." That being said, they fail to adequately highlight that in vitro somatic mosaicism during hiPSC neuronal differentiation mirrors the somatic mosaicism that occurs during corticodevelopment that is associated with neurodevelopmental disorder risk! ["In our study, mutated genes are enriched in published sets of somatic mutations originating early in development, mainly in the postzygotic stage (pAdj=9.46.10-4,

hypergeometric test) (Rodin et al. 2021) or in the stem cell and neural progenitor stage (pAdj=9.02.10-5) (Wei et al. 2016)."] This is an important finding that warrants repeating, as a major implication of the study is that in vitro somatic mosaicism could be used to model and understand the clinical impact of this process in vivo!

1. Mayshar, Y., et al., Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. Cell Stem Cell, 2010. 7(4): p. 521-31.
2. Hussein, S.M., et al., Copy number variation and selection during reprogramming to pluripotency. Nature, 2011. 471(7336): p. 58-62.
3. Gore, A., et al., Somatic coding mutations in human induced pluripotent stem cells. Nature, 2011. 471(7336): p. 63-7.
4. Ruiz, S., et al., Analysis of protein-coding mutations in hiPSCs and their possible role during somatic cell reprogramming. Nat Commun, 2013. 4: p. 1382.
5. Lu, J., et al., The distribution of genomic variations in human iPSCs is related to replication-timing reorganization during reprogramming. Cell Rep, 2014. 7(1): p. 70-8.
6. McConnell, M.J., et al., Mosaic copy number variation in human neurons. Science, 2013. 342(6158): p. 632-7.

Although still lacking functional validation, this meta-analysis of large existing datasets yields interesting insights of value not just to hiPSC biologists, but also human geneticists, and so is a resource worthy of sharing in Cell Genomics.

Authors' response to the second round of review

Reviewer #1:

In this revised manuscript, the authors have adequately addressed the majority of the issues raised in the first cycle of review. Here I will restrict myself to emphasizing the general interest in results reported in this work. The work reported here was initiated by analyzing four distinct differentiation experiments using human induced pluripotent stem cells (iPSC) but quickly focusses on one of these where iPSCs were directed to dopaminergic fates.

In the two decades that human embryonic and pluripotent stem cells have been extensively studied there are several reports that these cells acquire genetic change in cell culture. These genetic alterations include gain or loss of specific chromosomal sites and mutation in genes known to be involved in human cancer including TP53. These genetic alterations raise important concerns when considering the use of iPSC derived cells in cell therapeutic strategies. Parkinson's disease is one of the most extensively explored indications for iPSC derived cell therapy making the current study particularly relevant as it is based on the ability of hIPSCs to successfully generate dopaminergic neurons in vitro. This manuscript presents convincing evidence that mutations in the BCL6 corepressor BCOR occur in many human iPSC lines impairing their differentiation to dopaminergic fates and conferring a selective advantage on these differentiation-impaired lines.

These results are of specific interest to the development of protocols for efficient and safe differentiation of dopaminergic neurons. They will also be of interest to the wider community developing in vitro iPSC protocols to model and develop therapeutic strategies for neurological and psychiatric disorders. Here the focus on neuronal differentiation as the primary data set is unusual. In particular the analysis of the cell lines at 3 differentiation time points along the neural trajectory draws attention to the importance of defining genetic variation at different stages of fate acquisition aligning with ongoing work analyzing cell line competition in chimeric organoids.

We thank the reviewer for the positive comments on this work. We chose to focus primarily on dopaminergic neuron differentiation because the dataset had the largest sample size and the

best data (single-cell) to answer questions about mutation effects on differentiation. However, our approach and results are likely translatable to other differentiation systems, hence we appreciate the reviewer for highlighting the wider scope of our work.

Reviewer #2:
For the authors:
The authors robustly demonstrate that deleterious somatic mutations in key developmental genes are strongly associated with failure in dopaminergic neuron differentiation. Their revised manuscript now includes confirmation of reduced BCOR expression in failed lines with LoF mutations, correlation of genes affected by somatic mosaicism in the brain with genes with acquired somatic mutations in hiPSCs, and expanded enrichment analysis to additional disease-related gene sets.

Again, we thank the reviewer for the thorough review and all the comments that have helped improve the current version of the manuscript.

Can the authors speculate as to why they only detect the impact of somatic mutation on dopaminergic differentiation and not sensory neurons, endoderm, or macrophages?

Our ability to detect mutation effects only in dopaminergic neurons is likely due to the DN dataset having the largest sample size. The number of iPSC lines differentiated to DNs is almost two times larger than what was used for the other iPSC-derived cell types. Even when considering only those lines with paired iPSC-parental fibroblast WES data available, the difference in sample size remains significant (DN observed, 35 failed-91 successful; DN predicted, 33 failed-316 successful; sensory neurons, 3 failed-82 successful; macrophages, 22 failed-80 successful; endoderm, 86 lines). Also, in the DN dataset, the differentiation outcome is defined more accurately than in the other datasets thanks to the single-cell RNAsequencing readout and pooled differentiations. Specifically, neuron production per line can be quantified with less experimental and technical noise than those differentiated individually.

Likewise, why do they not detect p53 mutations, which have been repeatedly identified as somatic mutations in pluripotent stem cells and are expected to favor replication and impair differentiation across many lineages (PMID: 28445466).

The reviewer points out a very interesting observation. Despite previous reports of p53 mutations in PSCs, in the 384 iPSC lines for which we have detected somatic mutations acquired in vitro, we only identified three lines that contained a TP53 somatic mutation acquired or selected in vitro during reprogramming. In all cases (2 SNVs, 1 insertion), they were predicted to be either missense pathogenic or loss-of-function (LoF) and were reported to have oncogenic activity (Cosmic Tier 1, version 94). However, none of these cell lines were included in the DA differentiation experiment. Alternatively, two were present in the sensory neuron and macrophage differentiation studies, but the TP53 mutations did not affect the final differentiation outcome in these studies. As for the in silico predicted DA differentiation outcomes, only one of the three lines with a deleterious TP53 mutation (HPSI0513i-golb_1) was predicted to fail.

When we expanded our TP53 mutation search (either germline or somatic) to the whole set of 832 HipSci iPSC lines for which exome sequencing data was available, we identified a total of 26 deleterious mutations and 2 LoF mutations, of which only one was associated with oncogenic activity. In the paper cited by the reviewer (referenced also in our Introduction,

Merkle et al[10]), TP53 mutations were found in 14/257 hiPSC lines (5.4%). The fraction observed in HipSci lines (26/832) is somewhat lower (3.1%). 2 out of the 26 TP53 mutated lines were included in the DA dataset, with 1 successful and 1 failed outcome. Only 3 out of the 26 lines were predicted to fail DA differentiation.

Interestingly, Merkle et al[10] show that sub-clonal TP53 mutations carried by embryonic stem cells indeed provide growth advantage during differentiation, but they do not seem to impair the formation of pancreatic polyhormonal cells, i.e. differentiation, in contrast to what we observe with BCOR LoF mutations and dopaminergic neurons in the current study. Still, we agree with the reviewer that mutations in cancer driver genes, such as TP53, are highly likely to have practical implications for disease modeling, especially in pooled experiments, requiring frequent genetic screening to monitor the growth rate and proportional balance of cell lines. The degree to which these effects are cell type and lineage-specific remains open. An important step towards better characterizing the mutational processes during in vitro differentiation and their effect on proliferation and differentiation success will involve sequencing the DNA of cell lines in several stages, starting with iPSC, followed by progenitors and ending with mature cell types.

As written, it is too easy for the reader to miss the value of this analysis. It has been well established for more than ten years that somatic mosaicism occurs in stem cell lines and that this can impact replication and differentiation [1-6], which is why stem call labs perform extensive quality control of hiPSC lines to rule out lines where novel mutations have been selected for. Repeatedly, in the abstract, introduction, and discussion, the authors highlight "somatic mutations as an important source of variation in iPSC-based models and call for caution when interpreting differentiation-related phenotypes to understand disease." That being said, they fail to adequately highlight that in vitro somatic mosaicism during hiPSC neuronal differentiation mirrors the somatic mosaicism that occurs during corticodevelopment that is associated with neurodevelopmental disorder risk! ["In our study, mutated genes are enriched in published sets of somatic mutations originating early in development, mainly in the postzygotic stage (pAdj=9.46.10-4, hypergeometric test) (Rodin et al. 2021) or in the stem cell and neural progenitor stage (pAdj=9.02.10-5) (Wei et al. 2016)."] This is an important finding that warrants repeating, as a major implication of the study is that in vitro somatic mosaicism could be used to model and understand the clinical impact of this process in vivo!

1. Mayshar, Y., et al., Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. Cell Stem Cell, 2010. 7(4): p. 521-31.
2. Hussein, S.M., et al., Copy number variation and selection during reprogramming to pluripotency. Nature, 2011. 471(7336): p. 58-62.
3. Gore, A., et al., Somatic coding mutations in human induced pluripotent stem cells. Nature, 2011. 471(7336): p. 63-7.
4. Ruiz, S., et al., Analysis of protein-coding mutations in hiPSCs and their possible role during somatic cell reprogramming. Nat Commun, 2013. 4: p. 1382.
5. Lu, J., et al., The distribution of genomic variations in human iPSCs is related to replicationtiming reorganization during reprogramming. Cell Rep, 2014. 7(1): p. 70-8.
6. McConnell, M.J., et al., Mosaic copy number variation in human neurons. Science, 2013. 342(6158): p. 632-7.

We would like to thank the reviewer for emphasizing the relevance of our work to somatic mosaicism in stem cells. We agree this is a really interesting and highly relevant research area, and the advent of large-scale hiPSC studies, such as ours, is beginning to shed more light on the phenomenon at scale and specifically on its consequences on differentiation.

However, given some of the limitations of our data in the current study, we adopted a cautious writing style. We believe more controlled experiments are required to firmly establish how accurately mutational processes that affect human iPSCs reflect those that cause in vivo somatic mosaicism and e.g. neurodevelopmental diseases.

Regarding the second part of the reviewer's comment, we have modified the text at various sections to better highlight the overlap between the genes with somatic mutations in our study and those shown to be mutated in neurons. This is also accompanied with discussion of somatic mosaicism and the potential implications that this might have for disease modeling.

Although still lacking functional validation, this meta-analysis of large existing datasets yields interesting insights of value not just to hiPSC biologists, but also human geneticists, and so is a resource worthy of sharing in Cell Genomics.

We have added the missing functional validation as a limitation of the study.