Mapping origins of variation in neural trajectories of human pluripotent stem cells

Suel-Kee Kim[1,11,13], Seungmae Seo[1,13], Genevieve Stein-O'Brien[1,7,13], Amritha Jaishankar[1,13], Kazuya Ogawa[1], Nicola Micali[1,11], Yanhong Wang[1], Thomas M. Hyde[1,3,5], Joel E. Kleinman[1,3], Ty Voss[9], Elana J. Fertig[4], Joo-Heon Shin[1], Roland Bürli[10], Alan J. Cross[10], Nicholas J. Brandon[10], Daniel R. Weinberger[1,3,5,6,7], Joshua G. Chenoweth[1], Daniel J. Hoeppner[1], Nenad Sestan[11,12], Carlo Colantuoni[1,3,6,8,], Ronald D. McKay[1,2,*]*

[1]Lieber Institute for Brain Development, 855 North Wolfe Street, Baltimore, MD 21205

[2]Department of Cell Biology, [3]Department of Neurology, [4]Departments of Oncology, Biomedical Engineering, and Applied Mathematics and Statistics, [5]Department of Psychiatry, [6]Department of Neuroscience, [7]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD 21205

[8]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201

[9]Division of Preclinical Innovation, Nation Center for Advancing Translational Science / NIH, Bethesda, MD 20892

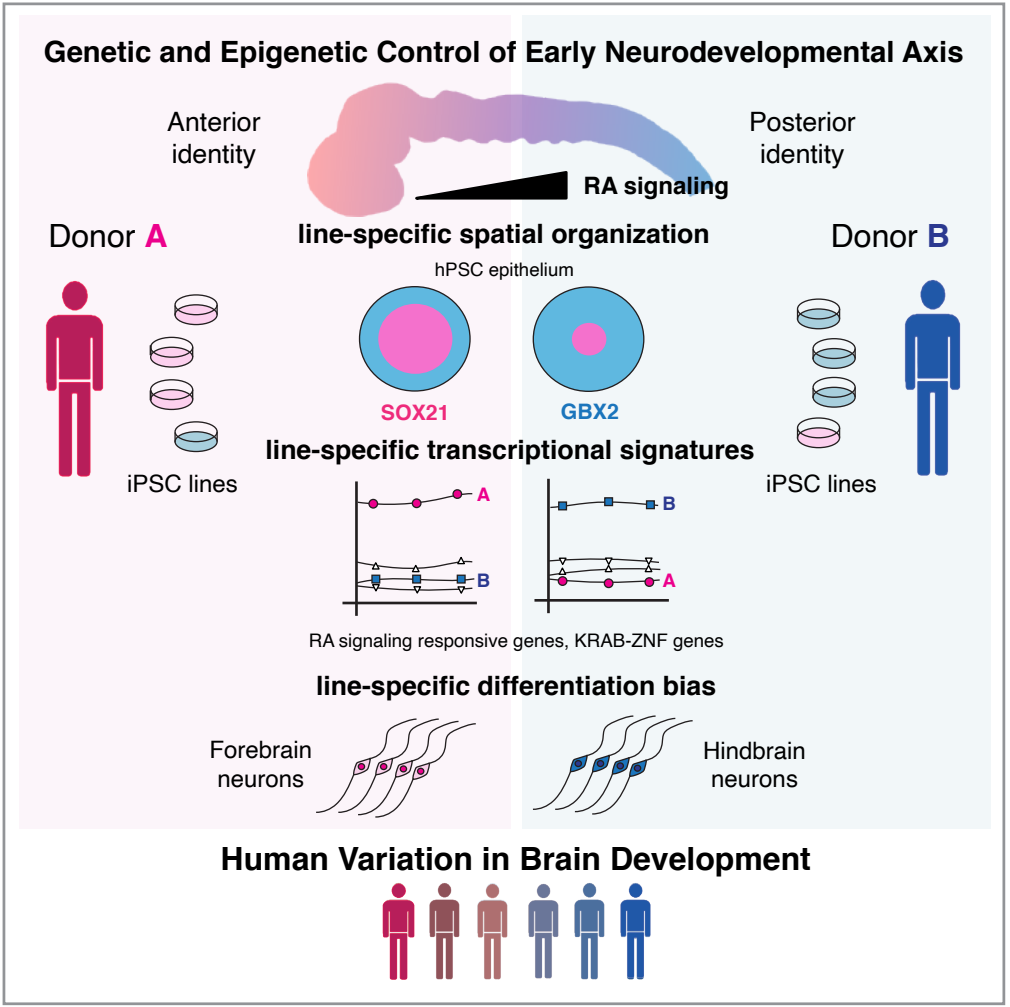
[10]Astra-Zeneca Neuroscience iMED., 141 Portland Street, Cambridge, MA 01239

[11]Department of Neuroscience, [12]Departments of Genetics, of Psychiatry and of Comparative Medicine, Kavli Institute for Neuroscience, Program in Cellular Neuroscience, Neurodegeneration and Repair, Child Study Center, Yale School of Medicine, New Haven, CT 06510, USA.

[13]Co-first author

*Corresponding authors: ccolantu@jhmi.edu (C.C.), ronald.mckay@libd.org (R.D.M.)

Lead contact: R. D. M.


Genetic and Epigenetic Control of Early Neurodevelopmental Axis
Anterior identity
Posterior identity
RA signaling
Donor A
line-specific spatial organization
Donor B
hPSC epithelium
SOX21
GBX2
iPSC lines
iPSC lines
line-specific transcriptional signatures
A
B
B
A
RA signaling responsive genes, KRAB-ZNF genes
line-specific differentiation bias
Forebrain neurons
Hindbrain neurons
Human Variation in Brain Development

SUMMARY

Variability between human pluripotent stem cell (hPSC) lines remains a challenge and opportunity in biomedicine. We identify differences in the spontaneous self-organization of individual hPSC lines during self-renewal that lie along a fundamental axis of in vivo development. Distinct stable and dynamic transcriptional elements were revealed by decomposition of RNA-seq data in pluripotency and early lineage emergence. Stable differences within pluripotency predicted regional bias in the dynamics of neural differentiation that were also observed in large collections of hPSC lines. Using replicate human induced PSC (hiPSC) lines and paired adult tissue, we demonstrate that cells from individual humans expressed unique transcriptional signatures that were maintained throughout life. In addition, replicate hiPSC lines from one donor showed divergent expression phenotypes driven by distinct chromatin states. These stable transcriptional states are under both genetic and epigenetic control and predict bias in subsequent forebrain and hindbrain neural trajectories. These results define mechanisms controlling transcription in pluripotency to first establish human neural diversity. Data availability: GSE164055; https://nemoanalytics.org/p?l=KimEtAL2021&g=GBX2.

Keywords: Pluripotent stem cells, RNA-seq, in vivo development, variation in neural fates, SOX21, GBX2, epigenetic control, human variation

INTRODUCTION

In mammalian embryos, the few hundred pluripotent cells of the epiblast execute spatially coordinated restrictions in cell state, generating distinct tissues (Arnold and Robertson, 2009; De Robertis, 2009). Human pluripotent stem cells (hPSCs) represent the epiblast state, acutely poised to establish multiple tissues (Bao et al., 2009; Brons et al., 2007; Guo et al., 2016; Tesar et al., 2007). Great attention is currently focused on using hPSCs to generate disease models that promise novel drug and cell therapies (Temple and Studer, 2017). Previous studies have focused on defining variation across hPSCs by exploring the genomes and transcriptomes of many lines (Carcamo-Orive et al., 2017; Choi et al., 2015; Cuomo et al., 2020; DeBoever et al., 2017; Kilpinen et al., 2017; Rouhani et al., 2014). However, we lack a detailed understanding of variation in the transitions that transform pluripotent cells into distinct neural stem cells (NSCs), complicating the application of stem cell technologies to neurological and psychiatric disorders.

The urgency to answer this question is illustrated by recent work reporting developmental differences between hPSC lines in the generation of neural precursors and their possible implications in the etiology of neurodevelopmental disorders (Burke et al., 2020; Kanton et al., 2019; Mariani et al., 2015; Micali et al., 2020; Strano et al., 2020; Wang et al., 2020). Here we employ cellular and bioinformatic approaches to define functional variation in hPSC lines during in vitro morphogenesis as they traverse the earliest events in forebrain development. High-resolution decomposition of transcription during hPSC differentiation revealed dynamic transcriptional change in lineage emergence that is shared between lines and stable gene expression phenotypes that are specific to individual hPSC lines and donors. We also show that these transcriptional signatures are regulated by both genetic and epigenetic mechanisms and demonstrate how they influence cellular trajectories in neural development.

RESULTS

Cell line variation in the emergence of neural fate from pluripotency

We used a monolayer culture system in which the emergence of early fates were explored over a short period of expansion following the passage of dissociated single cells (Figure 1A). We initially assessed the emergence of cellular heterogeneity by tracking the spatial distribution of the pluripotency regulators POU5F1 (OCT4) and NANOG in the epithelial sheets using human embryonic stem cell (hESC) line SA01. After 24 h, small groups of cells assembled, surrounded by cell-free regions. Automated cell mapping showed that over the following 3 days in self-renewal (SR) conditions, the pluripotency genes were preferentially expressed in cells on the edge of the epithelial sheets (Figure 1B), consistent with previous work on unconstrained (Hough et al., 2014), surface patterned (Warmflash et al., 2014), and microfluidic directed (Rifes et al., 2020) hPSC colony organization.

To determine if this spatial segregation was related to differentiation bias, cells were exposed to agonists or antagonists of BMP/TGFβ signaling, known to promote mesendodermal and neural fates, respectively (Chambers et al., 2009; Faial et al., 2015). Treatment of cells with BMP4 on day 0 (D0T) induced phosphorylation of SMAD1/5 and co-expression of its targets BRACHYURY (TBXT) and CDX2, which mediate differentiation to early embryonic and extraembryonic fates (Faial et al., 2015; Mendjan et al., 2014) across the entire cell population (Figure 1C). When BMP4 treatment was delayed until day 2 (D2T), the induction of these markers was restricted to the edge of the epithelial sheets (Figure 1C and S1A). However, cells in the core remained competent to respond to BMP4, as indicated by induced phosphorylation of SMAD2/3 (Figure S1A). Expression of SOX17 and GATA4, which mediate differentiation to primitive and definitive endoderm (Faial et al., 2015), suggested that these fates occurred in the core zone in response to BMP4 (Figure S1A). When cells were treated with Noggin and SB431542 (NSB), which induce neural differentiation, the neuroectodermal fate regulators SOX2, SOX21, and OTX2 (Kuzmichev

et al., 2012; Simeone et al., 2002) were expressed in the core of the epithelium while cells on the edge still expressed a high level of NANOG (Figure 1D and S1B). Recent studies have defined multiple lineage competent states that generate distinct embryonic and extraembryonic fates (Etoc et al., 2016; Guo et al., 2020; Nakanishi et al., 2019). Our results suggest that self-organizing processes in this 2-dimensional system direct germ layer emergence, recapitulating aspects of early neuroepithelial lineages.

To search for variation in these self-organizing processes between hPSC lines, the hESC line SA01 was compared with the hiPSC line i04 (Mallon et al., 2013). The two lines showed consistent differences in the formation of core and edge zones defined by larger core zones and greater expression of SOX21 and OTX2 in SA01 and larger edge zones and greater NANOG expression in i04 (Figure S1C). This cell line difference was invariant across doses of neuroectodermal inducers and initial cell plating densities, and was not due to differential proliferation rate between the lines (Figure S1D and S1E). When variation in BMP4-induced differentiation was assessed, i04 rapidly induced CDX2, while SA01 showed high TBXT expression (Figure S1F). These data demonstrate that functional variation in early fate bias can be defined in this system where dissociated pluripotent cells rapidly self-organize to form developmentally distinct fates.

Variation in the specification of early neural fates between multiple lines was demonstrated by consistent SOX21 protein expression differences between three hESC and three hiPSC lines during both SR and neuroectoderm (NSB) conditions (Figure 1E). RNA sequencing (RNA-seq) was performed at 2, 4, and 6 days of the SR, NSB, and BMP4 conditions for all lines (Tables S1). In principal component analysis (PCA), PC1 indicated that the major change in gene expression was associated with mesendodermal differentiation and PC2 showed ordered change with differentiation time in all conditions (Figure 1F).

All six lines exhibited similar general trajectories across PC1 and PC2, while cell line differences were also evident. Consistent with the SOX21 immunostaining, SA01 advanced the furthest along the NSB trajectory (Figure 1F and S1G). This cell line bias in differentiation could also be observed in the PCA of NSB samples alone (Figure 1G). To relate the dynamics described by NSB PC1 to expression within SR, we projected the SR data into the transcriptional space defined by NSB differentiation using projectR (Sharma et al., 2020). The same ranking of cell lines seen in NSB PC1 was present in SR, indicating that elements of gene expression differences defined in differentiation were already present in pluripotency, with SA01 leading the neural trajectory (Figure 1G). Similarly, the projection of SR data into BMP4 PC1 showed a common ranking of cell lines in both SR and differentiation, with i04 leading in this case (Figure S1H). These observations suggest that heterogeneity in gene expression within pluripotency is linked to bias in the early emergence of either neural or mesendodermal fates in hPSC lines.

Decomposing dynamic and stable transcription modules in early differentiation

To more thoroughly analyze the low dimensional transcriptional change across these cell lines and conditions, we employed the Genome-Wide Coordinated Gene Activity in Pattern Sets (GWCoGAPS) non-negative matrix factorization algorithm (Fertig and Favorov, 2010; Stein-O'Brien et al., 2017) to generate a set of 22 patterns that define gene expression dynamics across the samples (Figure S2A, GWCoGAPS-I, gene weights in Table S2). GWCoGAPS patterns are used in combination to represent the full expression pattern of each gene, thus decomposing multiple signals embedded in the expression of individual genes (Figure S2B). Two classes of pattern were identified. One class defined transcriptional trajectories that changed over time or condition, and the other defined transcriptional differences that were stable over time and condition but varied between cell lines (Figure 2A).

Of the 13 dynamic patterns (Figure 2A and S2A), three described aspects of pluripotency (P7, P17, and P22), six represented the response to BMP4 (P2, P3, P5, P6, P8, and P9), and three captured different temporal phases of the response to NSB (P4, P15, and P12). Top genes from

P7 contained the core pluripotency genes *POU5F1*, *SOX2*, and *NANOG*. Top genes from BMP4 patterns P3 and P9 included many early mesendodermal and extraembryonic fate regulators, such as *TBXT, EOMES*, and GATA family members. Top genes from NSB patterns P12 and P15 contained neuroectodermal regulators, such as *HES3*, *WNT1*, *OTX2*, *SOX21*, and *PAX6* (Figure 2B and Table S3).

To relate these transcriptional dynamics defined in vitro to gene expression change as lineage emerges during in vivo development, we projected single-cell RNA-seq (scRNA-seq) data from the developing mouse gastrula (Pijuan-Sala et al., 2019) into the GWCoGAPS-I patterns (Figure 2C). Consistent with the designation as a pluripotency module, pattern P7 showed the highest levels in cells of the pluripotent epiblast and decreased in all cell populations undergoing early germ layer commitment. In contrast, the BMP4 pattern P3 was not expressed in pluripotency and increased in later mesodermal lineages, including posterior primitive streak derivatives. Expression of *POU5F1* and *BAMBI*, the top-ranked genes in these patterns, respectively, are representative of these expression dynamics in lineage emergence (Figure S3Ai and S3Aii).

Patterns P15 and P12 showed the highest levels in neuroectodermal cells of the mouse embryo (Figure 2D) and parallel expression dynamics of the neural lineage drivers *SOX21* and *PAX6* (Figure S3Bi and S3Bii). The sequential induction of these early neural expression modules was also found in cortical neuron differentiation data from a different set of hiPSC lines (Burke et al., 2020) (Figure S3Biii). Projection of ChIP-seq data from differentiating hPSCs (Gifford et al., 2013) linked activating and repressing chromatin marks to these transcriptional modules, strengthening the view that they represent a dynamic readout of the transcriptional change regulating pluripotency and early germ layer lineages (Figure S3Av and S3Biv).

The GWCoGAPS decomposition also identified stable expression patterns that were invariant across time and treatment in each of the cell lines (Figure 2Ei, S2A, and S3Ci). Projection of microarray data from embryoid body differentiation of these same hPSC lines (Mallon et al., 2013)

showed that these line-specific patterns are stable features found in widely different culture conditions (Figure 2Eii and S3Cii). When RNA-seq data from other research groups using multiple hPSC lines, including the widely studied hESC line H9 (Choi et al., 2015; Kyttala et al., 2016; Rouhani et al., 2014), were projected into the H9-specific pattern P13, this cell line showed the strongest signal (Figure 2Eiii and S3D). Projection of DNA methylation data from these same lines showed that cell line-specific gene expression patterns were associated with hypomethylation at promoters over-expressed in the corresponding cell lines (Figure S3Ciii). These analyses suggest that dynamic patterns reflecting transcriptional change in differentiation are distinct from cell line-specific expression signatures, which are stable in different conditions.

SOX21 regulates early forebrain fate choice

Our data and the recent identification of time- and lineage-specific eQTLs that influence differentiation (Cuomo et al., 2020) suggest that the interaction between constitutive and dynamic control of gene expression is central to the lineage decisions that underlie cell line-specific differentiation bias. SOX21 is highly represented in the neuroectoderm initiating pattern P15 and is differentially expressed across hPSC lines in both SR and NSB conditions (Figure 1E and S1G), suggesting it may play a key role in generating telencephalic organizer states preceding neurogenesis (Micali et al., 2020). Previously, we have shown that SOX21 is required for reprogramming to the pluripotent state and regulates the anterior-posterior identity in the adult mouse intestine (Kuzmichev et al., 2012). Other laboratories have reported SOX21 roles in embryonic versus extraembryonic fate at the 4-cell stage (Goolam et al., 2016), later events in forebrain development (Fang et al., 2019), and adult hippocampal neurogenesis (Matsuda et al., 2012). These data suggest that SOX21 repeatedly interacts with other class B SOX genes (SOX1, 2, and 3) to regulate major morphogenetic decisions.

To better define the role of SOX21 in early neural specification, three SA01 SOX21-knockout (KO) lines were generated by CRISPR/Cas9 technology (Figure S4A). Immunostaining after 3 days of NSB treatment showed increased expression of NANOG, SOX2, and SOX3 in the

epithelial edge zone and increased NANOG in the core zone of SOX21-KO lines (Figure 3Ai and 3B). These results suggest the loss of SOX21 disrupts the normal spatial organization regulating the initial emergence of neural fates. RNA-seq data in these KO lines showed that while the overall pattern of gene expression conformed to the wild type (WT) cells, loss of SOX21 delayed neuroectodermal differentiation and accelerated mesodermal differentiation (Figure S4B). The expression of the top genes in the pluripotency and neuroectoderm patterns (P7, P15, and P12), along with NANOG and SOX1 protein levels, showed that even though loss of pluripotency genes occurred, the transition to neural fates was diminished in SOX21-KO cells (Figure 3C and S4C).

To further test the role of SOX21 in regulating the emergence of neural fates in the core zone, SOX21-KO cells were treated with BMP4 on day 2 when the core zone had already formed (Figure 3Aii). Consistent with a progressive restriction of responsiveness to BMP during the first days of SR observed in Figure 1C, WT cells showed only minimal induction of mesendodermal genes defined by BMP4 patterns P9 and P3 (Figure 3D). In contrast, many BMP4-induced genes were strongly upregulated in SOX21-KO cells. Immunostaining showed no difference between the WT and SOX21-KO lines in the expression of CDX2, GATA3, and ID1 in the edge zone, while primitive streak regulators TBXT and EOMES (Costello et al., 2011) were induced strongly in the core zone of SOX21-KO cells (Figure 3E). The mutually exclusive expression of *Sox21* in the rostral neuroectoderm versus *Cdx2* in the caudal mouse epiblast (Pijuan-Sala et al., 2019) is consistent with an early anteriorizing morphogenetic role of SOX21 (Figure S4Dv). Consistent with the recently defined role of CDX2 in neuromesodermal precursor fates (NMP) (Guibentif et al., 2020), the NMP transcriptional signature is mutually exclusive with SOX21 expression (Figure S4Dvi). Many genes in the NMP signature showed higher expression in SOX21-KO cells when treated with BMP4 at day 2 (Figure S4E). These results support a model where anterior neural fates arise by SOX21 acting to restrict the expression of anterior mesendodermal and neuromesodermal regulators in the core zone of the epithelium.

Cell line-specific transcriptional signatures underlie variation in forebrain versus hindbrain fate bias

In representing early forebrain differentiation, SOX21 expression and P15 captured consistent variation across the six cell lines both in SR and NSB conditions (Figure 1E, 2D, and S5A). To assess the source of this variation in neural differentiation, we focused on the distinct responses of SA01 and i04 to NSB treatment. The top genes defined by P15 included *OTX2* (Figure 2B and S1B), which is expressed in pluripotency and known to have critical functions in the anterior region of the early mouse embryo to pattern the forebrain (Tam and Steiner, 1999). The morphogenetic role of OTX2 is also seen in the formation of a precise boundary between the fore- and hind-brain, where it opposes a posteriorizing action of retinoic acid (RA) in a GBX2-dependent manner (Millet et al., 1999; Simeone et al., 2002). Interestingly, OTX2 and GBX2 were clearly differentially expressed between SA01 and i04 cells in both SR and neuroectoderm conditions (Figure 4A, S5B, and S5C). These results led us to investigate these specific genes and a larger group of genes known to be upregulated by RA (Balmer, 2002) in the cell line-specific transcriptional patterns (Figure 4B). The i04 line-specific pattern was significantly enriched for these RA response (RAR) genes, including *GBX2*. Consistent with the role of RA signaling in generating early hindbrain fates from hiPSCs (Frith et al., 2018), the RAR genes showed highest expression in the caudal epiblast and posterior neural fates of the mouse gastrula (Pijuan-Sala et al., 2019) (Figure S5D). This indicates that RA-driven regional lineage bias may be present at distinct levels in hPSC lines. The differential enrichment of RAR genes in the cell line-specific expression patterns suggested that these patterns could be used to predict the differentiation behavior of individual cell lines.

RA response in neural differentiation was interrogated across multiple cell lines by monitoring the expression of HOXB1, known to drive hindbrain cell fates (Krumlauf, 2016). In a RA dose-response study (Figure 4C), lines i04, i07, and i13 produced more HOXB1^{+} cells at day 8 compared to lines SA01, H9, and UC06 (Figure 4Di). RAR gene enrichment in the cell line-specific transcriptional signatures was strongly correlated with hindbrain fate potential

represented by HOXB1 levels (Figure 4Dii). Expression of OTX2, GBX2, HOXB1, and HOXB4 in neural precursors at day 8 shows that this differential response in lines i04 and SA01 is consistent with an early transition to anterior and posterior neural fates in vivo (Figure S5D and S5E). Later differentiation at day 28 showed a coordinated production of multiple hindbrain neuronal fates indicated by expression of OLIG2, ISL1, and PHOX2B (Figure 4E). Importantly, a comparison of WT and SOX21-KO lines following RA treatment demonstrated increased HOXB1 expression in the SOX21-KO line (Figure 4F). These data suggest that cell line-specific transcriptional signatures regulate the differential emergence of anterior versus posterior fates in these hPSC lines.

Genetic and epigenetic elements control transcriptional signatures that drive lineage bias

To further test the validity of these measures of hPSC variation, a new set of six hiPSC lines were generated from scalp fibroblasts of three donors (2053, 2063, and 2075) whose postmortem brain RNA-seq data was generated previously (Jaffe et al., 2018). RNA samples from each donor's replicate lines were collected at 2, 4, and 6 days under SR and neuroectoderm conditions for sequencing (Figure 5A). PCA and projection analysis demonstrated that these new cell lines followed the same overall transcription dynamics, observed in the previous lines studied (Figure S6A). In addition to these temporal lineage dynamics, GWCoGAPS analysis of this new RNA-seq data (GWCoGAPS-II, Table S4) identified transcriptional signatures that were stable across time and conditions for both replicate lines of each donor (Figure 5Bi), consistent with a genetic origin. Projection of RNA-seq data from the cerebral cortex of 260 individuals (Jaffe et al., 2018) into the donor-specific transcriptional signatures demonstrated that these expression traits were also present in the corresponding donor’s adult forebrain tissue (Figure 5Bii). To further explore the stability of these signatures, RNA-seq data from fibroblasts from these same donors were also obtained. Projection showed that the donor-specific transcriptional signatures identified in the pluripotent cells were also highest in the parental fibroblasts of the corresponding donors (Figure S6B). These expression studies across replicate hiPSC lines, fibroblasts, and mature

brain tissue suggest that donor-specific transcriptional signatures are stable gene expression traits of an individual human's cells throughout life and may have genetic origins.

Supporting this possibility, the gene amplitudes from the donor-specific signatures were enriched in genes involved in eQTLs that span multiple tissues in the GTEx Consortium data (p=1e-6 to p=5e-21) (GTEx_Consortium, 2015). To further explore possible genetic origins, we compared the strength of the donor-specific signatures across the brain RNA-seq collection with genetic similarity between donors using single nucleotide polymorphism (SNP) genotype data from the postmortem brain collection, which contains members of different ethnicities (Jaffe et al., 2018) (Figure 5C). The observed correlation of genetic similarity with the strength of projected donor-specific transcriptional signatures is consistent with a model in which genetic factors influence donor-specific expression signatures.

The modest magnitude of these correlations indicates that additional factors also play a role in determining these donor-specific transcriptional phenotypes. The GWCoGAPS-II patterns revealed a 2053-6 line-specific transcriptional signature in addition to the donor-specific signature (Figure 5D and S6C). Projection of the prefrontal cortex RNA-seq data showed that this signature was not present in this donor's brain, suggesting an epigenetic origin of this divergent line-specific transcriptional signature (Figure 5D, lower panel). Consistent with this possibility, projection of the 2053-6 line-specific signature into the postmortem brain RNA-seq data showed no correlation with genetic distance between donors (Figure 5E). These observations suggest that transcriptional signatures specific to hPSC lines can arise from distinct genetic and epigenetic origins.

To define variation in their neural differentiation trajectories, RNA-seq data from the new lines were projected into the GWCoGAPS patterns P15 and P12 from Figure 2D. In the neuroectoderm condition, all the new lines showed similar induction of these forebrain patterns, except 2053-6 (Figure 5F). Additionally, line 2053-6 showed less SOX21 induction in response to NSB than line

2053-2 (Figure 5G and S6D), similar to the differences previously observed between lines SA01 and i04 (Figure 1E). Consistent with the regional variation defined in Figure 4, RAR genes were enriched in the 2053-6 line-specific pattern, and this line generated more HOXB1^{hi} hindbrain cells in response to RA treatment (Figure 5H and S6E). This indicates that the decision to preferentially form forebrain versus hindbrain fates is being regulated epigenetically in these lines.

To find common transcriptional elements underlying this lineage bias across the two sets of lines, we compared differences in gene weights between the i04 and SA01 line-specific signatures with differences between the 2053 donor-specific and the 2053-6 line-specific signatures. No global correlation was observed, while GBX2 showed extremely high relative expression in lines i04 and 2053-6, which are both biased toward posterior neural fates (Figure 5I). This indicates that different RAR genes are highly expressed in lines i04 and 2053-6, but increased GBX2 in particular, could be responsible for their shared differentiation bias toward posterior fates.

To rule out the possibility that large-scale somatic or reprogramming-related genomic mutations underlie the observed differentiation bias between the 2053-2 and -6 lines, genome sequencing was performed on the two replicate lines and the mature postmortem brain tissue from donors 2053 and 2075. We identified variants gained and lost in each of these genomic DNA samples with respect to the reference genome (Figure S6F). For both donors, the majority of copy number variations (CNVs) were shared in all three samples, suggesting minimal genetic alternation during reprogramming. Furthermore, the numbers of CNVs observed uniquely in the 2053 lines did not differ significantly from those in the 2075 lines. These results suggest that the discordant expression traits and lineage bias between lines 2053-2 and -6 are not due to large scale genome differences.

Relevant to possible epigenetic mechanisms that could influence stable gene expression phenotypes, we found that KRAB-ZNF genes were significantly enriched in all line- and donor-specific transcriptional signatures that we have identified (Table S4 and S5). KRAB-ZNF genes

have been shown to repress the expression of transposable elements (TEs) during early development and structure lasting H3K9me3-mediated heterochromatin to regulate gene expression in mature tissues (Chuong et al., 2017; Ecco et al., 2016). These KRAB-ZNF genes are located in gene clusters that are stably expressed at distinct levels in hPSC lines across lineage and time - notably at different levels in lines derived from the same donor (Figure S6G and S6H). This pattern of expression suggests that KRAB-ZNF genes themselves are epigenetically co-regulated in pluripotent cells and their lineage-committed derivatives, consistent with the finding that KRAB-ZNF genes repress other KRAB-ZNF genes through this same H3K9me3 mechanism (Frietze et al., 2010). Surprisingly, the most differentially regulated ZNF genes in lines 2053-2 and 2053-6 (Figure S6H) do not show differential H3K9me3 levels (data not shown). These observations raise the possibility that early KRAB-ZNF-driven H3K9me3 and additional epigenetic mechanisms in pluripotency may shape lasting transcriptional phenotypes that influence cell function throughout life.

To explore this further, we generated H3K9me3 ChIP-seq data from hiPSC lines 2053-2 and 2053-6 in the SR condition. Projection of these data into the 2053-6 line-specific transcriptional signature indicated enrichment of this repressive epigenetic mark in line 2053-2, at promoters of genes over-expressed in line 2053-6 (Figure 5J). This is consistent with a model in which many specific genes have been de-repressed in line 2053-6 via loss of H3K9me3, leading to stable expression phenotypes. In particular, the GBX2 locus, which was highly represented in the 2053-6 line-specific signature, showed higher H3K9me3 levels in 2053-2 (Figure 5K). This distinct setting of heterochromatin between the lines in pluripotency indicates an early epigenetic origin for their divergent transcriptional phenotypes and resulting lineage bias.

Early developmental bias defines hPSC variation in the wider human population

To interrogate these genetic and epigenetic differences in a larger human population, we performed PCA on RNA-seq data from a hiPSC library of 317 lines from 101 donors generated by the NextGen Consortium (Carcamo-Orive et al., 2017). 91% of variance in PC1 is attributable to

differences across donors, indicating that this dominant feature of human transcriptional variation within pluripotency is under genetic control (Figure 6A). Importantly, PC1 was highly correlated with RAR gene expression, paralleling the differences we have observed here in smaller number of lines (Table S5). In addition, similar to the case of hPSC lines 2053-2 and 2053-6, some individual donor's iPSC lines occupied significantly divergent PC1 positions and RAR gene expression levels (Figure 6A, an example in blue circles), suggesting that epigenetic mechanisms also influence this transcriptional variation.

To define how differences between lines i04 and SA01 are distributed in the larger dataset, we compared the PC1 gene loadings with the differences in gene weights between the i04 and SA01 line-specific patterns (Figure 6B and Table S6). We observed a positive correlation when considering all genes and a stronger correlation within RAR genes. Consistent with this observation, line i04 and line 2053-6, which both showed posterior neural fate bias, exhibited highest levels of PC1 (Figure S7A). RAR genes showing the highest loadings in PC1 and stable high expression in line i04 (Figure 6B, in yellow) included genes shown to be highly expressed in self-renewing cells at the edge of pluripotent colonies (Hough et al., 2014) and in a recently defined pluripotent founder cell that shares transcriptional elements with the primitive endoderm (Nakanishi et al., 2019). GBX2 showed a very large difference between the stable line-specific patterns, but a more modestly high loading in PC1 (>98th%ile, Figure 6B, in green), indicating that the origins of heterogeneity in hPSC lines represented by this PC1 involve more than GBX2 alone.

Further insight into the specific role of RAR gene expression in relation to human neural variation was gained by comparing the NextGen hiPSC library PC1 loadings to gene weights in the early neuroectoderm pattern P15 (Figure 6C and Table S6). There was a strong negative correlation between the P15 weights and the PC1 loadings, corresponding to an effect opposing that seen in the i04 – SA01 differences. In addition, the bimodal distribution of RAR genes in P15 weights (Figure 6C, Y-axis, horizontal dashed line) defines the expression of two groups of RAR genes in

distinct neural and NMP (high in P15) or mesodermal and endodermal modules (low in P15) in the mouse gastrula (Figure 6D), corresponding to distinct RA-driven lineage biases in hPSC differentiation potential.

Projection of the NextGen hiPSC library PC1 onto the landscape of early mouse development reveals the lineage identity of variation in hPSC lines (Figure 6E). High PC1 represents mesodermal and endodermal signals in the mouse gastrula (in red to yellow). Further decomposition of these signals demonstrates that this lineage bias in hiPSC lines defined by high PC1 consists of distinct mesodermal (*COL3A1, PRRX1, ANXA1, INHBA*) and primitive endodermal (*RARB, APOA2, CER1, GATA4, GATA6, SOX17, EOMES*) identities (Figure 6E, S7B-D and Table S7). In striking contrast, low PC1 maps to the pluripotent epiblast and emerging neural lineages (Figure 6E, in purple to pink and Figure S7B-D) with high expression of *SOX3, LHX2, PAX6, EMX2, FEZF2,* and *RAX*. The expression patterns of these and other genes can be explored at nemoanalytics.org/p?l=KimEtAL2021&g=FEZF2, where we have used the gEAR and NeMO Analytics framework (Orvis et al., 2020) to construct an integrated gene expression data environment spanning the collection of in vivo and in vitro data examined in this study.

Our data indicate that in the epiblast, two cellular meta-states within pluripotency emerge into different lineages via distinct RAR gene induction routes: A high NANOG state initiates the primitive streak expressing MIXL1 and transits into mesendodermal lineages and, with a delay, a high SOX2/SOX3 state leads into neural lineages (Figure S7E). These states can be seen in the individual hPSC lines used here, in the 317 lines of the NextGen data, and in the individual cells of the mouse epiblast. These observations suggest that variation in hPSC lines lies along lineage-specific trajectories within the pluripotent epiblast that structure the earliest axes of the developing mammalian embryo.

DISCUSSION

Following the discovery of reprogramming factors (Takahashi and Yamanaka, 2006), functional heterogeneity in hiPSC differentiation potential has been a continuing focus of interest (Hu et al., 2010; Kanton et al., 2019; Mariani et al., 2015; Micali et al., 2020; Strano et al., 2020; Wang et al., 2020). Initial studies noted functional consequences of epigenetic memory of parental cell types in reprogrammed cells (Cahan and Daley, 2013; Kim et al., 2010; Kim et al., 2011). More recently, genetic factors have been shown to be central in determining transcriptional variation across hiPSC lines (Choi et al., 2015; Kyttala et al., 2016; Rouhani et al., 2014).

In this work, we define dynamic lineage-driving transcriptional modules, which are influenced by stable gene expression traits that hold pluripotent cells in distinct lineage-biased states. The bias we observed in the emergence of the neural lineage from pluripotency in a small number of hPSC lines was also seen in the primary dimension of transcriptional variation across hundreds of hiPSC lines prior to differentiation. These observations indicate that differences in the interplay of SOX21- and GBX2-dependent mechanisms during pluripotency regulate regional neural fate specification. The low variance within donors suggests aspects of this lineage bias are under genetic control. These findings contribute a novel perspective on a growing body of research that is beginning to reveal the genetic origins of heterogeneity in transcription and differentiation potential of hPSCs (Bonder et al., 2021; Carcamo-Orive et al., 2017; Cuomo et al., 2020; DeBoever et al., 2017; Jerber et al., 2021; Kilpinen et al., 2017; Ortmann et al., 2020; Strano et al., 2020; Vigilante et al., 2019).

In addition, our observation of bias toward divergent forebrain and hindbrain neural fates in replicate lines from the same donor suggests that epigenetic mechanisms also contribute significantly to variation in this early neural fate choice. Our data support a model where the key posterior fate regulator GBX2 is repressed in the forebrain biased state and the transcriptional signature driving hindbrain fates emerges in cells in which the GBX2 locus is not epigenetically repressed. This suggests the existence of alternate epigenetically predisposed states within

pluripotency prior to the high-level transcriptional output that implements these early regional neural fate choices. This sequential mechanism parallels recent findings in the mouse gastrula, demonstrating epigenetic structuring before transcriptionally defined cell identity (Grosswendt et al., 2020; Pijuan-Sala et al., 2020).

The transcriptional architecture of self-organization that we define here indicates that signatures of regional embryonic organizers exist within stem cells and precede classical anatomical organizer structures. We have previously reported a regional neural fate bias spanning the dorso-ventral (D/V) telencephalic axis, which was also divergent in hiPSC lines from a single donor (Micali et al., 2020). We showed that this lineage bias in neural stem cells was the result of the differential emergence of transcriptional signatures that lead to the formation of the cortical hem and other organizing centers. This D/V specification in neural development has been repeatedly linked to risk for neuropsychiatric disease (Chen et al., 2014; Madison et al., 2015; Marchetto et al., 2017; Mariani et al., 2015). Recent scRNA-seq data in cerebral organoids has demonstrated striking bias in this regional neural fate choice across human donors (Kanton et al., 2019), and variation in this D/V choice regulated by differential WNT signaling has also been observed across many hiPSC lines from numerous donors (Strano et al., 2020). In this study, we defined stable cell line-specific expression traits that regulate lineage bias in pluripotency. This understanding suggests constitutive control of gene expression results in the differential emergence of organizer signals (Micali et al., 2020) as well as intrinsically distinct levels of neural signaling pathways (Strano et al., 2020). Further exploration of stable expression phenotypes within pluripotency may reveal the earliest genetic and epigenetic architecture of human neural variation.

Our results showing KRAB-ZNF gene enrichment and H3K9me3-mediated regulation in cell line-specific signatures suggest possible mechanisms controlling these stable expression phenotypes. Evolutionarily recent transposable elements and their corresponding KRAB-ZNF repressor genes have been directly implicated in both human zygotic genome activation and later tissue-specific

gene regulation (Pontis et al., 2019), suggesting that these epigenetic events in early development determine lasting gene expression traits. The exploration of how and when these stable expression phenotypes and corresponding lineage-predisposing states are generated and stably set in vivo and in vitro will enable a deeper understanding of human variation and further the discovery potential of hPSCs. The genetics of neurodevelopmental disorders (Deneault et al., 2018; Klaus et al., 2019; Paulsen et al., 2020; Sawada et al., 2020) and human brain evolution (Kanton et al., 2019; Mora-Bermudez et al., 2016; Pollen et al., 2019; Trujillo et al., 2021) are being integrated into cerebral organoid models that parallel in utero and early postnatal brain development (Gordon et al., 2021; Trevino et al., 2020). The developmental perspectives we put forward here will inform molecular phenotyping and assay development as these in vitro models of brain development become increasingly complex and the implications of distant endpoint assays more uncertain.

Our observations raise the question of how many dimensions of hPSC variation can be mapped across in vivo development, and how they are mechanistically and functionally interrelated. Recent progress in the selection of specific hPSC lines for stem cell therapeutics (Kim et al., 2021; Piao et al., 2021) stresses the continuing value of further defining the genetic and epigenetic control of human cellular variation as iPSCs adopt neural fates. Novel approaches that define single-cell genome mosaicism are being combined with lineage tracing (Breuss et al., 2020; Chan et al., 2019; He et al., 2020; Kalhor et al., 2018; Rodin et al., 2020; Spanjaard et al., 2018) to provide a new understanding of the importance of cellular selection in neural development (Fasching et al., 2020; Wang et al., 2020). Together with these approaches, our findings layout a powerful perspective to guide the interrogation of large collections of hiPSC lines from many more genetically diverse donors to define human variation in early cellular states that impact brain development to modify complex human traits and disease risk.

AUTHOR CONTRIBUTIONS

S.-K.K., S.S., G.SO., A.J., J.G.C., D.J.H., C.C., and R.D.M. conceived the study. S.-K.K., S.S., and Y.W. performed cell culture and differentiation. Y.W. generated iPSC lines. K.O. generated SOX21-KO lines. S.-K.K., A.J., Y.W., and K.O generated RNA-seq data. S.S. generated ChIP-seq data. G.SO., S.S., E.J.F., C.C., and J.-H.S. developed and applied informatic methods to analyze RNA-seq data. S.S. and C.C. analyzed ChIP-seq data. T.M.H., J.K., and D.R.W. provided fibroblasts for iPSC line generation and human brain tissue data. S.-K.K., S.S., and N.M. performed immunocytochemistry. S.-K.K., S.S., A.J., T.V., and D.J.H. performed high-content image analysis. S.-K.K., S.S., G.SO., N.M, J.G.C., D.J.H., N.S., C.C., and R.D.M. interpreted the data. N.S., R.B., A.J.C., N.J.B., D.R.W., and R.D.M. directed the research. S.-K.K., S.S., C.C., and R.D.M. wrote the manuscript. All authors participated in the discussion of results and manuscript editing.

ACKNOWLEDGEMENTS

We thank the Lieber and Maltz families for their generous support of this work at the Lieber Institute for Brain Development (LIBD). This work was also supported by NCI/NIH grants R01CA177669, P30CA006973, U01CA212007, and U01CA253403 and the Johns Hopkins University Catalyst Award awarded to E.J.F., and R01NS095654, R01HG010898, MH116488, and U01MH124619 awarded to N.S. We thank J. Jessee, MTI-GlobalStem, for technical support of CRISPR/Cas9 plasmid transfection. We thank many members of LIBD and the Sestan lab for their helpful comments on this work.

REFERENCES

Arnold, S.J., and Robertson, E.J. (2009). Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. Nature reviews Molecular cell biology *10*, 91-103.

Balmer, J.E. (2002). Gene expression regulation by retinoic acid. The Journal of Lipid Research *43*, 1773-1808.

Bao, S., Tang, F., Li, X., Hayashi, K., Gillich, A., Lao, K., and Surani, M.A. (2009). Epigenetic reversion of post-implantation epiblast to pluripotent embryonic stem cells. Nature *461*, 1292-1295.

Bonder, M.J., Smail, C., Gloudemans, M.J., Fresard, L., Jakubosky, D., D'Antonio, M., Li, X., Ferraro, N.M., Carcamo-Orive, I., Mirauta, B.*, et al.* (2021). Identification of rare and common regulatory variants in pluripotent cells using population-scale transcriptomics. Nat Genet *53*, 313-321.

Breuss, M.W., Yang, X., Antaki, D., Schlachetzki, J.C.M., Lana, A.J., Xu, X., Chai, G., Stanley, V., Song, Q., Newmeyer, T.F.*, et al.* (2020). Somatic mosaicism in the mature brain reveals clonal cellular distributions during cortical development. bioRxiv, 2020.2008.2010.244814.

Brons, I.G., Smithers, L.E., Trotter, M.W., Rugg-Gunn, P., Sun, B., Chuva de Sousa Lopes, S.M., Howlett, S.K., Clarkson, A., Ahrlund-Richter, L., Pedersen, R.A.*, et al.* (2007). Derivation of pluripotent epiblast stem cells from mammalian embryos. Nature *448*, 191-195.

Burke, E.E., Chenoweth, J.G., Shin, J.H., Collado-Torres, L., Kim, S.K., Micali, N., Wang, Y., Colantuoni, C., Straub, R.E., Hoeppner, D.J.*, et al.* (2020). Dissecting transcriptomic signatures of neuronal differentiation and maturation using iPSCs. Nat Commun *11*, 462.

Cahan, P., and Daley, G.Q. (2013). Origins and implications of pluripotent stem cell variability and heterogeneity. Nat Rev Mol Cell Biol *14*, 357-368.

Carcamo-Orive, I., Hoffman, G.E., Cundiff, P., Beckmann, N.D., D'Souza, S.L., Knowles, J.W., Patel, A., Papatsenko, D., Abbasi, F., Reaven, G.M.*, et al.* (2017). Analysis of Transcriptional Variability in a Large Human iPSC Library Reveals Genetic and Non-genetic Determinants of Heterogeneity. Cell stem cell *20*, 518-532 e519.

Chambers, S.M., Fasano, C.A., Papapetrou, E.P., Tomishima, M., Sadelain, M., and Studer, L. (2009). Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nature biotechnology *27*, 275-280.

Chan, M.M., Smith, Z.D., Grosswendt, S., Kretzmer, H., Norman, T.M., Adamson, B., Jost, M., Quinn, J.J., Yang, D., Jones, M.G.*, et al.* (2019). Molecular recording of mammalian embryogenesis. Nature *570*, 77-82.

Chen, H.M., DeLong, C.J., Bame, M., Rajapakse, I., Herron, T.J., McInnis, M.G., and O'Shea, K.S. (2014). Transcripts involved in calcium signaling and telencephalic neuronal fate are altered in induced pluripotent stem cells from bipolar disorder patients. Transl Psychiatry *4*, e375.

Choi, J., Lee, S., Mallard, W., Clement, K., Tagliazucchi, G.M., Lim, H., Choi, I.Y., Ferrari, F., Tsankov, A.M., Pop, R.*, et al.* (2015). A comparison of genetically matched cell lines reveals the equivalence of human iPSCs and ESCs. Nature biotechnology *33*, 1173-1181.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet *18*, 71-86.

Costello, I., Pimeisl, I.M., Drager, S., Bikoff, E.K., Robertson, E.J., and Arnold, S.J. (2011). The T-box transcription factor Eomesodermin acts upstream of Mesp1 to specify cardiac mesoderm during mouse gastrulation. Nature cell biology *13*, 1084-1091.

Cuomo, A.S.E., Seaton, D.D., McCarthy, D.J., Martinez, I., Bonder, M.J., Garcia-Bernardo, J., Amatya, S., Madrigal, P., Isaacson, A., Buettner, F.*, et al.* (2020). Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. Nature Communications *11*, 810.

De Robertis, E.M. (2009). Spemann's organizer and the self-regulation of embryonic fields. Mechanisms of development *126*, 925-941.

DeBoever, C., Li, H., Jakubosky, D., Benaglio, P., Reyna, J., Olson, K.M., Huang, H., Biggs, W., Sandoval, E., D'Antonio, M.*, et al.* (2017). Large-Scale Profiling Reveals the Influence of Genetic Variation on Gene Expression in Human Induced Pluripotent Stem Cells. Cell stem cell *20*, 533-546 e537.

Deneault, E., White, S.H., Rodrigues, D.C., Ross, P.J., Faheem, M., Zaslavsky, K., Wang, Z., Alexandrova, R., Pellecchia, G., Wei, W.*, et al.* (2018). Complete Disruption of Autism-Susceptibility Genes by Gene Editing Predominantly Reduces Functional Connectivity of Isogenic Human Neurons. Stem Cell Reports *11*, 1211-1225.

Ecco, G., Cassano, M., Kauzlaric, A., Duc, J., Coluccio, A., Offner, S., Imbeault, M., Rowe, Helen M., Turelli, P., and Trono, D. (2016). Transposable Elements and Their KRAB-ZFP Controllers Regulate Gene Expression in Adult Tissues. Developmental Cell *36*, 611-623.

Etoc, F., Metzger, J., Ruzo, A., Kirst, C., Yoney, A., Ozair, M.Z., Brivanlou, A.H., and Siggia, E.D. (2016). A Balance between Secreted Inhibitors and Edge Sensing Controls Gastruloid Self-Organization. Dev Cell *39*, 302-315.

Faial, T., Bernardo, A.S., Mendjan, S., Diamanti, E., Ortmann, D., Gentsch, G.E., Mascetti, V.L., Trotter, M.W., Smith, J.C., and Pedersen, R.A. (2015). Brachyury and SMAD signalling collaboratively orchestrate distinct mesoderm and endoderm gene regulatory networks in differentiating human embryonic stem cells. Development *142*, 2121-2135.

Fang, Z., Liu, X., Wen, J., Tang, F., Zhou, Y., Jing, N., and Jin, Y. (2019). SOX21 Ensures Rostral Forebrain Identity by Suppression of WNT8B during Neural Regionalization of Human Embryonic Stem Cells. Stem Cell Reports *13*, 1-15.

Fasching, L., Jang, Y., Tomasi, S., Schreiner, J., Tomasini, L., Brady, M., Bae, T., Sarangi, V., Vasmatzis, N., Wang, Y.*, et al.* (2020). Early developmental asymmetries in cell lineage trees in living individuals. bioRxiv, 2020.2008.2024.265751.

Fertig, E.J.D., J., and Favorov, A.V.P., G. Ochs, M. F. (2010). CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data. Bioinformatics *26*, 2792-2793.

Frietze, S., O'Geen, H., Blahnik, K.R., Jin, V.X., and Farnham, P.J. (2010). ZNF274 recruits the histone methyltransferase SETDB1 to the 3' ends of ZNF genes. PLoS One *5*, e15082.

Frith, T.J., Granata, I., Wind, M., Stout, E., Thompson, O., Neumann, K., Stavish, D., Heath, P.R., Ortmann, D., Hackland, J.O.*, et al.* (2018). Human axial progenitors generate trunk neural crest cells in vitro. Elife *7*, e35786.

Gifford, C.A., Ziller, M.J., Gu, H., Trapnell, C., Donaghey, J., Tsankov, A., Shalek, A.K., Kelley, D.R., Shishkin, A.A., Issner, R.*, et al.* (2013). Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. Cell *153*, 1149-1163.

Goolam, M., Scialdone, A., Graham, Sarah J.L., Macaulay, Iain C., Jedrusik, A., Hupalowska, A., Voet, T., Marioni, John C., and Zernicka-Goetz, M. (2016). Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. Cell *165*, 61-74.

Gordon, A., Yoon, S.J., Tran, S.S., Makinson, C.D., Park, J.Y., Andersen, J., Valencia, A.M., Horvath, S., Xiao, X., Huguenard, J.R.*, et al.* (2021). Long-term maturation of human cortical organoids matches key early postnatal transitions. Nat Neurosci *24*, 331-342.

Grosswendt, S., Kretzmer, H., Smith, Z.D., Kumar, A.S., Hetzel, S., Wittler, L., Klages, S., Timmermann, B., Mukherji, S., and Meissner, A. (2020). Epigenetic regulator function through mouse gastrulation. Nature *584*, 102-108.

GTEx_Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science *348*, 648-660.

Guibentif, C., Griffiths, J.A., Imaz-Rosshandler, I., Ghazanfar, S., Nichols, J., Wilson, V., Gottgens, B., and Marioni, J.C. (2020). Diverse Routes toward Early Somites in the Mouse Embryo. Dev Cell *56*, 141-153.e146.

Guo, G., Stirparo, G.G., Strawbridge, S., Spindlow, D., Yang, J., Clarke, J., Dattani, A., Yanagida, A., Li, M.A., Myers, S.*, et al.* (2020). Human Naïve Epiblast Cells Possess Unrestricted Lineage Potential. bioRxiv, 2020.2002.2004.933812.

Guo, G., von Meyenn, F., Santos, F., Chen, Y., Reik, W., Bertone, P., Smith, A., and Nichols, J. (2016). Naive Pluripotent Stem Cells Derived Directly from Isolated Cells of the Human Inner Cell Mass. Stem Cell Reports *6*, 437-446.

He, Z., Gerber, T., Maynard, A., Jain, A., Petri, R., Santel, M., Ly, K., Sidow, L., Sanchís-Calleja, F., Riesenberg, S.*, et al.* (2020). Lineage recording reveals dynamics of cerebral organoid regionalization. bioRxiv, 2020.2006.2019.162032.

Hough, Shelley R., Thornton, M., Mason, E., Mar, Jessica C., Wells, Christine A., and Pera, Martin F. (2014). Single-Cell Gene Expression Profiles Define Self-Renewing, Pluripotent, and Lineage Primed States of Human Pluripotent Stem Cells. Stem Cell Reports *22*, 881-895.

Hu, B.Y., Weick, J.P., Yu, J., Ma, L.X., Zhang, X.Q., Thomson, J.A., and Zhang, S.C. (2010). Neural differentiation of human induced pluripotent stem cells follows developmental principles but with variable potency. Proc Natl Acad Sci U S A *107*, 4335-4340.

Jaffe, A.E., Straub, R.E., Shin, J.H., Tao, R., Gao, Y., Collado-Torres, L., Kam-Thong, T., Xi, H.S., Quan, J., Chen, Q.*, et al.* (2018). Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. Nat Neurosci *21*, 1117-1125.

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.

Kalhor, R., Kalhor, K., Mejia, L., Leeper, K., Graveline, A., Mali, P., and Church, G.M. (2018). Developmental barcoding of whole mouse via homing CRISPR. Science *361*, eaat9804.

Kanton, S., Boyle, M.J., He, Z., Santel, M., Weigert, A., Sanchis-Calleja, F., Guijarro, P., Sidow, L., Fleck, J.S., Han, D.*, et al.* (2019). Organoid single-cell genomic atlas uncovers human-specific features of brain development. Nature *574*, 418-422.

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.

Kim, K., Doi, A., Wen, B., Ng, K., Zhao, R., Cahan, P., Kim, J., Aryee, M.J., Ji, H., Ehrlich, L.I.*, et al.* (2010). Epigenetic memory in induced pluripotent stem cells. Nature *467*, 285-290.

Kim, K., Zhao, R., Doi, A., Ng, K., Unternaehrer, J., Cahan, P., Huo, H., Loh, Y.H., Aryee, M.J., Lensch, M.W.*, et al.* (2011). Donor cell type can influence the epigenome and differentiation potential of human induced pluripotent stem cells. Nat Biotechnol *29*, 1117-1119.

Kim, T.W., Piao, J., Koo, S.Y., Kriks, S., Chung, S.Y., Betel, D., Socci, N.D., Choi, S.J., Zabierowski, S., Dubose, B.N.*, et al.* (2021). Biphasic Activation of WNT Signaling Facilitates the Derivation of Midbrain Dopamine Neurons from hESCs for Translational Use. Cell Stem Cell *28*, 343-355 e345.

Klaus, J., Kanton, S., Kyrousi, C., Ayo-Martin, A.C., Di Giaimo, R., Riesenberg, S., O'Neill, A.C., Camp, J.G., Tocco, C., Santel, M.*, et al.* (2019). Altered neuronal migratory trajectories in human cerebral organoids derived from individuals with neuronal heterotopia. Nat Med *25*, 561-568.

Krumlauf, R. (2016). Hox Genes and the Hindbrain: A Study in Segments. Current topics in developmental biology *116*, 581-596.

Kuzmichev, A.N., Kim, S.K., D'Alessio, A.C., Chenoweth, J.G., Wittko, I.M., Campanati, L., and McKay, R.D. (2012). Sox2 acts through Sox21 to regulate transcription in pluripotent and differentiated cells. Curr Biol *22*, 1705-1710.

Kyttala, A., Moraghebi, R., Valensisi, C., Kettunen, J., Andrus, C., Pasumarthy, K.K., Nakanishi, M., Nishimura, K., Ohtaka, M., Weltner, J.*, et al.* (2016). Genetic Variability Overrides the Impact of Parental Cell Type and Determines iPSC Differentiation Potential. Stem Cell Reports *6*, 200-212.

Madison, J.M., Zhou, F., Nigam, A., Hussain, A., Barker, D.D., Nehme, R., van der Ven, K., Hsu, J., Wolf, P., Fleishman, M.*, et al.* (2015). Characterization of bipolar disorder patient-specific induced pluripotent stem cells from a family reveals neurodevelopmental and mRNA expression abnormalities. Mol Psychiatry *20*, 703-717.

Mallon, B.S., Chenoweth, J.G., Johnson, K.R., Hamilton, R.S., Tesar, P.J., Yavatkar, A.S., Tyson, L.J., Park, K., Chen, K.G., Fann, Y.C.*, et al.* (2013). StemCellDB: the human pluripotent stem cell database at the National Institutes of Health. Stem cell research *10*, 57-66.

Marchetto, M.C., Belinson, H., Tian, Y., Freitas, B.C., Fu, C., Vadodaria, K.C., Beltrao-Braga, P.C., Trujillo, C.A., Mendes, A.P., Padmanabhan, K.*, et al.* (2017). Altered proliferation and networks in neural cells derived from idiopathic autistic individuals. Mol Psychiatry *22*, 820-835.

Mariani, J., Coppola, G., Zhang, P., Abyzov, A., Provini, L., Tomasini, L., Amenduni, M., Szekely, A., Palejev, D., Wilson, M.*, et al.* (2015). FOXG1-Dependent Dysregulation of GABA/Glutamate Neuron Differentiation in Autism Spectrum Disorders. Cell *162*, 375-390.

Matsuda, S., Kuwako, K., Okano, H.J., Tsutsumi, S., Aburatani, H., Saga, Y., Matsuzaki, Y., Akaike, A., Sugimoto, H., and Okano, H. (2012). Sox21 promotes hippocampal adult neurogenesis via the transcriptional repression of the Hes5 gene. J Neurosci *32*, 12543-12557.

Mendjan, S., Mascetti, V.L., Ortmann, D., Ortiz, M., Karjosukarso, D.W., Ng, Y., Moreau, T., and Pedersen, R.A. (2014). NANOG and CDX2 Pattern Distinct Subtypes of Human Mesoderm during Exit from Pluripotency. Cell Stem Cell *15*, 310-325.

Micali, N., Kim, S.-K., Diaz-Bustamante, M., Stein-O'Brien, G., Seo, S., Shin, J.-H., Rash, B.G., Ma, S., Wang, Y., Olivares, N.A.*, et al.* (2020). Variation of Human Neural Stem Cells Generating Organizer States In Vitro before Committing to Cortical Excitatory or Inhibitory Neuronal Fates. Cell Reports *31*, 107599.

Millet, S., Campbell, K., Epstein, D.J., Losos, K., Harris, E., and Joyner, A. (1999). A role for Gbx2 in repression of Otx2 and positioning the mid_hindbrain organizer. Nature *401*, 161-164.

Mora-Bermudez, F., Badsha, F., Kanton, S., Camp, J.G., Vernot, B., Kohler, K., Voigt, B., Okita, K., Maricic, T., He, Z.*, et al.* (2016). Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. Elife *5*, e18683.

Nakanishi, M., Mitchell, R.R., Benoit, Y.D., Orlando, L., Reid, J.C., Shimada, K., Davidson, K.C., Shapovalova, Z., Collins, T.J., Nagy, A.*, et al.* (2019). Human Pluripotency Is Initiated and Preserved by a Unique Subset of Founder Cells. Cell *177*, 1-15.

Ortmann, D., Brown, S., Czechanski, A., Aydin, S., Muraro, D., Huang, Y., Tomaz, R.A., Osnato, A., Canu, G., Wesley, B.T.*, et al.* (2020). Naive Pluripotent Stem Cells Exhibit Phenotypic Variability that Is Driven by Genetic Variation. Cell Stem Cell *27*, 470-481.e476.

Orvis, J., Gottfried, B., Kancherla, J., Adkins, R.S., Song, Y., Dror, A.A., Olley, D., Rose, K., Chrysostomou, E., Kelly, M.C.*, et al.* (2020). gEAR: gene Expression Analysis Resource portal for community-driven, multi-omic data exploration. bioRxiv.

Paulsen, B., Velasco, S., Kedaigle, A.J., Pigoni, M., Quadrato, G., Deo, A., Adiconis, X., Uzquiano, A., Kim, K., Simmons, S.K.*, et al.* (2020). Human brain organoids reveal accelerated development of cortical neuron classes as a shared feature of autism risk genes. bioRxiv.

Piao, J., Zabierowski, S., Dubose, B.N., Hill, E.J., Navare, M., Claros, N., Rosen, S., Ramnarine, K., Horn, C., Fredrickson, C.*, et al.* (2021). Preclinical Efficacy and Safety of a Human Embryonic Stem Cell-Derived Midbrain Dopamine Progenitor Product, MSK-DA01. Cell Stem Cell *28*, 217-229 e217.

Pijuan-Sala, B., Griffiths, J.A., Guibentif, C., Hiscock, T.W., Jawaid, W., Calero-Nieto, F.J., Mulas, C., Ibarra-Soria, X., Tyser, R.C.V., Ho, D.L.L.*, et al.* (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. Nature *566*, 490-495.

Pijuan-Sala, B., Wilson, N.K., Xia, J., Hou, X., Hannah, R.L., Kinston, S., Calero-Nieto, F.J., Poirion, O., Preissl, S., Liu, F.*, et al.* (2020). Single-cell chromatin accessibility maps reveal regulatory programs driving early mouse organogenesis. Nat Cell Biol *22*, 487-497.

Pollen, A.A., Bhaduri, A., Andrews, M.G., Nowakowski, T.J., Meyerson, O.S., Mostajo-Radji, M.A., Di Lullo, E., Alvarado, B., Bedolli, M., Dougherty, M.L.*, et al.* (2019). Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. Cell *176*, 743-756 e717.

Pontis, J., Planet, E., Offner, S., Turelli, P., Duc, J., Coudray, A., Theunissen, T.W., Jaenisch, R., and Trono, D. (2019). Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. Cell Stem Cell *24*, 724-735 e725.

Rifes, P., Isaksson, M., Rathore, G.S., Aldrin-Kirk, P., Moller, O.K., Barzaghi, G., Lee, J., Egerod, K.L., Rausch, D.M., Parmar, M.*, et al.* (2020). Modeling neural tube development by differentiation of human embryonic stem cells in a microfluidic WNT gradient. Nat Biotechnol *38*, 1265-1273.

Rodin, R.E., Dou, Y., Kwon, M., Sherman, M.A., D'Gama, A.M., Doan, R.N., Rento, L.M., Girskis, K.M., Bohrson, C.L., Kim, S.N.*, et al.* (2020). The Landscape of Mutational Mosaicism in Autistic and Normal Human Cerebral Cortex. bioRxiv, 2020.2002.2011.944413.

Rouhani, F., Kumasaka, N., de Brito, M.C., Bradley, A., Vallier, L., and Gaffney, D. (2014). Genetic background drives transcriptional variation in human induced pluripotent stem cells. PLoS genetics *10*, e1004432.

Sawada, T., Chater, T.E., Sasagawa, Y., Yoshimura, M., Fujimori-Tonou, N., Tanaka, K., Benjamin, K.J.M., Paquola, A.C.M., Erwin, J.A., Goda, Y.*, et al.* (2020). Developmental excitation-inhibition

imbalance underlying psychoses revealed by single-cell analyses of discordant twins-derived cerebral organoids. Mol Psychiatry *25*, 2695-2711.

Sharma, G., Colantuoni, C., Goff, L.A., Fertig, E.J., and Stein-O'Brien, G. (2020). projectR: An R/Bioconductor package for transfer learning via PCA, NMF, correlation, and clustering. Bioinformatics *36*, 3592-3593.

Simeone, A., Puelles, E., and Acampora, D. (2002). The Otx family. Curr Opin Genet Dev *12*, 409-415.

Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., and Junker, J.P. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. Nat Biotechnol *36*, 469-473.

Stein-O'Brien, G.L., Carey, J.L., Lee, W.S., Considine, M., Favorov, A.V., Flam, E., Guo, T., Li, S., Marchionni, L., Sherman, T.*, et al.* (2017). PatternMarkers & GWCoGAPS for novel data-driven biomarkers via whole transcriptome NMF. Bioinformatics, 1-3.

Strano, A., Tuck, E., Stubbs, V.E., and Livesey, F.J. (2020). Variable Outcomes in Neural Differentiation of Human PSCs Arise from Intrinsic Differences in Developmental Signaling Pathways. Cell Reports *31*, 107732.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663-676.

Tam, P.P., and Steiner, K. (1999). Anterior patterning by synergistic activity of the early gastrula organizer and the anterior germ layer tissues of the mouse embryo. Development *126*, 5171-5179.

Temple, S., and Studer, L. (2017). Lessons Learned from Pioneering Neural Stem Cell Studies. Stem Cell Reports *8*, 191-193.

Tesar, P.J., Chenoweth, J.G., Brook, F.A., Davies, T.J., Evans, E.P., Mack, D.L., Gardner, R.L., and McKay, R.D. (2007). New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature *448*, 196-199.

Trevino, A.E., Sinnott-Armstrong, N., Andersen, J., Yoon, S.-J., Huber, N., Pritchard, J.K., Chang, H.Y., Greenleaf, W.J., and Pașca, S.P. (2020). Chromatin accessibility dynamics in a model of human forebrain development. Science *367*, eaay1645.

Trujillo, C.A., Rice, E.S., Schaefer, N.K., Chaim, I.A., Wheeler, E.C., Madrigal, A.A., Buchanan, J., Preissl, S., Wang, A., Negraes, P.D*., et al.* (2021). Reintroduction of the archaic variant of NOVA1 in cortical organoids alters neurodevelopment. Science *371*.

Vigilante, A., Laddach, A., Moens, N., Meleckyte, R., Leha, A., Ghahramani, A., Culley, O.J., Kathuria, A., Hurling, C., Vickers, A*., et al.* (2019). Identifying Extrinsic versus Intrinsic Drivers of Variation in Cell Behavior in Human iPSC Lines from Healthy Donors. Cell Rep *26*, 2078-2087 e2073.

Wang, M., Wei, P.C., Lim, C.K., Gallina, I.S., Marshall, S., Marchetto, M.C., Alt, F.W., and Gage, F.H. (2020). Increased Neural Progenitor Proliferation in a hiPSC Model of Autism Induces Replication Stress-Associated Genome Instability. Cell Stem Cell *26*, 221-233.e226.

Warmflash, A., Sorre, B., Etoc, F., Siggia, E.D., and Brivanlou, A.H. (2014). A method to recapitulate early embryonic spatial patterning in human embryonic stem cells. Nat Methods.

FIGURE LEGENDS

Figure 1. Cell line variation in the emergence of neural fate from pluripotency. (**A**) Experimental scheme. (**B**) Single-cell levels of OCT4 and NANOG expression in SA01. (i) Map of individual nuclei within a field. Colors correspond to distance from the nearest edge of the epithelium. Scale bar, 200 μm. Dashed lines indicate edge of colonies. (ii) Scatter plot of expression levels colored according to their distance from edge. ANOVA comparing protein levels on day 3 between core and edge zones: p=2.2e-16 for both OCT4 and NANOG. (**C**) CDX2, TBXT, and pSMAD1/5 levels at 24 hours after BMP4 treatments (D0T: treated on day 0, D2T: treated on day 2). (i) Representative images. Scale bar, 100 μm. (ii) Expression levels plotted against distance from the edge. (**D**) Spatial expression of NANOG, SOX21, and SOX2 on day 4 in SR and NSB. (i) Representative images. Scale bar, 100 μm. (ii) Spatial expression levels. (**E**) Variation in NANOG and SOX21 protein expression between lines. (i) Representative images on day 4. Scale bar, 100 μm. (ii) Expression levels in each line plotted across time (n=4, *: Comparison between SA01 and i04: p<0.001). (**F**) PCA of RNA-seq data showing differentiation trajectories. (**G**) PC1 in NSB data and projection of SR data into this NSB PC1.

Figure 2. Decomposing dynamic and stable transcription modules. (**A**) Hierarchical clustering of GWCoGAPS patterns and ANOVA p-values for association of line, time and condition to these patterns. (**B**) Gene expression patterns of pluripotency or differentiation regulators and their weights in dynamic GWCoGAPS patterns. (i) Heatmap of gene expression. (ii) Heatmap showing the gene weights. SR pattern: P7; NSB pattern: P12 and P15; BMP4 pattern: P9 and P3. (**C**) P7 represents loss of pluripotency and P3 represents induction of mesendoderm regulators. (i) SR pattern P7 and BMP4 pattern P3. (ii) Projections of mouse gastrula scRNA-seq data (Pijuan-Sala et al., 2019). (**D**) NSB patterns P15 and P12 represent distinct steps in the transition to neural fates. (i) P15 and P12 reveals early and later induction of neuroectoderm differentiation, respectively. (ii) Projection of mouse gastrula scRNA-seq data. (**E**) Cell line-specific signatures defined by GWCoGAPS analysis. (i) P13 defines the distinct

transcriptional signature of H9 from all other lines across time and conditions. (ii) Projection of embryoid body (EB) differentiation data from same 6 lines (p=3.0e-4). (iii) Projection of multiple hPSC line data (Rouhani et al., 2014). H9 samples circled in green.

Figure 3. SOX21 mediates forebrain fate choice. (**A**) Experimental design. (**B**) Spatial expression of NANOG, SOX2, and SOX3 on day 3 NSB. (**C**) Heatmaps showing top 100 genes from P7, P15, and P12 in NSB. (**D**) Heatmaps showing top 100 genes from P9 and P3 in day 2 SR (S2) and at 24 h after BMP4 D2T (B3). (**E**) Expression of early gastrulation (TBXT) and later mesendoderm (CDX2, GATA3, and ID1) regulators at 24 h after BMP4 D2T. (i) Representative images. Scale bar, 100 μm. (ii) Spatial expression. *, Comparison between WT (n=3) and SOX21-KO (n=3): $p<0.05$.

Figure 4. Cell line variation in hindbrain fate. (**A**) Differential GBX2 and OTX2 expression in SA01 and i04. (i) Representative images on day 6 in neuroectoderm differentiation conditions (LDN193189 and SB431542, LSB). Scale bar, 50 μm. (ii) OTX2 and GBX2 expression level in SR. *, Comparison between SA01 and i04 ($p<0.05$). (**B**) Distribution of gene-specific weights of 252 RA responsive genes (Balmer, 2002) compared to all 21,022 genes in SA01- and i04-specific transcriptional signatures. (**C**) Experimental protocol. (**D**) Correlation of RA response with RAR gene enrichment scores. (i) Proportion of $HOXB1^{hi}$ cells on day 8. *, Comparison between SA01 and i04 ($p<0.05$). (ii) Scatter plot of $HOXB1^{hi}$ cell proportions at 1 μM RA and RAR gene enrichment scores in each cell line-specific pattern. Linear fit ($R^2=0.7084$). (**E**) Differential production of hindbrain neurons in response to RA in SA01 and i04. Number of OLIG2, ISLET1, and PHOX2B expressing cells on day 28. *, Comparison between SA01 and i04 ($p<0.05$). (**F**) HOXB1 induction in WT and SOX21-KO lines after RA exposure. *, Comparison between WT and SOX21-KO lines ($p<0.05$).

Figure 5. Genetic and epigenetic elements contribute to donor- and line-specific transcriptional signatures. (**A**) Experimental design. (**B**) Donor-specific patterns and projections. (i) Patterns describing distinct transcriptional signatures shared between duplicate

lines from the same donor across time and condition. (ii) RNA-seq data of 260 human brain samples projected into each donor-specific pattern. Permutation analysis of these projections showed the significance of the specificity of these donor-specific transcriptional signatures in mature brain tissue: 2053, p=3.8e-6; 2075, p=1.7e-3; 2063, p=3.1e-5 (Figure S6C). (**C**) Scatter plots showing the association between genetic similarity and the strength of donor-specific transcription signatures across pair-wise comparisons with other donors. Each point is a comparison of the highlighted donor to another donor. Genetic distance between donors was quantified as the number of differing alleles across all genotyped SNPs (X-axis). Transcriptional distance was quantified as the strength of the donor-specific transcriptional signature in other donors' tissue (Y-axis; same data depicted in Y-axis of panel Bii, centered such that the source donor has a value of zero). Pearson's correlation and p-values were calculated omitting the donor data point at the origin. Solid black lines pass through the origin and each ethnicity's average Y-value, indicated as horizontal dashed lines. Ethnicity of donors is indicated by color: CAUC, Caucasian; AA, African American; HISP, Hispanic; AS, Asian. (**D**) Replicate line-specific variation within a donor. 2053-6 line-specific pattern and projection of brain RNA-seq data (as in panel B). (**E**) Scatter plot showing the lack of association between genetic similarity and the strength the 2053-6 line-specific transcription signature across pair-wise comparisons with other donors. (**F**) Projection of RNA-seq data into the neural patterns P15 and P12. (**G**) Representative images of SOX21 and NANOG expression in 2053 replicate lines on day 4. Scale bar, 200 μm. (**H**) Correlation of HOXB1^{hi} cell proportions and RAR gene enrichment scores in each cell line-specific pattern. Linear fit (R^2=0.75 in 1 μM RA and R^2=0.87 in 10 μM RA). (**I**) Scatter plot comparing differences in gene amplitudes between the i04 and SA01 line-specific patterns with differences in gene amplitudes between 2053 donor-specific (Figure 6C) and 2053-6 line-specific (Figure 6D) patterns. (**J**) Projection of H3K9me3 ChIP-seq data from lines 2053-2 and 2053-6 in SR into the 2053-6 line-specific transcriptional pattern. (**K**) H3K9me3 ChIP-seq data in lines 2053-2 and 2053-6 at 2, 4, and 6 days of SR at the GBX2 locus.

Figure 6. Early developmental bias defines hPSC variation in the wider human population. (**A**) PC1 of the NextGen Consortium hiPSC library (317 lines from 101 donors) RNA-seq data (Carcamo-Orive et al., 2017). Each individual line is colored by its average RAR gene expression level. Donors are ranked along the X-axis by the average PC1 level of all lines derived from that donor. All lines from one donor are shown vertically along the same X-axis position. Pearson's R and p-values indicate the correlation of PC1 with mean RAR gene expression. Blue circles indicate lines from one donor which show large variance in PC1 and RAR gene expression level. The PC1 value of zero intersects with an inflection point in the variation across donors (gray dashed lines – see Figure S7E for details). ICC=intraclass correlation coefficient. (**B**) Scatter plot comparing PC1 gene loadings from the NextGen hiPSC library data with differences in gene amplitudes between the i04 and SA01 line-specific patterns. (**C**) Scatter plot comparing PC1 gene loadings with gene amplitudes of neuroectoderm pattern P15 (Figure 2D). (**D**) UMAP plots of mouse gastrula scRNA-seq data (Pijuan-Sala et al., 2019) colored by the expression level of RAR genes with high or low gene amplitudes in P15. (**E**) Projection of NextGen PC1 onto the mouse gastrula scRNA-seq data (Pijuan-Sala et al., 2019), mapping transcriptional variation in hPSC lines onto in vivo lineage emergence.

Figure 1

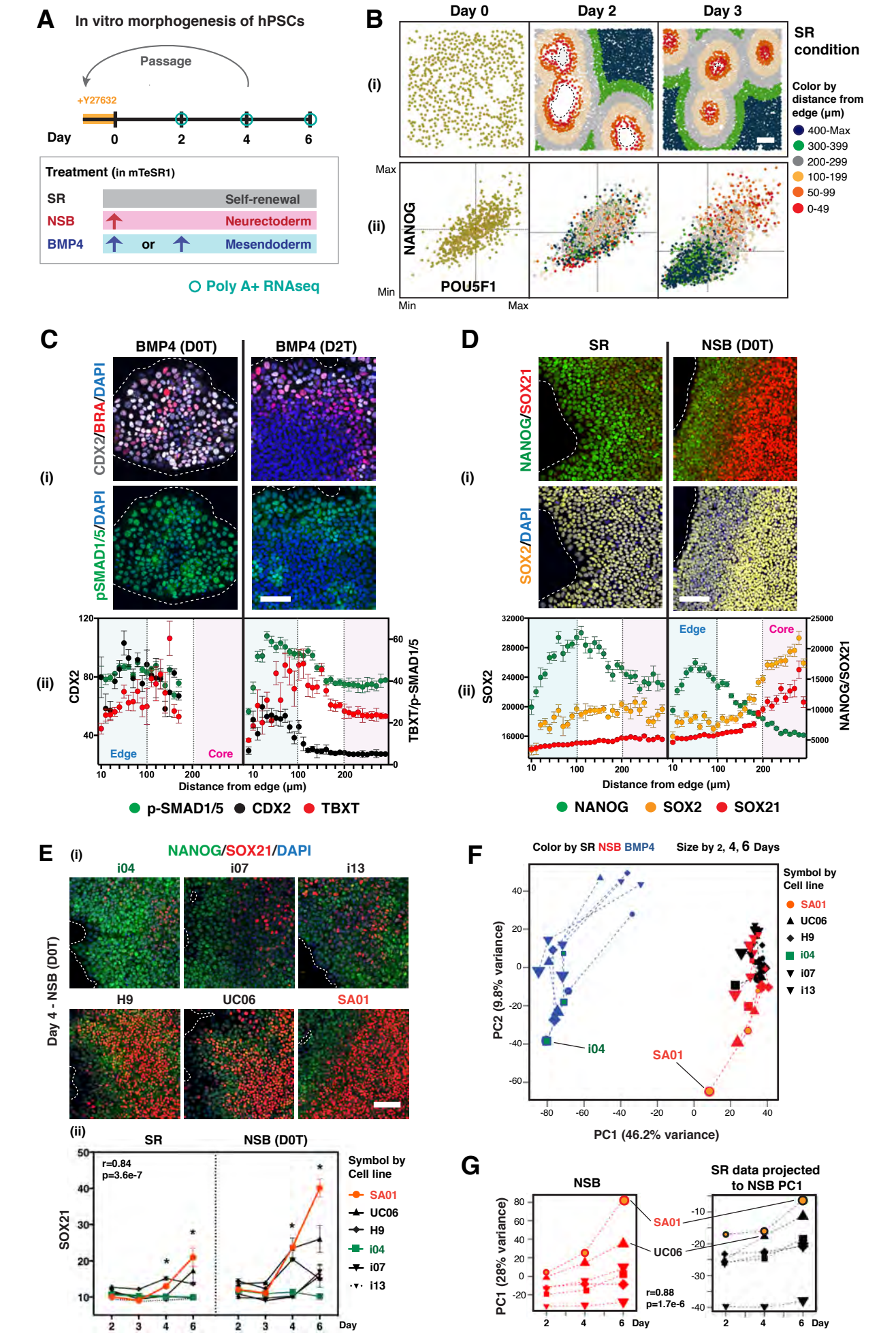

A
In vitro morphogenesis of hPSCs
Passage
+Y27632
Day
0
2
4
6
Treatment (in mTeSR1)
SR
Self-renewal
NSB
Neurectoderm
BMP4
or
Mesendoderm
Poly A+ RNAseq
B
Day 0
Day 2
Day 3
(i)
SR condition
Color by distance from edge (µm)
400-Max
300-399
200-299
100-199
50-99
0-49
(ii)
Max
Min
NANOG
POU5F1
Min
Max
C
BMP4 (D0T)
BMP4 (D2T)
(i)
CDX2/BRA/DAPI
pSMAD1/5/DAPI
(ii)
CDX2
TBXT/p-SMAD1/5
Edge
Core
Distance from edge (µm)
p-SMAD1/5
CDX2
TBXT
D
SR
NSB (D0T)
(i)
NANOG/SOX21
SOX2/DAPI
(ii)
SOX2
NANOG/SOX21
Edge
Core
Distance from edge (µm)
NANOG
SOX2
SOX21
E
(i)
NANOG/SOX21/DAPI
i04
i07
i13
H9
UC06
SA01
Day 4 - NSB (D0T)
(ii)
SR
NSB (D0T)
r=0.84
p=3.6e-7
SOX21
Day
Symbol by Cell line
SA01
UC06
H9
i04
i07
i13
F
Color by SR NSB BMP4
Size by 2, 4, 6 Days
Symbol by Cell line
SA01
UC06
H9
i04
i07
i13
PC2 (9.8% variance)
PC1 (46.2% variance)
i04
SA01
G
NSB
SR data projected to NSB PC1
PC1 (28% variance)
SA01
UC06
r=0.88
p=1.7e-6
Day

Figure 2

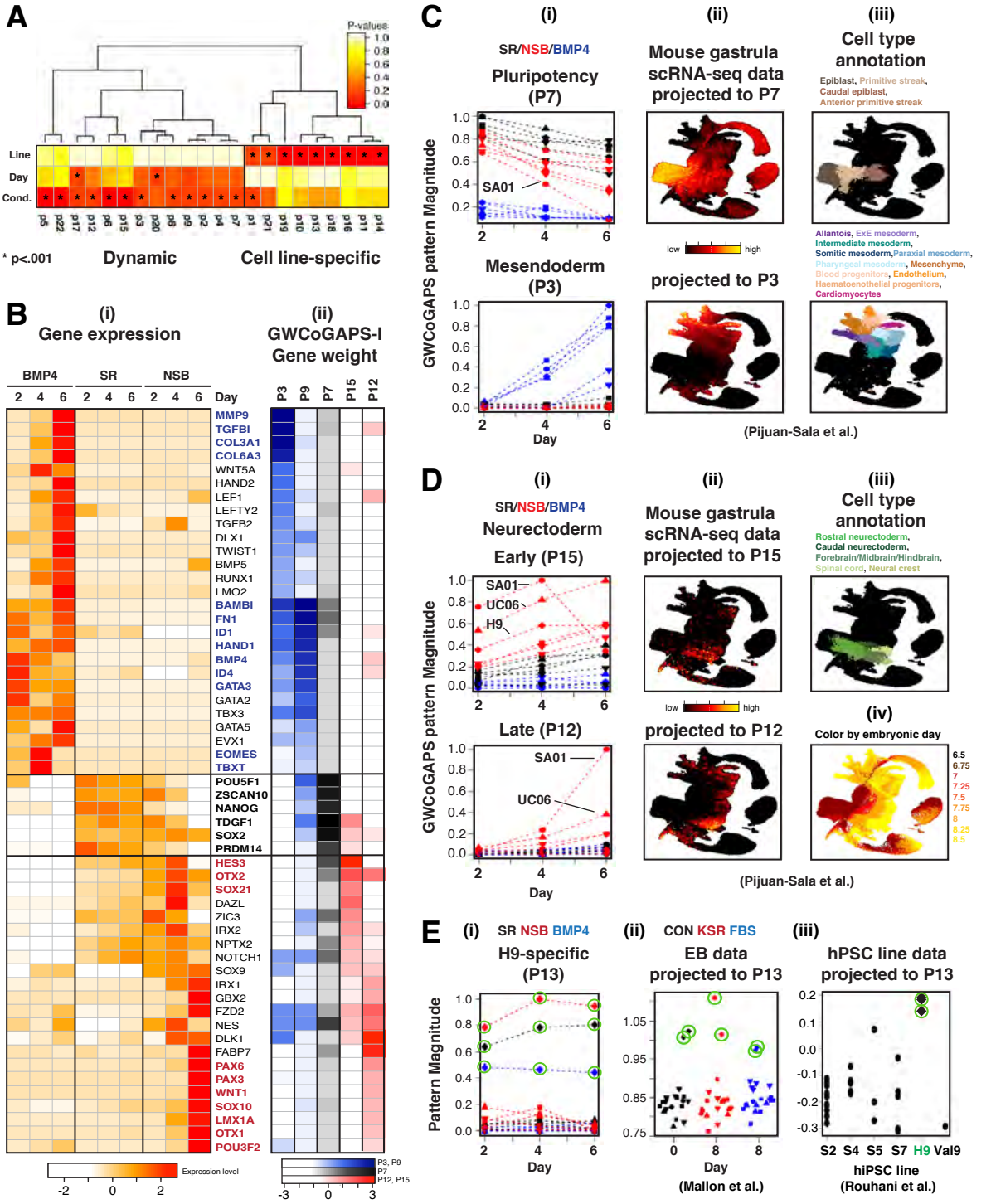

A
P-values
Line
Day
Cond.
* p<.001
Dynamic
Cell line-specific
B
(i) Gene expression
(ii) GWCoGAPS-I Gene weight
BMP4
SR
NSB
Day
MMP9
TGFBI
COL3A1
COL6A3
WNT5A
HAND2
LEF1
LEFTY2
TGFB2
DLX1
TWIST1
BMP5
RUNX1
LMO2
BAMBI
FN1
ID1
HAND1
BMP4
ID4
GATA3
GATA2
TBX3
GATA5
EVX1
EOMES
TBXT
POU5F1
ZSCAN10
NANOG
TDGF1
SOX2
PRDM14
HES3
OTX2
SOX21
DAZL
ZIC3
IRX2
NPTX2
NOTCH1
SOX9
IRX1
GBX2
FZD2
NES
DLK1
FABP7
PAX6
PAX3
WNT1
SOX10
LMX1A
OTX1
POU3F2
Expression level
P3, P9
P7
P12, P15
C
(i)
SR/NSB/BMP4
Pluripotency (P7)
Mesendoderm (P3)
GWCoGAPS pattern Magnitude
SA01
Day
(ii)
Mouse gastrula scRNA-seq data projected to P7
projected to P3
low
high
(iii)
Cell type annotation
Epiblast, Primitive streak, Caudal epiblast, Anterior primitive streak
Allantois, ExE mesoderm, Intermediate mesoderm, Somitic mesoderm, Paraxial mesoderm, Pharyngeal mesoderm, Mesenchyme, Blood progenitors, Endothelium, Haematoendothelial progenitors, Cardiomyocytes
(Pijuan-Sala et al.)
D
(i)
SR/NSB/BMP4
Neurectoderm
Early (P15)
Late (P12)
SA01
UC06
H9
(ii)
Mouse gastrula scRNA-seq data projected to P15
projected to P12
(iii)
Cell type annotation
Rostral neurectoderm, Caudal neurectoderm, Forebrain/Midbrain/Hindbrain, Spinal cord, Neural crest
(iv)
Color by embryonic day
6.5
6.75
7
7.25
7.5
7.75
8
8.25
8.5
(Pijuan-Sala et al.)
E
(i) SR NSB BMP4
H9-specific (P13)
Pattern Magnitude
(ii) CON KSR FBS
EB data projected to P13
(Mallon et al.)
(iii)
hPSC line data projected to P13
S2 S4 S5 S7 H9 Val9
hiPSC line
(Rouhani et al.)

Figure 3

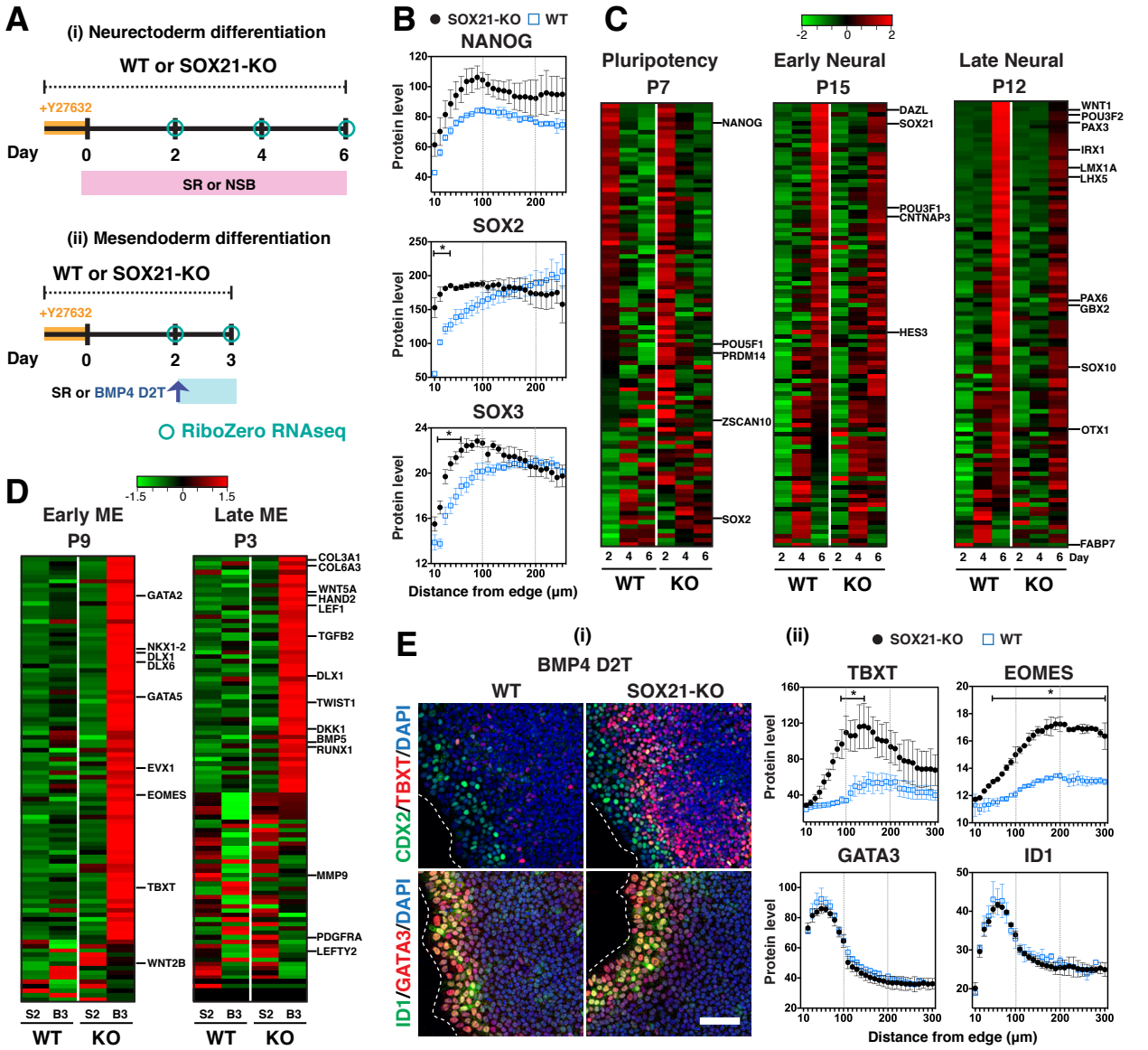
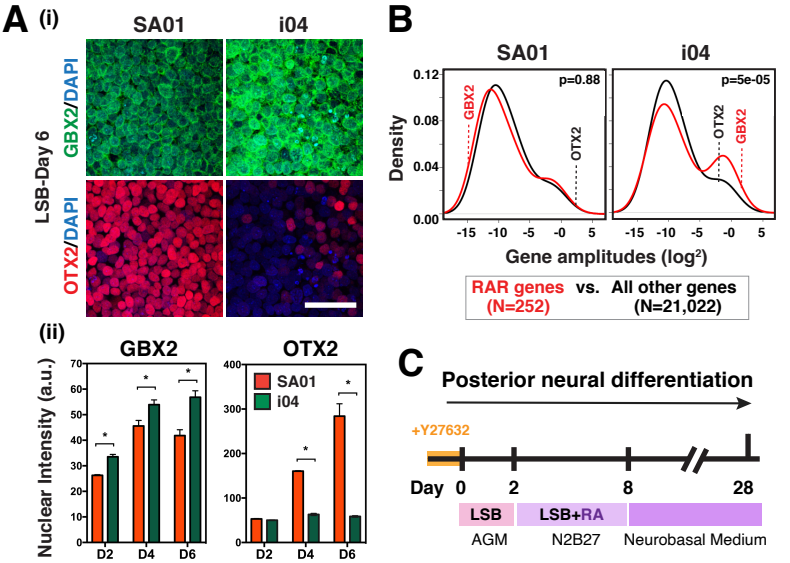
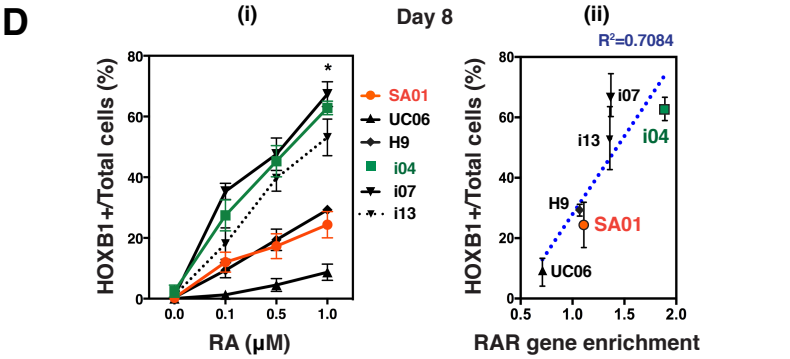

A
(i) Neurectoderm differentiation
WT or SOX21-KO
+Y27632
Day
0
2
4
6
SR or NSB
(ii) Mesendoderm differentiation
WT or SOX21-KO
+Y27632
Day
0
2
3
SR or BMP4 D2T
RiboZero RNAseq
B
SOX21-KO
WT
NANOG
SOX2
SOX3
Protein level
Distance from edge (μm)
C
-2
0
2
Pluripotency
P7
Early Neural
P15
Late Neural
P12
NANOG
POU5F1
PRDM14
ZSCAN10
SOX2
DAZL
SOX21
POU3F1
CNTNAP3
HES3
WNT1
POU3F2
PAX3
IRX1
LMX1A
LHX5
PAX6
GBX2
SOX10
OTX1
FABP7
Day
WT
KO
D
-1.5
0
1.5
Early ME
P9
Late ME
P3
GATA2
NKX1-2
DLX1
DLX6
GATA5
EVX1
EOMES
TBXT
WNT2B
COL3A1
COL6A3
WNT5A
HAND2
LEF1
TGFB2
DLX1
TWIST1
DKK1
BMP5
RUNX1
MMP9
PDGFRA
LEFTY2
S2
B3
WT
KO
E
(i)
BMP4 D2T
WT
SOX21-KO
CDX2/TBXT/DAPI
ID1/GATA3/DAPI
(ii)
SOX21-KO
WT
TBXT
EOMES
GATA3
ID1
Protein level
Distance from edge (μm)

Figure 4


A
(i)
SA01
i04
LSB-Day 6
GBX2/DAPI
OTX2/DAPI
(ii)
GBX2
OTX2
Nuclear Intensity (a.u.)
SA01
i04
D2
D4
D6
B
SA01
i04
Density
p=0.88
p=5e-05
GBX2
OTX2
Gene amplitudes (log2)
RAR genes (N=252) vs. All other genes (N=21,022)
C
Posterior neural differentiation
+Y27632
Day 0 2 8 28
LSB
LSB+RA
AGM
N2B27
Neurobasal Medium


D
(i)
Day 8
HOXB1+/Total cells (%)
RA (μM)
SA01
UC06
H9
i04
i07
i13
(ii)
R2=0.7084
HOXB1+/Total cells (%)
RAR gene enrichment
i07
i13
i04
H9
SA01
UC06

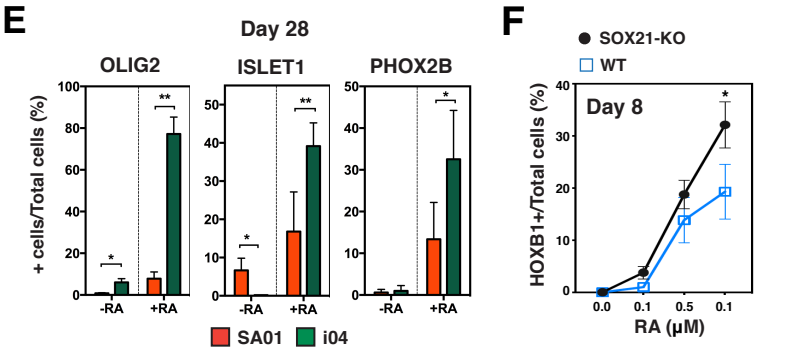

E
Day 28
OLIG2
ISLET1
PHOX2B
+ cells/Total cells (%)
-RA
+RA
SA01
i04
F
SOX21-KO
WT
Day 8
HOXB1+/Total cells (%)
RA (μM)

Figure 5

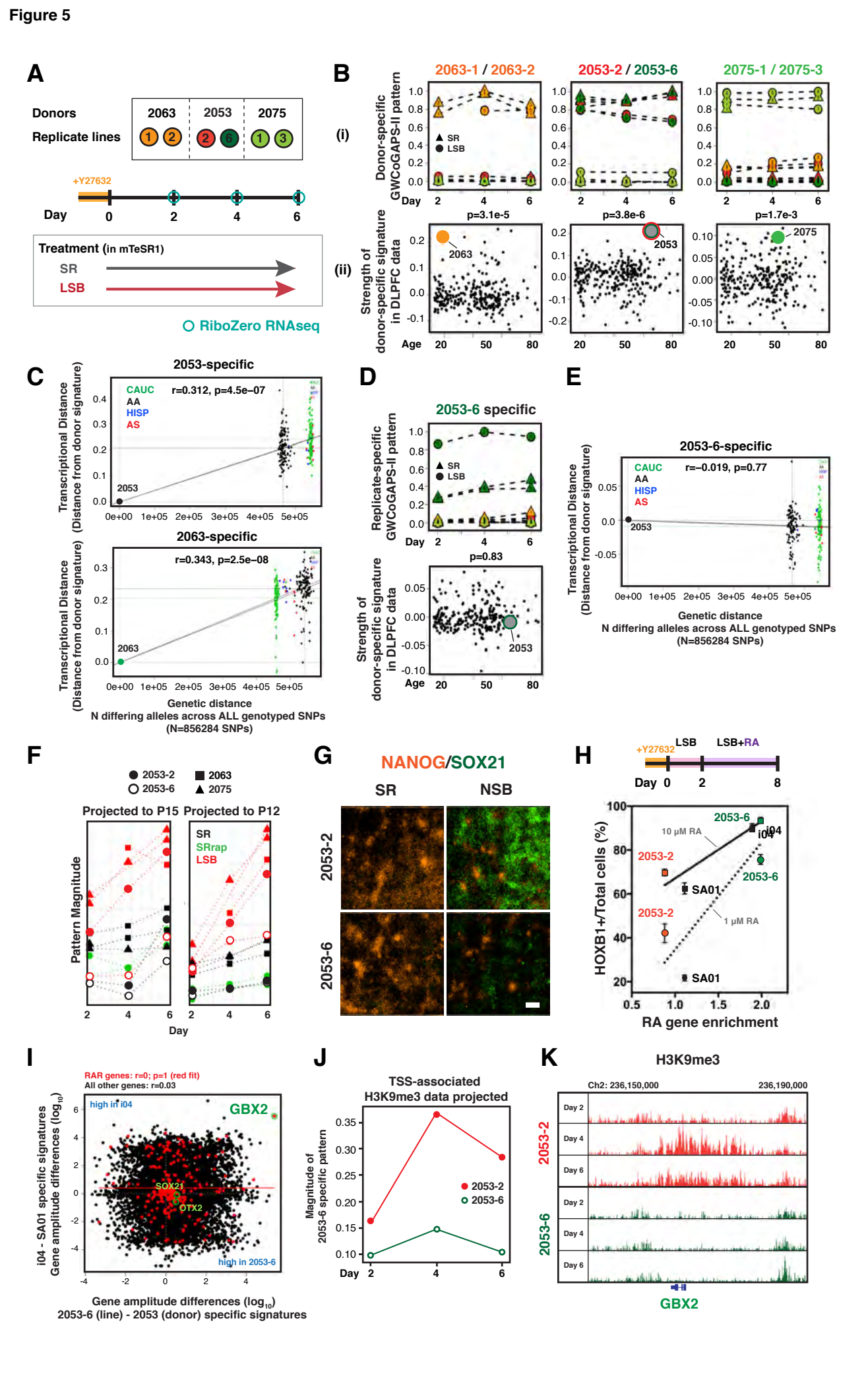
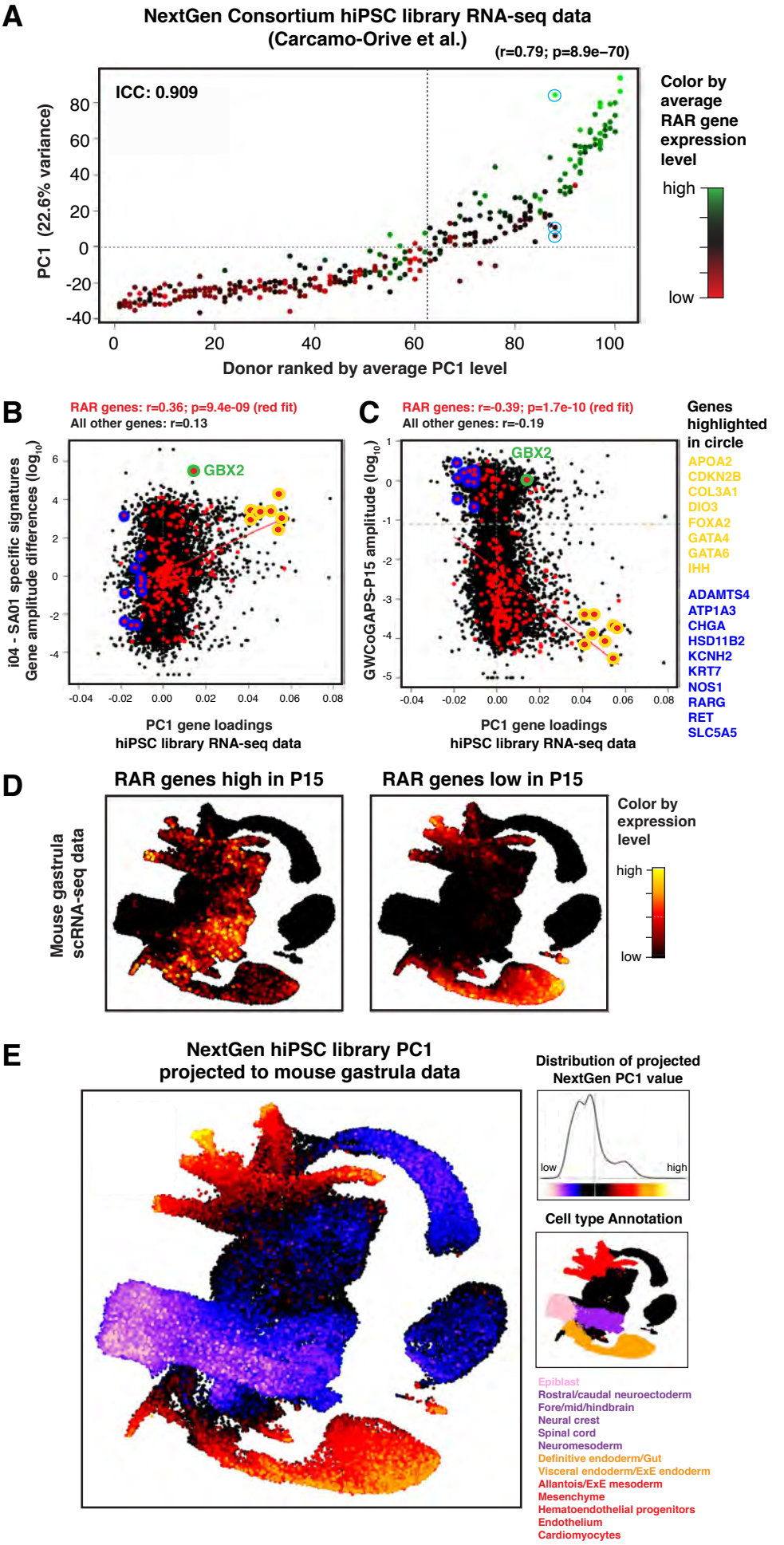

A
Donors
Replicate lines
2063
2053
2075
+Y27632
Day
0
2
4
6
Treatment (in mTeSR1)
SR
LSB
RiboZero RNAseq
B
2063-1 / 2063-2
2053-2 / 2053-6
2075-1 / 2075-3
(i)
Donor-specific GWCoGAPS-II pattern
SR
LSB
Day
(ii)
Strength of donor-specific signature in DLPFC data
p=3.1e-5
p=3.8e-6
p=1.7e-3
2063
2053
2075
Age
C
2053-specific
r=0.312, p=4.5e-07
CAUC
AA
HISP
AS
2063-specific
r=0.343, p=2.5e-08
Transcriptional Distance (Distance from donor signature)
Genetic distance
N differing alleles across ALL genotyped SNPs
(N=856284 SNPs)
D
2053-6 specific
Replicate-specific GWCoGAPS-II pattern
p=0.83
E
2053-6-specific
r=-0.019, p=0.77
F
2053-2
2053-6
2063
2075
Projected to P15
Projected to P12
Pattern Magnitude
SR
SRrap
LSB
Day
G
NANOG/SOX21
SR
NSB
2053-2
2053-6
H
+Y27632
LSB
LSB+RA
Day
HOXB1+/Total cells (%)
10 μM RA
1 μM RA
SA01
i04
RA gene enrichment
I
RAR genes: r=0; p=1 (red fit)
All other genes: r=0.03
high in i04
GBX2
SOX21
OTX2
high in 2053-6
i04 - SA01 specific signatures Gene amplitude differences (log10)
Gene amplitude differences (log10)
2053-6 (line) - 2053 (donor) specific signatures
J
TSS-associated H3K9me3 data projected
Magnitude of 2053-6 specific pattern
K
H3K9me3
Ch2: 236,150,000
236,190,000
Day 2
Day 4
Day 6
GBX2

Figure 6


A
NextGen Consortium hiPSC library RNA-seq data
(Carcamo-Orive et al.)
(r=0.79; p=8.9e−70)
ICC: 0.909
PC1 (22.6% variance)
Donor ranked by average PC1 level
Color by average RAR gene expression level
high
low
B
RAR genes: r=0.36; p=9.4e-09 (red fit)
All other genes: r=0.13
GBX2
i04 - SA01 specific signatures Gene amplitude differences (log10)
PC1 gene loadings hiPSC library RNA-seq data
C
RAR genes: r=-0.39; p=1.7e-10 (red fit)
All other genes: r=-0.19
GBX2
GWCoGAPS-P15 amplitude (log10)
PC1 gene loadings hiPSC library RNA-seq data
Genes highlighted in circle
APOA2
CDKN2B
COL3A1
DIO3
FOXA2
GATA4
GATA6
IHH
ADAMTS4
ATP1A3
CHGA
HSD11B2
KCNH2
KRT7
NOS1
RARG
RET
SLC5A5
D
RAR genes high in P15
RAR genes low in P15
Mouse gastrula scRNA-seq data
Color by expression level
high
low
E
NextGen hiPSC library PC1 projected to mouse gastrula data
Distribution of projected NextGen PC1 value
low
high
Cell type Annotation
Epiblast
Rostral/caudal neuroectoderm
Fore/mid/hindbrain
Neural crest
Spinal cord
Neuromesoderm
Definitive endoderm/Gut
Visceral endoderm/ExE endoderm
Allantois/ExE mesoderm
Mesenchyme
Hematoendothelial progenitors
Endothelium
Cardiomyocytes