Finding human genetic variation in whole genome expression data with applications for “missing” heritability:

The GWCoGAPS algorithm, the PatternMarkers statistic, and the ProjectoR package

by
Genevieve Stein-O’Brien

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
June, 2017

© 2017 Genevieve Stein-O’Brien
All Rights Reserved

Abstract

Starting from a single fertilized egg, the compendium of human cells is generated via stochastic perturbations of earlier generations. Concurrently, canalization of developmental pathways limits the type and degree of variation to ensure viability; thus, it is unsurprising that deviations early in life have been linked to late manifesting diseases. Human pluripotent stem cells (hPSCs) are a highly robust and uniquely human experimental system in which to model the sources and consequences of this variability. Further, variation in hPSCs' transcriptomes has been directly linked to both genomic background and biases in differentiation efficiency. Taking advantage of this link between genomic background and developmental phenotypes, we developed Genome-Wide CoGAPS Analysis in Parallel Sets (GWCoGAPS), the first robust whole genome Bayesian non-negative matrix factorization (NMF), to find conserved transcriptional signatures representative of the functional effect of human genetic variation. Time course RNA-seq data obtained from three human embryonic stem cells (ESC) and three human induced pluripotent stem cells (IPSC) in three different experimental conditions was analyzed. GWCoGAPS distinguished shared developmental trajectories from unique transcriptional signatures of each of the cell lines. Further analysis of these "identity" signatures found they were predictive of lineage biases during neuronal differentiation. Additionally, lineage biases were consistent with early differences in morphogenetic phenotypes within monolayer culture, thus, linking transcriptional genomic signatures to stable quantifiable cellular features. To test whether the cell line signatures were genome specific, we next developed the projectoR algorithm to assess a given signatures robustness in independent data sets. By using the identity signatures as inputs to projectoR, we were able to identify samples from the same donor genome in datasets from multiple tissues and across technical platforms, including RNA-seq results from post-mortem brain, micro arrayed embryoid bodies, and publicly available datasets. The identification of signatures that define the functional rather than physical background of an individual's genome has the potential to profoundly influence our view of human variation and disease.

Preface

To my Father, who taught me to be proud of who I was and cultivated a life long love of learning.
To my Mother, who made me the woman that I am more than any other factor in my life.
To Shaughn, who builds with me so that I will be supported and happy in all that has yet to be.

Definition of the Gene

For this construct of what makes us
unique and, yet, predictable
the gene is but a thesis

To evolve with time a must,
if both are to be viable,
for this construct of what makes us

Merged and passed through coitus
--ask what is inheritable--
the gene is but a thesis

From biochemical corpus
arises the sequencable
for this construct of what makes us

With a panacean promise
for mutated and mutable
the gene is but a thesis

Questions are ubiquitous
though more is well definable
for a construct of what makes us
the gene is but a thesis

Intended to be blank.

Table of Contents

ABSTRACT .... **II**

PREFACE .... **III**

Definition of the Gene **iv**

TABLE OF CONTENTS.. **VI**

LIST OF FIGURES.. **VIII**

THESIS **1**

Introduction **1**

PatternMarkers & Genome Wide CoGAPS Analysis in Parallel Set (GWCoGAPS) for novel data-driven biomarkers via whole transcriptome NMF **3**

projectoR: Integrative Analysis of Low-Dimensional Molecular Dynamics across High-Dimensional Multi-Omic Data Sets **7**

Defining the origins of variation in human pluripotent stem cells **8**

APPENDICES **17**

Supplementary Figures **18**

Supplementary Experimental Procedures **19**

Permission to Reprint **21**

BIBLIOGRAPHY.. **22**

CURRICULUM VITAE **23**

Intended to be blank.

List of Figures

Fig. 1 Flow chart of GWCoGAPS to obtain patternMarkers from NMF solutions ii

Fig. 2 ProjectoR links BMP4 pathway activity in in vitro hPSCS and in vivo embryos iii

Fig. 3 ProjectoR overcomes technical artifacts to characterizes commonalities of divergent BMP4 and Activin signaling iv

Fig. 4 ProjectoR connects related expression dynamics to epigenetic regulation vi

Fig. 5 ProjectoR maps spatial localization of GWCoGAPS gene expression patterns viii

Fig. 6. PatternMarkers from GWCoGAPS patterns associated with Mesendoderm, Endoderm, and, Mesoderm 1

Fig. 7 ProjectoR associates alternative pathway regulation with differences in cancer survival..... 1

Fig. 8 Spatiotemporal dynamics of pluripotency, lineage specification, and growth signaling regulators reveal in vitro morphogenesis of human PSCs. 3

Fig. 9 Cellular and transcriptional dynamics during self-renewal and differentiation show variation between human PSC cell lines. 7

Fig. 10 Decomposition of transcription during in vitro morphogenesis. 8

Fig. 11 CoGAPS reveals stable transcriptional signatures of individual humans. 17

Fig. 12 Cell line-specific transcriptional signatures predict response to retinoic acid. 18

Figure S1. Spatiotemporal dynamics of neural fate bias, Related to Figure 1. 19

Figure S2. Cell line variation in transcriptional dynamics during self-renewal and differentiation, Related to Figure 2 and Figure 3. 21

Figure S3. Decomposing dynamic & stable transcription modules, Related to Figure 3................ 22

Figure S4. Cell line variation in hindbrain fate bias, Related to Figure 4. 23

Figure S5. Resetting the pluripotency cycle, Related to Figure 5. 24

Figure S6. SOX21 mediates forebrain fate choice, Related to Figure 6. 37

Figure S7. Cell line variation in neural fate trajectories, Related to Figure 7. 38

Thesis

Introduction

Convention speaks of high throughput genetics yielding "high dimensional" data, but I would argue that what has increased isn't the dimensionality of the data but the number of tests performed on one, perhaps two dimensions. In 2000, the Human Genome Project announced the sequencing of the entire human genome and since then it is this physical map the field of human genetics has used to guide their exploration. Linkage and GWAS exist in one dimension--that of the physical structure of the genome. Even with eQTLs while the phenotype is continuous the SNP is discrete and belongs to a single linear world--it must, the assumptions of linkage disequilibrium and independent assortment upon which both tests are based requires it to be so.

Perhaps the furthest that this methodology has been advanced was by Seymore Benzer in 1959 when he used the same math involved in linkage and association to test function as well as position via an elegant complementation test in t4 phage. As sequencing becomes cheaper, the rise of whole exome and whole genome seq has the potential to keep us solidly grounded in this physical plane. Yet, all of the cells in the a body have the same DNA, but are as diverse as the populations of people that they construct. Any population geneticists will tell you that at most only 20% of the overall variance between two individuals can be accounted for by divergence in their race. The various iterations of HapMap and now 1000 Genomes have delved farther and farther into this terrain. Yet again and again the ancient treasure at the end of the map has been a common variant. The public databanks are a wealth of common variants, yet the inflation of our understand of their functional consequence has failed to keep pace. If we want functional findings than we need a different map.

In 1999, matrix decomposition techniques were pioneered for high-dimensional biology. It was a technique before its time. Since then technological advances in high throughput transcriptomics, epigenomic, proteomic, and single-cell have risen as significant tributaries to the data stream. In 2010, the data deluge was reported as the increase of “high dimensional” DNA sequencing data at a rate outpacing Moore’s law. Data-driven pattern learning techniques, including matrix decomposition methods, have advanced to power systems level analyses of multiple data types. A major benefit of these techniques is that, as unsupervised methods, they can highlight the questions that we have not yet thought to ask.

As datasets grow to tens of thousands of samples, humans can no longer fully consider the composition of each sample in a dataset simultaneously. Unsupervised techniques that identify such patterns and summarize them into human-interpretable results will power the next age of discovery. As this era begins, I have found guidance in voices from the past. In his "A Mathematical Theory of Natural Selection", Haldane established the method for the construction of genetic models of selection. Beyond his own models or even those of population genetics, his method provides the framework for many of the models required for parametric linkage analysis and informs some of the Bayesian models used in association studies. His relevance is his reasoning. At the beginning of the first of these ten papers, Haldane explicitly characterized what was necessary to construct a satisfactory model of selection, and I would argue can be generalized to provide the framework for any genetic model-- must be quantitative, account for the facts, explain the rate of change of the relevant characteristics, and ultimately, it must be dynamic. To navigate a dynamic system you need dynamics map. To take full advantage of high throughput data, we must truly increase its dimensionally by layering the levels of annotation and mapping the interactions.

The work that follows consists of two novel unsupervised pattern-detection methods whose unifying goal is to provide tools for building that dynamic map. The first, Genome Wide CoGAPS Analysis in Parallel Set (GWCoGAPS) is also the first robust whole genome Bayesian non-negative matrix factorization (NMF) algorithm. By accounting for pleiotrophy and the inherent dependence of biological systems, GWCoGAPS is able to find conserved transcriptional signatures of an individual's genomic background. The second, ProjectoR enables unprecedented *in silico* experimentation across genomic technologies, model systems, and species by using relationships defined within a given data set to interrogate related biological phenomena in entirely new data. Together these algorithms were able to define and probe the functional effect of human genetic variation using global patterns of gene expression instead of DNA sequence.

PatternMarkers & Genome Wide CoGAPS Analysis in Parallel Set (GWCoGAPS) for novel data-driven biomarkers via whole transcriptome NMF

Introduction

Numerous high-throughput studies link gene expression changes to biological processes (BPs) including regulatory networks and the cell signaling processes. Previously shown effective at deconvoluting multiplexed regulation and gene reuse in BPs [1,2], NMF algorithms have identified genes associated with yeast cell cycle and metabolism, cancer subtypes, and perturbations to cellular signaling in cancer[3-10]. However, the continuous and interdependent nature of many NMF results can make biological inference challenging especially when searching for biomarkers or genetic drivers. A method to obtaining genes that uniquely identify NMF solutions would eliminate these challenges.

Here, we develop patternMarkers, a statistic to take the relative gene weights output from NMF algorithms and to return only those genes that are strongly associated with a particular pattern or with a linear combination of patterns. Identifying unbiased biomarkers using patternMarkers requires genome-wide transcriptional data. To maximize the potential for novel marker detection, we set out to expand the O(1,000) gene limit, which is typical to achieve convergence in NMF, to the O(10,000) genes comprising the entire human transcriptome. Currently, NMF methods are highly dependent upon the genes selected or compaction methods to limit the size of the data matrices used for analysis[11]. Therefore, we developed GWCoGAPS, a whole genome implementation of CoGAPS (Fertig, et al. 2010), a Markov chain Monte Carlo (MCMC) NMF that encodes sparsity in the decomposed matrices with an atomic prior[12]. Previously, we demonstrated that CoGAPS analysis of datasets containing representative subsets of the genes converge with similar patterns. These patterns can then be fixed to a consensus pattern across the datasets to provide a robust whole-genome NMF, without the prohibitively large computational cost of NMF factorization of a single matrix containing the entire genome. GWCoGAPS takes advantage of parallel computing to massively cut runtime and ensure genome-wide convergence. We also include a Shiny web application, patternMatcher, to compare patterns across parallel runs to increase robustness and interpretability of the resulting patterns. Using patternMarkers with GWCoGAPS to analyze tissues from the Genotype-Tissue Expression

Project[13], we parsed patterns of expression specific to brain regions and cell types to demonstrate the power of these algorithms for biomarker discovery.

Methods

NMF decomposes a data matrix of D with N genes as rows and M samples as columns, into two matrices, as D ~ AP. The pattern matrix P has rows associated with BPs in samples and the amplitude matrix A has columns indicating the relative association of a given gene, where the total number of BPs (k) is an input parameter. CoGAPS is a Bayesian NMF that incorporates both non-negativity and sparsity in A and P as described in (Fertig, et al. 2010). Both patternMarkers and GWCoGAPS are in the CoGAPS Bioconductor package as of version 3.5 and are generalized for other NMF algorithms.

The patternMarkers statistic (s_{ij}) scores the association of the *i*th gene's values in the amplitude matrix (A_i) with the j^{th} pattern or linear combination of patterns by computing

$$s_{ij}\left(\overline{w}_j\right) = \sqrt{\sum_k \left(\frac{A_{ik}}{\max A_i} - \overline{w}_{jk}\right)^2} \tag{1}$$

where *i* indices all the genes in the original data matrix, k indices all the patterns in the NMF solution, and is a vector $\overline{w}_j$ of components specifying the j^{th} linear combination of patterns that is constrained to sum to 1, and *j* indices the total number of linear combinations for which patternMarkers statistics are computed. The default setting for Eq. (1) sets *j={1, … , k}*, such that $\overline{w}$ is a set containing a unit vector for each pattern and $s_{ij}(\overline{w_j})$ is an l_2 norm indicating the exclusivity of the contribution of gene i to the pattern j and the corresponding BP. Scaling by the maximum value of each gene in the NMF solution (max A_i) decouples the effect of overall gene expression level without impacting the quality of the factorization. Genes are ranked by increasing $s_{ij}(\overline{w_j})$ such that the higher the rank of the gene, the less it is associated with the considered pattern. Users can output a list of data frames containing the scores and ranks for every gene using the "All" option of the "threshold" argument. Alternatively, unique gene sets can be generated by either subsetting each gene by its lowest ranking $s_{ij}(\overline{w_j})$. In the case where j >1, the ranked list for each pattern can also be thresholded by the highest value for which $s_{ij}(\overline{w_j})$ is the lowest.

The GWCoGAPS function automates and parallelizes the whole-genome CoGAPS analysis from (Fertig, et al. 2013) in a single R function. GWCoGAPS has three parameters: the number of sets for partitioning the whole genome data, the seed for each Markov Chain, and the method for determining the consensus patterns. A new modification to CoGAPS, setting the seed both ensures that each set of genes is run with a different set of random numbers and that runs on any dataset are reproducible. A default pattern matching function is provided along with a Shiny-based web application patternMatcher for recompiling the parallelized results (Fig. 1A). Additional runtime options, input, and manual implementations are described in the GWCoGAPS vignette.

RPKM RNAseq data for the seven samples with most brain regions was downloaded from dbGaP. GWCoGAPS was run for a range of k patterns with k=10 selected and uncertainty as 10% of the data (Fertig, et al. 2013). The code to reproduce this analysis and the GWCoGAPS results are available in the online supplemental files of (Stein-O'Brien, et al. 2017).

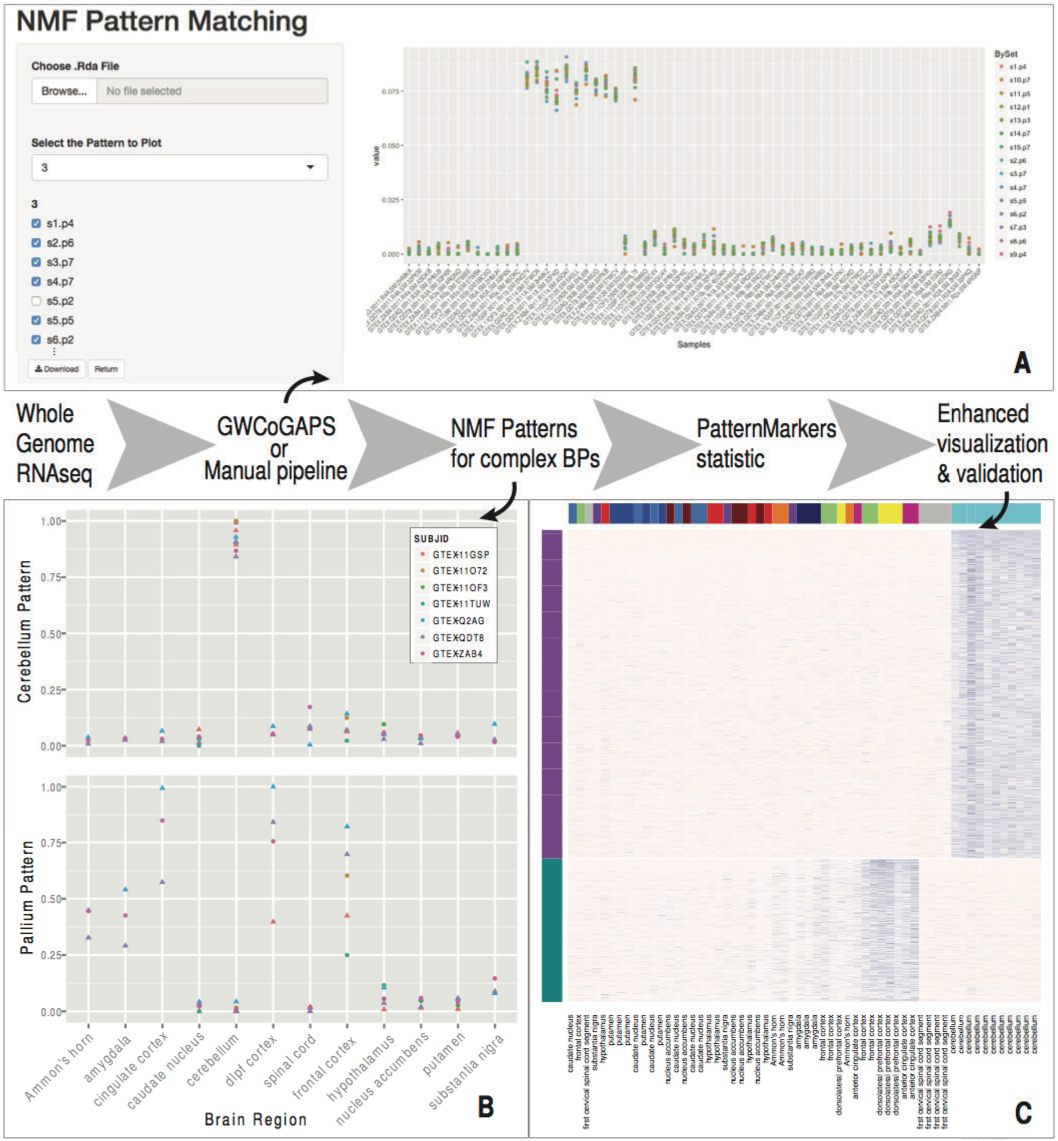
Results

We apply GWCoGAPS to analyze patterns related to brain regions for different individuals in GTEx. The GWCoGAPS solutions for the initial parallel runs of the patterns is used to illustrate the strong association between patterns identified from the subsets using patternMatcher (Fig. 1A). The first pattern highlights GWCoGAPS' ability to deconvolute tissue specific signatures (Fig. 1B). This pattern uniquely identifies the cerebellum, determined to be the most distinct region by the consortium[13]. GTEx found that strong individual specific effects increase with tissue relatedness as illustrated by their inability to achieve tissue specific clusters of the different brain regions by expression alone[13,14]. By allowing for gene reuse across different patterns, GWCoGAPS is able to overcome these effects to isolate the cerebellums signature as confirmed by gene set enrichment[15] in cerebellum development and morphogensis (GO:0021549 and GO:0021587 FWER p-value <1.0E-03 and 2.6E-03, respectively, described in Supplemental File 5) on these patternMarkers scores.

The second pattern illustrates patternMarkers' power as inference is difficult from the GWCoGAPS result alone (Fig. 1B). This pattern depicts subpopulations of cells in multiple brain regions derived from common pallium precursors. Progeny of the pallium are specified by transcription factors TBr1 and Emx1[16] ranked second and fourth by the patternMarkers statistic. Gene set analysis on these patternMarkers scores confirms enrichment for pallium development (GO:0021543 FWER p-value <1.0E-03, Online Supplemental File 5).

Deconvolution of cell type and tissue specific signatures from aggregate data represent a major technical challenge. We have illustrated the unique ability of GWCoGAPS, the first whole genome Bayesian NMF, to accomplish this. The manual pipeline and Shiny App, patternMatcher, also expanded this methodology to a variety of NMF techniques. Finally, the patternMarkers statistic derives gene sets uniquely representative of BPs from the continuous weights of NMF solutions. Together, patternMarkers and GWCoGAPS find data-driven biomarkers and genetic drivers in whole genome transcriptomic data.

Fig. 1 Flow chart of GWCoGAPS to obtain patternMarkers from NMF solutions



Flow chart of the GWCoGAPS to obtain pattern markers from NMF solutions. A) The patternMarker Shiny App illustrating high concordance of the GWCoGAPS solutions for the initial unsupervised parallel runs for the patterns associated with the cerebellum. B) Two of the ten final GWCoGAPS patterns for the GTex data associated with the cerebellum and dorsal pallium, respectively. C) Visualization of unique biomarkers for the cerebellum and dorsal pallium from the patternMarkers statistic.

projectoR: Integrative Analysis of Low-Dimensional Molecular Dynamics across High-Dimensional Multi-Omic Data Sets

Abstract

Technological advances continue to spur the exponential growth of omics data. To fully leverage these vast databases, methods using previously learned knowledge to improve new analysis must be developed. Transfer learning methodologies (TLMs) are agnostic to distribution or feature space making them particularly well suited for integrating different omics data. Thus, we developed TLMs for integrating high-dimensional analyses across multi-omic data in the R package ProjectoR. Using public data, we apply ProjectoR to 1) link BMP4 pathway activity in vitro hPSCS and in vivo embryos, 2) characterize commonalities of divergent BMP4 and Activin signaling, 3) connect related expression dynamics to epigenetic regulation, and 4) associate alternative regulation of these pathways in tumors with significant differences in cancer survival. Thus, ProjectoR enables unprecedented in silico experimentation across genomic technologies, model systems, and species by using relationships defined within a given data set to interrogate related biological phenomena in entirely new data.

Introduction

When "data deludge" is used to describe the exponential growth of biological data, the first challenge is accessibility, curation, and storage of the data being produced [17-20]. In 2009, the deluge was the result of DNA sequencing becoming faster and cheaper at a rate outpacing Moore's law[21]. Since then technological advances in high throughput techniques for transcriptomics, epigenomic, proteomic, and single-cell techniques have risen as significant tributaries to the data stream[22]. The numerous resources and archive databases now available successfully store growing datasets and make them available to researchers for analysis. Performing analyses of datasets from numerous sources and across high throughput omics technologies enables unprecedented inquiry of the regulatory relationships in complex biological systems. Thus, the current challenge is now to develop computational methods to integrate this data to power systems level analyses in data that span omics technologies and experimental systems.

It is not trivial to obtain interpretable biological knowledge from disparate data sources. Even datasets collected with a single omic platform and common study design can be heterogeneous. Non-biological artifacts such as batch effects{Leek:2010jq}, library preparation[23], and antibody quality{Park:2009gl} can dominate signal. This problem becomes all the more complex if datasets use different technologies to measure molecular features. Given that many data mining and machine learning algorithms require that all datasets have the same distribution and/or feature space, data must often be heavily manipulated to allow for integration across different technologies. Thus, accounting for undesired technical variation can easily grow to be prohibitively complex when integrating across both biological and technical mechanism[24].

In contrast, transfer learning methods do not require that training and future data have the same distribution, domain, or feature space[25,26]. Instead these algorithms use previously learned knowledge from one or more sources to improve learning of a new target. In particular, statistical{Raykar:2008js}, clustering{Dai:2008de, Dai:2007is}, and dimension reduction{Wang:2008fk, Pan:2011ev} TLMs have been successful in computer aided diagnosis, natural language processing, image recognition, Wi-fi localization, and text classification tasks. TLMs are able to relax many of the constraints of other methods by using the fact that if two datasets are related, there may exist mappings or features to connect the samples and relationships[25]. As a result, these transfer learning methods are uniquely suited for integrating -omic analysis across data modalities and studies.

Thus, we implemented TLMs to developed ProjectoR, an R package for integrated unsupervised analysis of high dimensional omic data from disparate studies. Projection can roughly be defined as a mapping or transformation of points from one space to another often lower dimensional space. Mathematically, this can be described as a function $\varphi(x) = y: \Re^{D} \mapsto \Re^{d}$ s.t $d \leq D$ for $x \in \Re^{D}$, $y \in \Re^{d}$. The projectoR function is S3 class coded for specific analyses including regression, PCA, NMF, clustering as described in the methods and package vignette. ProjectoR uses the relationships (e.g. principal components, clusters, metagenes, modules, etc) defined within a given high dimensional data set, to interrogate related biological phenomena in an entirely new data set. By leveraging relative comparisons within data type, ProjectoR is able to extract shared low-dimensional molecular dynamics, while circumventing many issues arising from technological variation. Specifically, meaningful relationships/features will stratify new data consistent with their underlying biological processes while artifacts or data specific relationships/features result in little to no information content.

Here we begin with a target RNAseq data set from in vitro differentiation of human pluripotent stem cells (hPSC). Using projectoR, we use PCA, correlation, clustering, and NMF analyses to explore this target data set in the context of 4 additional publicly available data sets spanning species, model systems, and omics data types. In a final analysis, we use RNAseq coupled to patient survival data from TCGA to elucidate how cancer prognosis is connected to specific tumor signaling mechanisms common to early development. Together, these results demonstrate that ProjectoR can use the vast repositories of public data for novel hypothesis generation and discovery through *in silico* experimentation.

RESULTS

Target data

Although projectoR is generalized for output from any gene wise analysis, the target dataset was specifically chosen for several important features: (i) RNA's key role as intermediary between the genome, epigenome, and the proteome engenders a single degree of separation between the biological mechanism in this data and any other data. This, we hypothesized, would make biological inference easier, as compared to higher degrees of separation, i.e. methylation projected onto tandem mass spectrometry, where increased disconnect between data types could require more assumptions and abstraction. (ii) Time course data provides the unique opportunity to interrogate biological processes as they unfold as well as the methods used to investigate them. Integrated analyses of this type create the opportunity to work out temporal and regulatory dynamics between different biological processes which in turn may aid in identification of epistatic and/or causal relationships related to cellular phenotypes. This is especially relevant to development, where differential timing of a pathways activation has been directly linked to differential fate specification[27]. (iii) Developmental processes are often the consequence of relative as opposed to absolute level of gene expression making them excellent systems to test the sensitivity and power of the ProjectoR algorithim[27-29]. (iv) The dataset contained a large number of samples, including widely used cell lines, sequenced at extremely high coverage level. Thus, the quality and quantity of the data made it optimal for comparing different analytical methods.

Specifically, the target data is comprised of RNASeq (>79 M reads per sample; mean=127M) from 3 human embryonic stem cell (hESCs) lines and 3 human induced pluripotent cell (hiPSC) lines at 2, 4, and 6 days of a pluripotency, neurectodermal differentiation, or mesendodermal differentiation conditions. A high-level overview of the target dataset is provided by three classes of unsupervised analytical techniques commonly used on –omic data: principal component analysis (PCA), non-negative matrix factorization (NMF), and clustering. Each of these techniques simplifies the data in ways that imparts additional information about its structure. Regardless of method, the end result is a set of features and/or relationships that can be used for TLMs.

ProjectoR links BMP4 pathway activity in in vitro hPSCS and in vivo embryos

Principal component 1 (PC1) captures the direction of the maximal spread of the data reflecting the major reorganization of gene expression by BMP4 treatment. PC2 is correlated with time in all conditions and illustrates variation between cell lines in their differentiation rate (Fig. 2B). By projecting single cell data from human embryos[30], we observed that PC1 clearly separated tissues by level of BMP induction and PC2 continued to be associated with temporal progression (Fig. 2C). It is important to note, the ability of projectoR to related the shared features—BMP4 response and temporal progression—in the two data sets despite of the fact that the embryonic tissues are not the lineages targeted by the in vitro differentiation protocols. The induction of a trophoblast phenotype by BMP4 treatment in hPSC is a well-established paradox in stem cell biology[29] and is reflected in the relationship illustrated by projectoR. Thus, to further refine our understanding of the dynamics captured in PC1, we next turned to an in vitro developmental a datasets specifically designed to parse out the role of BMP4 in hESCs differentiation[27,29,31].

In this study, RNAseq was collected from a single cell line at a common baseline and following treatment with BMP4 or Activin for 36 hours to induce mesoderm or endoderm, respectively[29]. In accordance with what was observed in the target dataset, the projected PC1 splits the conditions in this new dataset by degree of BMP4 pathway activation with the Activin treated samples halfway between the SR controls and the mesoderm counterparts (Fig. 2D). This effect of BMP4 treatment has been shown to be time dependent such that within 24 hrs the same BMP4 signaling necessary for inducing the mesendodermal precursor switches to repressing the endodermal lineage and advancing the mesodermal lineage[27] directly corresponding to the distance traveled in the projection.

To test whether Activin pathway activation was limited in the target data to only the early mesendoderm induction, another microarray dataset looking at BMP4 or Activin treatment in this same cell line for 72 hours was projected into this same PCA space[32]. While the 72 hour BMP4 treated samples moved a significant distance in PC1 (two-sided t-test, p-value = 5.353e-05), the 72 hour Activin treated samples occupied the same space as the 36 hour Activin treated samples (two-sided t-test, p-value=.3519) illustrating that progression along PC1 terminates early in Activin treatment and remains off.

Conversely, separation between the projected 36 hour and 72 hour Activin treated samples was observed in PC2 of the target data. The orthogonal movement of Activin's temporal progression is reflective of the relationship between PC1 and PC2 of the 36 hour microarray data (Fig. 3A). Interestingly, the projected Activin samples move in the opposite temporal direction than the target data samples that define PC2. Together these data suggest that the correlation of the target PC2 with time may be convoluted with progression in underlying signaling pathways that are not obvious from the experimental design. Further, PC2 of the target data may also be associated with Activin related signaling.

ProjectoR characterizes commonalities of divergent BMP4 and Activin signaling

To pin point the location of Activin activity in the target experiment, we correlated the positions of the target data and both microarray datasets projected into entire target PCA space. Hierarchical clustering of the resulting correlation coefficients placed the 72 hour (3 days) BMP microarray samples directly between the day 2 and day 4 BMP target data (Fig. 3A). Similarly, the 36 hour (1.5 days) BMP microarray samples clustered between the day 2 BMP4 target samples and 72 hour Activin microarray samples. Interestingly, the 36 hour Activin samples cluster much closer to target samples in late pluripotency and NSB treatment than to any of the other samples from the microarray including the other Activin treated samples. This may reflect Activin's role as an inhibitor and further illustrates the ability of projectoR to overcome technical artifacts to reveal biological relationships. By comparison, correlation of all gene expression values followed by hierarchical clustering of the two microarray dataset and the target data segregated by technical batch (Fig. 3B).

To deconvolute the signaling pathways and subpopulations captured in PCA of the target data, we next turned our attention to the output of Genome Wide CoGAPS Analysis in Parallel Sets (GWCoGAPS) [33]. The self-organization of differentiation patterns within multi-cellular systems is an iconic feature of developmental systems both in vivo and in vitro[34-36]. By accounting for gene reuse via Bayesian non-negative matrix factorization (NMF), GWCoGAPS has previously been successful in parsing apart the activity of highly related pathways and identifying subpopulations in bulk RNAseq data[33,37].

Consistent with the PCA analysis, the majority of the GWCoGAPS patterns captured the effect of BMP4 on gene expression. We then projected the same BMP4 and Activin treatment microarray datasets used to interrogate PCA space into the GWCoGAPS patterns. Since the second of these datasets was developed to investigating the role of T (Brachyury, BRA) in hESC differential toward mesendoderm, we were able to transfer knowledge of T biology by observing the relationship between the knockdown to the controls in the target data patterns.

Specifically, mechanistic studies reveal that T+ EpiSCs have an earlier and faster responses to BMP4 stimulation than T- EpiSCs[31]. This difference in response rates was captured in the projection of a GWCoGAPS pattern associated with genes induced by BMP4 treatment continuously from the second day of treatment (Fig. 4B). This pattern included high weights for HOX gene clusters that specify position along the body axis (GO:0009952: anterior/posterior pattern specification FDR=5.1e-16, IPR020479: Homeodomain, metazoan FDR=3.34e-08). Since

T is a T-box transcription factor in all nascent mesoderm that plays a known role in axial elongation, we will refer to this as the mesoderm pattern (Fig. 4A).

ProjectoR connects related expression dynamics to epigenetic and spatial regulation

To see if we could relate this mesoderm pattern to the corresponding epigenetic state, we used ProjectoR to interrogate a dataset containing RNAseq and CHiPSeq data for six different histone marks in pluripotency, mesoderm, endoderm, and ectoderm[31]. Not only did the mesoderm RNAseq affirm the mesoderm designation, but also the ProjectoR result showed corresponding enrichment in activating H3K27ac, H3K4me1, and H3K4me3 histone markers and depletion in repressive H3K27me3 histone modification in the mesoderm lineage (Fig. 4C). Thus, the pattern of epigenetic markers output from ProjectoR recapitulated the induction of mesoderm seen in the gene expression pattern.

Taking advantage of the known topological biases of different histone modifications, we decided to test the sensitivity of the projectoR by projecting vectors of binned CHiPSeq reads mapped to the 10KB flanking each genes transcription start site. The resolution of resulting mappings are remarkable and recapitulate known topological biases of different histone modifications (Fig. 4D). Peaks in activating H3K27ac and H3K4me3 ChIPSeq reads localize to the transcription start site, while HSK36me3 enrichment is restricted to the gene body. As strong indicator of the specificity of ProjectoR, the WCE ChIPSeq measures of baseline did are not enriched for any lineage or structure.

Having confirmed this approach for a known lineage in the target data, we next sought GWCoGAPS pattern(s) associated with the endoderm lineage, Activin activity and/or the common mesendoderm precursor. Projection of the RNAseq and CHiPSeq data yielded one pattern associated with endoderm (Fig. 5B) and one pattern associated with both endoderm and mesoderm (Fig. 5A). Using the PatternMarkers statistic[33], we confirmed the identity of these patterns as genes uniquely associated with the endoderm pattern included the markers Sox7, Sox17, Fox11, and PDGFR while those uniquely associated with the mesendoderm pattern included the markers EOMES, T, and ID1 (Fig. 6).

To further investigate the temporal and spatial relationship of these three patterns, we projected iTranscriptome, a spatially transcriptome describing the regionalization of gene expression and cell fates in the mid-gastrulating mouse embryo[38]. While all both the mesendoderm and mesoderm patterns localized to the proximal portion previously reported as associated with BMP signaling[38](Fig. 5D,F), the endoderm pattern was strikingly absent from this region (Fig. 5E). Conversely, both endoderm and mesoderm were enriched in the posterior region of the embryo previously mapped as primitive streak[38].

ProjectoR associates alternative pathway regulation with differences in cancer survival

To find additional evidence for alternative response to BMP in the mesoderm and endoderm patterns, we projected gene expression from a comprehensive glioma study in the Cancer Genome Atlas (TCGA) into the 22 CoGAPS patterns{TheCancerGenomeAtlasResearchNetwork:2015ga}. BMP4-induced differentiation of glioma stem cells (GSCs) is a well establish phenomena and, subsequently, the use of BMPs as biomarkers and potential targeted therapeutics has been well studied[39,40]. Recapitulating several studies establishing BMP4 expression as a prognostic indicator[41,42], the projected mesoderm pattern was significantly associated with tumor grade by Wilcoxon test (1.28e-08, FDR=1.415e-07) and tumor histology by ANOVA (8.15e-103, Fig. 7A,C). Strikingly, the projected endoderm pattern yielded almost perfectly inverted results and was also significant associated with tumor grade by Wilcoxon test (6.4e-12, FDR=9.4-06) and tumor histology by ANOVA (8.01e-123, Fig. 7B,D). Stratifying samples by top 25% vs. bottom 75% in each pattern yielded significantly different survival curves (Fig. 7E,F). The projected mesoderm pattern has increased mortality in the top 25% (p-value = 7.58e−06). Conversely, the upper quartile of the projected endoderm pattern has significantly higher survival rates (p-value = 2.07e−04). Taken together this strongly suggests that differential regulation of BMP signaling in early development is associated with significant differences in cancer survival.

Discussion

Here, we use Projector to illustrate the power of transfer learning to perform integrated analysis and in silico experimentation using public data. An R package, ProjectoR, contains methods for regression, correlation, clustering, principal component analysis (PCA), and non-negative matrix factorizations (NMF) techniques. ProjectoR uses learned weights to transfer features or relationships across datasets. In this manner, basis vectors corresponding to meaningful biological variation can be compared directly, independent of laboratory of origin or technical artifacts. Projection of artefactual basis vectors result in little to no information content. Conversely, biological basis vectors stratify samples consistent with their underlying biological processes. Thus, ProjectoR enable rapid comparisons of multiple data types, tissues, and even across species.

In addition to testing biological hypothesis, a priori knowledge of the biological relationship between datasets can be used to contrast analytical methods or technical platforms. The assumptions and biases of a given analysis or technology can strongly effect its results. Artifacts from sample preparation and processing techniques are pervasive to nearly every comprehensive database. Further, the impact of technical variation on genomics data is highly variable within each experiment. By relying on relative comparisons within data type, ProjectoR is able to circumvent many issues arising from technological variation. Thus, ProjectoR can also be used as a potentially powerful tool for constructing models including multiple data types and levels of regulation. Furthermore, the ability to model relationships between data could be extended to aid efforts towards reproducible research. Thus *in silico* experimentation via projectoR, while not a replacement for bench science, allows for rapid hypothesis testing and development.

Fig. 2 ProjectoR links BMP4 pathway activity in in vitro hPSCs and in vivo embryos

(a) PCA of gene expression values from all four (b) PCA of target data RNAseq. PC1 contains the greatest amount of variance and is strongly associated with BMP4 treatment. PC2 is associated with time in culture. (c) Projection of scRNAseq from human embryos projected into PCs from (b). Using the projected scRNAseq samples, PC1 is associated with in vivo BMP4 activity and PC2 is associated with embryonic age, suggesting a concordance with developmental time. (d) Projection of wild-type samples from two microarray experiments on hPSC differentiation in BMP4 and Activin treatments. Using the projected sample values, PC1 is associated with in BMP4 treatment with advancement in PC1 associated with increased BMP4 exposure time. PC2 is associate with time in culture and an interaction between time in culture and Activin treatment.

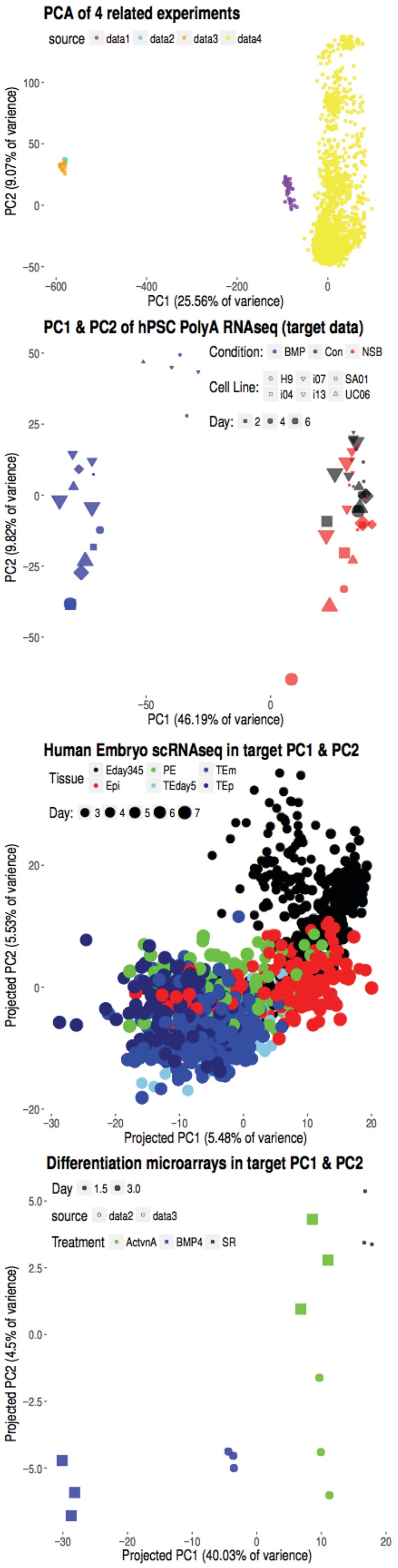
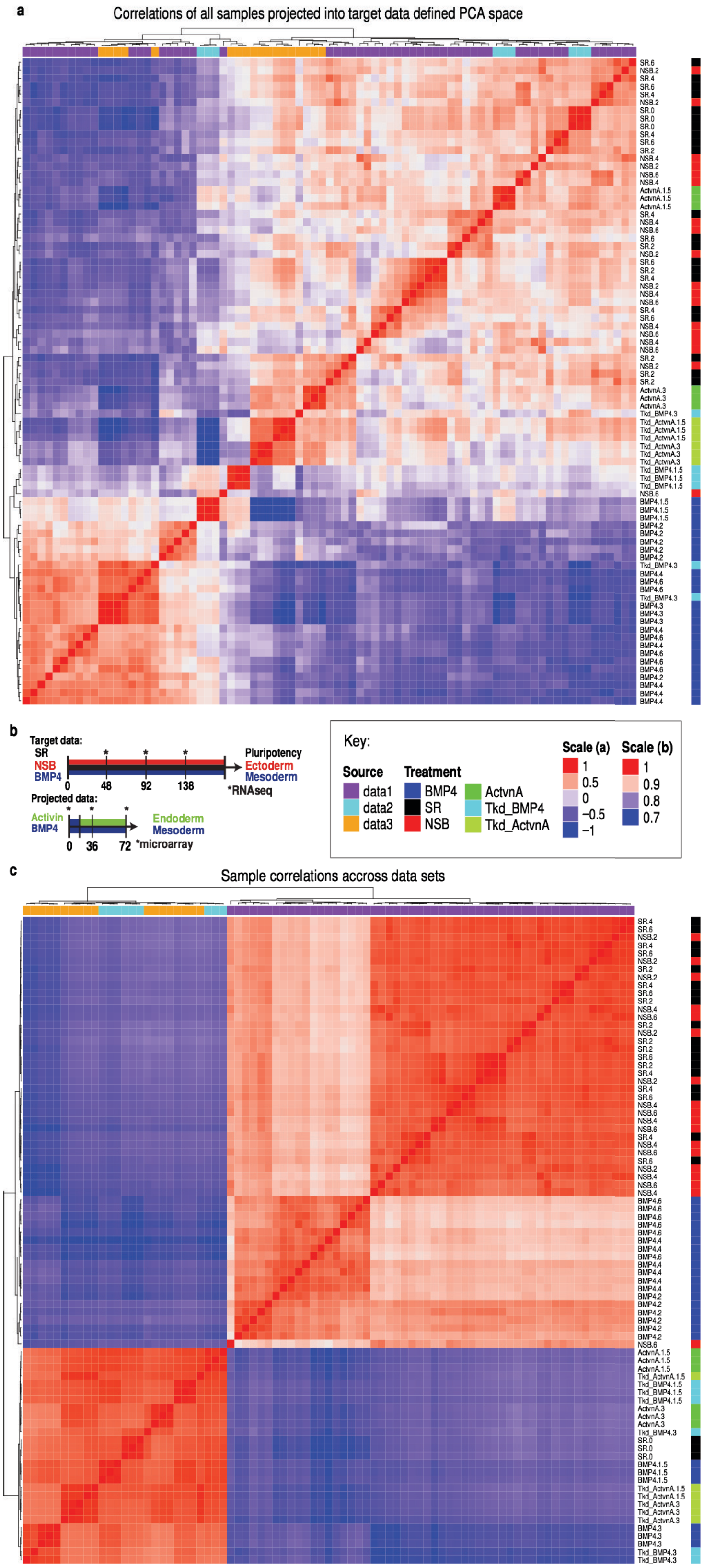


Fig. 3 ProjectoR overcomes technical artifacts to characterizes commonalities of divergent BMP4 and Activin signaling



(a) Heatmap and hierarchical clustering of sample correlation matrix for projected PC loading of from the two microarray dataset and the target data. (b) Diagram of relationship between target and microarray samples. (c) Heatmap and hierarchical clustering of sample correlation matrix for gene expression values from the two microarray dataset and the target data demonstrates clear segregation by technical batch.

Fig. 4 ProjectoR connects related expression dynamics to epigenetic regulation

(a) GWCoGAPS pattern for rise of mesoderm lineage in response to prolonged BMP4 treatment in target hPSC differentiation dataset. (b) Microarray data of hPSCs treated with BMP4 and Activin projected into (a) reveals graded rate response between control and T KD concordant with T's role in mesoderm induction. (c) Projection of normalised gene-wise aggregated ChIPseq and RNAseq from the four primary embryonic lineages into the transcription signature of (a) yields a pattern of epigenetic regulation coinciding with the mesoderm lineage. (d) Expansion of (c) into additional histone modification and along 20kb fragements of the genome (+/- 10kb from the TSS) illustrates the resolution and sensitivity of ProjectoR to capture the epigenetic landscape of this projection.

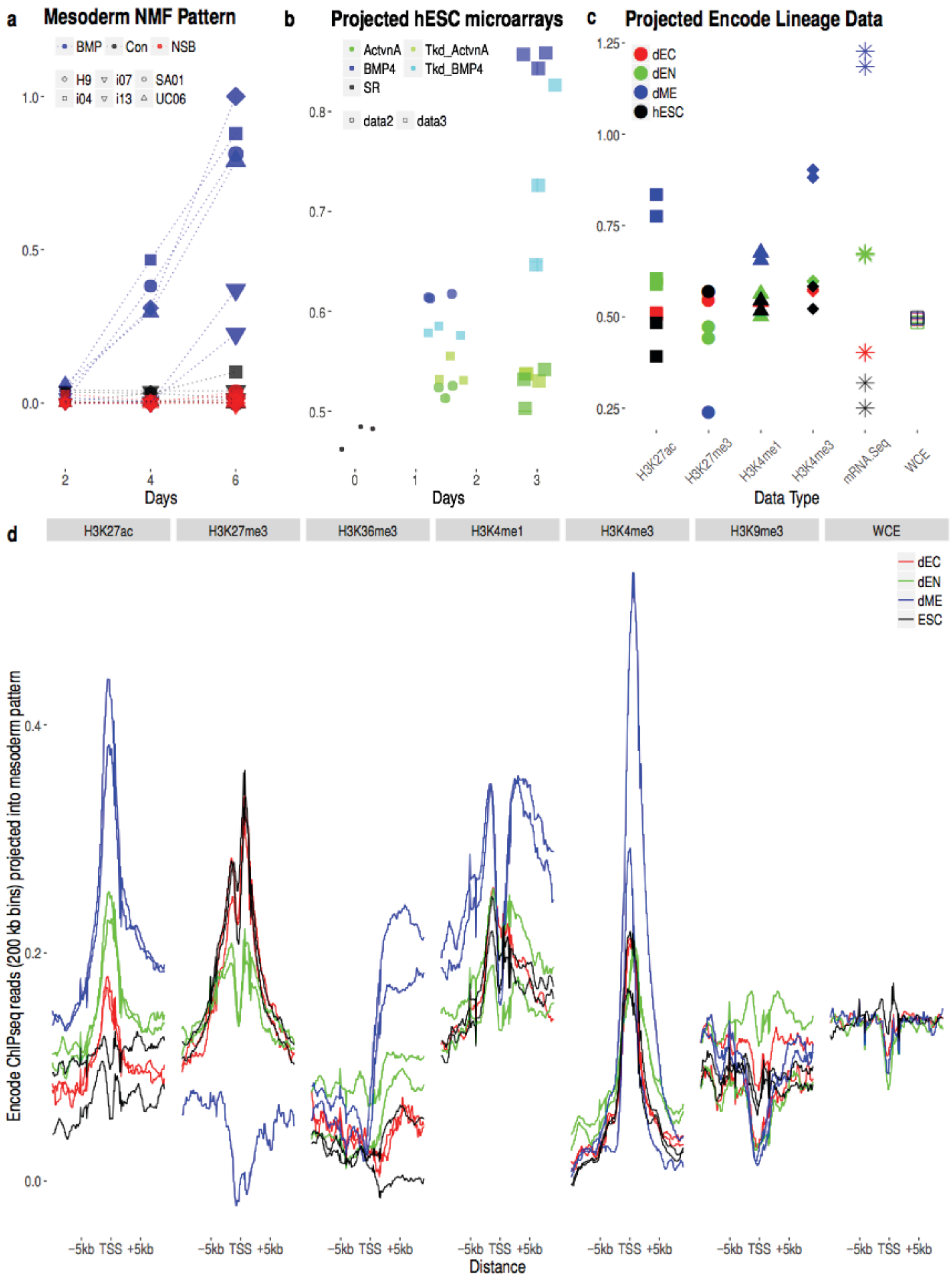
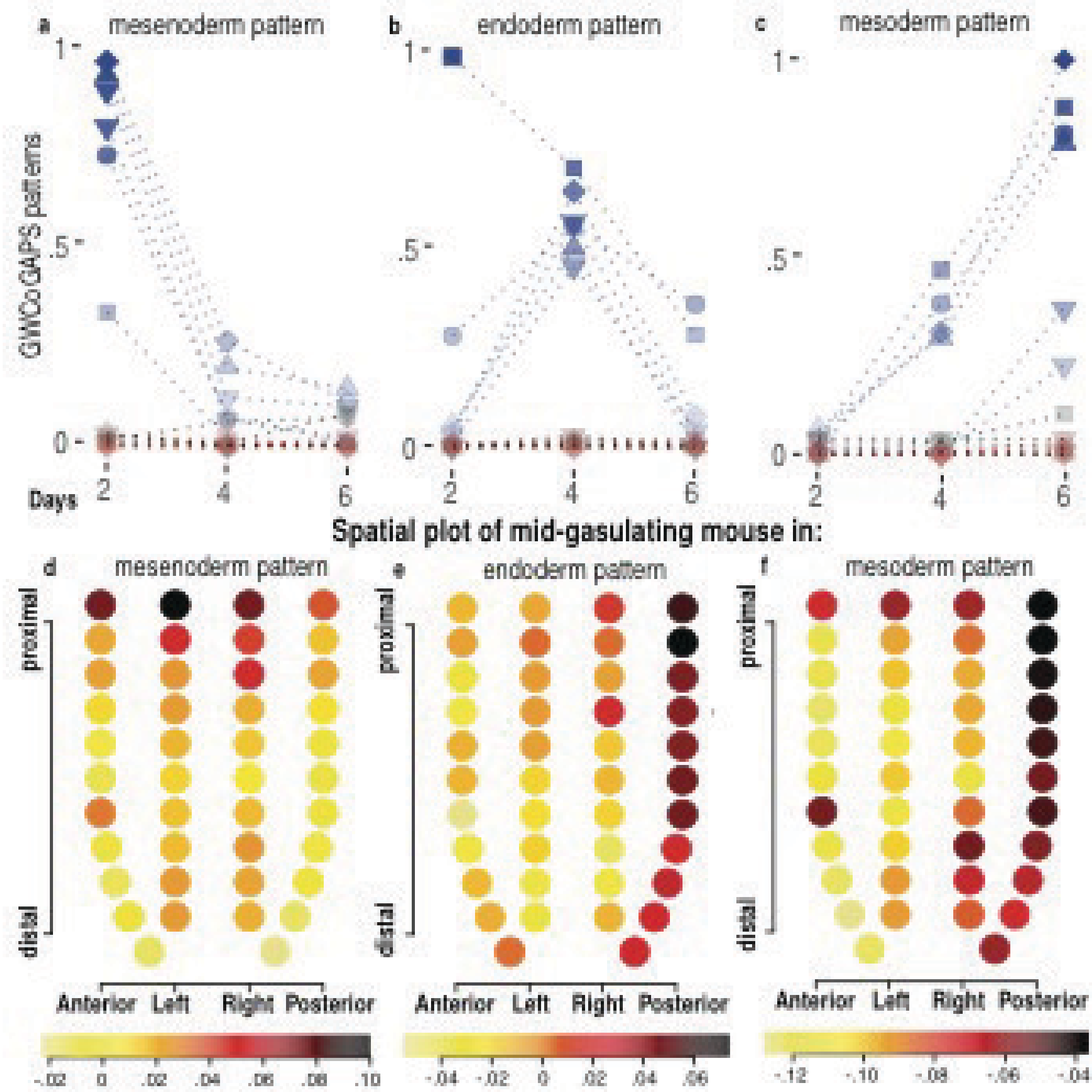

a
Mesoderm NMF Pattern
BMP
Con
NSB
H9
i07
SA01
i04
i13
UC06
1.0
0.5
0.0
2
4
6
Days
b
Projected hESC microarrays
ActvnA
Tkd_ActvnA
BMP4
Tkd_BMP4
SR
data2
data3
0.8
0.7
0.6
0.5
0
1
2
3
Days
c
Projected Encode Lineage Data
1.25
1.00
0.75
0.50
0.25
dEC
dEN
dME
hESC
H3K27ac
H3K27me3
H3K4me1
H3K4me3
mRNA.Seq
WCE
Data Type
d
H3K27ac
H3K27me3
H3K36me3
H3K4me1
H3K4me3
H3K9me3
WCE
dEC
dEN
dME
ESC
Encode ChIPseq reads (200 kb bins) projected into mesoderm pattern
0.4
0.2
0.0
−5kb TSS +5kb
Distance

Fig. 5 ProjectoR maps spatial localization of GWCoGAPS gene expression patterns



(a) Projections of activiating marks including H3K4me3 are enriched in both DE and VE patterns in endoderm samples. (b-d) A planar representation of the cylindrical mid-gasulating mouse embryo. The position of each cell sample is defined by the coordinate position in the cross-sectional plane (A, P, R, L) and the distal-proximal axis (1–11). The relative magnitude of each projected cell is presented in a color-coded format. Enrichment in the mesoderm (b) and VE patterns (d) co-localized in the proximal portion previously reported as associated with BMP signaling, while enrichment in the DE pattern (c) occurs in the distal portion of the embryo previously mapped as endoderm. Enrichment in all three patterns is also localized to the posterior region of the embryo.

Fig. 6. PatternMarkers from GWCoGAPS patterns associated with Mesendoderm, Endoderm, and, Mesoderm

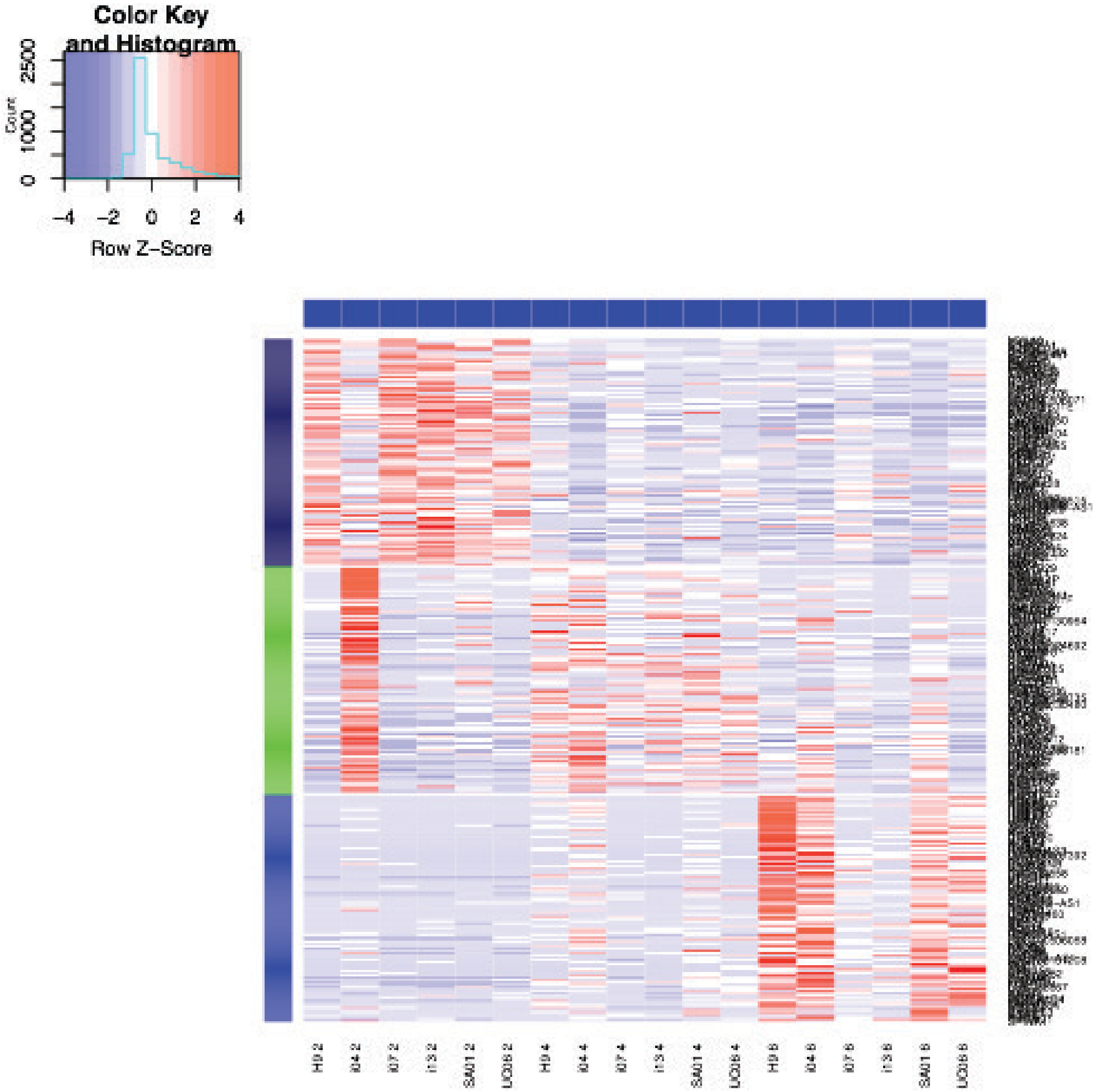
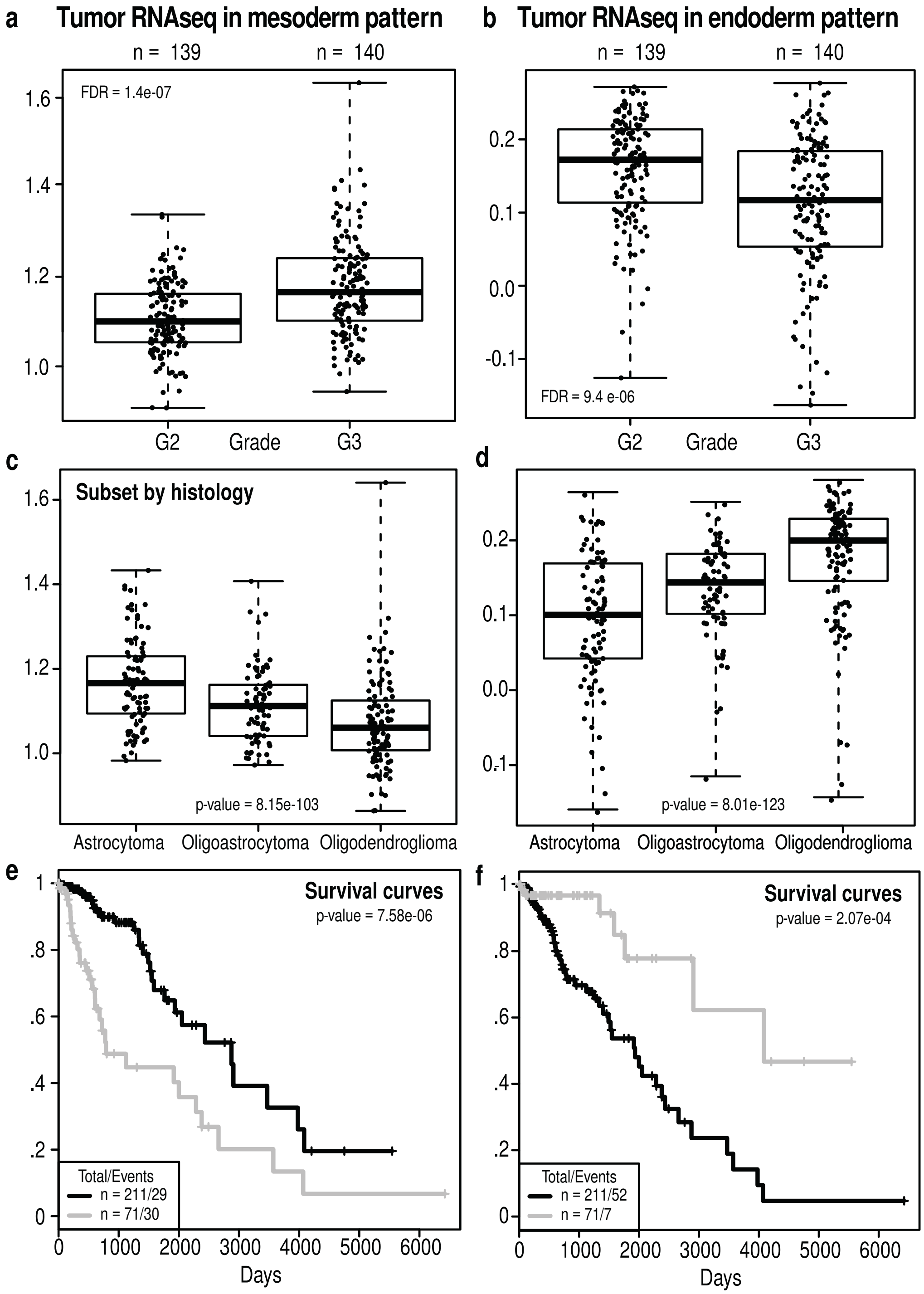


FIG. 7 ALTERNATIVE PATHWAY REGULATION ASSOCIATED WITH DIFFERENCES IN CANCER SURVIVAL



Projection of TCGA glioma associates histology and survival associated with different lineages and BMP4 pathway activity. (a,b,c) RNAseq of low grade glioma from TCGA projected into mesoderm and endoderm patterns, respectively. (c,d,e) Survival curves for top 25% vs. bottom 75% of projected samples indicate significantly different prognosis for tumors stratified by these genetic signatures.

Defining the origins of variation in human pluripotent stem cells

Abstract

Defining and controlling variation in early differentiation is necessary for the optimal use of human pluripotent stem cells (hPSCs) in cell therapy, disease modeling and therapeutic discovery. To achieve this goal, we require new tools to quantify differences between pluripotent lines as they first reveal their differentiation potential. We report that hPSCs spontaneously self-organize to form an epithelium with distinct zones representing major embryonic axes and differences between cell lines were seen in these morphogenic mechanisms that specify early fates. In addition, transcriptional differences between lines were defined in the early stages of stem cell development that were sustained in adult tissues from the same donor. These signatures provide a strategy to define the mechanistic basis for differentiation bias between pluripotent cell lines and to determine how this bias influences cellular behavior throughout the life of an individual.

Introduction

The transition from a single cell, the zygote, through a few hundred pluripotent cells, to several thousand committed cells forms the basic architecture of a human embryo [43,44]. hPSCs share defining features with cells of the first epithelium, the epiblast, that forms a new individual[45,46] and great attention is currently focused on the epigenetic mechanisms that regulate the generation of adult cell types from hPSCs[47,48]. However, the use of hPSCs in bio-medicine is limited by the difficulty in defining the consequences of variation between hPSC lines in the first steps of development **[49,50]. Previous studies using transplantation or dissociation of the cells of an embryo emphasize the primary role of cell interactions in embryonic patterning. Here we test the hypothesis that interactions between hPSCs as they form an epithelium can recapitulate key aspects of the architecture of the human embryo and that stable variation between lines will be evident in these morphogenic processes. Novel high-content imaging tools showed that hPSCs generated spatial domains representing major axes of embryonic development. Remarkably, stable quantitative differences between lines were observed in these processes that model early events in human embryogenesis. High-resolution analysis of RNA-seq data defined stable transcriptional traits that regulated early fate decisions and were sustained for the entire lifetime of an individual. This work shows that the dynamics of hPSC self-organization provides a powerful tool to define the biology of individual human genomes.

The spatiotemporal dynamics of epithelial morphogenesis in vitro

In amniotes, the epiblast is an epithelial sheet comprised of pluripotent cells that is first specified along the primary embryonic axis to generate the posterior and anterior parts of a new individual[51]. To search for variation between hPSC lines in these first morphogenic events, we established a cell culture system where dissociated hPSCs spontaneously generated a two dimensional epithelium (Fig. 8A, see methods for details of this procedure). To define the initial emergence of cellular heterogeneity within a cell line, the spatial distribution of the pluripotency regulators POU5F1 (OCT4) and NANOG were tracked over time in the human embryonic stem cell (hESC) line SA01. Automated mapping of the location of every cell relative to the nearest epithelial edge defined rapid emergence of cellular heterogeneity. While initially uniformly expressed, a clear distinction in expression levels rapidly emerged as cells on the edge of the epithelium showed higher levels of POU5F1 (OCT4) and NANOG (Fig. 8A). To monitor temporal dynamics of POU5F1 and NANOG gene expression, the mRNA levels of these transcription factors and the class B SOX genes (SOX2, SOX3, and SOX21) were examined on different days within a passage. NANOG and POU5F1 mRNA levels were initially high and decreased with time while the SOX genes showed the inverse pattern (Fig. 8B). These observations show that precise spatio-temporal rules constrain the spontaneous emergence of cellular heterogeneity during the cycle of hPSC self-renewal.

Activation of the PI3K/AKT/mTOR pathway has been shown to play an important role in regulating protein translation, cell growth and the maintenance of self-renewal in hPSCs[52]. To determine the levels of protein synthesis across the edge and core zones, incorporation of O-propargyl-puromycin (Op-PURO) into nascent polypeptide chains during protein synthesis was assessed[53]. Op-PURO showed an enhanced signal in the edge zone indicating variation in protein synthesis across the epithelium was already present on Day1 (Fig. 8C). Preferential activation of

PI3K/AKT/mTOR pathway was also observed in the edge zone, using antibodies and dose-response analysis of small molecule inhibitors (fig. S1). Neuregulin 1β (NRG1β) acting through the ERBB2/ERBB3 receptor dimer promotes hPSC growth through the PI3K/AKT/mTOR pathway[52]. These receptors were enriched in cells in the edge zone and these cells also showed a rapid preferential response to addition of exogenous NRG1β monitored by phosphorylation of AKT (Ser473), the target of mTOR complex 2 (mTORC2; fig. S1B and C). Inhibition of AKT (Ser473) phosphorylation following treatment with a range of kinase inhibitors including the mTORC1/2 inhibitor AZD80552, showed elevated base-line activation and greater dynamic range of inhibition in the edge zone (fig. S1D). These results show that exogenous and endogenous inputs to the self-renewing AKT signaling pathway define the rapid emergence of distinct edge and core zones in the self-renewing pluripotent epithelium.

To define the role of this spatial segregation on fate specification, hPSCs were treated with agonists and antagonists of BMP/TGFb signaling that induce distinct differentiation trajectories to mesendodermal and neurectodermal fates, respectively[29,54]. When this signaling pathway was inhibited by treatment with NSB (Noggin, a BMP antagonist plus SB431542, a TGFb type 1 ALK receptor-selective inhibitor) on Day0, cells in the core expressed high levels of neurectodermal fate regulators, SOX21 and OTX2 on Day4 (Fig. 8D and fig. S2A and B;[55,56]. The cells at the edge of the epithelium remained in the NANOGhi state suggesting that they were resistant to neural induction. When BMP4 treatment at Day0 was used to activate this pathway, all the cells rapidly induced expression of Brachyury (T) or CDX2, transcription factors that mediate differentiation toward mesendoderm[29]. Importantly, when BMP4 was introduced at a later time point when distinct edge and core zones had been established (Day2), only cells in the edge expressed these mesendodermal fate regulators (Fig. 8E). A previous report showed gastrulation-like events when cells were plated on micro-patterned surfaces[576]. Another recent analysis of single cell RNASeq and ChIPSeq data defined two similar states in self-renewing mouse ES cells[58]. Our data extend these findings by showing that when self-renewing hPSCs spontaneously and rapidly self-organized into two dominant states or domains; an early state at the edge expressing high self-renewal signaling and biased to generate mes-endodermal fates followed by a state that formed in the core of the epithelium biased to form neurectoderm (Fig. 8F).

Cell line variation in morphogenesis

Next we determined if spontaneous morphogenesis could reveal differences between cell lines. To test this hypothesis, nuclear SOX21 protein expression was monitored through time in 3 ES (SA01, UC06, H9) and 3 iPS (i04, i07, i13) cell lines during self-renewal (SR) and differentiation. When neural differentiation was induced by NSB treatment, SOX21 expression increased in the different lines to varying levels over the 6 day period of culture with the ES line SA01 showing the highest level and the iPS line i04 showing the lowest level of this neural fate regulator and core marker (Fig. 9A). The similar relative levels of SOX21 in these cell lines across SR and NSB conditions suggested cell line-specific differences in the potential for neurectodermal differentiation were already present in pluripotency before neural induction (comparing mean SOX21 levels across condition within line and day: r=0.84, p=3.6e-7, fig. S5; see methods for further statistical analysis on multiple replicate cultures showing that variation in SOX21 expression is a robust discriminating feature of cell lines).

To determine if this variation in SOX21 expression results from the generation of distinct morphogenic zones, SOX21 and NANOG expression were imaged in the two male lines, SA01 and i04 on Day6 of treatment with NSB (Fig. 9B). This analysis revealed that the SOX21hi core zone was prominent in the SA01 line and the NANOGhi edge zone was favored in the i04 line. When these two cell lines were exposed to increasing doses of neurectodermal inducers, SA01 cells showed increased induction of SOX21 while i04 cells were resistant to neural induction at all doses (fig. S2C). This result suggests that the variation between lines is a consequence of robust differences in their ability to generate the edge and core zones that give rise to distinct cellular fates.

The nervous system and anterior mes-endodermal fates are generated from the anterior domain of the epiblast[51]. To determine if there were cell line differences in the generation of regionally-distinct zones, expression of anterior and posterior mes-endodermal genes were assessed by immunocytochemistry [59] by treatment with BMP4 for 24 hours at day 2 when distinct edge and core zones had been established (Fig. 9C). Posterior mesendodermal regulators (GATA3, ID1 and p-SMAD1/5) were upregulated within 100μm in the edge of the epithelium in both cell lines. In contrast, the anterior mesendodermal regulators (NANOG, SOX17 and GATA4) were expressed in the core zone but at lower levels in i04 compared to SA01. This analysis demonstrates stable phenotypic variation between hPSC lines in mechanisms that establish distinct morphogenic zones.

Variation in molecular mechanisms of anterior neural differentiation

Transcriptional differences between hPSCs with different genomes has been reported but the structure and functional impact of this variation in gene expression has not been defined[49,60]. To determine differences in the transcriptional dynamics of pluripotent cell lines as they establish morphogenic zones, RNASeq was performed on the 3 ES and 3 iPS cell lines at 2, 4, and 6 days of the SR, NSB, and BMP4 (>79 M reads per sample; mean=127M; see Methods and fig. S3 and tables S1 and S2 for details on cell derivation, sequence quality and assembly). The changing gene expression in these samples was first assessed by principal component analysis (Fig. 9D and fig. S4). Principal component 1 (PC1) indicated a major reorganization of gene expression following treatment with BMP4. PC2 defined trajectories of gene expression change over time in all conditions. PC2 also showed variation between cell lines in the timing of their differentiation to mes-endodermal or neurectodermal fates. The advanced position of the cell line SA01 at every time point in NSB showed that this cell line undergoes relatively rapid neural differentiation.

To more fully explore the differences in the efficiency of neural differentiation across lines, PCA was conducted on the NSB differentiation data alone (Fig. 9E). PC1 within the NSB condition provided a metric for genome-wide transcriptional change where each gene has a weight defining its contribution to this developmental trajectory. The ranking and slope of transcriptional change captured by PC1 defined differences between cell lines in the first steps of neural differentiation. To investigate the origins of these different efficiencies in response to NSB, we projected SR data into the individual gene weights of NSB PC1 (see Methods). The recapitulation of the cell line rankings showed that the transcriptional differences in neurectodermal differentiation were already present in pluripotent cells (Fig. 9E; comparing NSB

PC1 with the projection of SR data into this PC within line and day: r=0.88, p=1.7e-6, fig. S5B). Correlation of nuclear SOX21 protein levels in NSB (Fig. 9A) with this global measure of transcriptional change in NSB PC1 (fig. S5C, r=0.89, p=1e-6) was a further demonstration of the stability of this morphogenic difference across lines. This same analysis within the BMP4 treated samples revealed similar predictive expression dynamics in the self-renewing state, with line i04 showing the most pronounced bias towards mesendodermal differentiation during SR (fig. S5D). The concept of lineage priming defines a central problem in stem cell biology, how different cell types emerge from a common precursor[61]. This demonstration of stable variation in global transcription is consistent with a lineage priming model where cell lines show distinct expression of genes in SR that interact with morphogenic mechanisms to bias their subsequent differentiation (Fig. 9F).

To more precisely define the variation in the transcriptome, we employed a Bayesian non-negative matrix factorization (NMF) method, CoGAPS[62]. Unlike clustering methods, CoGAPS allows a gene to contribute to more than one pattern capturing the pleiotropy that results from relative levels of gene expression that are known to play a central role in defining distinct pluripotent states[63-65]. PCA and clustering analyses of these data are provided for comparison (fig. S4, S8 and 9). CoGAPS assigns gene-specific amplitudes that quantify how much of each pattern is needed to model the full expression of each gene (examples of CoGAPS mediated decomposition of gene can be inspected in fig. S6). Here we analyze 22 patterns that were robustly identified with the CoGAPS algorithm (patterns in fig. S7, see Methods for details). ANOVA assessing the effects of condition, day, and line in these patterns identified two types of pattern defining (1) dynamic transcriptional change over time or condition (p-values of <.001 for Condition and/or Day) and (2) transcriptional differences between cell lines that were stable over time (p-values of <.001 for Line) (Fig. 10A).

The CoGAPS patterns that change over time captured transcriptional modules that regulated self-renewal and differentiation. The core pluripotency genes POU5F1, SOX2 and NANOG were ranked 6, 8 and 41 of 21,174 genes in pattern 7 (Fig. 10B). Consistent with the large impact of BMP4 on gene expression (Fig. 9D), mesendodermal fates were described in 6 of the patterns (2, 3, 5, 6, 8 and 9). One of these, pattern 3, identified a graded response to BMP4 treatment and HOX gene clusters that specify position along the body axis, [66] were highly enriched in this pattern (p=1e-6; Fig. 10B). To test how well these patterns capture conserved transcriptional modules in mammalian development, RNASeq data derived from differentiating mouse embryos was projected into the gene amplitudes from these 22 CoGAPS patterns (Fig. 10B; see Methods). The self-renewal pattern (# 7) defined transcriptional activity that was down-regulated during early mouse embryo development, delineating the loss of pluripotency. In contrast, BMP-induced genes in hPSCs (pattern 3) increased between days 6 and 8 of mouse embryo development, paralleling gastrulation. The individual gene amplitudes underlying these patterns clearly intersect with epigenetic mechanisms controlling lineage-specific gene expression as observed by projection of ChIPseq data in differentiating hPSCs (fig. s10; (5)). This analysis demonstrates that the CoGAPS decomposition defines conserved transcriptional mechanisms in pluripotency and early differentiation.

To further validate this informatic analysis of the dynamics of gene expression, we focused on the first steps in neurectodermal differentiation defined by 2 CoGAPS patterns (15 and 12). Consistent with the PCA, the SA01 line showed the most rapid differentiation in both of these patterns (Fig. 10C). The top transcription factors ranked in pattern 15 were HES3, OTX2, POU3F1 and SOX21 (ranked 3, 12, 26 and 46) and all are known to be important regulators of early neurectodermal differentiation[55,65,67]. Many transcription factors known to specify neurectodermal lineage commitment including PAX6, SOX1, SOX10, SOX11, EMX2, FEZF2, WNT1, NEUROG2, and HES5[68] were represented primarily in subsequent steps in neural differentiation captured in pattern 12 (Fig. 10C and table S3). Immunocytochemistry confirmed that these neural fate regulators were induced in the core zone (fig. S11). This analysis suggests that an early step in neural specification induced by NSB was regulated by the transcription factors highly ranked in pattern 15.

To determine if transcription defined by pattern 15 has a specific role in the transition from pluripotency to neural lineage commitment, we initially focused on SOX21 as this gene was highly ranked (46 of 21,174) only in this dynamic pattern. The sequential gene expression change in neural differentiation was most evident in the cell line SA01 where pattern 15 was already elevated by Day2 and was diminished by Day6, while pattern 12 was induced first on

Day4 and increased further on Day6. In SA01, knockdown and overexpression of SOX21 regulated expression of the pattern 12 genes OTX2, SOX1 and PAX6 (Fig. 10D and fig. S12). To further map the spatial and temporal role of SOX21, three independent frame-shift mutations in the DNA binding HMG domain of both alleles were generated by CRISPR/Cas9 technology (fig. S13). In NSB, the average of the 3 SOX21-KO cell lines generated with different guide RNAs to minimize off-target effects showed showed elevated levels of the pluripotency genes NANOG and SOX2 in a region extending up to 300μm from the edge of the epithelial sheet (Fig. 10E). SOX3, another class B SOX activator that interacts with SOX2 and SOX21 was induced in a restricted zone close to the epithelial edge. Expression of both SOX3 and the fore-brain master regulator OTX2 were repressed in the core zone when SOX21 was absent. These data show that SOX21 regulated the spatially ordered transition from pluripotency to anterior neurectoderm that occurs in the self-organized epithelium established by hPSCs (Fig. 10G).

The morphogenic analysis suggested that anterior fates in both the neural and mes-endodermal lineages were more efficiently generated in SA01 than i04. To determine the effects of loss of SOX21 in the putative anterior mes-endoderm, BMP4 treatment was applied on day 2 when the edge and core zones have already formed. Under these conditions the average of the 3 SOX21-KO cell lines generated with different guide RNAs showed elevated expression of anterior mes-endodermal regulators T and GATA4 in the core zone (Fig. 10F). T (Brachyury, BRA) is a T-box transcription factor that plays a conserved role in defining the spatial organization of bilaterian embryos[29]. Genes marking the posterior endoderm (CDX2, GATA3, and ID1) were elevated in the edge zone but showed no change when SOX21 expression was absent. Consistent with the early appearance of pattern 15, these data show that SOX21 regulates fundamental aspects of the emergence of spatial patterning in hPSCs (Fig. 10G).

Transcriptional signatures of individual cell lines and genomes

In addition to the dynamic patterns defining transcriptional change in self-renewal and differentiation, the CoGAPS decomposition identified patterns that were invariant across time and treatment for each cell line. In a hierarchical clustering of the CoGAPS patterns, the transcriptional signatures of individual hPSC lines formed a distinct, tightly clustered branch

(Fig. 11A). When gene expression datasets from other institutions that included these cell lines[60,69] were projected into the cell line specific signatures defined by CoGAPS, the same lines were identified (Fig. 11B and fig. S14). These findings extend the identification of genetic differences as a determinant of transcriptional variation in human pluripotency[49] by demonstrating that these transcriptional signatures were robust, present in both the pluripotent and differentiated states and that these signatures share a common gene expression structure.

To test the generality of the dynamic and stable patterns defined by CoGAPS, we generated duplicate iPS cell lines from 3 donors whose brain tissue was also obtained and RNA sequenced post-mortem (see Methods). Importantly, CoGAPs analysis of RNAseq data from these new lines and projection analyses demonstrated that the differentiation dynamics defined by CoGAPs in the original 6 lines were also present in the 6 new lines (fig. S15 and S16). In addition, single transcriptional ID signatures stable across conditions and duplicate lines were obtained for each of the 3 new donors studied (Fig. 11C and fig. S17). Projection of RNASeq data from adult prefrontal cerebral cortex samples from these and many more donors into these donor-specific patterns demonstrated that these transcriptional signatures observed in differentiating pluripotent cells were also present in the mature tissue of these same donors (Fig. 11C and fig. S17). The presence of an ID signature in multiple clones of pluripotent cells and in differentiated tissues derived from the same donor indicates a significant genetic contribution to these transcriptional ID signatures. A cell line pattern specific to only a single clone from an individual donor was also identified (fig. S15; pattern 14 in the second CoGAPS analysis). Projection of the prefrontal cortex RNASeq data into this pattern showed this signature was not present in differentiated tissue of the donor (fig. S17D), suggesting an epigenetic origin of this inter-clonal variation. These data show that genetic and epigenetic components of cell line specific transcriptional signatures can be distinguished by the CoGAPS decomposition.

The GTEx Project is a multi-site consortium established to enable an understanding of the relationship between genetic variation and gene expression in individual humans[70]. The GTEx analysis of RNASeq data from tissues obtained from 175 donors illustrated the need for additional tools to define gene expression differences between individuals. When CoGAPS decomposition was applied to the GTEx gene expression dataset spanning many mature tissues sampled from the same individual donors along with patterns defining different tissues, transcriptional signatures of individual donors across all tissues were clearly identified (Fig. 11D and fig.S18). The GTEx study also defined a distinct set of eQTLs associated with transcriptional variation in multiple tissues from the same individuals. Genes involved in these multi-tissue eQTLs were highly enriched in the CoGAPS ID signatures (genes in top 2% of GTEx multi-tissue eQTLs: p=1.3e-3 to 1e-6; bottom 2%: p=0.08 to 0.83). This analysis shows that ID signatures are a stable aggregate of genetic effects on gene expression in multiple tissues from the same donor from pluripotency to old age. To our knowledge this is the first demonstration of transcriptional signatures that may contribute to cellular phenotypes throughout the life of an individual.

Functional impact of transcriptional signatures

The identification of transcriptional signatures that identify pluripotent, differentiating, and mature cells of an individual human donor raises the important question of the functional consequences of this transcriptional diversity. In contrast to the variance analysis and clustering methods that have been used to determine that genetic variation is the greatest contributor to transcriptional variance between hPSC lines[49,60], the decomposition of expression data by CoGAPS defines the structure of genetically driven transcriptional differences for every gene between each donor. Here we explored the predictive power of ID signatures to explain the morphogenic differences between cell lines, specifically the difference in neural differentiation between the cell lines SA01 and i04.

OTX2, a bicoid homeobox transcription factor responsible for forebrain differentiation[71,72] was highly ranked (102) in the SA01 ID signature and this amplitude was more than 3-fold greater than its level in the i04 ID signature (Fig. 12A). The ability of OTX2 to specify forebrain is opposed by the action of the homeobox GBX2 that correctly positions the boundary between the fore-brain and the hind-brain[73]. GBX2 was highly ranked in the i04 ID signature and was minimally represented in the SA01 ID signature (Fig. 12A). Along with GBX2, other genes associated with posterior regions of the epiblast, including EOMES, FN1 and NR5A2, were highly-ranked in the i04 ID signature[51,74]. GBX2 was first identified as a RA response gene expressed in the posterior neurectoderm[75,76]. The role of retinoic acid (RA) in the induction of hindbrain neural stem cells has been intensively studied in powerful models of early neural development in vertebrates[77]. RA acts through canonical sets of target genes[78] that were enriched in the i04 but not in the SA01 ID signature (Fig. 12A). These observations show that key regulators of anterior and posterior neural fates are differentially represented in the ID signatures of SA01 and i04 cell lines and predict that the i04 cell line should preferentially differentiate to hind-brain fates.

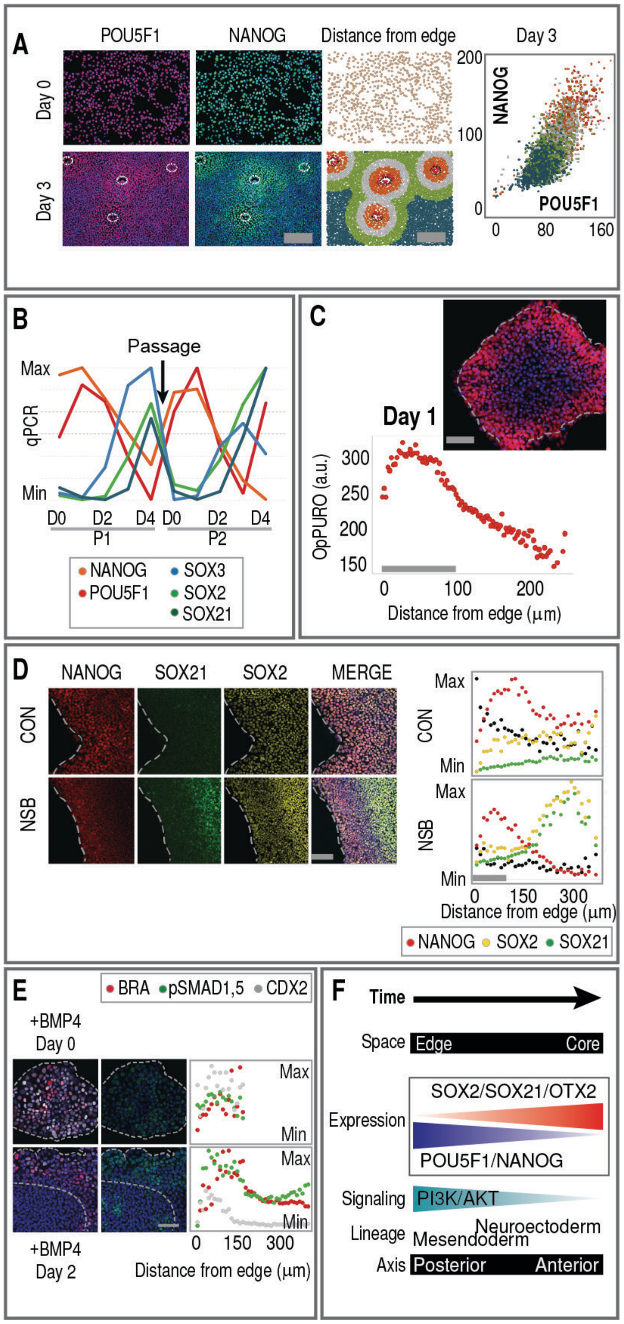
To test this possibility, SA01 and i04 cells were treated with RA during neural induction with NSB. Expression of the homeotic genes HOXB1 and HOXB4 report on a highly conserved transcriptional signaling system controlling key features of hind-brain development in vertebrates[79]. The transcription factors OLIG2, ISL1, and PHOXB2 report on the differentiation of hindbrain motor neurons[42,80]. A dose response study revealed that i04 and SA01 cells were differentially sensitive to RA treatment (Fig. 12B and C). In the absence of RA, the relative levels of OTX2 and GBX2 protein were consistent with their transcriptional signatures (Fig. 12A). In response to 0.1μm and 0.5 μm RA, OTX2 expression was inhibited in both cell lines but elevated GBX2 and HOXB1 expression was only observed in i04. HOXB4 positive cells were observed in i04 at 1μm RA, consistent with the known scalar response of HOX gene clusters. The spatial analysis showed that GBX2 was induced in the core region of the epithelia while HOX positive cells were located at the edge of the epithelial sheet, consistent with the migratory behavior of hindbrain cells expressing HOX genes (Fig. 12C). Further differentiation showed i04 cells more efficiently generated cranial motor neurons compared to SA01, demonstrating that ID signatures predict a bias in terminal fates (Fig. 12D). This demonstrates that stable transcriptional differences between hPSCs predict variation in both early morphogenesis and terminal differentiated fates

A recent report used the generation of inter-species chimeras with hPSCs to propose that a cell line could exist in alternate states with a bias towards either anterior or posterior fates[50]. In this study, relative levels of OTX2 and GBX2 expression were correlated with the change between positional epigenetic states defined by introduction of hPSCs into non-human embryos. Projection of the SA01 and i04 transcriptional signatures into RNAseq from these regionally specified pluripotent cells demonstrates the anterior bias of SA01 and the posterior bias of i04 that we have described here (fig. S19). GBX2 expression is thought to play an important role in regulating the naive pluripotent state found in mouse ES cells and OTX2 regulates, at least in part the transition from the naïve to primed pluripotent state in the mouse[81,82]. The data presented here show that OTX2 and GBX2 play important roles in regulating alternate pluripotent states in the human system and relate to an interesting contemporary debate about on mechanisms controlling pluripotent states in humans and other species. Because of ethical concerns around using inter-species chimeras to analyze variation between hPSCs, the allocation of US federal funds for this purpose has been restricted (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-158.html).

The important role of cell interactions in embryology has been known since the discovery by Spemann of the organizer in the amphibian embryo and emphasized by extensive recent work in the mammalian system[43,44,51]. Here we set out to test the hypothesis that interactions between hPSCs as they form an epithelium can recapitulate aspects of the architecture of the human embryo and address a central issue in the use of hPSCs, the lack of phenotypic consistency within individual cell lines and unexplained variation across lines[83]. The imaging and informatics strategies we present here showed stable differences between hPSC lines in the mechanisms that execute the early steps of embryonic fate choice leading to distinct developmental paths. Remarkably, these cell line specific transcriptional signatures were sustained in differentiated adult cells suggesting that stable transcriptional variation may continuously influence the biology of an individual (Figure 12E). The future use of stem cell technologies in developing new cell and pharmacological approaches to human disease will require systematic definition of this variation in early differentiation.

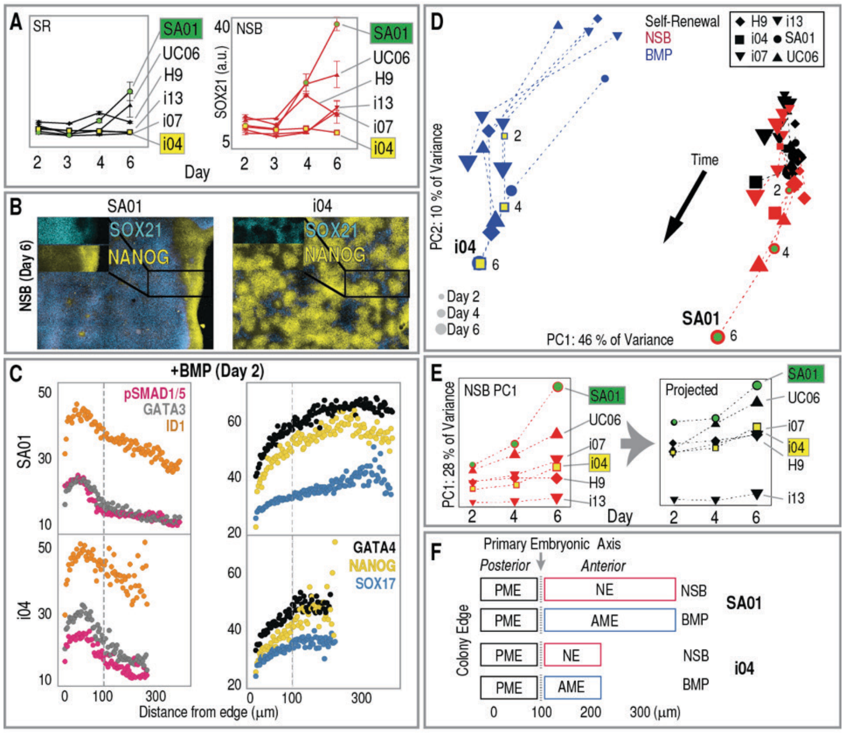
The spectrum of germline and somatic variation in human brain development is the focus of intense current interest[84,85]. The definition of stable variation across human pluripotent cells presented here encourages the expanded use of in vitro systems to determine how many hPSC states with distinct developmental properties can be obtained from a single genome. It will also be important to define how genetic variation influences the range of hPSC states with distinct developmental potentials. We propose that reconstructing the early stages of development in the laboratory using larger numbers of replicate hPSC lines across multiple donors will provide a strategy to systematically map genetic and epigenetic origins of variation in hPSCs and the impact of this variation on cellular function and disease risk throughout the lifetime.

FIG. 8 SPATIOTEMPORAL DYNAMICS OF PLURIPOTENCY, LINEAGE SPECIFICATION, AND GROWTH SIGNALING REGULATORS REVEAL IN VITRO MORPHOGENESIS OF HUMAN PSCS.



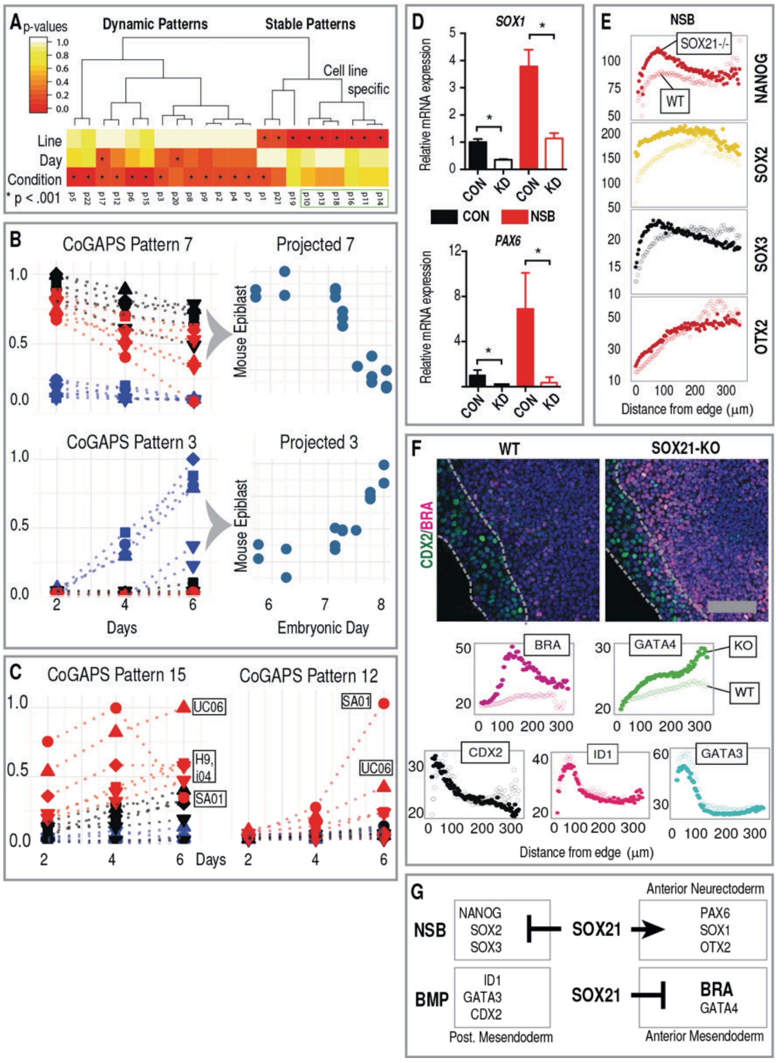
(**A**) Distribution of POU5F1 and NANOG in SA01 over time showing emergence of distinct zones during formation of an epithelium. Transcription factor expression in relation to the distance of each cell from the nearest edge (dotted line). Colors correspond to those in distance map (indicating percentiles from edge: 0-5%=red, 5-25%=orange, 25-50%=grey, 50-75%=green, and 75-100%=blue). ANOVA comparing day 3 protein expression across domains: p<2.2e-16 for both POU5F and, NANOG. (**B**) Cyclical expression of pluripotency regulators during passage. Relative qPCR expression levels for each gene (correlation of average expression of POU5F1, NANOG with SOX: r=-0.90, p=3.3e-4). (**C**) OP-puro incorporation shows enhanced global translation at edge. (**D**) Spatial expression of NANOG, SOX21, and SOX2 on Day 4 indicating induction of neuroectoderm in the core (left). NANOG, SOX21 and SOX2 levels and nuclear area (μm^2,black) of individual cells plotted against their distance from the edge (right). (**E**) BMP4 treatment at different days (Day 0-top, Day 2-bottom) shows that once distinct zones are established, induction of mesendoderm occurs only in the edge. (**F**) Model shows spatiotemporal dynamics of morphogenesis regulating lineage bias and cell growth signaling in self-renewal.

FIG. 9 CELLULAR AND TRANSCRIPTIONAL DYNAMICS DURING SELF-RENEWAL AND DIFFERENTIATION SHOW VARIATION BETWEEN HUMAN PSC CELL LINES.



(**A**) SOX21 levels in SR and NSB in 6 hPSC lines. Variation between cell lines is evident within each condition. The effects of cell line assessed using mixed effects models incorporating this data and 4 additional independent growth experiments: p=1e-6. (**B**) NANOG and SOX21 expression in NSB, Day 6 show SA01 and i04 have distinct morphogenic behaviors. (**C**) Spatial expression of anterior (right) and posterior (left) mesendoderm regulators in BMP4, Day 2. Significantly increased expression of anterior markers in line SA01 compared to i04: GATA4, p=2e-6; SOX17, p=3e-5, NANOG, p=6e-6. (**D**) PCA of RNAseq shows each cell line differentiates with varying efficiency. SA01 (green circle) moves fastest in neurectodermal differentiation while i04 (yellow square) leads the mesendodermal trajectory. (**E**) PC1 of PCA in NSB (left). SR data projected into NSB PC1 reveals global lineage priming (right). Fig. S5B contains multiple analyses of data from panels A and E demonstrating the stability of the morphogenic differences across lines. (**F**) Model depicting differences in morphogenesis and lineage bias between SA01 and i04 cell lines.

FIG. 10 DECOMPOSITION OF TRANSCRIPTION DURING *IN VITRO* MORPHOGENESIS.



(**A**) Dendrogram of CoGAPS patterns. Cell line specific patterns boxed. Heatmap of ANOVA p-values for effects of Line, Condition, and Day (**B**) 2 CoGAPS patterns delineating gene expression changes across time (left). Projection of microarrays of the developing mouse embryo into CoGAPS patterns (right) correspond to loss of pluripotency and initiation of gastrulation *in vivo*. (**C**) CoGAPS.1 patterns 15 and 12 represent the sequential differentiation to neurectodermal fates. (**D**) siRNA-mediated SOX21-KD prevents induction of SOX1 and PAX6 in NSB. (**E**) Dysregulation of NANOG, SOX2, OTX2 and SOX3 expression in SOX21-KO cells (solid) in day 3, NSB. SOX21KO-effect assessed at <100«M, >200«M for each gene: <100«M: SOX3 p=9.434e-09, OTX2 p=0.30, SOX2 p<2.2e-16, NANOG p=0.013; >200«M: SOX3 p=8.441e-08, OTX2 p<2.2e-16, SOX2 p=0.0088, NANOG p=0.66). (**F**) Immunofluorescence of anterior (T/BRA) and posterior (CDX2) mesendodermal regulators in day 2, BMP for WT and SOX21-KO (top). Spatial plots (lower) illustrate induction of anterior regulators (T and GATA4) in SOX21-KO (solid; n=3) compared to WT (open; n=2). At >100«M from the colony edge, GATA4 and T are differentially expressed (GATA4, p=1.774e-05; T, p<2.2e-16), while CDX2, GATA3, and ID1 are unchanged. (**G**) Model of SOX21 in specification of anterior fates.

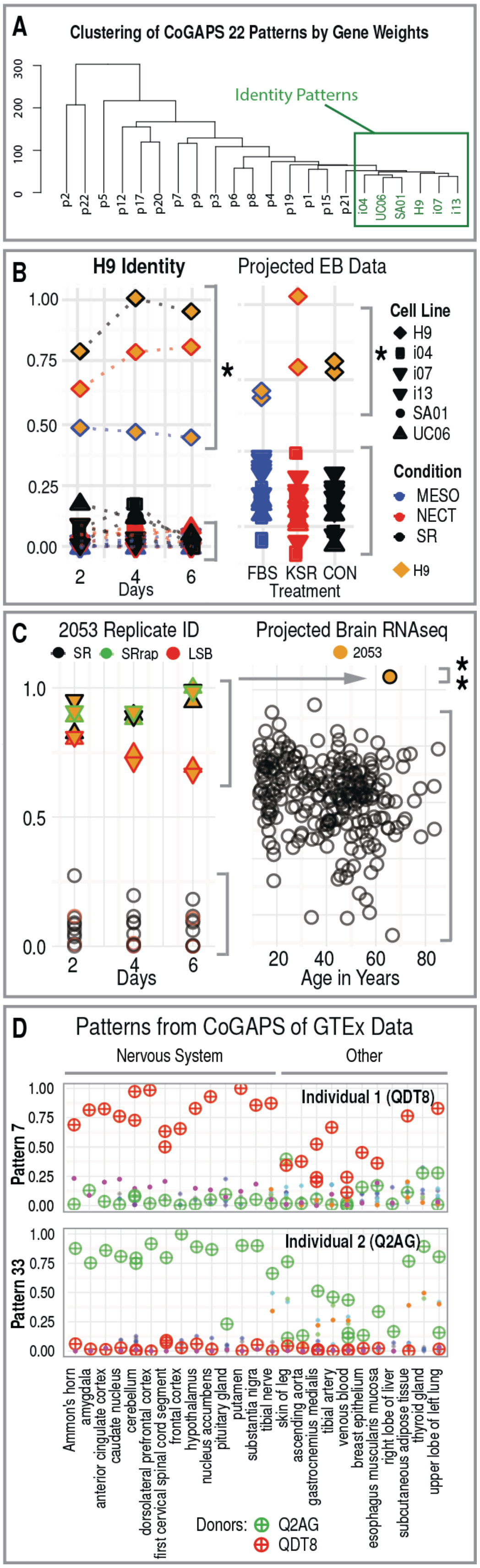


FIG. 11 COGAPS REVEALS STABLE TRANSCRIPTIONAL SIGNATURES OF INDIVIDUAL HUMANS.
(**A**) Hierarchical clustering of all gene amplitudes for the 22 CoGAPS patterns. ID signatures highlighted in green. (**B**) A CoGAPS pattern that defines the transcriptional signature of H9, distinguishing it from all other lines across time and differentiation conditions (left). Projection of microarray data using the same 6 lines differentiated by embryoid body formation methods using FBS to induce mesendodermal differentiation and KSR to induce neural differentiation into this signature (right, p=3e-4) (30). (**C**) ID signature for 2 replicate lines from donor 2053 across 3 conditions (SR, LD193189 plus SB431542: LSB, rapamycin: SRrap) in the context of 9 other lines from 7 donors (left). Projection of RNASeq data from 260 human brain samples into this transcriptional signature identifies donor 2053 (right). (**D**) ID signatures of 2 individuals defined across 22 differentiated tissues from 10 donors in the Genotype-Tissue Expression (GTEx) project dataset.

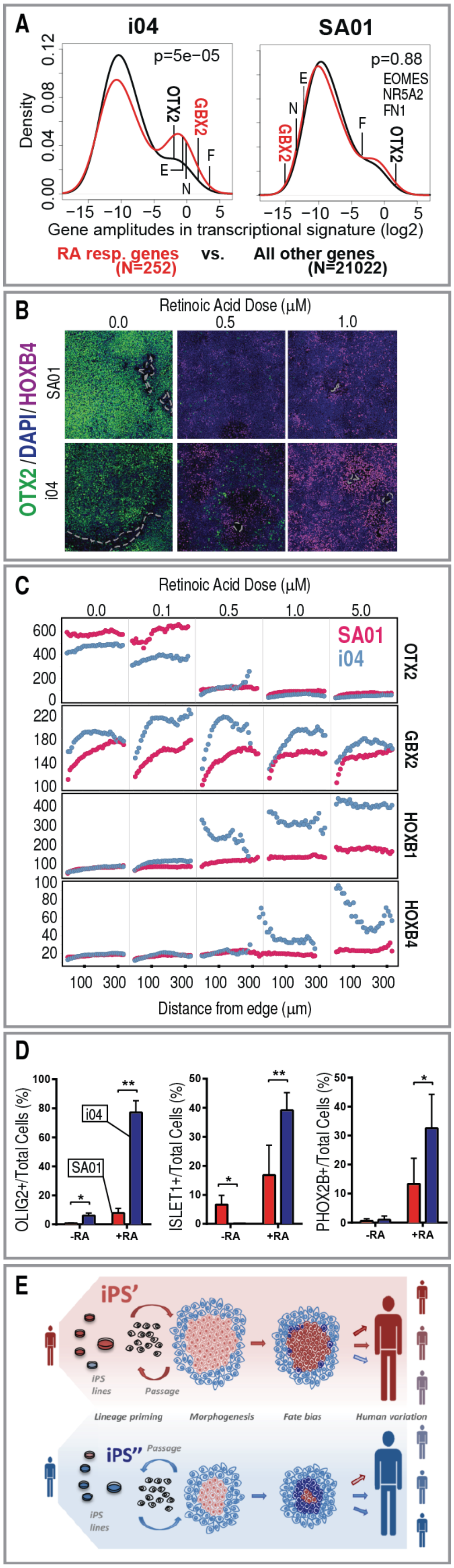


FIG. 12 CELL LINE-SPECIFIC TRANSCRIPTIONAL SIGNATURES PREDICT RESPONSE TO RETINOIC ACID. (**A**) The distribution of gene-specific amplitudes for retinoic acid (RA) responsive genes (red) and all other genes (black) in SA01 and i04 transcriptional signatures. Amplitudes for specific genes are marked in each plot (GBX2, OTX2, N=NR5A2, E=EOMES, F=FN1). RA response genes are enriched in the i04 signature (p=5e-5). (**B**) OTX2 and HOXB4 expression on day 8 of LSB+RA-induced differentiation shows differential posteriorization of neural precursors between SA01 and i04 cells. (**C**) Dose-response analysis of RA on spatial expression of anterior/posterior neural regulators shows a sequential posteriorization from the core to edge zones and a cell-line specific RA-response. (**D**) RA-induced motor neuron differentiation occurs with varying efficiency between SA01 and i04 cells. Quantitation of Olig2+, Islet1+, and PHOX2B+ expressing motor neuron precursors and differentiated motor neurons at 28 days of differentiation in SA01 and i04. (**E**) A diagram depicting how conserved morphogenic events in the epiblast can be modeled in vitro revealing stable variation between human stem cells that raises many important questions; including, How are these differences sustained over many passages? What is the range of stable developmental variation shown by hPSCs from a single donor? and How do specific transcriptional signatures influence the physiology of individual humans?

Appendices

Supplementary Figures

FIGURE S1. SPATIOTEMPORAL DYNAMICS OF NEURAL FATE BIAS, RELATED TO FIGURE 1.
(**A**) Dynamics of SOX21 and OTX2 expression in SA01 NSB conditions show induction of these neural fate transcription factors in the core zone through time. (i) Representative images are shown. Scale bar, 100 μm. Dashed lines indicate edge of colonies. (ii) Scatter plot showing single-cell levels of SOX21 and OTX2 expression in SR and NSB conditions through culture time demonstrates positive correlation between them during neurectodermal differentiation. (**B**) Changes of SOX3 and SOX21 expression in SA01 NSB conditions show reduction of SOX3 expression during neurectodermal differentiation. (i) Representative images are shown. Scale bar, 100 μm. Dashed lines indicate edge of colonies. (ii) Scatter plot showing single-cell levels of SOX3 and SOX21 expression in SR and NSB conditions through culture time demonstrates inverse correlation between them during neurectodermal differentiation. (**C**) Different cell lines vary in their formation of core zone. SOX21 and NANOG expression at day 6 of NSB treatment in SA01 and i04 lines show more efficient formation of core zone in SA01 compared to i04. Scale bar, 200 μm. Dashed lines indicate edge of colonies. (**D**) Dose-response curve of BMP/TGFβ signaling inhibitor LDN193189 and SB431542 on SOX21 induction show differential responses between SA01 and i04 lines. *, $p<0.05$ between lines. (**E**) Cell numbers in SA01 and i04 lines are similar throughout time in all conditions suggesting the differential morphogenetic composition in epithelium is not a result of different colony size or overall proliferation rate. *, $p<0.05$ between lines.

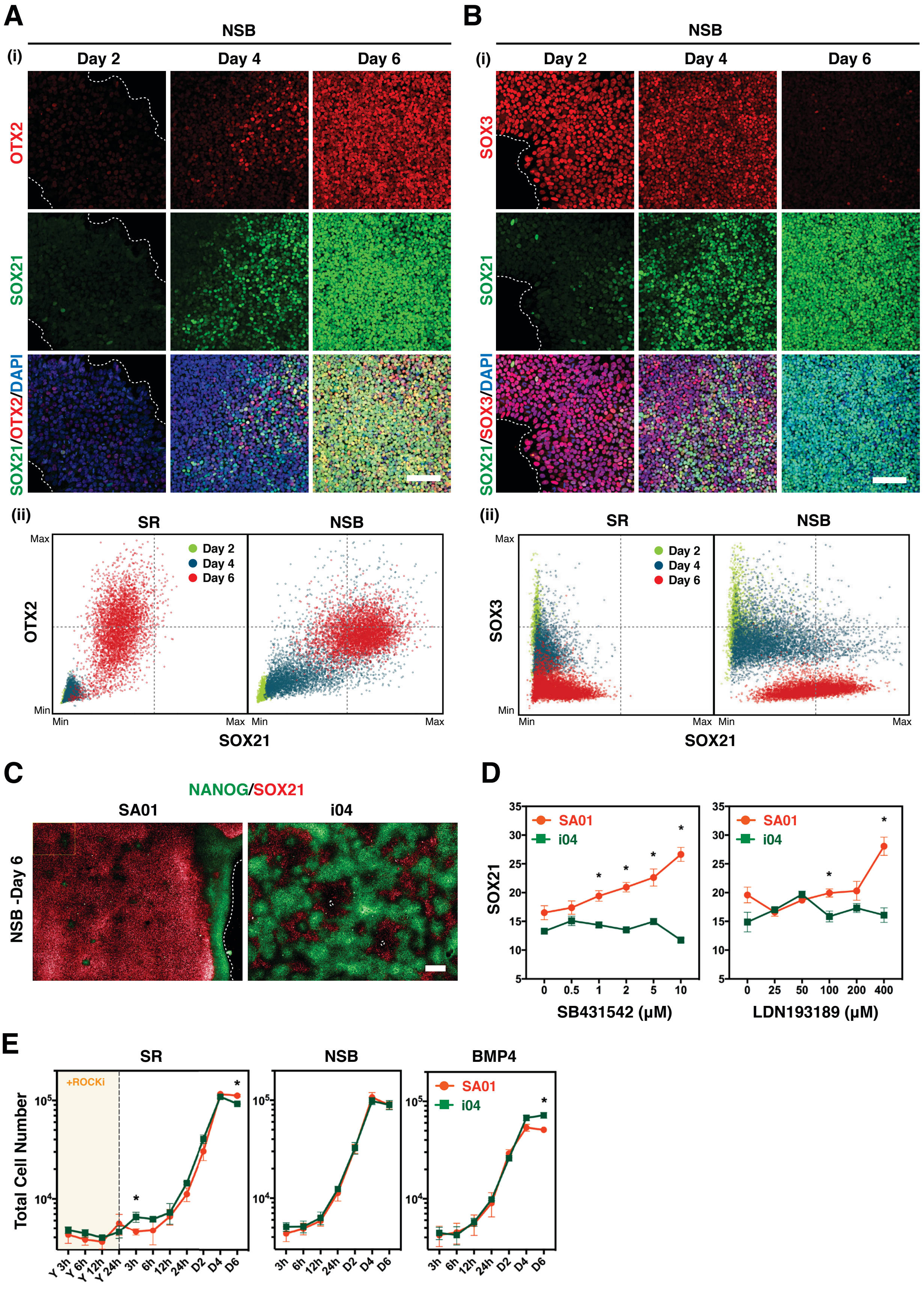

A
NSB
(i)
Day 2
Day 4
Day 6
OTX2
SOX21
SOX21/OTX2/DAPI
(ii)
SR
NSB
Day 2
Day 4
Day 6
OTX2
Max
Min
Min
Max
SOX21
B
NSB
(i)
Day 2
Day 4
Day 6
SOX3
SOX21
SOX21/SOX3/DAPI
(ii)
SR
NSB
SOX3
SOX21
C
NANOG/SOX21
SA01
i04
NSB -Day 6
D
SA01
i04
SOX21
SB431542 (μM)
LDN193189 (μM)
0
0.5
1
2
5
10
25
50
100
200
400
35
30
25
20
15
10
5
*
E
SR
NSB
BMP4
+ROCKi
Total Cell Number
10^5
10^4
Y 3h
Y 6h
Y 12h
Y 24h
3h
6h
12h
24h
D2
D4
D6

FIGURE S2. CELL LINE VARIATION IN TRANSCRIPTIONAL DYNAMICS DURING SELF-RENEWAL AND DIFFERENTIATION, RELATED TO FIGURE 2 AND FIGURE 3.

(**A**) The selected 25 genes among 100 genes most strongly contributing to PC1 (left) and PC2 (right). Well-known pluripotency and early fate regulators are highlighted in colors. (**B**) The selected 25 genes among 100 genes most strongly contributing to NSB PC1. Neurectodermal regulators including SOX21, OTX2 and PAX6 strongly contribute to the up-regulation while pluripotency regulators including POU5F1 and NANOG contribute to the down-regulation in this PC (highlighted in colors). (**C**) Scatter plot showing correlation between SOX21 levels in NSB (Figure 2A) and NSB PC1 (Figure 2C). (**D**) The use of each GWCoGAPS pattern across genes can be precisely defined by gene-specific amplitudes for all patterns (Table S4). Two examples of individual genes whose expression patterns are represented by the combination of multiple GWCoGAPS patterns are shown. (Left) The complete expression of POU5F1 is represented by two dynamic GWCoGAPS patterns P7 and P9. (Right) The complete expression of OTX1 is represented by a dynamic GWCoGAPS pattern P12 and a UC06 line-specific GWCoGAPS pattern P11. (**E**) The 22 patterns generated by GWCoGAPS decomposition across 6 cell lines, 3 conditions and 3 times. Patterns are presented in groups of similar characteristics; dynamic patterns (pluripotency, BMP4-response mesendoderm, and NSB-response neurectoderm) or cell line-specific patterns (6 cell line-specific patterns and 2 combined cell line-specific patterns).

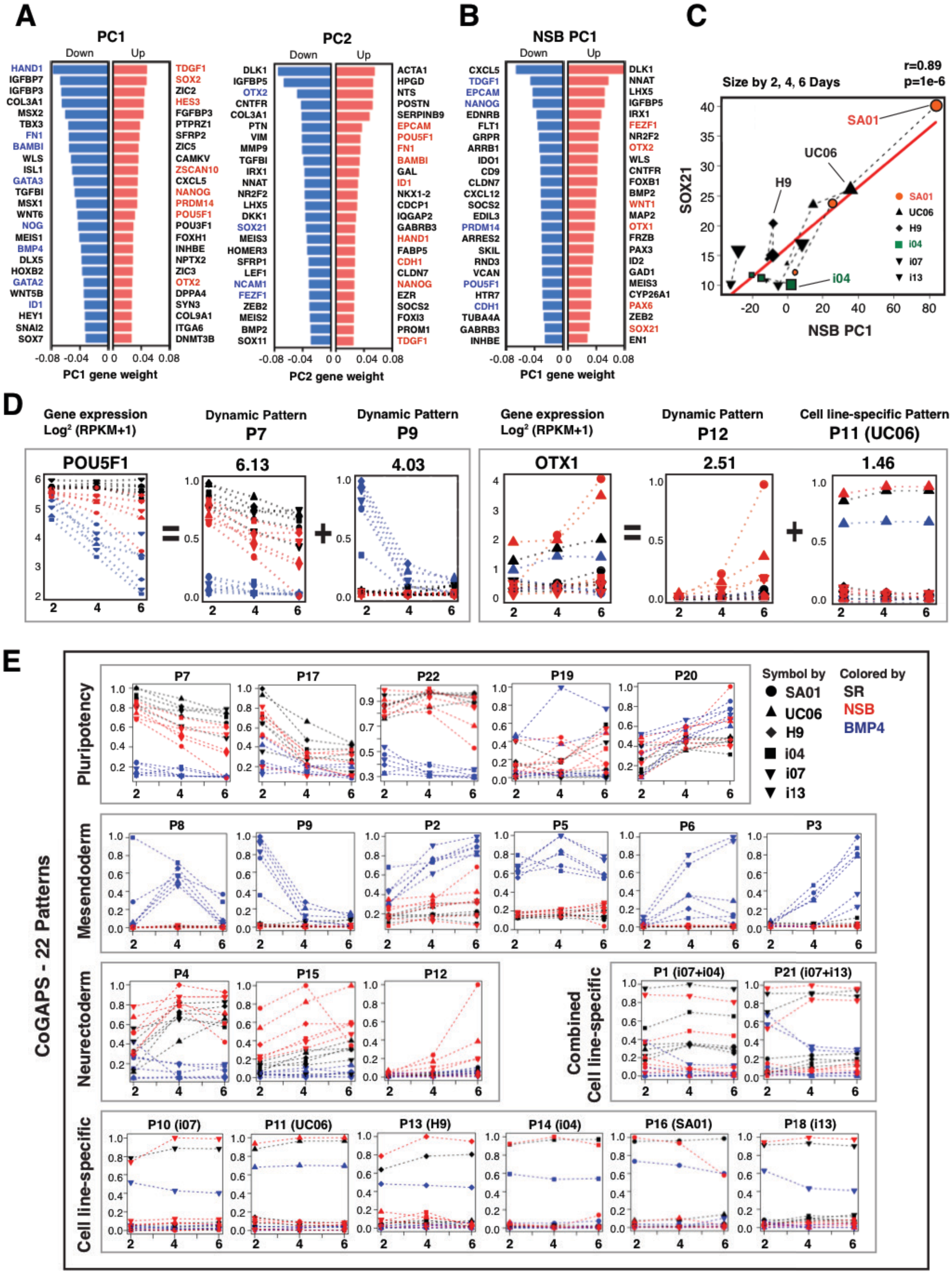

A
PC1
Down
Up
HAND1 IGFBP7 IGFBP3 COL3A1 MSX2 TBX3 FN1 BAMBI WLS ISL1 GATA3 TGFBI MSX1 WNT6 NOG MEIS1 BMP4 DLX5 HOXB2 GATA2 WNT5B ID1 HEY1 SNAI2 SOX7
TDGF1 SOX2 ZIC2 HES3 FGFBP3 PTPRZ1 SFRP2 ZIC5 CAMKV ZSCAN10 CXCL5 NANOG PRDM14 POU5F1 POU3F1 FOXH1 INHBE NPTX2 ZIC3 OTX2 DPPA4 SYN3 COL9A1 ITGA6 DNMT3B
-0.08 -0.04 0 0 0.04 0.08
PC1 gene weight
PC2
DLK1 IGFBP5 OTX2 CNTFR COL3A1 PTN VIM MMP9 TGFBI IRX1 NNAT NR2F2 LHX5 DKK1 SOX21 MEIS3 HOMER3 SFRP1 LEF1 NCAM1 FEZF1 ZEB2 MEIS2 BMP2 SOX11
ACTA1 HPGD NTS POSTN SERPINB9 EPCAM POU5F1 FN1 BAMBI GAL ID1 NKX1-2 CDCP1 IQGAP2 GABRB3 HAND1 FABP5 CDH1 CLDN7 NANOG EZR SOCS2 FOXI3 PROM1 TDGF1
PC2 gene weight
B
NSB PC1
CXCL5 TDGF1 EPCAM NANOG EDNRB FLT1 GRPR ARRB1 IDO1 CD9 CLDN7 CXCL12 SOCS2 EDIL3 PRDM14 ARRES2 SKIL RND3 VCAN POU5F1 HTR7 CDH1 TUBA4A GABRB3 INHBE
DLK1 NNAT LHX5 IGFBP5 IRX1 FEZF1 NR2F2 OTX2 WLS CNTFR FOXB1 BMP2 WNT1 MAP2 OTX1 FRZB PAX3 ID2 GAD1 MEIS3 CYP26A1 PAX6 ZEB2 SOX21 EN1
PC1 gene weight
C
Size by 2, 4, 6 Days
r=0.89
p=1e-6
SA01
UC06
H9
i04
SA01 UC06 H9 i04 i07 i13
SOX21
NSB PC1
D
Gene expression Log2 (RPKM+1)
Dynamic Pattern P7
Dynamic Pattern P9
Gene expression Log2 (RPKM+1)
Dynamic Pattern P12
Cell line-specific Pattern P11 (UC06)
POU5F1
6.13
4.03
OTX1
2.51
1.46
=
+
E
CoGAPS - 22 Patterns
Pluripotency
P7 P17 P22 P19 P20
Mesendoderm
P8 P9 P2 P5 P6 P3
Neurectoderm
P4 P15 P12
Combined Cell line-specific
P1 (i07+i04) P21 (i07+i13)
Cell line-specific
P10 (i07) P11 (UC06) P13 (H9) P14 (i04) P16 (SA01) P18 (i13)
Symbol by
SA01 UC06 H9 i04 i07 i13
Colored by
SR
NSB
BMP4

FIGURE S3. DECOMPOSING DYNAMIC & STABLE TRANSCRIPTION MODULES, RELATED TO FIGURE 3. (**A**)Protein expression of genes highly ranked in 5 dynamic patterns. The experimental protocols of mesendoderm and neurectoderm differentiation are shown in boxes. (i) Representative images show expression of MMP9, CK14 and HOXB9, genes that are highly weighted in a mesendoderm pattern P3. (ii) Representative images show expression of pluripotency and neurectodermal regulators that are highly weighted in patterns P7, P15, and P12 during the accelerated neural differentiation. The neural differentiation is accelerated by FGF2 withdrawal and conversion to N2B27 medium supplemented with LDN193189 and SB431542. Pluripotency and early neurectodermal fate regulator SOX2 show down-regulation on day 6 in the core zone in the accelerated neural differentiation condition. An early neurectoderm pattern P15 gene HES3 is expressed higher at the edge while another P15 gene SOX21 is expressed at the core on day 2 and gradually expands over time. Under this accelerated differentiation condition, a P12 gene HES5 is expressed as early as on day 2 at the core and other P12 genes including PAX6, PAX2, WNT1, LMX1A and EN1 are induced later. Dashed lines indicate edge of colonies. Scale bar, 100 μm. (**B**) Projections of microarray and DNA methylation data generated from multiple hPSC lines into the cell line-specific CoGAPS patterns demonstrate stability of the cell line-specific transcriptional signatures within a cell line. (i) H9-, SA01-, and i04-specific GWCoGAPS patterns that define distinct transcriptional signature from all other lines across time and condition. Each cell line sample circled in green. (ii) Projections of microarray dataset (Mallon et al., 2013) containing the same 6 lines under embryoid-body (EB) differentiation conditions (SR for self-renewal in ES medium plus FGF2, KSR for ectodermal differentiation, and FBS for mesendodermal differentiation) into the each cell line-specific patterns discriminate the corresponding cell line samples from all other lines. (iii) Projections of DNA methylation data (Rouhani et al., 2014) show promoters of genes expressed specifically in each cell line are hypomethylated in the corresponding cell line.

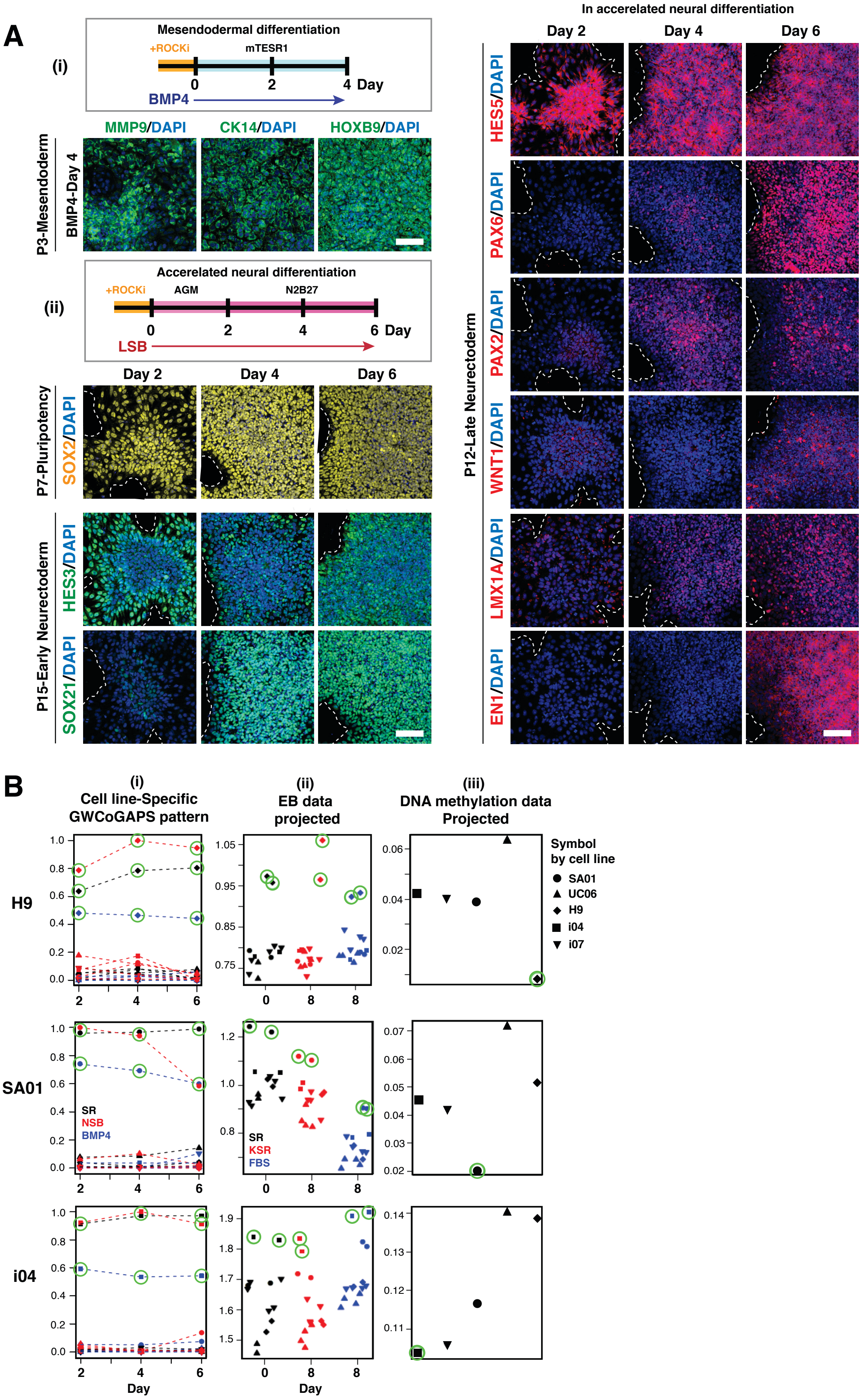

A
(i)
Mesendodermal differentiation
+ROCKi
mTESR1
0
2
4 Day
BMP4
MMP9/DAPI
CK14/DAPI
HOXB9/DAPI
P3-Mesendoderm
BMP4-Day 4
Accerelated neural differentiation
(ii)
+ROCKi
AGM
N2B27
0
2
4
6 Day
LSB
Day 2
Day 4
Day 6
P7-Pluripotency
SOX2/DAPI
P15-Early Neurectoderm
HES3/DAPI
SOX21/DAPI
In accerelated neural differentiation
Day 2
Day 4
Day 6
P12-Late Neurectoderm
HES5/DAPI
PAX6/DAPI
PAX2/DAPI
WNT1/DAPI
LMX1A/DAPI
EN1/DAPI
B
(i)
Cell line-Specific
GWCoGAPS pattern
(ii)
EB data
projected
(iii)
DNA methylation data
Projected
Symbol
by cell line
● SA01
▲ UC06
◆ H9
■ i04
▼ i07
H9
SA01
i04
SR
NSB
BMP4
SR
KSR
FBS
1.0
0.8
0.6
0.4
0.2
0.0
2
4
6
Day
1.05
0.95
0.85
0.75
1.2
1.0
0.8
1.9
1.8
1.7
1.6
1.5
0
8
8
Day
0.06
0.04
0.02
0.07
0.05
0.03
0.14
0.13
0.12
0.11

FIGURE S4. CELL LINE VARIATION IN HINDBRAIN FATE BIAS, RELATED TO FIGURE 4.

(**A**) Differential OTX2 and GBX2 expression in SA01 and i04 cell lines during neurectoderm differentiation. Scatter plot of single-cell level OTX2 and GBX2 expression on day 6 in SR and LSB condition. Dashed lines indicate the average intensity of protein level in each condition. (**B**) Dose-response analysis of retinoic acid (RA) shows differential posterization of neural precursors between SA01 and i04 lines. (i) Representative images show differential OTX2 and HOXB4 expression on day 8 of LDN193189+SB431542 (LSB)+RA-induced differentiation between SA01 and i04 lines. Scale bar, 100 μm. Dashed lines indicate edge of colonies. (ii) Dose-response analysis of RA on spatial expression of anterior/posterior neural regulators shows a sequential posteriorization from the core to edge zones and a cell line-specific RA-response. Single-cell levels of OTX2, GBX2, HOXB1 and HOXB4 expression are measured on day 8 of LSB+RA-induced differentiation and plotted against the distance from the edge.

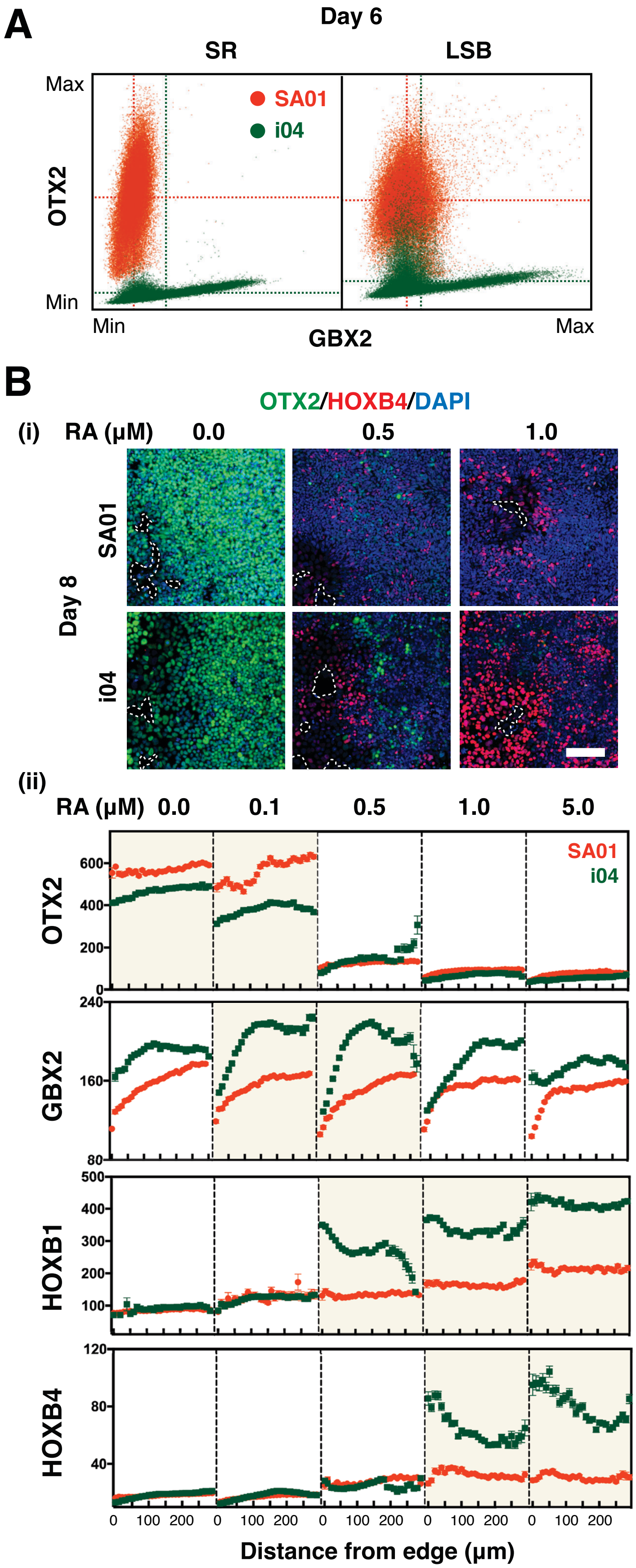
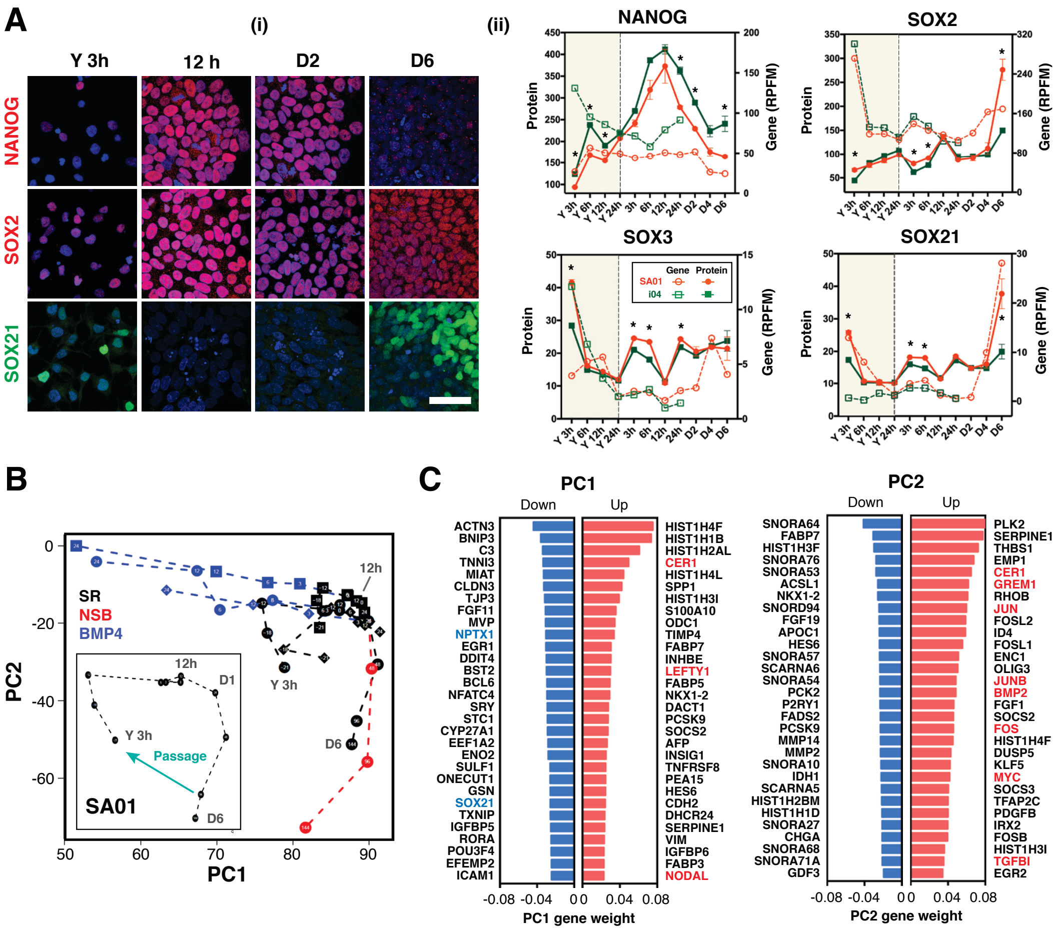

A
Day 6
SR
LSB
Max
Min
OTX2
● SA01
● i04
Min
GBX2
Max
B
OTX2/HOXB4/DAPI
(i) RA (µM) 0.0 0.5 1.0
Day 8
SA01
i04
(ii)
RA (µM) 0.0 0.1 0.5 1.0 5.0
SA01
i04
OTX2
600
400
200
0
GBX2
240
160
80
HOXB1
500
400
300
200
100
HOXB4
120
80
40
0 100 200
Distance from edge (µm)

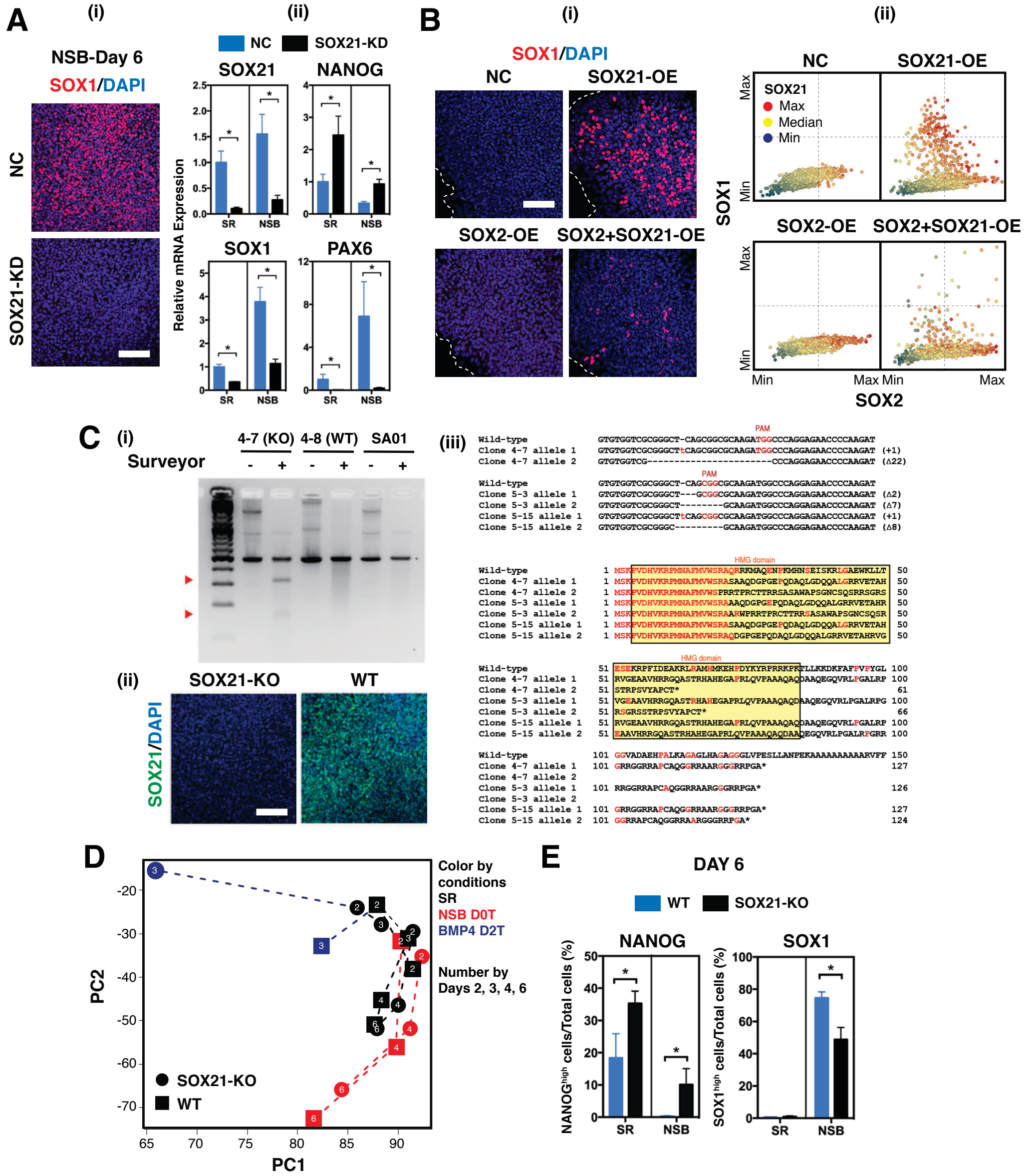
FIGURE S5. RESETTING THE PLURIPOTENCY CYCLE, RELATED TO FIGURE 5.



(**A**) Dynamics of NANOG, SOX2, SOX3, and SOX21 expressions reveal that cell states are reset within the first 36 hours after passaging. (i) Immunofluorescence images of NANOG, SOX2 and SOX21 expression in SA01 show loss of heterogenic expression in cell population during early times after passage in SR. Scale bar, 50 μm. (ii) Protein and RNA levels of NANOG, SOX2, SOX3, and SOX21 expression are measured in SA01 and i04 lines over time. *, Comparison in protein levels between SA01 and i04 ($p<0.05$). (**B**) Projection of early time data into PC1 and PC2 of days 2, 4, and 6 data (Figure 2B) shows continuum between the two datasets. Inset shows cycling pattern of SA01 SR samples upon passaging as described in the PC space. (**C**) 30 selected genes among 100 genes most strongly contributing to PC1 (left) and PC2 (right) of SA01 SR shown in Figure 5B. Well-known pluripotency and neurectoderm regulators, signaling genes for early embryo patterning, and immediate early genes are highlighted in colors.

FIGURE S6. SOX21 MEDIATES FOREBRAIN FATE CHOICE, RELATED TO FIGURE 6.

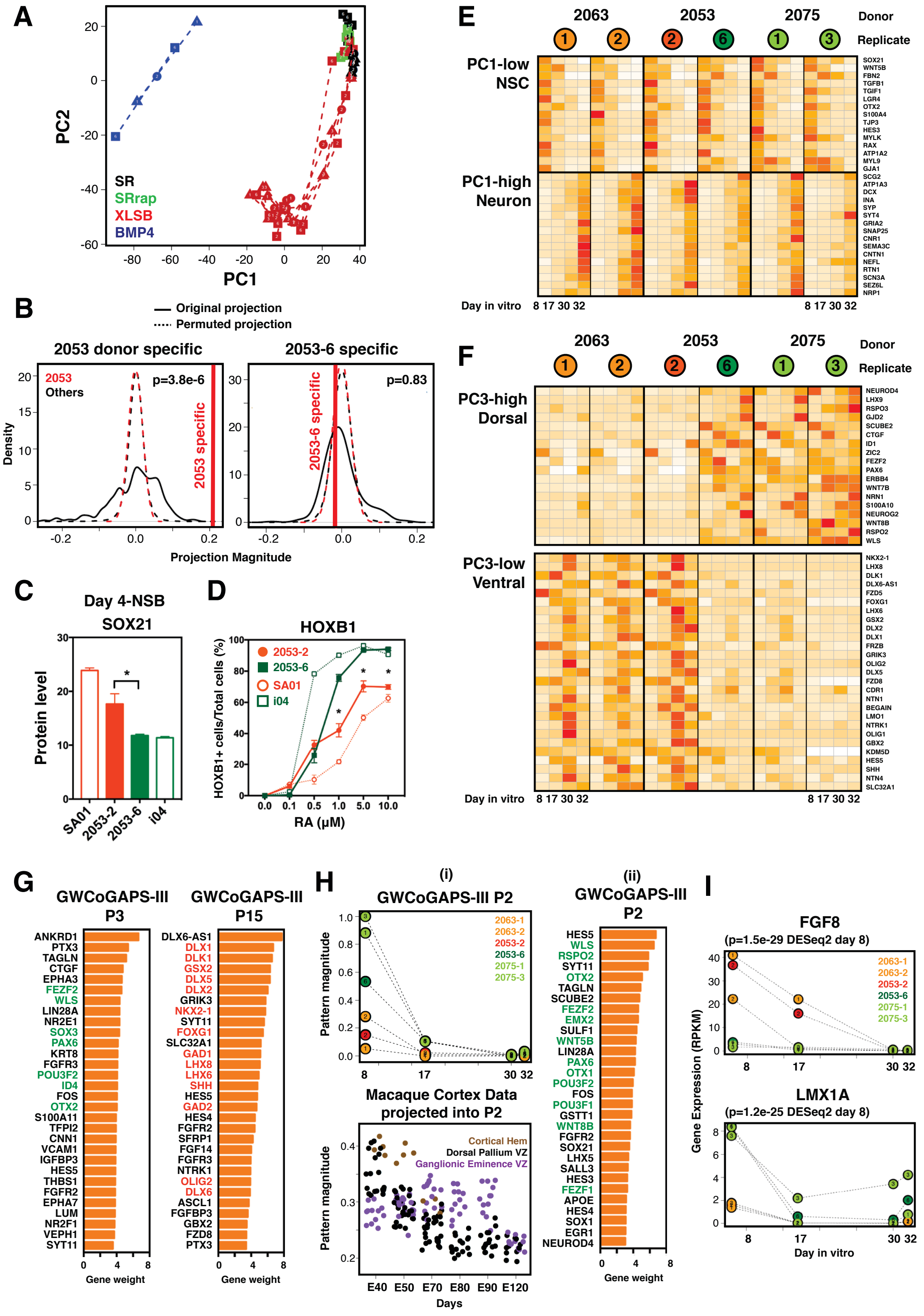
Supple. F6



(**A**) Knockdown (KD) of SOX21 by siRNA treatment prevents induction of neurectodermal regulators in NSB condition. (i) Representative images show SOX1 reduction after SOX21-KD in day 6 NSB condition. Scale bar, 100 μm. (ii) Relative mRNA expression level analysis shows reduction of neurectodermal regulators SOX1 and PAX6, and induction of pluripotency regulator NANOG after SOX21-KD. Values are normalized to negative control siRNA treatment (NC) within each condition. *, Comparison between NC and SOX21-KD ($p<0.05$). (**B**) Overexpression of SOX21 induces SOX1 expression. (i) Representative images show induction of SOX1 expression after SOX21 overexpression in day 4 NSB condition. Cells were treated with mRNAs for 4 days after ROCK inhibitor removal. Scale bar, 100 μm. (ii) Scatter plot illustrates relative expression of SOX1, SOX2, and SOX21 after mRNA transfection. Colors represent SOX21 levels. Overexpressing SOX2 and SOX21 together reduced SOX1 induction compared to overexpressing SOX21 alone. (**C**) Establishment of SOX21-KO lines by CRISPR/Cas9 technology. All SOX21-KO ESC lines were screened using Surveyor and immunofluorescence assays and verified by DNA sequencing. (i) Analysis of SOX21-KO clones using Surveyor assay. The gel image shows modification at the SOX21 locus in a clone 4-7. Red arrowheads indicate expected fragment sizes for SOX21 locus. (ii) Immunostaining of WT and SOX21-KO clones shows complete loss of SOX21 expression in the clone 4-7. The cells were cultured in the presence of NSB for 6 days. Scale bar, 100 μm. (iii) Amino acid sequence of SOX21 alterations by CRISPR/Cas9 confirms knockout of SOX21. In three clones (clones 4-7, 5-3, and 5-15) frame shift mutation, premature stop codon mutation, or mutation that disrupts HMG domain were confirmed in both SOX21 alleles. These three clones were used for the functional assays shown in Figure 6. (**D**) Projection of WT and SOX21-KO line data into PC1 and PC2 of days 2, 4, and 6 data (Figure 2B) reveals delayed neurectodermal differentiation in NSB and accelerated early mesendodermal differentiation under BMP4 D2T in SOX21-KO compared to WT. (**E**) SOX21 regulates transition from pluripotency to neurectoderm. NANOG expression was induced while SOX1 expression was reduced in SOX21-KO cells compared to WT on day 6 in SR and NSB conditions. *, Comparison between WT and SOX21-KO ($p<0.05$).

FIGURE S7. CELL LINE VARIATION IN NEURAL FATE TRAJECTORIES, RELATED TO FIGURE 7.
(**A**) Projection of the new 6 hiPSC forebrain neural differentiation RNAseq data into PC1 and PC2 of days 2, 4, and 6 data (Figure 2B) demonstrates the generality of differentiation dynamics defined by early differentiation data and further progress in neural fate trajectory. (**B**) Permutation analysis demonstrates the statistical significance of the magnitude of the donor-specific transcriptional signatures in the corresponding donor's adult brain tissue. The 2053 donor-specific signatures, but not the 2053-6 replicate-specific signature was significantly enriched in the corresponding adult brain tissue. The wider distribution of the original projection values (solid line) compared to the permuted projection values (dashed line) indicates that the expression of the gene subset in the cell line-specific transcriptional signatures is more variable across individuals than randomly selected gene subset. (**C**) Differential SOX21 levels in SA01, i04 and two 2053 lines. *, Comparison between 2053-2 and 2053-6 ($p<0.05$). (**D**) Two replicate lines from donor 2053 show differential responsiveness to RA. Number of HOXB1 expressing cells on day 8 in response to varying doses of RA in SA01, i04 and two 2053 lines. *, Comparison between 2053-2 and 2053-6 ($p<0.05$). (**E**) Heatmap showing expression levels of neural precursor- and neuron-related genes that are highly represented in PC1 demonstrates transition from neural precursors to neurons in all 6 hiPSC lines during forebrain neural differentiation. (**F**) Heatmap showing expression levels of dorsal and ventral fate specification-related genes that are highly represented in PC3 demonstrates distinct brain regional biases dominated by each line. (**G**) 30 selected genes among top 50 genes most strongly contributing to the GWCoGAPS-III P3 (left) and P15 (right) shown in Figure 7D. Well-known dorsal (green) and ventral (red) fate specification-related genes are highlighted in colors. (**H**) Genes involved in cortical hem are highly expressed in the lines with dorsal lineage bias at day 8. (i) GWCoGAPS-III P2 reveals distinct dorsal to ventral trajectories between the lines at day 8 (top). Projection of primate cortex data (Bakken et al., 2016) into P2 distinguished cortical hem samples from dorsal pallium ventricular zone (VZ) and ganglionic eminence VZ samples (bottom). (ii) 30 selected genes among top 100 genes most strongly contributing to the GWCoGAPS-III P2. (**I**) Gene expression (RPKM) of FGF8 and LMX1A during the forebrain neural differentiation in 6 lines. $p=1.5e-29$ in DESeq2 analysis of differential LMX1A expression on day 8. $p=1.2e-25$ in DESeq2 analysis of differential FGF8 expression on day 8.

Supple. F7


A
PC2
PC1
20
0
-20
-40
-60
-80
-60
-40
-20
0
20
40
SR
SRrap
XLSB
BMP4
B
Original projection
Permuted projection
2053 donor specific
2053-6 specific
2053
Others
p=3.8e-6
p=0.83
2053 specific
2053-6 specific
Density
Projection Magnitude
-0.2
0.0
0.2
C
Day 4-NSB
SOX21
Protein level
*
SA01
2053-2
2053-6
i04
D
HOXB1
2053-2
2053-6
SA01
i04
HOXB1+ cells/Total cells (%)
RA (μM)
0.0
0.1
0.5
1.0
5.0
10.0
E
2063
2053
2075
Donor
Replicate
PC1-low
NSC
PC1-high
Neuron
SOX21
WNT5B
FBN2
TGFB1
TGIF1
LGR4
OTX2
S100A4
TJP3
HES3
MYLK
RAX
ATP1A2
MYL9
GJA1
SCG2
ATP1A3
DCX
INA
SYP
SYT4
GRIA2
SNAP25
CNR1
SEMA3C
CNTN1
NEFL
RTN1
SCN3A
SEZ6L
NRP1
Day in vitro
8 17 30 32
F
PC3-high
Dorsal
PC3-low
Ventral
NEUROD4
LHX9
RSPO3
GJD2
SCUBE2
CTGF
ID1
ZIC2
FEZF2
PAX6
ERBB4
WNT7B
NRN1
S100A10
NEUROG2
WNT8B
RSPO2
WLS
NKX2-1
LHX8
DLK1
DLX6-AS1
FZD5
FOXG1
LHX6
GSX2
DLX2
DLX1
FRZB
GRIK3
OLIG2
DLX5
FZD8
CDR1
NTN1
BEGAIN
LMO1
NTRK1
OLIG1
GBX2
KDM5D
HES5
SHH
NTN4
SLC32A1
G
GWCoGAPS-III P3
ANKRD1
PTX3
TAGLN
CTGF
EPHA3
FEZF2
WLS
LIN28A
NR2E1
SOX3
PAX6
KRT8
FGFR3
POU3F2
ID4
FOS
OTX2
S100A11
TFPI2
CNN1
VCAM1
IGFBP3
HES5
THBS1
FGFR2
EPHA7
LUM
NR2F1
VEPH1
SYT11
Gene weight
GWCoGAPS-III P15
DLX6-AS1
DLX1
DLK1
GSX2
DLX5
DLX2
GRIK3
NKX2-1
SYT11
FOXG1
SLC32A1
GAD1
LHX8
LHX6
SHH
HES5
GAD2
HES4
FGFR2
SFRP1
FGF14
FGFR3
NTRK1
OLIG2
DLX6
ASCL1
FGFBP3
GBX2
FZD8
PTX3
Gene weight
H
(i)
GWCoGAPS-III P2
Pattern magnitude
2063-1
2063-2
2053-2
2053-6
2075-1
2075-3
8
17
30 32
Macaque Cortex Data projected into P2
Cortical Hem
Dorsal Pallium VZ
Ganglionic Eminence VZ
E40
E50
E70
E80
E90
E120
Days
(ii)
GWCoGAPS-III P2
HES5
WLS
RSPO2
SYT11
OTX2
TAGLN
SCUBE2
FEZF2
EMX2
SULF1
WNT5B
LIN28A
PAX6
OTX1
POU3F2
FOS
POU3F1
GSTT1
WNT8B
FGFR2
SOX21
LHX5
SALL3
HES3
FEZF1
APOE
HES4
SOX1
EGR1
NEUROD4
Gene weight
I
FGF8
(p=1.5e-29 DESeq2 day 8)
LMX1A
(p=1.2e-25 DESeq2 day 8)
Gene Expression (RPKM)
2063-1
2063-2
2053-2
2053-6
2075-1
2075-3
Day in vitro

Supplementary References

Bakken, T.E., Miller, J.A., Ding, S.L., Sunkin, S.M., Smith, K.A., Ng, L., Szafer, A., Dalley, R.A., Royall, J.J., Lemon, T., *et al.* (2016). A comprehensive transcriptional map of primate brain development. Nature *535*, 367-375.

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.

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.

Supplementary Experimental Procedures

Human pluripotent stem cell (hPSC) culture and differentiation

Single cell-based monolayer culture of hPSC in feeder-free condition was done following the previously established protocol (Chen et al., 2012). The hPSCs were dissociated to single cells with accutase (A11105, Life Technologies), plated at a density of 1 X 10^5 cells/cm^2 in a Matrigel (354277, BD)-coated plate and cultured with mTeSR1 (05850, Stem Cell Technology). Cells were plated in medium containing 5 μM Y27632, ROCK inhibitor (Y0503, Sigma-Aldrich) to increase cell survival upon dissociation. ROCK inhibitor was removed from the medium at 24 hours after plating and cells were cultured for another 4 days before next passage. For differentiation, cells were plated at a density of 18 X 10^3 cells/cm^2 in a Matrigel-coated plate and Noggin (500 ng/ml, 719-NG, R&D Systems) and SB431542 (2 μM, S4317, Sigma-Aldrich) were added to mTeSR1 medium for neurectodermal differentiation while BMP4 (100 ng/ml, 314-BP, R&D Systems) was added for mesendodermal differentiation upon ROCK inhibitor removal and cultured for 6 days. For accelerated neurectodermal differentiation to check the expression patterns of the genes defined by dynamic CoGAPS patterns, cells were cultured with Aggrewell medium (05893, Stem Cell Technology) for 2 days after ROCK inhibitor removal and then cultured with N2B27 medium supplemented with LDN193189 (100 nM, 04-0074, Stemgent) and SB431542 (2 μM, S4317, Sigma-Aldrich) for another 4 days. For posterior neural differentiation, retinoic acid (R2625, Sigma-Aldrich) was added on day 4 in the accelerated neurectodermal differentiation protocol to induce the specification toward neural crest and posterior neurectoderm. For the further differentiation to spinal neurons, cells were cultured with Neurobasal medium (21103-049, Life Technologies) supplemented with bovine Insulin (25 μg/ml, I6634, Sigma-Aldrich), B27 (17504-044, Life Technologies), recombinant human BDNF (10 ng/ml, 248-BD, R&D Systems) and recombinant human NT-3 (10 ng/ml, 267-N3, R&D Systems) for another 20 days. For the forebrain neural differentiation, cells were differentiated as previously described (Maroof et al., 2013); cells were cultured in N2B27 medium supplemented with XAV939 (2 μM, Stemgent), LDN193189 (100 nM) and SB431542 (10 μM) for 12 days and the differentiation medium was switched to Neurobasal medium supplemented with B27 at day 17 and cells were further cultured until day 32.

Generation of human induced pluripotent stem cell (hiPSC) lines

The hiPSC line i04, i07 and i13 (NIH-i4, NIH-i7, NIH-i13) have been reported previously (Mallon et al., 2013). The hiPSC lines reprogrammed with synthetic mRNAs were generated using mRNA reprogramming kit (00-0071, Stemgent) and microRNA Booster kit (00-0073, Stemgent) with modifications. Human fibroblasts (Donor 2075, 2053, and 2063) were seeded at 5 X 10^3 cells/cm^2 in a Matrigel-coated plate and cultured with DMEM medium supplemented with 10% FBS (Life Technologies) and 2 mM L-glutamine. After 24 hours (day 1), the medium was changed to Pluriton human NUFF conditioned media with 300 ng/ml B18R protein. On day 1 and 5, the microRNA booster kit was used with the StemFect RNA transfection reagent kit (00-0069, Stemgent) to enhance reprogramming. On day 2-12, the OSKML RNAs were transfected. The mRNA reprogramming process was performed at 37°C in 5% O_2 and CO_2 incubator.

Knockdown and overexpression of SOX21

Silencing endogenous SOX21 expression was performed by siRNA transfection using DharmaFECT 1 reagent (T-2001-02, Thermo Scientific). Cells were transfected with non-targeting negative control siRNA (#4390843, Life Technologies) or siRNAs targeting SOX21 (sc-38433, Santa Cruz Biotechnology) at a final concentration of 50 nM in mTeSR1 medium for 4 days after ROCK inhibitor removal. Human SOX21 synthetic mRNA was custom produced by Stemgent. Transfection was performed using the StemFect RNA transfection reagent kit at a final concentration of 0.5 μg/ml for 24 hours after ROCK inhibitor removal.

Generation of CRISPR/Cas9 mediated SOX21-KO human embryonic stem cell (hESC) line

The SOX21-KO hESC lines were generated by CRISPR/Cas9 mediated genome deletion system. SOX21 specific gRNAs were designed using the CRISPR Design Tool, Optimized CRISPR Design - MIT for Sox21NHEJ4 (http://crispr.mit.edu/) (Ran et al., 2013) and CHOPCHOP for Sox21NHEJ5 (https://chopchop.rc.fas.harvard.edu/) (Montague et al., 2014). The oligonucleotides (CACCGCGGGCTCAGCGGCGCAAGA –top for Sox21NHEJ4; AAACTCTTGCGCCGCTGAGCCCGC –bottom for Sox21NHEJ4;

CACCGGGTGTGGTCGCGGGCTCAG –top for Sox21NHEJ5; AAACCTGAGCCCGCGACCACACCC –bottom for Sox21NHEJ5) were cloned into pSpCas9(BB)-2A-Puro (px459; Addgene) and designated the plasmid as pX459-Sox21NHEJ4 and pX459-Sox21NHEJ5. All oligonucleotides were synthesized by Integrated DNA Technologies. SA01 hESCs were transfected with 2.5 µg pX459-Sox21NHEJ4 plasmid or pX459-Sox21NHEJ5 using DNA-In Stem (MTI-Global stem, gifted from Dr. Jessee). Transfected cells were dissociated and plated into 10 cm culture dish. After 48 hours of 0.5 µg/ml puromycin selection, hESC colonies were maintained for 10 days. Individual colonies were isolated and clonally expanded. Genomic DNA was isolated from each hESC clonal line using Wizard Genomic DNA Purification Kit (Promega). The genomic region surrounding the CRISPR target site for SOX21 was amplified by PCR (KOD Xtreme Hot Start DNA Polymerase; EMD Millipore), and products were treated with SURVEYOR nuclease (SURVEYOR Mutation Detection Kit for Standard Gel Electrophoresis, Transgenomic) to detect CRISPR/Cas9 -induced indel mutations. The PCR products were cloned into pGEM® -T Easy Vector (Promega) and sequenced to confirm the genotypes. Knockout validation of SOX21 protein was performed by immunostaining.

Immunofluorescence

Cells were fixed with 4% paraformaldehyde for 10 min and permeabilized for 40 min using 0.1% Triton X-100 (Sigma-Aldrich) in PBS. Subsequently, cells were blocked with 10% donkey serum (Sigma-Aldrich) and incubated with primary antibodies overnight. Following primary antibodies and dilutions were used: Antibodies for CK14 (ab7800, 1:200), HOXB4 (ab133621, 1:400), HOXB9 (ab66765, 1:400) and WNT1 (ab85060, 1:200) were from Abcam. Antibody for CDX2 (AM392) was from Biogenex. Antibodies for p-SMAD1/5 (9516, 1:200) and p-SMAD2/3 (8828, 1:200) were from Cell Signaling Technology. Antibodies for PAX2 (PRB-276P, 1:400), PAX6 (PRB-278P, 1:500), and TUJ1 (PRB-435P, 1:1000) were from BioLegend. Antibody for Engrailed-1 was from Developmental Studies Hybridoma Bank. Antibody for SOX3 (GT15119, 1:200) was from Neuromics. Antibodies for BRACHYURY (AF2085, 1:500), GATA3 (MAB6330, 1:200), GATA4 (AF2606, 1:400), GBX2 (AF4638, 1:200), HOXB1 (AF6318, 1:200), ID1 (AF4377, 1:200), ISLET1 (AF1837, 1:200), MMP9 (AF911, 1:200), NANOG (AF1997, 1:200), OCT4A (MAB17591, 1:200), OLIG2 (AF2418, 1:200), OTX2 (AF1979, 1:200), PAX6 (AF8150, 1:200), PHOX2B (AF4940, 1:200), SOX1 (AF3369, 1:400), SOX17 (AF1924, 1:500), SOX2 (AF2018, MAB2018, 1:200), SOX21 (AF3538, 1:200), and TUJ1 (MAB1195, 1:400) were from R&D Systems. Antibody for NANOG was from Reprocell. Antibodies for HES3 (sc-323948, 1:200), HES5 (sc-13859, 1:200) and LMX1A (sc-54273) were from Santa Cruz Biotechnology. Secondary antibody incubation was performed with Alexa flour conjugated antibodies at dilution of 1:400 (Life Technologies). For direct immunostaining, primary antibodies were conjugated using Alexa fluor monoclonal antibody labeling kits (A20181, A20184, A20186, Life Technologies). Nuclei were counterstained with DAPI (Life Technologies).

High-content analysis of colony morphology

Images were acquired with the Operetta (Perkin Elmer), analyzed in batch mode with custom building blocks on a Columbus server (Perkin Elmer) and visualized with Spotfire (Perkin Elmer). Colony morphology analysis ('distance from the edge' measurement) was achieved using a custom Acapella script (Perkin Elmer) run in Columbus with the following commands; 1) stitch a montage from 3x3 user-defined contiguous overlapping fields captured with the 20x objective, 2) segment and binarize DAPI signal from individual nuclei to create nuclear objects, 3) segment and binarize DAPI signal from the cytoplasm surrounding each nucleus to create cytoplasmic objects (note that hPSCs show strong blue fluorescence arise from sequestration of retinyl esters in cytoplasmic lipid bodies (Muthusamy et al., 2014), 4) dilate nuclear objects to eliminate gaps between neighboring objects, 5) create super objects by filling holes containing less than 30 pixels, 6) segment super objects, 7) create a perimeter line at the edge of each super object, 8) calculate the minimum distance between the centroid of each nucleus and the closest super object perimeter, 9) report fluorescence signal from nucleus and cytoplasm for each object. For each cell this script reports nuclear and cytoplasmic signals for all channels and a single measure of minimum distance to the closest perimeter of the epithelium. Using data visualization in Spotfire, median fluorescence signals from all cells within 10 μm was plotted corresponding to distance from an edge of epithelium.

Quantitative RT-PCR

Reverse transcription was performed using SuperScript III Reverse Transcriptase (Life Technologies). Quantitative RT-PCR was performed using Taqman gene expression assays (Life Technologies). Following Taqman probes were used: NANOG (Hs02387400_g1), PAX6 (Hs00240871_m1), and SOX1 (Hs01057642_s1). Relative RNA levels were calculated using the ΔΔCt method with human GAPDH as reference gene.

RNAseq library preparation

Total RNA was extracted using mirVana kit (Ambion) according to manufacturer's protocol. RNA quality control was performed using the Agilent 2100 Bioanalyzer System. RNAseq libraries were constructed using Illumina mRNA sequencing sample Prep Kit (for Poly-A libraries) or TruSeq Stranded Total RNA RiboZero sample Prep Kit (for strand-specific libraries) following the manufacturer's protocol. Briefly, poly-A containing mRNA molecules were purified or ribosomal RNAs were removed using RiboZero beads from ~ 800 ng DNase treated total RNA. Following purification, the resulting RNA was fragmented into small pieces using divalent cations under elevated temperature at 94°C for 2 min. Under this condition, the range of the fragment length obtained was 130-290 bp, with a median length of 185 bp. Reverse transcriptase and random primers were used to copy the cleaved RNA fragments into first strand cDNA. The second strand cDNA was synthesized using DNA Polymerase I and RNase H. These cDNA fragments went through an end repair process using T4 DNA polymerase, T4 PNK and Klenow DNA polymerase, the addition of a single 'A' base using Klenow exo (3' to 5' exo minus) and the ligation of Illumina PE adapters using T4 DNA Ligase. An index was inserted into Illumina adapters so that multiple samples can be sequenced in one lane of 8-lane flow cell if necessary. The concentration of RNA was measured by Qubit (Life Technologies). Quality of RNAseq library was measured by LabChipGX (Caliper) using HT DNA 1K/12K/HiSens Labchip. The final cDNA libraries were sequenced using HiSeq 2000 (for samples with Poly-A library preparation) or HiSeq 3000 (for samples with RiboZero library preparation) for high-throughput DNA sequencing.

RNAseq data processing

After sequencing run the Illumina Real Time Analysis (RTA) module was used to perform image analysis, base calling, and the BCL Converter (CASAVA v1.8.2) were followed to generate FASTQ files which contain the sequence reads. The current sequencing depth is over 80 million (40 million paired-end) mappable sequencing reads (Table S1). Read-level Q/C was performed by FastQC (v0.10.1). Pair-end reads of cDNA sequences are aligned back to the human genome (UCSC hg19 from Illumina iGenome) by the spliced read mapper TopHat (v2.0.4) with default option with "--mate-innder-dist 160" based on known transcripts of Ensembl Build GRCh37.75. For stranded RiboZero samples, TopHat used "--library-type fr-firststrand" option. The alignment statistics and Q/C was achieved by samtools (v0.1.18) and RSeQC (v2.3.5) to calculate quality control metrics on the resulting aligned reads, which provides useful information on mappability, uniformity of gene body coverage, insert length distributions and junction annotation, respectively. To achieve gene-level expression profile, the properly paired and mapped reads are achieved by "samtools sort –n" option, and these reads are counted by htseq-count v0.5.3 (with intersection-strict mode and stranded option for RiboZero samples) according to gene annotation (Illumina iGenome) and RPKM is calculated. This provides 23,368 gene-level expression profiles.

Statistics for SOX21 protein level

To determine the effect of cell line of origin on nuclear SOX21 protein levels in Figure 2A, we used a mixed model comparing the mean expression levels across cell lines while accounting for the correlation of expression levels within replicate experiments (a total of 5 independent growth experiments were conducted) with a random intercept, implemented in R using the lme4 library and the lmer() function: expression~ as.factor (line)*condition*day+(1| replicate). To specifically test the effect of line of origin, this model was compared to a second model with no line effect, using anova().

Bioinformatic analyses

Principle component analysis was done using the prcomp() function in R. Agglomerative hierarchical clustering of genes using gene-level RPKM from RNAseq data was performed using hclust() and cutree() with correlational distance (dist=1-r) in the R statistical language. Genome-wide CoGAPS Analysis in Parallel Sets (GWCoGAPS) was run using default parameters as previously described[2,62,86,87], for a range of k patterns (k=22 selected) and uncertainty as 10% of the data. Briefly, whole transcriptomic data was parallelized into seven sets. GWCoGAPS decomposes a matrix of experimental observations, **D**—here, log2 RNAseq RPKMs—with genes as rows and samples as columns, into two matrices, by the following equation.

$$\mathbf{D} \sim \mathrm{N}(\mathbf{AP},\Sigma) \tag{1}$$

Where, **A** is the amplitude matrix indicating the strength of involvement of a given gene in each pattern, **P** is the pattern matrix defining relationships (i.e. patterns) between samples. N and Σ are both functions of each element of **AP** and represent the Normal distribution and the standard deviation, respectively. Projection of principal components and GWCoGAPS gene weights defines patterns of relationships between samples in a new data associated with the gene expression signatures of the patterns from the primary data. These were achieved using the default projectR function in the projectR package as previously described (projectR at: https://github.com/genesofeve). Enrichment was calculated via either the calcCoGAPSStat function in the CoGAPS Bioconductor package or the geneSetTest function in the limma Bioconductor package in R. ANOVAs were used to assess the association of each GWCoGAPS pattern with treatment, time, and cell line of origin in Figure 3A, using lm() and summary() in R: lm(pattern~treatment*day+line).

Supplementary References

Chen, K.G., Mallon, B.S., Hamilton, R.S., Kozhich, O.A., Park, K., Hoeppner, D.J., Robey, P.G., and McKay, R.D. (2012). Non-colony type monolayer culture of human embryonic stem cells. Stem cell research *9*, 237-248.

Fertig, E.J., Ren, Q., Cheng, H., Hatakeyama, H., Dicker, A.P., Rodeck, U., Considine, M., Ochs, M.C., and Chung, C.H. (2012). Gene expression signatures modulated by epidermal growth factor receptor activation and their relationship to cetuximab resistance in head and neck squamous cell carcinoma. BMC genomics *13*.

Fertig, E.J., Stein-O'Brien, G., Jaffe, A., and Colantuoni, C. (2014). Pattern identification in time-course gene expression data with the CoGAPS matrix factorization. Methods in molecular biology *1101*, 87-112.

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.

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.

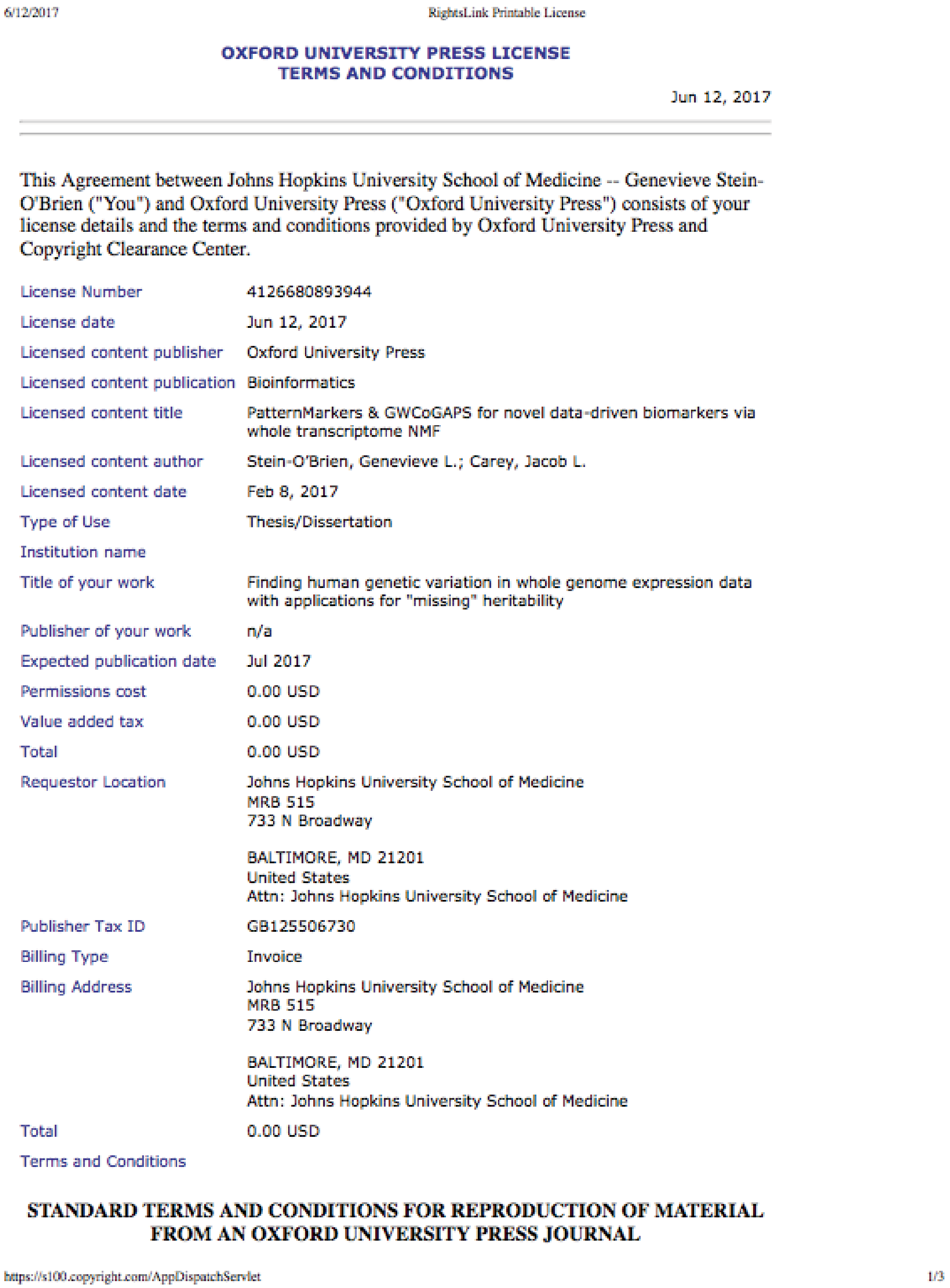
Maroof, A.M., Keros, S., Tyson, J.A., Ying, S.W., Ganat, Y.M., Merkle, F.T., Liu, B., Goulburn, A., Stanley, E.G., Elefanty, A.G., *et al.* (2013). Directed differentiation and functional maturation of cortical interneurons from human embryonic stem cells. Cell stem cell *12*, 559-572.

Montague, T.G., Cruz, J.M., Gagnon, J.A., Church, G.M., and Valen, E. (2014). CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. Nucleic acids research *42*, W401-407.

Muthusamy, T., Mukherjee, O., Menon, R., Megha, P.B., and Panicker, M.M. (2014). A method to identify and isolate pluripotent human stem cells and mouse epiblast stem cells using lipid body-associated retinyl ester fluorescence. Stem Cell Reports *3*, 169-184.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nature protocols *8*, 2281-2308.

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.



6/12/2017 RightsLink Printable License

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Jun 12, 2017

This Agreement between Johns Hopkins University School of Medicine -- Genevieve Stein-O'Brien ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

License Number	4126680893944
License date	Jun 12, 2017
Licensed content publisher	Oxford University Press
Licensed content publication	Bioinformatics
Licensed content title	PatternMarkers & GWCoGAPS for novel data-driven biomarkers via whole transcriptome NMF
Licensed content author	Stein-O'Brien, Genevieve L.; Carey, Jacob L.
Licensed content date	Feb 8, 2017
Type of Use	Thesis/Dissertation
Institution name	
Title of your work	Finding human genetic variation in whole genome expression data with applications for "missing" heritability
Publisher of your work	n/a
Expected publication date	Jul 2017
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Requestor Location	Johns Hopkins University School of Medicine MRB 515 733 N Broadway BALTIMORE, MD 21201 United States Attn: Johns Hopkins University School of Medicine
Publisher Tax ID	GB125506730
Billing Type	Invoice
Billing Address	Johns Hopkins University School of Medicine MRB 515 733 N Broadway BALTIMORE, MD 21201 United States Attn: Johns Hopkins University School of Medicine
Total	0.00 USD
Terms and Conditions	

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

https://s100.copyright.com/AppDispatchServlet 1/3

Permission to Reprint

1. Use of the material is restricted to the type of use specified in your order details.
2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.
3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.
4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.
5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.
6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oup.com
7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.
8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.
9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.
10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
12. Other Terms and Conditions:

v1.4

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Bibliography

1. Trendafilov, N. T. & Unkel, S. Exploratory factor analysis of data matrices with more variables than observations. *Journal of Computational and Graphical Statistics* **20,** 874–891 (2011).
2. Ochs, M. F. & Fertig, E. J. Matrix Factorization for Transcriptional Regulatory Network Inference. *... Bioinformatics and Computational Biology ...* 1–10 (2012).
3. Li, Y. & Ngom, A. The non-negative matrix factorization toolbox for biological data mining. *Source Code Biol Med* **8,** 10 (2013).
4. Brunet, J. P. *et al.* Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U.S.A.* **101,** 4164–4169 (2004).
5. Mejía-Roa, E. *et al.* bioNMF: a web-based tool for nonnegative matrix factorization in biology. **36,** W523–W528 (2008).
6. Fertig, E. J. *et al.* Gene expression signatures modulated by epidermal growth factor receptor activation and their relationship to cetuximab resistance in head and neck squamous cell carcinoma. *BMC Genomics* **13,** 160 (2012).
7. Ochs, M. F. *et al.* Detection of Treatment-Induced Changes in Signaling Pathways in Gastrointestinal Stromal Tumors Using Transcriptomic Data. *Cancer Res* **69,** 9125–9132 (2009).
8. Fertig, E. J. *et al.* Preferential Activation of the Hedgehog Pathway by Epigenetic Modulations in HPV Negative HNSCC Identified with Meta-Pathway Analysis. *PLoS ONE* **8,** e78127 (2013).
9. Kossenkov, A. V. & Ochs, M. F. in **467,** 59–77 (Elsevier, 2009).
10. Wang, G., Kossenkov, A. V. & Ochs, M. F. LS-NMF: A modified non-negative matrix factorization algorithm utilizing uncertainty estimates. *BMC Bioinformatics* **7,** 175 (2006).
11. de Campos, C. P. *et al.* Discovering Subgroups of Patients from DNA Copy Number Data Using NMF on Compacted Matrices. *PLoS ONE* **8,** e79720 (2013).
12. Sibisi, S. & Skilling, J. Prior Distributions on Measure Space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59,** 217–235 (1997).
13. Consortium, T. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).
14. Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348,** 660–665 (2015).
15. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 15545–15550 (2005).
16. Remedios, R. *et al.* A stream of cells migrating from the caudal telencephalon reveals a link between the amygdala and neocortex. *Nat Neurosci* **10,** 1141–1150 (2007).
17. Bell, G., Hey, T. & Szalay, A. Beyond the data deluge. *Science* (2009).
18. Wiley, H. S. At the Tipping Point. *The Scientist, 25(2):28* (2011).
19. Tyers, M. & Mann, M. From genomics to proteomics. *Nature* **422,** 193–197 (2003).
20. Green, E. D., Guyer, M. S., National Human Genome Research Institute, Manolio, T. A. & Peterson, J. L. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470,** 204–213 (2011).
21. Pollack, A. DNA sequencing caught in deluge of data. *New York Times* (2011).
22. Sagoff, M. Data deluge and the human microbiome project. *Issues in Science and Technology* (2012).
23. Li, S. *et al.* Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat Biotechnol* – (2000). doi:doi:10.1038/nbt.2972
24. Gamazon, E. *et al.* Integrative Genomics: Quantifying Significance of Phenotype-Genotype Relationships from Multiple Sources of High-Throughput Data. *Frontiers in Genetics* **3,** (2013).
25. Pan, S. J. & Yang, Q. A Survey on Transfer Learning. **22,** 1345–1359 (2010).
26. Torrey, L. & Shavlik, J. in *Handbook of Research on Machine Learning Applications and Trends Algorithms, Methods, and Techniques* (ed. Olivas, E. S.) 242–264 (2009).
27. Loh, K. M. *et al.* Efficient endoderm induction from human pluripotent stem cells by logically directing signals controlling lineage bifurcations. *Cell Stem Cell* **14,** 237–252 (2014).
28. Buganim, Y. *et al.* Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150,** 1209–1222 (2012).
29. Bernardo, A. S. *et al.* BRACHYURY and CDX2 Mediate BMP-Induced Differentiation of Human and Mouse Pluripotent Stem Cells into Embryonic and Extraembryonic Lineages. **9,** 144–155

(2011).

30. Petropoulos, S. *et al.* Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* **165,** 1012–1026 (2016).
31. Song, L., Chen, J., Peng, G., Tang, K. & Jing, N. Dynamic Heterogeneity of Brachyury in Mouse Epiblast Stem Cells Mediates Distinct Response to Extrinsic Bone Morphogenetic Protein (BMP) Signaling. *J. Biol. Chem.* **291,** 15212–15225 (2016).
32. Faial, T. *et al.* Brachyury and SMAD signalling collaboratively orchestrate distinct mesoderm and endoderm gene regulatory networks in differentiating human embryonic stem cells. *Development* **142,** 2121–2135 (2015).
33. Stein-O'Brien, G. L. *et al.* PatternMarkers & GWCoGAPS for novel data-driven biomarkers via whole transcriptome NMF. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx058
34. Nazareth, E. J. P. *et al.* High-throughput fingerprinting of human pluripotent stem cell fate responses and lineage bias. *Nat Meth* **10,** 1225–1231 (2013).
35. Poh, Y.-C. *et al.* Generation of organized germ layers from a single mouse embryonic stem cell. *Nat Comms* **5,** 470 (2014).
36. Etoc, F. *et al.* A Balance between Secreted Inhibitors and Edge Sensing Controls Gastruloid Self-Organization. *Developmental Cell* **39,** 302–315 (2016).
37. Fertig, E. J. *et al.* Gene expression signatures modulated by epidermal growth factor receptor activation and their relationship to cetuximab resistance in head and neck squamous cell carcinoma. *BMC Genomics* **13,** 160 (2012).
38. Peng, G. *et al.* Spatial Transcriptome for the Molecular Annotation of Lineage Fates and Cell Identity in Mid-gastrula Mouse Embryo. *Developmental Cell* **36,** 681–697 (2016).
39. González-Gómez, P., Anselmo, N. P., Anselmo, N. P., Mira, H. & Mira, H. BMPs as Therapeutic Targets and Biomarkers in Astrocytic Glioma. *BioMed Research International* **2014,** 1–8 (2014).
40. Videla Richardson, G. A. *et al.* Specific Preferences in Lineage Choice and Phenotypic Plasticity of Glioma Stem Cells Under BMP4 and Noggin Influence. *Brain Pathology* **26,** 43–61 (2015).
41. Wu, Q. & Yao, J. BMP4, a new prognostic factor for glioma. *World J Surg Onc* **11,** 264 (2013).
42. Bao, Z. *et al.* BMP4, a strong better prognosis predictor, has a subtype preference and cell development association in gliomas. *J Transl Med* **11,** 100 (2013).
43. De Robertis, E. M. Spemann's organizer and the self-regulation of embryonic fields. *Mech Dev* **126,** 925–941 (2009).
44. Gerhart, J. Changing the axis changes the perspective. **225,** 380–383 (2002).
45. Brons, I. G. *et al.* Derivation of pluripotent epiblast stem cells from mammalian embryos. **448,** 191–195 (2007).
46. Tesar, P. J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448,** 196–199 (2007).
47. Gifford, C. A. *et al.* Transcriptional and Epigenetic Dynamics during Specification of Human Embryonic Stem Cells. *Cell* **153,** 1149–1163 (2013).
48. Stergachis, A. B. *et al.* Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. **154,** 888–903 (2013).
49. Choi, J. *et al.* A comparison of genetically matched cell lines reveals the equivalence of human iPSCs and ESCs. **33,** 1173–1181 (2015).
50. Wu, J. *et al.* An alternative pluripotent state confers interspecies chimaeric competency. *Nature* **521,** 316–321 (2015).
51. Arnold, S. J. & Robertson, E. J. Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. *Nat Rev Mol Cell Biol* **10,** 91–103 (2009).
52. Singh, A. M. *et al.* Signaling network crosstalk in human pluripotent cells: a Smad2/3-regulated switch that controls the balance between self-renewal and differentiation. *Cell Stem Cell* **10,** 312–326 (2012).
53. Liu, J., Xu, Y., Stoleru, D. & Salic, A. Imaging protein synthesis in cells and tissues with an alkyne analog of puromycin. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 413–418 (2012).
54. Chambers, S. M. *et al.* Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nat Biotechnol* **27,** 275–280 (2009).
55. Simeone, A. Genetic control of dopaminergic neuron differentiation. *Trends in Neurosciences* **28,** 62–65 (2005).
56. Kuzmichev, A. N. *et al.* Sox2 Acts through Sox21 to Regulate Transcription in Pluripotent and

Differentiated Cells. *Current Biology* **22,** 1705–1710 (2012).

57. Warmflash, A., Sorre, B., Etoc, F., Siggia, E. D. & Brivanlou, A. H. A method to recapitulate early embryonic spatial patterning in human embryonic stem cells. **11,** 847–854 (2014).
58. Rotem, A. *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol* **33,** 1165–1172 (2015).
59. Zorn, A. M. & Wells, J. M. Vertebrate endoderm development and organ formation. *Annu Rev Cell Dev Biol* **25,** 221–251 (2009).
60. Rouhani, F. *et al.* Genetic background drives transcriptional variation in human induced pluripotent stem cells. *PLoS Genet* **10,** e1004432 (2014).
61. Nimmo, R. A., May, G. E. & Enver, T. Primed and ready: understanding lineage commitment through single cell analysis. *Trends Cell Biol* **25,** 459–467 (2015).
62. Fertig, E. J., Ding, J., Favorov, A. V., Parmigiani, G. & Ochs, M. F. CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data. (2010).
63. Niwa, H., Miyazaki, J. & Smith, A. G. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat Genet* **24,** 372–376 (2000).
64. Factor, D. C. *et al.* Epigenomic comparison reveals activation of 'seed' enhancers during transition from naive to primed pluripotency. *Cell Stem Cell* **14,** 854–863 (2014).
65. Boroviak, T. *et al.* Lineage-Specific Profiling Delineates the Emergence and Progression of Naive Pluripotency in Mammalian Embryogenesis. *Dev Cell* **35,** 366–382 (2015).
66. Denans, N., Iimura, T. & Pourquie, O. Hox genes control vertebrate body elongation by collinear Wnt repression. **4,** (2015).
67. Kuzmichev, A. N. *et al.* Sox2 acts through Sox21 to regulate transcription in pluripotent and differentiated cells. *Curr Biol* **22,** 1705–1710 (2012).
68. Nord, A. S., Pattabiraman, K., Visel, A. & Rubenstein, J. L. Genomic perspectives of transcriptional regulation in forebrain development. *Neuron* **85,** 27–47 (2015).
69. Mallon, B. S. *et al.* StemCellDB: the human pluripotent stem cell database at the National Institutes of Health. *Stem Cell Res* **10,** 57–66 (2013).
70. Mele, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348,** 660–665 (2015).
71. Hoch, R. V., Lindtner, S., Price, J. D. & Rubenstein, J. L. OTX2 Transcription Factor Controls Regional Patterning within the Medial Ganglionic Eminence and Regional Identity of the Septum. *Cell Rep* **12,** 482–494 (2015).
72. Acampora, D., Gulisano, M., Broccoli, V. & Simeone, A. Otx genes in brain morphogenesis. *Prog Neurobiol* **64,** 69–95 (2001).
73. Waters, S. T. & Lewandoski, M. A threshold requirement for Gbx2 levels in hindbrain development. **133,** 1991–2000 (2006).
74. Labelle-Dumais, C., Jacob-Wagner, M., Pare, J. F., Belanger, L. & Dufort, D. Nuclear receptor NR5A2 is required for proper primitive streak morphogenesis. *Dev Dyn* **235,** 3359–3369 (2006).
75. Onai, T. *et al.* Retinoic acid and Wnt/beta-catenin have complementary roles in anterior/posterior patterning embryos of the basal chordate amphioxus. *Dev. Biol.* **332,** 223–233 (2009).
76. Bouillet, P. *et al.* A new mouse member of the Wnt gene family, mWnt-8, is expressed during early embryogenesis and is ectopically induced by retinoic acid. *Mechanisms of Development* **58,** 141–152 (1996).
77. Simoes-Costa, M. & Bronner, M. E. Establishing neural crest identity: a gene regulatory recipe. **142,** 242–257 (2015).
78. Balmer, J. E. & Blomhoff, R. Gene expression regulation by retinoic acid. *J Lipid Res* **43,** 1773–1808 (2002).
79. Parker, H. J., Bronner, M. E. & Krumlauf, R. A Hox regulatory network of hindbrain segmentation is conserved to the base of vertebrates. **514,** 490–493 (2014).
80. Brook, F. A. & Gardner, R. L. The origin and efficient derivation of embryonic stem cells in the mouse. *Proc. Natl. Acad. Sci. U.S.A.* **94,** 5709–5712 (1997).
81. Dunn, S. J., Martello, G., Yordanov, B., Emmott, S. & Smith, A. G. Defining an essential transcription factor program for naive pluripotency. **344,** 1156–1160 (2014).
82. Buecker, C. *et al.* Reorganization of enhancer patterns in transition from naive to primed pluripotency. *Cell Stem Cell* **14,** 838–853 (2014).
83. Wu, J. & Izpisua Belmonte, J. C. Dynamic Pluripotent Stem Cell States and Their Applications. **17,**

509–525 (2015).

84. Lodato, M. A. *et al.* Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350,** 94–98 (2015).
85. Erwin, J. A., Marchetto, M. C. & Gage, F. H. Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat Rev Neurosci* **15,** 497–506 (2014).
86. Stein-O'Brien, G. *et al.* PatternMarkers & GWCoGAPS for novel data-driven biomarkers via whole transcriptome NMF. 083717 (2016). doi:10.1101/083717
87. Fertig, E. J., Favorov, A. V. & Ochs, M. F. Identifying context-specific transcription factor targets from prior knowledge and gene expression data. *IEEE Trans Nanobioscience* **12,** 142–149 (2013).

Curriculum Vitae

GENEVIEVE STEIN-O'BRIEN

115 W. MONUMENT ST. APT 2F BALTIMORE, MD 21201
267.319.5019 GSTEINO1@JHMI.EDU

EDUCATION

Johns Hopkins University, School of Medicine **Baltimore, Maryland**
Ph.D. Human Genetics 2017
Thesis: Finding human genetic variation in whole genome expression data with applications for "missing" heritability: The GWCoGAPS algorithm, the PatternMarkers statistic, and the ProjectoR package

Johns Hopkins University, School of Public Health **Baltimore, Maryland**
MHS Biostatistics expected 2017

Bryn Mawr College **Bryn Mawr, Pennsylvania**
A.B., Mathematics; minor Chemistry 2007
Senior Honors Thesis: Spectral Theorem for Compact Hermitian Operators

PUBLICATIONS

Stein-O'Brien, G., Carey, J., Lee, W.-S., Considine, M., Favorov, A., Flam, E., Guo, T., Li, L., Marchionni, L., Sherman, T., Gaykalova, D., McKay, R., Ochs, M., Colantuoni, C., Fertig, E. **PatternMarkers & GWCoGAPS for novel data-driven biomarkers via whole transcriptome NMF.** ***Bioinformatics*** (2017). doi: 10.1093/bioinformatics/btx058

Fertig E.J., Ozawa H., Thakar M., Howard J.D., Kagohara L.T., Krigsfeld G., Ranaweera R.S., Hughes R.M., Perez J., Jones S., Favorov A.V., Carey J., **Stein-O'Brien, G.**, Gaykalova D.A., Ochs M.F., Chung C.H., **CoGAPS matrix factorization algorithm identifies transcriptional changes in AP-2alpha target genes in feedback from therapeutic inhibition of the EGFR network.** ***Oncotarget*** (2016). doi:10.18632/oncotarget.12075

Sun, G., Zhou, Y., Ito S., Bonaguidi M., **Stein-O'Brien, G**., Kawasaki, N., Modak, N., Zhu, Y., and Ming, G. **Latent tri-lineage potential of adult hippocampal neural stem cells revealed by Nf1 inactivation**. ***Nature Neuroscience*** (2015). doi:10.1038/nn.4159

Fertig E.J., **Stein-O'Brien, G.L.**, Jaffe, A., and Colantuoni, C. **Pattern identification in time course gene expression data with the CoGAPS matrix factorization**. ***Gene Function Analysis, Methods in Molecular Biology, 2nd Ed.*** Edited by MF Ochs. Springer Verlag, 2012.

POSTERS AND PRESENTATIONS (Selected)

G. Stein-O'Brien, A. Jaishankar, S.K. Kim, S. Seo, J. Heon Shin, D. Hoeppner, J. Chenoweth1, T. Hyde, J. Kleinman, D. Weinberger, E. Fertig, C. Colantuoni, R. McKay, ***Genome specific transcriptional signatures predict differentiation biases in Human ES/IPS cells***. **Gordon Conference:** Stochastic Physics in Biology, January 2017, Ventura, Ca.

G. Stein-O'Brien, A. Jaishankar, S.K. Kim, S. Seo, J. Heon Shin, D. Hoeppner, J. Chenoweth1, T. Hyde, J. Kleinman, D. Weinberger, E. Fertig, C. Colantuoni, R. McKay, ***Genome specific transcriptional signatures predict differentiation biases in Human ES/IPS cells****; (Abstract #3074). Presented at the 65th Annual Meeting of The American Society of Human Genetics, October 8, 2015, Baltimore, MD. *Reviewers' Choice Abstract scoring in the top 10% of all abstracts reviewed

C. Colantuoni, A. Jaishankar, **G. Stein-O'Brien**, E. Fertig, J. Heon Shin, S.K. Kim, S. Seo, Y. Wang, D. Hoeppner, J. Chenoweth, R. McKay. ***Transcriptional signatures of lineage bias and human genomic identity in the pluripotent state.*** Presented at the Society for Neuroscience Annual Conference, Washington D.C., 2014.

V.S. Caviness, K.Y. Araki, **G. L. Stein-O'Brien**, Li Cai, P.G. Bhide1, R.S. Nowakowski. ***Gene Onotology and Patterns of Transcription in Murine Neocortical Progenitors.*** Presented at Society for Neuroscience Conference, San Diego, CA, 2004.

K.Y. Araki, **G. L. Stein-O'Brien**, Li Cai, R.S. Nowakowski, P.G. Bhide, V.S. Caviness. ***Large Scale Regulation of Neocortical Progenitor Transcription.*** Presented at Society for Neuroscience Conference, New Orleans, LA, 2003.

K.Y. Araki, **G. L. Stein-O'Brien**, Li Cai, R.S. Nowakowski, P.G. Bhide, V.S. Caviness. ***Transcriptional Profiles of Neocortical Progenitor Cells Change with Cell Cycle Progression.*** Presented at Society for Neuroscience Conference, Orlando, Fl, 2002.

SKILLS

COMPUTATIONAL	LABORATORY
Languages: Unix, R, Mathlab, Perl Design: Photoshop, InDesign, Illustrator Standard: Excel, Powerpoint, Keynote	PCR, Gel Electrophoresis, Western Blot, Cellular laser capture, double IdU-BrdU, Immunohistochemistry techniques, BLAST, cell culture, HIPAA procedures

RESEARCH EXPERIENCE

Johns Hopkins University, MD: Institute of Genetic Medicine, Graduate Student *2012 – 2017*

- Built the R package, projectoR to interrogate shared sources of variation across independent data sets, types, and technical platforms using gene weights derived from multiple high dimensional analytic methods
- Derived distance metric to extract unique biomarkers from continuous weights of non-negative matrix factorization (NMF) solutions for biological validation and enhanced visualization
- Developed Genome-Wide CoGAPS Analysis in Parallel Sets (GWCoGAPS), the first robust whole genome Bayesian NMF using the sparse, MCMC algorithm, CoGAPS to infer dynamics in high-throughput "omics" data.
- Applied analytical skills to study of human development and pluripotent stem cells (hPSC) systems resulting in methods to predict the functional effect of genomic bias in model systems

University of Pennsylvania, PA: Computational Memory Lab, Research Assistant *2009 – 2010*

- Collected and analyzed scalp EEG and behavioral data from human participants.
- Compiled background for and drafted NIH grant submissions and progress reports.
- Edited and composed content for Principle Investigator's text book, talk slides, and class slides.
- Organized and maintained files, daily administrative operations, and the physical lab.

Massachusetts General Hospital/Harvard Medical School: Neurology, Intern *2002 – 2005*

- Created pipeline using Affymetrix, SAM, and d-chip to analysis microarray data with users manual
- Performed laser capture, Immunohistochemistry, and double IdU-BrdU labeling
- Designed diagrams for articles in Cerebral Cortex and posters for the Society for Neuroscience Conference.
- Presented one of the resulting posters at the Annual SfN Conference in Orlando, Fl in 2002.

Rutgers University, NJ: Molecular Ecology Department, Research Assistant *Summer 2001*

- Independently carried out the daily procedures including PCR, DNA sequencing, and SNP genotyping.
- Offered 200% raise to return the following year.

TEACHING EXPERIENCE

Teaching Assistant & Grader, JHSPH: Statistical Methods in Public Health	*2015 - 2016*
Teaching Assistant, JHMI: Bioinformatics	*2015 - 2016*
Teaching Assistant, JHMI: Practical Genomics Workshop	*2015*
Teaching Assistant & Grader, Bryn Mawr: Linear Algebra	*2006 - 2007*
Teaching Assistant & Grader, Bryn Mawr: Multivariable and Basic Calculus	*2004 - 2006*
Tutor: Mathematics, Chemistry, Physics, SAT, & GRE	*2001 - 2010*

PROFESSIONAL EXPERIENCE (Selected)

Video Vértité, PA: Production Assistant, *Journey into Dyslexia* *2010*

- Worked with Oscar winning film makers Susan and Alan Raymond on HBO documentary released 2011.

Princeton, NJ: Freelance Editor and Research Assistant *2008*

- Edited and researched text for Poison Pills: The Untold Story of the Vioxx Drug Scandal.
- Compiled timeline, bibliography, and cast of characters to accompany text for publication.

ADDITIONAL EXPERIENCE (Selected)

Co-leader of Genomics for Students Discussion Group

Election Judge, Elections Board of the City of Baltimore

Machine Inspector, Elections Board of the 8th division of the 46th Ward of Philadelphia

- Responsible for operation and integrity of voting machines at polling place on election day.

Block Captain and Community Organizer, Philadelphia Streets Department

Editor, The Bi-College News – the student newspaper of Bryn Mawr and Haverford Colleges

Editor-In-Chief, (2005-2006); *Production Manager, Managing Editor*, (2004-2005).

- Oversaw production of this two-college newspaper with 3000+ readers.
- Designed and implemented a revolutionary management system improving the quality of the newspaper and the efficiency and experience of its staff. Overhauled style guide and newsroom.
- Created first annual "Bi-Co Boot Camp" training and recruiting retreat.
- Grew the Bryn Mawr staff from 8 to over 30 and the entire staff from 20 people to over 60.

Haverford Representative, Self-Governance Association of Bryn Mawr College

- Drafted and successfully lobbied for constitutional changes which restructured the party policy.
- As a voting member of the student government, worked to build the relationship between Bryn Mawr and Haverford College. As well as, increased awareness of cross-campus opportunities.

Bi-College Media Manager for WHRC, Bryn Mawr and Haverford Student Radio