TITLE

Analysis of regulatory network modules in hundreds of human stem cell lines reveals complex epigenetic and genetic factors contribute to pluripotency state differences between subpopulations

AUTHORS

Timothy D. Arthur[1,2], Jennifer P. Nguyen[2,3], Agnieszka D'Antonio-Chronowska[4], Hiroko Matsui[5], Nayara S. Silva[6], Isaac N. Joshua[5], iPSCORE Consortium[*], André D. Luchessi[6,7], William W. Young Greenwald[3], Matteo D'Antonio[2, 5], Martin F. Pera[8], and Kelly A. Frazer[4,5%]

[1]Biomedical Sciences Graduate Program, University of California, San Diego, La Jolla, CA 92093, USA.

[2]Division of Biomedical Informatics, University of California, San Diego, La Jolla, CA 92093, USA.

[3]Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, CA 92093, USA.

[4]Department of Pediatrics, University of California San Diego, La Jolla, CA 92093, USA.

[5]Institute of Genomic Medicine, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093, USA.

[6]Northeast Biotechnology Network (RENORBIO), Graduate Program in Biotechnology, Federal University of Rio Grande do Norte, Natal, Brazil

[7]Department of Clinical and Toxicological Analysis, Federal University of Rio Grande do Norte, Natal, Brazil

[8]The Jackson Laboratory, Bar Harbor, ME 04609, USA

[*] A full list of Consortium members and their affiliations appears at the end of the manuscript

[%] Correspondence to: Kelly A. Frazer (kafrazer@health.ucsd.edu).

SUMMARY

Stem cells exist *in vitro* in a spectrum of interconvertible pluripotent states. Analyzing hundreds of hiPSCs derived from different individuals, we show the proportions of these pluripotent states vary considerably across lines. We discovered 13 gene network modules (GNMs) and 13 regulatory network modules (RNMs), which were highly correlated with each other suggesting that the coordinated co-accessibility of regulatory elements in the RNMs likely underlied the coordinated expression of genes in the GNMs. Epigenetic analyses revealed that regulatory networks underlying self-renewal and pluripotency have a surprising level of complexity. Genetic analyses identified thousands of regulatory variants that overlapped predicted transcription factor binding sites and were associated with chromatin accessibility in the hiPSCs. We show that the master regulator of pluripotency, the NANOG-OCT4 Complex, and its associated network were significantly enriched for regulatory variants with large effects, suggesting that they may play a role in the varying cellular proportions of pluripotency states between hiPSCs. Our work captures the coordinated activity of tens of thousands of regulatory elements in hiPSCs and bins these elements into discrete functionally characterized regulatory networks, shows that regulatory elements in pluripotency networks harbor variants with large effects, and provides a rich resource for future pluripotent stem cell research.

KEYWORDS

Induced pluripotent stem cells, regulatory network, pluripotency cell states, gene co-expression, regulatory element co-accessibility, epigenomics, genome regulation, stem cell genetics

INTRODUCTION

Characterizing the molecular mechanisms underlying self-renewal, pluripotency, and transitions between closely related cellular states is essential for advancing our understanding of development and disease processes. Human pluripotent stem cells (hiPSCs) exhibit a spectrum of interconvertible pluripotent cell states that have different transcriptional and epigenetic profiles and hence present an excellent model to study how regulatory networks govern these biological processes[1,2]. However, the molecular characterization of self-renewal and pluripotent cell states in hiPSCs has been impeded because most studies examine a limited number of lines, and each line is composed of subpopulations of cells in different pluripotency states[1,3]. Therefore, examining the transcriptome-wide gene co-expression and genome-wide regulatory element co-accessibility across hundreds of hiPSC lines using network and unsupervised community detection algorithms could provide novel insights into pluripotency cell state transitions (i.e., differences between closely related states), and how genetic background contributes to variability in self-renewal and pluripotency cell state-specific network activity across cell lines.

While it is clear that under conventional culturing conditions, each hiPSC line is composed of subpopulations of cells in different pluripotency states, how these proportions vary from cell line to cell line has not yet been investigated in depth. Most cells in a conventional hiPSC culture resemble late post-implantation epiblast stem cells, referred to as 'primed' pluripotency (Figure S1). Primed pluripotent cells are poised for rapid transition from a bivalent to an active state of lineage-specific genes. Additionally, an intermediate cellular state, referred to as 'formative' pluripotency, exists as a subpopulation in conventional hiPSC cultures[1,4]. Formative pluripotency represents the early post-implantation epiblast (EPE) and is characterized by an enrichment of self-renewal processes, an absence of lineage priming, and the capacity for direct conversion into somatic or germline lineages[1]. Under specific culture conditions, hiPSCs can be maintained in the 'naïve' state resembling the preimplantation epiblast[5], or a totipotent state resembling the 8-cell morula[3]. Understanding how pluripotent subpopulations vary across hiPSC lines under conventional culturing could improve their utility for studying developmental processes and regenerative medicine.

Gene co-expression network analysis is a powerful method that has been widely used to analyze RNA-seq data to identify modules of co-expressed genes that are members of the same biological pathways[6–8]. But thus far, applications of these algorithms to analyze ATAC-seq data have been more limited. For example, studies have used paired bulk ATAC-seq and RNA-seq to examine how regulatory elements impact gene expression networks under dynamic conditions, including comparing cells at baseline and post-stimulation[9] or undergoing differentiation[10]. There has also been considerable effort studying patterns of co-accessibility in single-cell ATAC-seq under static conditions to explore local *cis* interactions (~500kb) between regulatory elements both in single tissues[11] and simultaneously across multiple tissues[12]; and a recent study analyzed paired single-cell ATAC-seq and RNA-seq data to study local *cis*-regulation of gene expression under environmental stimuli[13]. However, genes that are members of the same biological pathways are frequently encoded on different chromosomes, and there have been limited studies aimed at examining co-accessible ATAC-seq peaks distributed across the human genome to understand regulatory processes underlying the co-expression of gene modules. For example, in hiPSCs, the maintenance of pluripotency relies on the expression of pluripotency-related transcription factors, such as

NANOG, OCT4, and SOX2, which create global epigenomic regulatory networks that enable self-renewal through the repression of developmental genes[14], regulation of cell cycle transitions[1,15], and promotion of autoregulation[16,17]. The ability to bin the precise locations and epigenetic profiles of all regulatory elements distributed across the genome into discrete regulatory networks in hiPSCs would provide a valuable resource for further investigations of these important cellular processes.

In this study, we sought to determine if regulatory network modules in hiPSC lines could be identified by examining genome-wide chromatin co-accessibility across samples from hundreds of individuals in the iPSCORE cohort. We hypothesized that these regulatory network modules could be identified due to heterogeneity arising from variation in the proportions of pluripotent subpopulations (Figure S1) across hiPSC lines, and would underlie the coordinated expression of gene network modules. We generated 150 ATAC-seq samples from 143 hiPSC lines in iPSCORE and applied an unsupervised machine learning algorithm that enabled us to discover genome-wide regulatory network modules (RNMs) comprised of co-accessible regulatory elements. We integrated these data with gene network modules (GNMs) that we similarly generated from the RNA-seq dataset[18,19] for 213 hiPSC lines also in iPSCORE. We demonstrated that both the RNMs and GNMs were associated with the differential expression of marker genes defining hiPSC pluripotency cell states, and showed that their discovery is due to differences in the proportion of these transitory pluripotency states across the hiPSC samples. The analyses of these datasets coupled with whole-genome sequencing data enabled us to identify and functionally characterize the TF binding and chromatin state profiles of elements in high-resolution hiPSC regulatory networks associated with distinct pluripotency states; and characterize mechanisms by which regulatory variants affect TF binding and regulatory networks.

RESULTS

Relative proportions of pluripotent subpopulations vary across 213 hiPSC lines

Previous studies have demonstrated that human induced pluripotent stem cell (hiPSC) lines are composed of subpopulations of interconvertible pluripotent cell states (Figure 1A, Figure S1)[1,4]. We set out to determine if the relative proportion of these cell states varied across 213 hiPSC lines with bulk RNA-seq data (Table S1, Table S2) using cellular deconvolution[20]. Deconvolution typically uses gene signatures obtained from cell type clusters in single-cell data; however, due to the high similarity between hiPSC pluripotency states, they don't separate into distinct clusters. Therefore, we generated gene signatures using bulk RNA-seq data for FACS-sorted formative and paired unsorted (e.g. primed) cells[1]. We applied the CIBERSORT deconvolution algorithm with a signature matrix containing the 100 most differentially expressed genes between the two populations (Table S3) and observed that the estimated fraction of cells in the formative state exhibited a wide range across the 213 hiPSC lines (Figure S2A). We examined the expression of formative-specific and primed-specific genes in hiPSCs with the highest and lowest estimated proportions of formative cells and observed notable expression differences of key regulators of pluripotency, such as *DUSP5*, *LEFTY1*, *FST*, and *FZD5* (Figure 1B-E, Figure S2B-C). Taken together, these results show that remarkable variation in pluripotency cell state composition exists across the 213 hiPSC lines under conventional culture conditions.

Identification of gene networks associated with pluripotency state

Pluripotent cell states have traditionally been defined by the expression of genes underlying self-renewal, pluripotency, and cell cycle regulation[16,21,22]. To identify co-expressed gene modules for these key biological processes, we analyzed the 213 hiPSC RNA-seq samples (Figure 1) and calculated the pairwise correlation between 16,110 expressed autosomal genes (TPM ≥ 1 in at least 20% of samples) using a linear mixed model (LMM)[19,23,24]. We identified 3,533,609 co-expressed gene pairs (adjusted P-value < 0.05) with positive associations (FigShare XXX), created a global gene network (GN) using these gene pairs as edges and identified gene co-expression network modules (GNMs) by applying the unsupervised Leiden community detection algorithm[25,26] to the GN (Supplemental Note 1 shows our approach for detecting gene modules works better for the iPSCORE cohort, which contains related individuals, than the commonly used WGCNA[6] approach; Figure S3, Figure S4). In total, we identified 13 gene co-expression network modules (GNMs) consisting of between 118 to 1,854 genes (Table S2, Table S4). Biological networks are scale-free[27] and follow the Pareto Principle[28], which states that 20% of the nodes are responsible for 80% of a network's connectivity. Therefore, we examined the expression of 2,964 GNM-specific Pareto genes (top 20 percentile of intra-GNM degree connectivity) in UMAP space (Figure 1F), and observed that the genes within the same GNM clustered together, indicating that they were highly co-expressed across the 213 hiPSC lines. To further validate that the GNMs captured genes enriched for being co-expressed across the 213 hiPSC lines, we first determined the number of co-expressed genes between each pairwise combination of GNMs. We then performed Fisher's Exact tests on the number of co-expressed genes within each GNM compared with the number of co-expressed genes between each set of paired GNMs, which showed that genes within a GNM were significantly more likely to be co-expressed with one another (Figure 1G). These results show that the Leiden community detection algorithm identified 13 GNMs consisting of genes enriched for being co-expressed across the 213 hiPSC lines.

We next sought to determine if the GNMs were enriched for different pluripotency cell states. We first examined whether certain GNMs were associated with the estimated fraction of formative state cells (Figure S2A). For each of the 213 RNA-seq samples, we calculated a GNM score for each of the 13 GNMs by summing the inverse normal transformed TPM expression of the corresponding GNM-specific Pareto genes (each RNA-seq sample had 13 GNM scores). To identify associations between the estimated fraction of formative state cells and GNM scores across the 213 hiPSC lines, we ran a linear model and observed that GNM 5 was positively associated with the estimated proportion of formative state cells, while GNM 10 exhibited a strong negative association (Figure 1H-I, Figure S5). Next, we utilized marker genes for hiPSC pluripotency states defined in previous studies (Figure 1J, Table S5). We discovered that three GNMs (1, 10, and 11) were associated with genes upregulated in the 8-cell like cells (8CLC) totipotent cell state[3]; GNM 5 was enriched for gene sets upregulated in the naïve[29], formative (EPE)[1], epiblast[29] states; GNM 9 and GNM 13 were both enriched for genes upregulated in the primitive endoderm-primed founder cells (PrE)[29]; while GNM 10 was strongly depleted for genes associated with the formative state[1] and enriched for genes associated with the primed[1], 8CLC[3], and trophectoderm[29] states (Figure 1J). Examining GNM 5 and 10 networks, we observed that they consisted of genes in molecular pathways characteristic of different stem cell states. For example, GNM 5 consists of genes in Nodal signaling, ERK signaling, self-renewal processes, extracellular (ECM) matrix formation, and Wnt signaling pathways (Figure 1K), which are characteristic of formative and naïve pluripotent stem cell states[1]. While GNM 10 consists of frizzled receptors, and GTPase Coupled

Receptors (Figure 1L), characteristic of the primed pluripotent stem cell state and neural lineage priming [1,4]. These findings show that certain GNMs are enriched for genes upregulated or downregulated in specific pluripotency states and that their discovery is due to differences in the proportion of the pluripotency cellular states across the hiPSC samples.

Altogether, we identified 13 modules of co-expressed genes differentially enriched for the expression of marker genes defining the continuum of hiPSC pluripotency cell states.

Identification of genome-wide regulatory network underlying the co-expression of gene modules

After discovering gene co-expression modules associated with self-renewal and interconvertible pluripotency cell states, we sought to identify the underlying regulatory modules (Figure 2A). We generated 150 independent ATAC-seq libraries from 143 hiPSC lines derived from 133 iPSCORE individuals (7 lines were cultured independently twice and 5 individuals each had two or three independent clones) (Table S6).

To identify modules of co-accessible regulatory elements, we calculated the accessibility of 56,978 peaks across the 150 ATAC-seq samples (Figure S6A-C, Table S8, FigShare XXX) and applied an LMM to identify pairs of co-accessible ATAC-seq peaks. We created a genome-wide regulatory network (RN) using 8,696,814 co-accessible autosomal ATAC-seq peak pairs (P-value < $5x10^{-8}$, Effect Size > 0) as edges. We applied the unsupervised Leiden community detection algorithm[24,25] to the RN and identified 13 major regulatory network modules (RNMs) each consisting of at least 500 ATAC-seq peaks (mean = 3,673.9 ± 1,261.9 peaks) (Figure 2B). Of the 56,978 reference peaks, 47,761 were present in these 13 RNMs. We analyzed the accessibility of the most interconnected peaks within each module (top 10% intramodular degree) and observed RNM-specific clustering in the UMAP space (Figure 2C), indicating that the RNMs capture ATAC-seq peaks with similar varying accessibility across the hiPSC lines. We also calculated the intramodular co-accessibility enrichment between each pairwise combination of the 13 RNMs, which showed that peaks within an RNM were significantly more likely to be co-accessible compared with peaks in different RNMs (Figure 2D, Table S13). These findings validate that the RNMs capture highly co-accessible peaks.

We sought to determine whether the RNMs were associated with specific pluripotency states. We initially performed cellular deconvolution on the 150 ATAC-seq libraries using the most differentially accessible ATAC-seq peaks from FACS-sorted formative and primed cells[1]. Only a small fraction of cells (range 0-36.6%) in each ATAC-seq sample were estimated to be in the formative state (Figure 2E, Figure S2A, Table S6, Table S7). We calculated RNM scores by summing the inverse normal transformed accessibility of 9,545 Pareto peaks (top 20% intramodular connectivity) for each ATAC-seq sample (each sample had 13 RNM scores) (Table S8). We then ran a linear model to test for associations between the estimated proportion of formative state cells and RNM scores (Figure 2F-H). We observed that RNM 3 had a strong positive correlation with the estimated formative proportion, while RNMs 2 and 8 had weaker positive correlations (Figure 2F-H). Finally, we examined whether genome-wide ATAC-seq peak co-accessibility is associated with the coordinated gene expression in different pluripotency states. We annotated 32,327 ATAC-seq peaks in the 13 RNMs with 12,078 neighboring candidate target genes and then performed a Fisher's Exact test (see Methods) to calculate enrichments of RNMs in GNMs. We found that all RNMs had an association with at least one GNM (Figure 2I). For example, five RNMs (2, 3, 5, 8, and 13) were positively enriched for the formative GNM 5 (Figure 1G-H, J), suggesting that co-accessibility of ATAC-seq peaks across the genome mechanistically underlie the differential expression of genes between pluripotency states.

Altogether, we discovered 13 regulatory network modules composed of highly co-accessible peaks across 143 hiPSCs. We show that the RNMs are associated with different pluripotency states and demonstrate considerable variability in their proportions between hiPSC lines. Additionally, we show that these regulatory modules were strongly associated with, and likely mechanistically underlie, the coordinated expression within gene network modules.

Functional annotation of hiPSC ATAC-seq peaks

We hypothesized that the co-accessible ATAC-seq peaks within a module have similar epigenetic profiles and molecular functions. However, unlike genes, which have been annotated with regard to their expression profiles and biological functions, regulatory elements have not been well characterized. Therefore, in order to functionally characterize the 13 RNMs, we annotated the ATAC-seq peaks with three epigenetic annotations (Figure S7A): 1) hiPSC-specific chromatin states[30–32], 2) TF binding sites[33–35], and 3) formative (EPE)-associated ATAC-seq peaks[1].

We initially annotated each of the 56,978 reference ATAC-seq peaks with chromatin states collapsed into five main categories (See Methods, Table S8) and observed that 46.8% of the ATAC-seq peaks were in enhancers, 22.0% were in active promoters, 13.9% were in bivalent or poised chromatin, 5.8% were in repressed polycomb regions, and 11.45% were in transcribed regions (Figure S7B). Our annotations were consistent with previous characterizations of hiPSC regulatory elements[1,36–38], specifically; 1) the relatively large fraction of the peaks in bivalent chromatin indicating open but inactive regulatory elements, and 2) the presence of peaks in polycomb regions.

We identified 187 TFs that were expressed in the hiPSCs and predicted their binding across the 56,978 reference ATAC-seq peaks using TOBIAS, a digital footprinting method[33] (Figure S7C, Table S10, Table S11, Figshare XXX). To validate the TOBIAS predictions, we used ENCODE ChIP-seq data for 18 TFs generated using H1 embryonic stem cells (ESCs)[35,39,40] (See Methods, Figure S8, Table S9). We observed that TOBIAS-predicted binding sites were strongly enriched in corresponding TF ChIP-seq peaks (Fisher's Exact test; maximum nominal $p = 5.5 \times 10^{-188}$), indicating that TOBIAS accurately predicted TF binding across the epigenome. Since TOBIAS often predicted that TFs with similar motifs bound at the same site, we collapsed the 187 TFs based on their predicted overlap of bound sites (See Methods, Figure S10) into 92 TF groups (49 single motifs, 24 same TF family, 19 contained TFs from different families and are referred to as "Complexes") (Table S10, Table S11). Approximately 31% of the 56,978 reference ATAC-seq peaks were predicted to be not bound by TFs.

Finally, we annotated 938 ATAC-seq peaks as associated with the formative state because they overlapped peaks specific to the FACs sorted hiPSC formative subpopulation[1] (Figure S7D, Table S8). We then annotated 2,981 peaks as associated with the primed state because they overlapped peaks specific to the FACs sorted hiPSC primed subpopulation[1].

Functional characterization of regulatory network modules

To further characterize the 13 RNMs we calculated their enrichments for formative and primed state ATAC-seq peaks (Figure 3A), the 5 collapsed chromatin states (Figure 3B), and the 92 TF groups including "Not Bound" peaks (Figure 3C; Figure S10, Table S12).

We initially examined the three RNMs (2, 3, and 8) (Figure 3A) enriched for the formative-specific peaks. RNM 2 was enriched for pluripotency TFs (NANOG-OCT4 complex, POU2F2/OCT2, SOX-LEF1 complex, TEAD Family) in enhancers, suggesting that it represents the hallmark hiPSC pluripotency regulatory network active in the formative state (Figure 3A-C). RNM 3 was highly enriched for enhancers and the TEAD Family (TEAD1 and TEAD4) (Figure 3A-C). TEAD signaling is strongly implicated in the differentiation of hiPSC to the trophoblast lineage[41]; and recently Dattani et al.[42] showed that suppression of YAP/TEAD signaling was critical to the insulation of naive hiPSC from trophoblast differentiation. Though the capacity for conventional hiPSC to undergo trophoblast differentiation has been the subject of considerable controversy[41,43], it is certainly possible that formative state cells, closer to the naïve-state, might be capable of entering this differentiation pathway, and that regulatory elements in RMN 3 are in some way primed for activation in naïve and formative state cells [44]. RNM 8 was strongly enriched for G_1 cell cycle-associated TFs (E2F Family, E2F2, E2F5, SP-E2F Complex) which is consistent with the fact that the formative state has a highly proliferative phenotype and an abbreviated G_1 phase[1,45] (Figure 3A-C). This suggests that RNM 8 represents a network that may underlie cell cycle regulation in formative pluripotency.

The three primed RNMs (6, 7, and 10) exhibited similar enrichments for E2F cell cycle-associated TFBSs (Figure 3A-C), which is consistent with the primed state being in the G_1 phase[1]. Interestingly, the primed associated RNMs displayed distinct chromatin state enrichments (Figure 3A-B). RNMs 6 and 7 were strongly enriched with active promoters and actively transcribed regions, respectively, (Figure 3B), further supporting that the primed state is more transcriptionally active than the formative state[1]. RNM 10 is enriched with both bivalent chromatin and repressed polycomb complexes (Figure 3B) and likely captures a primed-specific regulatory mechanism for the repression of developmental genes. These observations suggest that repressed polycomb complexes are likely activated at the transition from the formative to the primed state. Overall, these analyses show that the RNMs encompass well-known epigenomic and cell state-specific features present in formative to primed states and importantly captured the coordinated activity of the majority of regulatory elements underpinning the spectrum of pluripotency traits in hiPSCs.

Seven RNMs (1, 4, 5, 9, 11, 12, and 13) were not enriched with formative or primed-associated peaks. While RNM 1 is not cell state associated, its sole enrichment is with NANOG-OCT4 and OTX2 TFs (Figure 3C). The fact that the loss of NANOG-OCT4 mediated self-renewal initiates neuronal differentiation[46,47], and that REST (RE1-Silencing Transcription factor), which is involved in the repression of neural genes in non-neuronal cells, was depleted[48], suggests RNM 1 could represent a regulatory network underlying neural specification. RNM 9 has similar chromatin state and TF enrichments as the canonical pluripotency RNM 2 but is not enriched with the formative or primed-associated peaks (Figure 3A-C), suggesting that there may be independent NANOG-OCT4-mediated regulatory networks in different pluripotent states.

To further demonstrate the utility of annotating the ATAC-seq peaks in the RNMs (pluripotency state, chromatin state, and predicted TFBSs), we plotted a 2 MB interval on chromosome 4 surrounding *SMAD1*, which is a transcription factor involved in early cell fate decisions during specification of trophoblast and amnion, and later, gastrulation[41,49,50] via regulation of bone morphogenic proteins (BMPs). We observed nine formative-state associated peaks and multiple peaks binding SMAD Family (SMAD2 and SMAD4) and SOX-LEF1 Complex (SOX2, SOX3, SOX4, LEF1) TFs (Figure 3D). Focusing on the ~15kb interval harboring the *SMAD1* promoter, we observed an RNM 8 ATAC-seq peak overlapping a formative-state associated peak, bound by a SOX-LEF1 Complex TF (Figure 3E). RNM 8 was enriched with formative-state peaks and bivalent chromatin (Figure 3A-B), suggesting that the regulatory elements in this network are not only associated with G_1 cell cycle-associated TFs (see above) but also repressing developmental processes and differentiation. Altogether these observations demonstrate the utility of annotating RNM ATAC-seq peaks with epigenomic and cell state-specific features.

In summary, these findings present novel predictions for hiPSC regulatory networks and biology. Foremost we demonstrate that it is possible to capture the coordinated activity of the majority of the regulatory elements in hiPSCs and bin these elements into discrete functionally characterized regulatory networks that most likely mechanistically underlie self-renewal, pluripotency, and cell state transitions.

hiPSC regulatory networks conserved in fetal cell types

After establishing the architecture of regulatory networks in hiPSCs, we sought to examine if the co-expression of genes in other cell types could use the same regulatory networks (i.e., the same ATAC-seq peaks would show co-accessibility) or if new regulatory networks would be established. To determine if the hiPSC regulatory networks were conserved in other cell types we examined 54 sets of fetal cell type-specific ATAC-seq peaks from the Descartes Human Chromatin Accessibility Atlas[51]. Of the 47,761 hiPSC reference ATAC-seq peaks in the 13 major RNMs we identified 11,830 that overlapped cell type-specific peaks for at least one of the 54 fetal cell types (Figure 4A-B, Table S13). For each RNM, we then examined if the overlapping peaks were enriched in one or more of the 54 fetal cell types (Figure 4C).

Based on the RNM functional annotations (Figure 3), we hypothesized that RNM 3, and RNM 1 could represent regulatory networks respectively underlying the early emergence of fate commitment to the placental trophoblast and neural lineages, respectively. We observed that RNM 3 was enriched for cell type-specific peaks in villous cytotrophoblasts, extravillous trophoblasts, trophoblast giant, and eight other cell types (Figure 4C) consistent with its enrichment in TEAD elements and their role in trophoblast differentiation[63]. On the other hand, RNM 1 was enriched in multiple neural cell types including astrocytes, granule neurons, limbic system neurons, ENS glia, and ENS neurons, consistent with observations above on its likely occupancy by NANOG-OCT4, factors that probably play a role in suppressing regulatory elements important in the differentiated neuronal state[52]. These findings suggest that some of the RNMs present in the hiPSCs could represent networks important for lineage specification and are conserved in the derived cell types.

We reasoned that one possible mechanism underlying RNM conservation during development would involve binding of the same TF group to the same regulatory elements in both hiPSCs and the derived fetal cell type. We directly calculated the enrichment of the 92 hiPSC TF groups in the 54 fetal cell-type specific ATAC-seq peaks (Figure 4D, Table S14). TEAD Family hiPSC TFBSs exhibited a strong enrichment in the three trophoblast cell types cells (Figure 4D), recapitulating their importance in the development of placental lineages (Figure 4C)[42,53]. Trophoblast giant cells and villous cytotrophoblasts were also enriched with GRHL2, which has been reported to be a master regulator of placental branching morphogenesis[54]. NANOG-OCT4 Complex and POU2F2 were strongly enriched in astrocyte-specific peaks, which further supports the NANOG-mediated repression of neural tissue differentiation during pluripotency[52].

Altogether, these results suggest that some hiPSC regulatory networks are conserved in fetal cell types and that the molecular mechanism underlying conservation can either be the binding of the same TF Family to the same regulatory elements in both stages of development or NANOG-mediated repression of regulatory elements active in early tissue differentiation.

Characterization of allele-specific chromatin accessibility SNPs (ASCA-SNPs)

Previous studies have shown that both genetic and epigenetic variation affect hiPSC phenotypes[55–57], however, the underlying mechanisms are poorly understood. One likely mechanism through which genetic variation influences hiPSC phenotypes[55–57] is by affecting chromatin accessibility. We examined 105,055 SNPs (MAF ≥ 0.05, HWE p-value $> 1x10^{-6}$) present in 35,614 ATAC-seq peaks in the 13 major RNMs and determined that 6,323 displayed allele-specific chromatin accessibility (ASCA, Table S15). To determine if specific RNMs harbored ASCA SNPs with large effects, we performed Mann-Whitney U tests on the allelic imbalance fraction (AIF) of ASCA SNPs in each of the 13 RNMs using the other 12 RNMs as background (Figure 5A). We observed that four RNMs 1, 2, 3, and 9 (all enriched for enhancers Figure 3B), contained ASCA SNPs with a greater allelic imbalance fraction (AIF) than the ASCA SNPs in the other 9 RNMs (Figure 5A). Of note, RNMs 2 and 3 were associated with the formative state (Figure 3A) and represent the NANOG-OCT4 and TEAD-mediated (Figure 3C) regulatory networks, respectively. While RNMs 1 and 9 were depleted for formative state regulatory elements, they represented NANOG-OCT4-mediated regulatory networks most likely in different pluripotent states. Our findings show that compared with the other regulatory networks active in hiPSCs, the regulatory elements in the pluripotency-associated networks harbor genetic variants that exert large effects on chromatin accessibility.

Genetic variation can influence chromatin accessibility by affecting the binding specificity of TF motifs in regulatory elements[58–61]. We examined if any specific group of TF groups were enriched for having ASCA SNPs overlapping their binding sites. Of the 6,323 ASCA SNPs, 1,933 (30.6%) overlapped at least one TFBS, while the majority did not overlap any TFBSs (n = 4,299, Figure 5B). We calculated the enrichment of the ASCA SNPs in TFBSs by performing a Fisher's Exact test for all 92 TF groups (Figure 5C). Regulatory elements predicted to bind 23 TF groups were enriched with ASCA SNPs (adjusted P-value < 0.05 and Odds Ratio > 1). These 23 groups included many pluripotency-associated TFs, such as NANOG-OCT4, TEAD Family, POU2F2, and SOX-LEF1 Complex (Figure 5C). We examined the allelic imbalance fraction (AIF) of the ASCA SNPs in each of the 23 TF groups and observed that NANOG-OCT4 contained ASCA SNPs with significantly higher AIF indicating that they had large effects on chromatin accessibility (Figure 5D). Six other TF groups (AR-FOXJ3, TEAD Family, CTCF Family, SALL4-ZIC1 Complex, and SOX-LEF1 Complex) exhibited increased AIF, albeit not significantly.

Taken together, these results show that pluripotency-associated TF binding sites and regulatory networks are enriched with genetic variants that have large effects on chromatin accessibility. While previous studies have shown that pluripotency TFs have a remarkably high degree of evolutionary constraint[62], our findings show that the regulatory elements to which they bind have a high amount of variability suggesting that they could play an important role in the observed pluripotent state differences between hiPSCs.

DISCUSSION

We sought to determine if regulatory network modules in hiPSCs underlying self-renewal and pluripotency could be discovered by examining genome-wide chromatin co-accessibility across samples from hundreds of iPSCORE individuals. We generated bulk ATAC-seq samples and employed cellular deconvolution to estimate the proportion of cells in the formative and primed states across 143 hiPSCs, which showed that the lines were composed of varying proportions of cells in these two pluripotency states. Using the bulk ATAC-seq data we calculated accessibility for each hiPSC at 56,978 reference peaks, and then using data from the 143 lines we determined genome-wide co-accessibility between all pairwise combinations of the reference peaks. Likewise, for 213 RNA-seq samples, we calculated co-expression between 16,110 genes. We applied an unsupervised community detection algorithm independently to each dataset which detected 13 regulatory network modules (RNMs) and 13 gene network modules (GNMs). We demonstrated that the RNMs and GNMs were strongly correlated with each other, suggesting that the coordinated co-accessibility of regulatory elements in the RNMs most likely underlie the coordinated expression of genes in the GNMs.

To functionally characterize the RNMs we annotated the individual ATAC-seq peaks with chromatin states, transcription factor (TF) binding sites[33–35], and for association with the formative and primed pluripotency subpopulations. We showed that regulatory elements in each of the 13 RNMs tended to share similar predicted TF binding, chromatin state profiles, and pluripotency state enrichments. Additional analyses suggest that some of the hiPSC regulatory networks are shared with derived fetal cell types (i.e., the same ATAC-seq peaks show co-accessibility).

Our findings highlight certain understudied regulatory networks that may be critical to hiPSC pluripotent state transitions. The formative cell state is enriched for self-renewal processes[1], compared to primed cells which can exhibit emergent features of lineage commitment[21]. Historically, epigenomic regulation in pluripotent cells has been characterized by a single regulatory network primarily mediated by NANOG, OCT4, and SOX2 binding (RNM 2)[22]. Our study suggests that there are several regulatory networks with distinct epigenetic profiles (chromatin state, TF binding), that mediate biological processes that are indispensable for maintaining the formative and primed pluripotency states (Figure 3). For example, a TEAD4-mediated regulatory network (RNM 3, Figure 3A-C) is present in the early postimplantation epiblast-like formative state and absent in the late postimplantation epiblast-like primed state. We demonstrate that trophoblast-specific peaks are enriched with RNM 3 peaks and TEAD Family binding sites (Figure 4C-D), which is consistent with previous observations that TEAD4 is a pioneering factor for the placental lineage commitment[63]. Recently much effort has been put into developing efficient trophoblast differentiation protocols to study common pregnancy-related diseases, such as preeclampsia[41,43,44,53]. RNM 3 may serve as a resource to identify regulatory elements that are early determinants of trophoblast cell fate commitment. Bivalent chromatin and repressed polycomb complexes are well-known pluripotency-associated epigenomic features that repress the expression of genes involved in differentiation[14,38,64]. We show that the primed-associated RNM 10 is enriched in both polycomb repressed regions and bivalent chromatin, while the formative-associated RNM 8 is primarily enriched with bivalent chromatin, indicating that the two pluripotency states have discrete regulatory networks involving different epigenomic mechanisms for the repression of developmental genes (Figure 3A-B). These analyses revealed regulation in hiPSCs is more complex than the canonical NANOG, OCT4, and SOX2 regulatory network and that leveraging large sample sizes can resolve independent functional distinct pluripotency networks.

Our study also shows that RNMs are differentially enriched with 49 fetal cell type-specific peaks. These results indicate that distinct types of early embryonic regulatory elements are reused in later stages of fetal development. It also indicates that lineage-specific regulatory elements are regulated by distinct mechanisms during pluripotency.

Our study also addresses a dearth of information about the role of genetic regulatory variation in stem cells because there are only a handful of labs actively investigating the topic. In addition to heritable, complex diseases that have known developmental origins (i.e. autism spectrum disorder, schizophrenia), there is mounting evidence that common diseases, such as type 2 diabetes[65] and cardiovascular disease[18], are influenced by regulatory variation that is active during fetal development. It is paramount to expand exploration into these areas of research to characterize developmental regulatory variation. In this study, we identify thousands of SNPs with allele-specific chromatin accessibility (ASCA). Despite the importance of maintaining pluripotency, we observed that the regulatory elements in the pluripotency-associated networks harbor genetic variants that exert large effects on chromatin accessibility. We also observed that compared with the other TF groups the binding sites for the NANOG-OCT4 complex were enriched for ASCA SNPs. Our findings suggest that variability in the regulatory elements in the pluripotency networks could play an important role in the observed varying proportions of pluripotency states between hiPSCs. These observations have remarkable implications for evolution and speciation.

In summary, our study suggests that epigenomic regulation of pluripotency and self-renewal processes are more complex than previously thought. It classifies hiPSC ATAC-seq peaks into 13 networks, 6 of which are associated with the formative or primed cell states, and 7 of which are likely associated with different pluripotent states. It also proposes potential mechanisms for how genetic variation influences TF binding and pluripotency regulatory mechanisms. These network classifications and ASCA SNPs could be a useful resource for researchers investigating, various aspects of pluripotency and development, such as; 1) epigenomic regulation in stem cells, 2) cell fate commitment, 3) fetal developmental processes, 4) embryonic-specific regulatory variation, 5) influence of regulatory variation on regulatory element co-accessibility, and 6) the development iPSC tissue derivation protocols.

METHODS

Subject Information

We used hiPSC lines from 219 individuals (Table S1) recruited as part of the iPSCORE project[19,23,66]. There were 140 individuals belonging to 40 families composed of two or more subjects (range: 2–14 subjects) and 134 genetically unrelated individuals (some individuals were in the same family but only related by marriage). The iPSCORE_ID (i.e, iPSCORE_4_1) indicates family (4) and individual number (1). Each subject was assigned a Universal Unique Identifier (UUID). Sex, age, and self-reported race/ethnicity were reported at the time of enrollment of each subject. We previously estimated the ancestry of each participant by comparing their genomes to those of individuals in the 1000 Genomes Project (KGP)[67]. Recruitment of these individuals was approved by the Institutional Review Boards of the University of California, San Diego, and The Salk Institute (project no. 110776ZF).

Molecular Data Sources

We used the following datasets from the iPSCORE resource:

- Previously published[19,23] 50X WGS (Illumina; 150-bp paired-end) generated from the blood or skin fibroblasts of the 219 individuals in this study. Of the 219 individuals, 127 had both RNA-seq and ATAC-seq libraries, 86 only had RNA-seq and 6 only had ATAC-seq.

- 150 ATAC-seq libraries generated from 143 hiPSC lines (collected after expanding in mTeSR1 medium with ROCK inhibitor on Matrigel) from 133 individuals, Table S6;

- Previously published[19,23,66] RNA-seq data from 213 hiPSC lines (collected after culturing in mTeSR1 medium on Matrigel) from 213 individuals, Table S2.

- Previously published[23] RNA-seq data from 3 hiPSC lines (collected after expanding in mTeSR1 medium containing ROCK inhibitor on Matrigel) from 3 individuals. Technical triplicates were collected for each hiPSC line.

Of note, both the RNA-seq and ATAC-seq data were generated from the same hiPSCs in the iPSCORE collection. However, the RNA-seq data and ATAC-seq data were generated from different passages of the hiPSC lines cultured under different experimental conditions. The RNA-seq data was generated from earlier passage ROCK inhibitor-naïve hiPSCs and the ATAC-seq data was generated from later passage hiPSCs that had been cultured with ROCK inhibitor. All iPSCORE resource molecular data is publicly available: RNA-seq at dbGaP (phs000924) and GEO; ATAC-seq at GEO (GSE_XXX); gVCF files at dbGaP (phs001325). The narrowPeak file for ATAC-seq is also available at GEO, while the TMM matrix and inverse normal transformed TMM matrix for ATAC-seq is available at GEO and at Figshare (XXX).

We used the following publicly available datasets:

- 28 RNA-seq samples corresponding to either the formative state (early postimplantation epiblast-like; EPE) or the general population[1].

- 6 ATAC-seq samples corresponding to either the formative state (GCTM-2^{high}CD9^{high}EPCAM^{high}) or the primed state (GCTM-2^{mid}-CD9^{mid})[1] .
- ChIP-seq data for 18 Transcription Factors from ENCODE[35,39,68] (Table S9);
- Seven hiPSC cell state gene sets obtained and curated from 3 published studies:
  - One gene set for 8-cell like cells[3]
  - Four gene sets for epiblast, conventional PSCs, naïve PSCs, primitive endoderm[29]
- Fifty-four sets of fetal tissue-specific ATAC-seq peaks from the Descartes; Human Chromatin Accessibility During Development Atlas[51]
- ChromHMM chromatin states in the ROADMAP Epigenomics hiPSC-18 line[31];

Human iPSC (hiPSC) generation

Generation of the 219 hiPSC lines has previously been described in detail[23]. Briefly, cultures of primary dermal fibroblast cells were generated from a punch biopsy tissue, expanded for approximately 3 passages and cryopreserved. In batch, the fibroblasts were thawed and plated at a density of 250K cells/well of 6-well plate and infected with the Cytotune Sendai virus (Life Technologies) per manufacturer's protocol to initiate reprogramming. The Sendai infected cells were maintained with 10% FBS/DMEM (Invitrogen) for Days 4-7 until the cells recovered and repopulated the well. These cells were then enzymatically dissociated using TrypLE (Life Technologies) and seeded onto a 10-cm dish pre-coated with mitotically inactive-mouse embryonic fibroblasts (MEFs) at a density of 500K/dish and maintained with hESC medium, as previously described[69]. Emerging hiPSC colonies were manually picked after Day 21 and maintained on Matrigel (BD Corning) with mTeSR1 medium (Stem Cell Technologies). Multiple independently established hiPSC clones (i.e. referred to as lines) were derived from each individual, which were cultured typically to passage 12, and at least ten stock vials were frozen from each cell line. hiPSC pellets collected from each cell line were frozen in RTL plus buffer (Qiagen) and used for total RNA isolation. Sendai virus clearance typically occurred at or before P9 and was not detected in the hiPSC lines at the P12 stage of cryopreservation. A subset of the hiPSC lines were evaluated by flow cytometry for expression of two pluripotent markers: Tra-1-81 (Alexa Fluor 488 anti-human, Biolegend) and SSEA-4 (PE anti-human, Biolegend). Pluripotency was also examined using PluriTest-RNAseq [70]. This iPSCORE resource was established as part of the Next Generation Consortium of the National Heart, Lung and Blood Institute and is available to researchers through the biorepository at WiCell Research Institute (www.wicell.org; NHLBI Next Gen Collection). For-profit organizations can contact the corresponding author directly to discuss line availability.

iPSCORE data generation and processing

***Generation of RNA-seq data for 213 hiPSC lines*:** We analyzed previously published[19] RNA-seq data from 213 hiPSC lines from 213 iPSCORE individuals (Table S2). Briefly, using the AllPrep DNA/RNA Mini Kit (QIAGEN) we extracted total RNA from lysed pellets frozen in RLT Plus buffer (collected after culturing hiPSC lines in mTeSR1 medium on

Matrigel), assessed quality to make sure RNA integrity number (RIN) was 7.5 or greater, prepared libraries using the Illumina TruSeq stranded mRNA kit and sequenced with 100bp paired end reads on HiSeq2500 (~22M reads/per sample).

Generation of RNA-seq for 3 hiPSC lines expanded with ROCK inhibitor: While the RNA-seq and ATAC-seq data were generated from same hiPSC lines, there were batch effects because of differences in culture conditions (ie, the ATAC-seq data were generated from hiPSCs that had been cultured with ROCK inhibitor). Therefore, the following data were used to examine the correlation between ATAC-seq peak co-accessibility network modules and gene co-expression network modules (see below: *Identifying Associations Between Gene and Regulatory Networks*). As previously described[70], for three individuals (iPSCORE_2_2, iPSCORE_2_3, and iPSCORE_2_9) one vial of a frozen hiPSC line was thawed into mTeSR1 medium containing 10 μM Y27632 (ROCK Inhibitor), plated onto one well of a Matrigel-coated six-well plate and incubated overnight. The media was replaced daily with mTeSR1 until the hiPSCs were visually estimated to be 80% confluent. The hiPSCs were then expanded by passaging from one well onto three wells of a six-well plate using Versene (Lonza) in mTeSR1 medium containing 5 μM ROCK inhibitor. When cells reached 80% confluency, the three wells of hiPSC line were individually dissociated using Accutase (Innovative Cell Technologies Inc.) in mTeSR1 medium containing 5 μM ROCK inhibitor, collected, counted, and 1x10^6 cells frozen in RLT plus buffer (three technical replicates per hiPSC line). Total RNA was extracted using AllPrep DNA/RNA Mini Kit (QIAGEN), RNA quality was assessed based on RNA integrity number (RIN) and nine libraries (three per hiPSC line) were prepared using the Illumina TruSeq stranded mRNA kit and sequenced with 100-bp paired end reads on a HiSeq2500.

RNA-seq data processing: We obtained transcript per million bp (TPM) as previously described[19]. Briefly, FASTQ files were aligned to the hg19 reference genome using STAR 2.5.0a[71] and Gencode V.34lift37[72] with parameters: *outFilterMultimapNmax 20*, *–outFilterMismatchNmax 999*, *–alignIntronMin 20*, *–alignIntronMax 1000000*, *–alignMates-GapMax 1000000.* We sorted and indexed the BAM files using Sambamba 0.6.7[73] and marked duplicates using biobambam2 (2.0.95) bammarkduplicates[74] and then re-indexed. To quantify gene expression (TPM), we used RSEM v1.2.20[75] with the following parameters: *--rsem-calculate-expression*, *--bam*, *--num-threads 16, --no-bam-output, --seed 3272015*, *--estimate-rspd*, *--paired-end*, *--forward-prob 0*. We identified 16,110 expressed genes on autosomes (TPM ≥ 1 in at least 20% of the 213 hiPSC lines) using rowQuantiles from the *matrixStats R package* (version 0.52.2).

RNA-seq gene expression transformation: To normalize gene expression and usage across samples, we performed inverse normal transformation using normalize.quantiles (preprocessCore package) and qnorm functions in R, in order to obtain mean expression = 0 and standard deviation = 1, as we previously described[18,76].

Generation of ATAC-seq data for 143 hiPSC lines: 150 ATAC-seq libraries were prepared from 143 hiPSC lines (for 7 lines we prepared two libraries) from 133 individuals (5 individuals each had two or three independent clones). The hiPSCs were harvested on Day 0 of our previously published large-scale iPSC-derived cardiovascular progenitor cells study prior to the initiation of WNT activation[77]. In brief, for each of the 150 ATAC-seq libraries, one vial of a frozen hiPSC line was thawed into mTeSR1 medium containing 10 μM ROCK Inhibitor (Sigma), plated on one well of a six-well plate coated with Matrigel, and incubated overnight. When hiPSCs were visually estimated to be 80% confluent they were passaged 1-2 times using Dispase II (2mg/ml; Gibco/Life technologies) in mTeSR1 and plated onto Matrigel coated plates.

The hiPSCs were then expanded by passaging using Versene (Lonza) in mTeSR1 medium containing 5 µM ROCK inhibitor. Finally, the hiPSCs were dissociated using Accutase (Innovative Cell Technologies Inc.) in mTeSR1 medium containing 5 µM ROCK inhibitor, collected, counted, and frozen as nuclear pellets of $2.5x10^4$ cells for the ATAC-seq assay.

We performed ATAC-seq on the 150 hiPSC samples using a modified version of the Buenrostro et al.[78] protocol to: 1) selectively permeabilize the nuclear envelope of the cells without disrupting mitochondrial membrane, and 2) increase the number of reads that do not contain sequences spanning multiple nucleosomes. Frozen nuclear pellets were thawed on ice and tagmented in total volume of 25µl in permeabilization buffer containing digitonin (10mM Tris-HCl pH 7.5, 10mM NaCl, 3mM MgCl2, 0.01% 0.01% digitonin) and 2.5µl of Tn5 from Nextera DNA Library Preparation Kit (Illumina) for 45-75min at 37°C in a thermomixer (500 RPM shaking). Inclusion of digitonin, which is a mild detergent capable of selective permeabilization of cholesterol-rich bilayers[79], permeabilized the nuclear membranes (containing 20-35% cholesterol and cholesterol esters[79,80] without disrupting the mitochondrial membranes (0.5-3% cholesterol)[81]. To enrich for reads that span a single nucleosome, we included a double size selection step during purification using AMPure XP DNA beads (Beckman Coulter). This step enriched for reads containing inserts under 140 bp in length (shorter than the 146 bp wrapper around a single nucleosome). To eliminate confounding effects due to index hopping, all libraries within a pool were indexed with unique i7 and i5 barcodes. Libraries were amplified for 12 cycles using NEBNext® High-Fidelity 2X PCR Master Mix (NEB) in total volume of 25µl in the presence of 800nM of barcoded primers (400nM each) custom synthesized by Integrated DNA Technologies (IDT).

***Sequencing of ATAC-seq*:** The 150 ATAC-seq libraries were batched and sequenced with 100-bp paired-end reads on a HiSeq4000. To improve overall read depth 113 ATAC-seq libraries were sequenced twice (the second time with 150-bp paired end reads) resulting in 263 FASTQ files (Table S6). The 263 FASTQ files were aligned to the hg19 reference genome with STAR[71] with the following flags: *--outFilterMultimapNmax 20*, *--outFilterMismatchNmax 999*, *--outFilterMismatchNoverLmax 0.04*, *--seedSearchStartLmax 20*, *--outFilterScoreMinOverLread 0.1*, *outFilterMatchNminOverLread 0.1*. We then filled in mate coordinates using samtools fixmate, marked duplicates using samtools markdup[82]. We used samtools merge to combine BAM files from the same library and indexed the merged BAM files with samtools index. We then filtered poorly mapped reads (MAPQ < 20%), duplicates and reads less than 38bp and greater than 2000bp with samtools view to obtain reads passing filters (PF). We re-indexed the filtered merged BAM file with samtools index (version v6.7)[82]. After quality control, all BAM files from the same library were combined (1-2 paired files per hiPSC line) resulting in 150 BAM files (Table S6).

ATAC-seq Peak Calling and Quality Control: MACS2 v.2.2.7[83] was used to call broad peaks for all 150 ATAC-seq libraries individually with settings: *--nomodel --shift -100 --extsize 200 -f BAMPE -g hs --broad*. To obtain one single set of high-quality peaks, 34 libraries were selected from unrelated individuals with 20-35 million reads passing filters, 60,000-90,000 broad peaks, and 100-225 bp mean fragment size to establish discrete regions of accessible chromatin (i.e., a reference set of ATAC-seq peaks) (Figure S6B, Table S6). MACS2 v.2.2.7 was used to call narrow peaks on the 34 reference libraries jointly with settings: *-f BAMPE -g hs -t --nomodel --shift -100 --extsize 200 –narrow*. Narrow peaks were filtered by MACS2 score (< 100), resulting in 136,333 peaks (including 132,225 autosomal peaks and 4,108 peaks on sex chromosomes). A NarrowPeak file with the coordinates of all 136,333 peaks was uploaded to GEO (GSE_XXX) and Figshare (XXX). For each of the 150 individual ATAC-seq samples coverage on the 136,333 peaks were obtained using *featureCounts* in Subread package v1.5.0[84] . Next, counts were trimmed mean of M value (TMM)-normalized for each peak across all 150 individual samples using the *cpm* function from edgeR v3.30.3[85].

ATAC-seq peak transformation: TMM values for all 136,333 ATAC-seq peaks were inverse normal transformed using the *normalize.quantiles* from the *preprocessCore* package and *qnorm* functions in R, in order to obtain mean expression = 0 and standard deviation = 1 for each peak across all 150 ATAC-seq samples. TMM and inverse normal transformed TMM matrices for the 136,333 peaks were uploaded to GEO (GSE_XXX) and Figshare (XXX).

Identification of 56,978 autosomal reference peaks*:** To reduce the computational burden in downstream analyses, we selected a subset of 56,978 reference peaks based on their MACS2 peak score and enrichment in active chromatin. We filtered ATAC-seq peaks that overlapped blacklisted regions[86], peaks in ZNF genes & repeats, heterochromatin, quiescent chromatin states from the ChromHMM model for iPSC-18[87], and peaks on sex chromosomes. We binned the remaining 87,659 ATAC-seq peaks into 20 MACS2 score quantiles each containing 4,383 peaks, where quantile 20 consisted of the peaks with the highest scores. We then examined the enrichment of the 87,659 ATAC-seq peaks in iPSC-18 ChromHMM chromatin states by MACS2 score quantile (Figure S6C). We created 20 bed files containing the coordinates of the peaks in each quantile and calculated their enrichment (Odds Ratio) in each of the 12 chromatin states for iPSC-18, using bedtools fisher. The higher quantiles tended to be enriched for active (TssA, TssAFlnk, TxFlnk), bivalent (TssBiv, BivFlnk, EnhBiv), and repressed polycomb (ReprPC, ReprPCWk) chromatin. Based on these observations, we calculated pairwise co-accessibility for the 56,978 ATAC-seq peaks in MACS2 quantiles 8-20 (see *Identifying co-accessible peaks* section below), which alleviated the computational burden by eliminating $7x10^9$ pairwise tests between peaks with low accessibility and peaks in inactive regions of the genome. For downstream analyses, we resolved a single chromatin state for each peak by intersecting the maximum summits with the iPS-18 chromatin states (described below in ***Collapsing into 5 hiPSC chromatin states) (Table S8).

Cellular deconvolution of pluripotency states using gene expression signatures

From GEO (GSE119324)[1] we downloaded 28 paired RNA-seq FASTQ files, including, fourteen from five unique human embryonic stem cells (hESCs) sorted (GCTM-2^{high}CD9^{high}EPCAM^{high}) for the formative (EPE) population and fourteen files for five unique samples generated from the primed population (unsorted). We aligned and filtered the 28 FASTQ files, with STAR 2.5.0a[71] and RSEM v1.2.20[75] as described above in ***RNA-seq data processing*** section. We performed differential expression analysis between the formative (EPE) and general population RNA-seq samples using DESeq2 (v.1.34) *R* package[88] and created a signature matrix containing 100 of the most differentially expressed genes (50 upregulated in the GCTM-2^{high}CD9^{high}EPCAM^{high} cells and 50 downregulated relative to the unfractionated cells; Table S3). We deconvoluted the 213 hiPSC RNA-seq samples by supplying the signature matrix and TPM expression matrix as input to the *CIBERSORT.R* script[20].

Covariates for co-expression and co-accessibility analyses

To perform the gene co-expression analysis, we used the following covariates: 1) sex; 2) age; 3) hiPSC passage number; 4) number of properly paired reads; 5) 20 genotype principal components to account for global ancestry; and 6) kinship matrix. All covariates are available in Tables S1-2.

To perform the ATAC-seq peak co-accessibility analysis, we used the following covariates: 1) sex; 2) age; 3) hiPSC passage number; 4) number of reads passing filters; 5) the mean fragment size; 6) the number of broad peaks; 7) the ratio of 100bp reads to 150bp reads; 8) 20 genotype principal components to account for global ancestry; and 9) kinship matrix. All covariates are available in Tables S1 and S6.

Sex: To account for sex-dependent chromatin, we assigned binary values to each sex and included it as a covariate.

Biological and Technical Covariates: We scaled age, hiPSC passage number, number of reads passing filters, and mean fragment size for each library to the mean across libraries and included these normalized values as covariates.

Broad peaks: As described above in the *ATAC-seq Peak Calling and Quality Control section*, we used MACS2 v.2.2.7 to call peaks on all 150 ATAC-seq libraries independently to assess sample quality. We normalized the number of broad ATAC-seq peaks for each sample to mean across all 150 samples.

Read length ratio: As described above in the *Sequencing of ATAC-seq section*, several ATAC-seq libraries were sequenced multiple times with 100 or 150bp paired end reads and merged. To account for the different proportions of read lengths in merged libraries, we included the ratio of reads passing filters from 100 and 150 bp sequencing runs for each library as a covariate.

Genotype Principal Component Analysis: We previously performed principal component analysis (PCA) on WGS variants to determine the global ancestry of each individual in this study[18]. Briefly, we used the genotypes of 1,634,010 SNPs that had allele frequencies between 30% and 60% in the 1000 Genomes Phase 3 Project and genotyped in both iPSCORE and GTEx. We merged the VCF files from 1000 Genomes, iPSCORE, and GTEx, and performed PCA using the *pca* function in *plink 1.90b3x*[89]. The top 20 genotype principal components used as covariates to account for global ancestry for all 219 individuals in this study can be found in Table S1.

Kinship Matrix*:* We included a kinship matrix generated for a previous iPSCORE study[18] as a random effects term to account for the genetic relatedness between individuals. The matrix was constructed using the kinship function in plink 1.90b3x[89] and the same set of 1,634,010 SNPs employed in the genotype PCA. The kinship matrix is available in GEO (GSE_XXX).

Gene Co-expression Analyses

We leveraged the previously published RNA-seq dataset of 213 hiPSC lines from iPSCORE individuals[19,66,70] to identify gene networks that are active in stem cells. We integrated and curated gene sets of defined pluripotency cell states from external sources[1,3,29,70] to annotate gene networks with stem cell states.

Identification and Characterization of Gene Co-expression: To identify pairwise correlations between the 16,110 expressed autosomal genes across the 213 hiPSC lines, we performed a gene co-expression analysis using a Linear Mixed-effects Model (LMM) with the *lmekin* function in the *coxme* R package v.2.2-17[24], which incorporates a kinship matrix to control for random effects from genetic relatedness. We included normalized values for age, hiPSC passage number, sex, number of properly paired reads, and the top 20 PCs from ancestry as fixed effect covariates (see above: *Covariates for co-expression and co-accessibility analyses*).

Formula #1:

$$Y_{ij} = \beta_k Y_{ik} + \sum_{m=1}^{M} \gamma_m PC_{im} + \sum_{p=1}^{P} \gamma_p C_{ip} + u_i + \epsilon_{ik}$$

Where Y_{ij} is the normalized expression value for gene *j* in sample *i*, Y_{ik} is the normalized expression value for gene *k* in sample *i*, β_k is the effect size (fixed effect) of gene *k*, PC_{im} is the value of the *m*^th genotype principal component for the individual associated with sample *i*, γ_m is the effect size of the *m*^th genotype principal component, *M* is the number of genotype principal components used (M =20), C_{ip} is the covariate of the *p*^th covariate for sample *i*, γ_p the effect size of the *p*^th covariate, *P* is the number of covariates used (P=5), u_i is a vector of random effects for the individual associated with sample *i* defined from the kinship matrix, and ϵ_{ik} is the error term for individual *i* at gene *k*.

Construction of the Genome-wide Gene Co-expression Network (GN): We created the genome-wide co-expression network (GN) using the *graph_from_data_frame* function from the *igraph* R package[26] by assigning the 16,110 expressed genes as nodes, the 3,533,609 co-expressed genes (Bonferroni-corrected p-value < 0.05 and β > 0) as edges. We extracted the genome-wide degree of each gene using the *degree* function on the GN.

To identify gene co-expression network modules (GNMs), we applied the unsupervised Leiden community detection algorithm (from the *igraph* R package[26]) to the GN. We optimized module detection by analyzing 1,700 combinations of three parameters: *resolution (range: 0-5), beta (range: 0-0.1), and n_iterations (range 5-50).* For each combination of parameters, we permuted the nodes to confirm the Leiden algorithm was clustering co-expressed genes better than NULL background. We calculated modularity, fractions of genes in major GNMs (membership > 100) and the number of modules for each combination of parameters. We selected the modules obtained by resolution =2.25, beta = 0.05, n_iterations = 45, which exhibited a decent modularity (Q=0.41) and a high fraction of genes captured by major GNMs (91.8%). Under these parameters, genes within the same GNM were significantly more co-expressed than genes randomly connected through permutation (P-value = 0). As there is no consensus on modularity thresholds or network validation methods[90], we used downstream analyses, such as co-expression and gene set enrichment for validation. For each gene, we calculated its intramodular degree (the number of co-expressed genes within the module), using the same functions described above, for each GNM independently, excluding inter-module edges (Table S4).

Gene module identification using weighted correlation network analysis (WGCNA): To benchmark our gene module detection approach (Supplemental Note 1), we processed the 213 RNA-seq samples using the standard workflow (https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/) for the WGCNA R package[6], which cannot account for covariates and kinship. We used WGCNA to calculate the associations with the 24 covariates used in the co-accessibility LMM, the estimated formative fraction (Figure S2A), and binary annotations for samples from 6 families with 5 or more individuals (Figure S3). We then compared our 13 GNMs with the 17 WGCNA modules by calculating the enrichment of the 7 cell state gene sets (Figure 1J, Figure S3).

***PCA and UMAP Analyses of GNMs**:* To assess whether genes within a GNM had similar expression profiles across hiPSCs, we performed PCA and UMAP analyses on the 213 RNA-seq samples. We first identified the most interconnected genes within each GNM by ranking the intramodular degree for each GNM independently. Biological networks are scale-free[27] and follow the Pareto Principle[28], which states that 20% of the nodes are responsible for 80% of a network's connectivity, therefore we defined intramodular Pareto genes as the top 20% interconnected genes within each GNM. We performed a PC analysis on the inverse normal transformed TPM expression of the intramodular Pareto genes in the 13 major GNMs (membership > 100 genes) with the *prcomp* function in base R. We performed UMAP dimensionality reduction using the *umap* function from the *umap* R package.

Calculation of GNM Score: For each of the 213 RNA-seq samples we calculated a GNM score for each of the 13 GNMs by summing the inverse normal transformed TPM expression of the corresponding GNM-specific Pareto genes (i.e., each RNA-seq sample had 13 GNM scores).

Functional Characterization: Gene Co-expression Network Modules

Calculation of GNM Co-expression Enrichment: To assess whether genes within each GNM were more co-expressed with each other across the 213 hiPSC lines than with genes in different GNMs, we enumerated the number of co-expressed genes shared between all pairwise combinations of GNMs. We then performed a Fisher's Exact test to calculate the enrichment of genes showing co-expression within each GNM, using the genes co-expressed between each set of paired GNMs as background.

GNM Enrichment in Stem Cell States**:** We determined if the GNMs were enriched for expressing marker genes from three published studies describing 7 hiPSC subpopulations representing different pluripotency cell states including: (1) 199 8-cell like cell (8CLC)-associated genes from Mazid et al[3]; (2) 123 primitive endoderm (PrE)-associated genes, 248 epiblast-associated genes, 175 trophectoderm genes ; 96 Naïve-associated PSCs genes from Stirparo et al.[29]; and (3) 266 EPE-associated genes and 452 general population-associated genes that we identified by reprocessing data from Lau et al.[1] (See **Cellular deconvolution of pluripotency states using gene expression signatures**). We performed a Fisher's Exact test to calculate the enrichment of hiPSC subpopulation or cell state associated genes in each GNM, using the genes in the remaining 12 major GNMs as background.

Cellular deconvolution of pluripotency states using ATAC-seq peak accessibility signatures

We downloaded 6 ATAC-seq samples that were performed on human embryonic stem cells (hESCs) that had been sorted for the formative (GCTM-2^{high}CD9^{high}EPCAM^{high}) and the primed (GCTM-2^{mid}-CD9^{mid}) cell states from GEO

(GSE147338)[1]. We aligned and filtered the FASTQ files, with STAR 2.5.0a[71] as described in ***Sequencing of ATAC-seq*** section. MACS2 v.2.2.7 was used to establish a reference set of narrow peaks on the 6 ATAC-seq samples simultaneously, using the parameters; *-f BAMPE -g hs -t --nomodel --shift -100 --extsize 200 –narrow.* For each of the 6 individual ATAC-seq samples, read counts in the 193,147 reference peaks were obtained using *featureCounts* in Subread package v1.5.0[84] . We performed differential accessibility analysis on the counts using DESeq2 (v.1.34) *R* package[88]. We used *bedtools intersect -r 0.25* to identify 938 formative-associated peaks that overlapped iPSCORE ATAC-seq peaks with a 25% reciprocal overlap (Table S8). We used the ATAC-seq peaks that were differentially accessible in the GCTM-2^{mid}-CD9^{mid} hiPSC subpopulation[1] to annotate 2,981 non-formative associated peaks using the same approach. We then deconvoluted the 150 ATAC-seq samples using the *CIBERSORT.R* script[20] with a signature matrix containing 200 of the most differentially accessible formative and primed peaks (Table S7).

ATAC-seq Co-Accessibility Analyses

We leveraged our newly generated 150 ATAC-seq samples to profile co-accessibility of open chromatin across the hiPSC epigenome and discovery regulatory networks that are active in different pluripotency cell states.

Identifying co-accessible peaks: Since pairwise calculations for all 132,225 autosomal ATAC-seq peaks would have been computationally intensive, requiring ~8.74 x 10^9 tests, we focused our analyses on the 56,978 reference peaks (see above: *Identification of 56,978 autosomal reference peaks*) which reduced the number of tests to 1.62 x 10^9. To identify pairwise correlations of accessibility between the 56,978 peaks across the 150 ATAC-seq samples, we performed a genome-wide analysis using a Linear Mixed-effects Model (LMM) with the *lmekin* function in the *coxme* R package v.2.2-17[24], which incorporates a kinship matrix to control for random effects from genetic relatedness. The following covariates were included: 1) sex; 2) age; 3) hiPSC passage number; 4) number of reads passing filters; 5) the mean fragment size; 6) the number of broad peaks; 7) the ratio of 100bp reads to 150bp reads; 8) 20 genotype principal components to account for global ancestry (see above: Covariates for co-expression and co-accessibility analyses).

<u>Formula #2:</u>

$$X_{ij} = \beta_k X_{ik} + \sum_{m=1}^{M} \gamma_m PC_{im} + \sum_{p=1}^{P} \gamma_p C_{ip} + u_i + \epsilon_{ik}$$

Specifically, we utilized inverse normal transformed TMMs across the 150 ATAC-seq samples for each of the 56,978 peaks. In Formula #2 co-accessibility, X_{ij} is the normalized accessibility value for peak *j* in sample *i*, X_{ik} is the normalized accessibility value for peak *k* in sample *j*, β_k is the effect size (fixed effect) of peak *k* and the remaining terms were consistent with co-expression variable. Of the 56,978 reference peaks, 47,761 were present in the 13 major RNMs. In total, we identified 8,696,814 pairs of co-accessible peaks (P-value < 5x10^{-8}, Effect Size > 0).

Construction of the Genome-wide Regulatory Co-accessibility Network (RN): We created the genome-wide co-accessibility network (RN) using the *graph_from_data_frame* function from the *igraph* R package[26] by assigning the 56,978 accessible peaks as nodes, the 8,696,814 co-accessible peaks (P-value < 5x10^{-8}, Effect Size > 0) as edges. We extracted the genome-wide degree of each peak using the *degree* function on the RN. To find the optimal Leiden community detection model, we followed the same approach as described in the ***Construction of the Genome-wide Gene Co-expression Network*** section. We selected modules obtained from the model using the following parameters; resolution = 2.5, beta = 0.09, n_iterations = 30, which had a modularity of 0.36 and 47,761 peaks (83.8%) in 13 major RNMs (membership ≥ 500 ATAC-seq peaks). For each combination of parameters, we permuted the nodes to confirm the Leiden algorithm was clustering co-expressed genes better than the null background. For each peak, we calculated its intramodular degree (the number of co-accessible peaks within the module), using the same functions described above, excluding intermodule edges and considering each RNM independently (Table S8).

PCA and UMAP Analyses of RNMs: To assess whether ATAC-seq peaks within an RNM had similar accessibility profiles across hiPSCs, we performed PCA and UMAP analyses on the 150 ATAC-seq samples. We first identified the most interconnected ATAC-seq peaks within each RNM by ranking the intramodular degree for each RNM independently. We performed a PC analysis on the inverse normal transformed TMM of the peaks with the top 10% intramodular degree (top 10%) from the 13 major RNMs (membership ≥ 500 peaks) with the *prcomp* function in base R. We performed UMAP dimensionality reduction using the *umap* function from the *umap* R package.

Calculation of ATAC-seq peak Co-accessibility Enrichment: To assess whether peaks within each RNM were more co-accessible with each other than peaks in different RNMs, we enumerated the number of co-accessible peaks shared between all pairwise RNMs. We then performed a Fisher's Exact test to calculate the enrichment of co-accessibility between RNM pairs.

Calculation of RNM Score: For each of the 150 ATAC-seq samples we calculated an RNM score for each of the 13 RNMs by summing the inverse normal transformed TPM matrix of the corresponding RNM-specific Pareto (top 20% intramodular degree) peaks (i.e., each ATAC-seq sample had 13 RNM scores).

Correlation of GNMs and RNMs

We examined the correlation between ATAC-seq peak co-accessibility network modules and gene co-expression network modules.

Identifying Associations Between Gene and Regulatory Networks: The GNMs and RNMs were identified using overlapping hiPSC lines from the iPSCORE collection. However, the RNA-seq data and ATAC-seq data were generated from different passages of the hiPSCs cultured under different experimental conditions. The RNA-seq data was generated from earlier passage ROCK inhibitor-naïve hiPSCs and the ATAC-seq data was generated from later passage hiPSCs after culturing with ROCK inhibitor. Therefore, to annotate peaks with putative gene targets we: 1) identified genes expressed after culturing with 3 hiPSC lines with ROCK inhibitor (see above: *Generation of RNA-seq for 3 hiPSC lines expanded with ROCK inhibitor*), and then 2) annotated each peak with a single gene (distance < 100 kb and highest expressed gene). Specifically, to identify candidate target genes for the 47,761 ATAC-seq peaks in the 13 major RNMs, we generated a bed file of the TSSs for the 16,110 autosomal genes expressed (TPM > 1 in at least one of the nine samples (triplicates of each hiPSC line) cultured with 5μM ROCK inhibitor and performed *bedtools closest* to identify the closest TSS within 100 kb of each ATAC-seq peak. For ATAC-seq peaks that overlapped the TSSs of multiple genes, we calculated the maximum TPM expression across all 3 hiPSC lines cultured with ROCK inhibitor and annotated the ATAC-peak with the gene with the maximum expression. Finally, to identify associations between the GNMs and RNMs we only used genes: 1) expressed in both ROCK inhibitor-naïve hiPSCs and ROCK inhibitor-exposed hiPSCs, 2) in one of the 13 major GNMs, and 3) annotated as a putative target in one of the 13 major RNMs. In total, 12,078 unique genes corresponding to 32,327 peaks were used for the association test (Table S8). We calculated the number of genes in common between all GNM-RNM pairwise combinations and performed Fisher's Exact tests to calculate enrichments.

Functional Annotation of hiPSC ATAC-seq peaks

To functionally characterize the hiPSC epigenome, we annotated the 47,761 ATAC-seq peaks in the 13 major RNMs with three epigenetic annotations: 1) hiPSC-specific chromatin states, 2) TF binding, 3) pluripotency cell state.

Collapsing into 5 hiPSC chromatin states: Using the single chromatin state annotation (described above in ***Identification of 56,978 autosomal reference peaks*)**, we binned the 12 chromatin states into 5 collapsed states by molecular similarities (Figure S7B). "Active promoters" (TssA) were not collapsed, the collapsed "Enhancer" annotation consisted of peaks in enhancers (Enh), genic enhancers (EnhG), and flanking active promoters (TssAFlnk), "Bivalent Chromatin" consisted of bivalent promoters (TssBiv), bivalent enhancers (EnhBiv), and regions flanking bivalent chromatin (BivFlnk), "Transcription" consists of strong (Tx) and weak (TxWk) transcription, and flanking transcription (TxFlnk), and "Repressed Polycomb" consists of both polycomb states (ReprPC, ReprPCWk, Table S8).

Prediction of transcription binding with TOBIAS: The TOBIAS algorithm[33] leverages distribution of reads across the genome for a given sample, therefore to profile TF occupancy, we ran TOBIAS to predict binding at 187 motifs across all 136,333 ATAC-seq peaks (Figshare XXX, GEO: GSE_XXX, Table S11). First, we identified 187 transcription factors with experimentally validated high-confidence motifs (Quality A and B) included in the HOCOMOCO V 11 collection[34] that were expressed (TPM > 1 in >20% samples) in the 213 hiPSC lines. We used samtools 1.9 to merge, sort and index the 34 BAM files of the reference hiPSC samples (see above: *ATAC-seq Peak Calling and Quality Control*). We then ran *TOBIAS ATACorrect* on the merged reference BAM file to correct for cut site biases introduced by the Tn5 transposase within the 136,333 ATAC-seq peaks, using the following parameters: --genome hg19 fasta (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/) and –blacklist hg19-blacklist.v2.bed (http://github.com/Boyle-Lab/Blacklist/tree/hg19-blacklist.v2.bed.gz). Next, we calculated footprints scores with TOBIAS *ScoreBigwig*, using the narrowPeak file with the 136,333 ATAC-seq peaks (MACS2 score > 100) for –regions. Finally, to identify the predicted transcription factor binding sites, we ran *TOBIAS BINDetect* with all 187 expressed TFs, using hg19 fasta file and narrowPeak file as the genome and regions, respectively. TOBIAS predicted a total of 2,349,030 TFBSs across all 187 motifs on 49,070 ATAC-seq peaks which represented 37.1% of the 132,225 peaks on autosomes (FigShare XXX). We annotated the 14,208 ATAC-seq peaks without a TFBS association as "Not Bound".

Validation of Predicted TF Binding Sites Using Experimental TF ChIP-seq data: To validate the TOBIAS TF binding predictions, we evaluated the concordance with binding profiles that have been experimentally validated via TF ChIP-seq[35,39,68] (Table S9). We obtained peaks from 19 sets of transcription factor ChIP-seq peak files for 18 TFs in H1 embryonic stem cells from ENCODE[35,39,68] (Table S9). The REST TF had two independent ChIP-seq experiments, which we merged using *bedtools merge*. We intersected the TOBIAS predicted TF binding sites for the 18 TFs with corresponding TF ChIP-seq peaks using *bedtools intersect -loj*. We performed a Fisher's Exact test to calculate whether TOBIAS predicted TF binding sites were enriched in corresponding TF ChIP-seq peaks (Figure S8).

Collapsing into 92 transcription factor groups: Many different transcription factors have very similar binding motifs, and TOBIAS often predicted that TFs with similar motifs bound at the same site. For example, NANOG.0.A and PO5F1.0.A are the same length and have nearly identical position weight matrices (PWMs, Figure S9B); hence, the identity of the exact TF bound could not be resolved. To reduce the effects of motif sequence similarity, we collapsed the 187 motifs into TF groups by hierarchically clustering the Euclidean distances based on the overlap of bound sites generated by TOBIAS thresholds using *cutree* (Table S10, FigShare XXX). We selected a 0.75 threshold and obtained 92 TF groups, of which 49 consisted of a single motif, 24 were composed of motifs from the same TF family (n= 73 TFs) and 19 consisted of multiple TFs from different families (n=65 TFs), hereby referred to as complexes (Figure S9, Table S10). The 187 TF motifs and their corresponding collapsed TF groups are in Table S10.

Functional characterization of RNMs

***Epigenetic Feature Enrichment within hiPSC Regulatory Networks*:** To molecularly characterize hiPSC regulatory networks, for each of the 13 RNMs we calculated enrichment of the formative and primed cell states, 5 collapsed chromatin states and 93 TF groups (including "Not Bound"). We performed a Fisher's Exact test to calculate enrichment for each epigenetic annotation in the Pareto peaks for each major RNM by using the Pareto peaks for the 12 remaining major RNMs as background (Table S12). We considered enrichments with a Bonferroni-corrected p-value < 0.05 significant.

hiPSC regulatory networks conserved in fetal cell types

Annotating hiPSC ATAC-seq peaks with REs with Fetal Cell Type-Specific Peaks: To identify hiPSC ATAC-seq peaks that correspond to active chromatin in fetal tissue, we integrated single cell ATAC-seq peaks from 54 fetal cell types in the Descartes; Human Chromatin Accessibility During Development Atlas[51]. To obtain cell type-specific peaks, we used the Z-score corrected single cell ATAC-seq peaks for the 54 fetal cell types (n=~9,500 peaks on autosomes per cell type) (https://atlas.brotmanbaty.org/bbi/human-chromatin-during-development/). We performed *bedtools intersect -f -r 0.25* on the 47,761 reference ATAC-seq peaks in the 13 major RNMs and identified 11,830 hiPSC ATAC-seq peaks with a 25% reciprocal overlap with at least one fetal cell type-specific peak. To calculate the enrichment of each RNM for each of the 54 fetal cell types, we performed Fisher's Exact tests of the overlap with these 11,830 hiPSC ATAC-seq peaks, using the remaining 12 RNMs as background (Table S13). To calculate the enrichment of hiPSC TFBSs in the 54 fetal cell type-specific ATAC-seq peaks, we performed Fisher's Exact tests with *bedtools fisher* (Table S14). For background, we merged the bed files for all 54 fetal cell type-specific peaks and calculated the number of base pairs in the merged peaks for each chromosome.

Allele-Specific Chromatin Accessibility (ASCA) analyses

To identify regulatory variants that impact transcription factor binding we performed allele-specific chromatin accessibility (ASCA) using the 150 ATAC-seq samples from the 133 iPSCORE individuals.

Calculation of Allele-Specific Chromatin Accessibility (ASCA): A VCF file from WGS data of 273 iPSCORE individuals[66] was obtained from dbGaP (phs001325). We extracted SNPs in the 47,761 ATAC-seq peaks in the 13 major RNMs: 1) with minor allele frequency > 0.05 in all 273 iPSCORE individuals[66] using bcftools view with parameters: *--types snps*, *--f PASS*, *-q 0.05:minor*; and 2) in Hardy-Weinberg equilibrium ($p > 1x10^{-6}$) in the 133 iPSCORE individuals with ATAC-seq data using vcftools --hwe 0.000001. To increase the power to detect allele-specific chromatin accessibility, we performed phasing on these variants using the Michigan Imputation Server with the 1000 Genomes Phase 3 as a reference panel and converted them into the hdf5 format using snp2hd5 in WASP, as suggested by the original developers. We realigned the BAM files after WASP correction and applied the same filters as described above in the ***Sequencing of ATAC-seq*** section, except for removing duplicates. Specifically, we excluded poorly mapped reads (MAPQ < 20%), and reads less than 38bp and greater than 2000bp with samtools view. We then identified allele mapping bias at heterozygous sites in each sample using the WASP mapping pipeline with default parameters[91] and duplicates were removed in a non-biased manner using the rmdup_pe.py script in WASP. Coverage of bi-allelic heterozygous variants was calculated using GATK ASEReadCounter with parameters: *-overlap COUNT_FRAGMENTS_REQUIRE_SAME_BASE, -U ALLOW_N_CIGAR_READ*[92]. To maximize detection of ASCA, we aggregated allele read counts for each SNP across all heterozygous individuals and required: 1) a minimum of 5 heterozygous individuals were tested, and 2) at least 50 total reads mapped to the position, and 3) at least 10 reads mapped to each of the reference and alternate alleles. This resulted in 105,055 bi-allelic SNPs in 35,614 accessible ATAC-seq peaks used for ASCA analysis. We annotated 104,938 SNPs with their corresponding rsid from the gnomAD database (v2)[93], and used the chromosome, position, reference, and alternate allele to annotate the 117 SNPs that were not in gnomAD. ASCA was determined using a two-sided binomial test with equal probability (probability = 0.5) for each allele being accessible. P-values were corrected using Benjamini-Hochberg. SNPs with an adjusted p-value < 0.05 were considered to display allele-specific chromatin accessibility (ASCA-SNPs) (Table S15). We annotated the 6,323 ASCA SNPs with the RNMs associated with the ATAC-seq peak and the overlapping TF group(s). To identify RNMs and TF groups with enriched allelic imbalance fractions (AIFs), we performed a Mann-Whitney U test, using the 12 other RNMs or 91 TF groups as background. To calculate the enrichment of ASCA SNPs in predicted TFBSs, we performed a Fisher's Exact test using bedtools fisher.

DATA AVAILABILITY

The 222 iPSCORE hiPSC lines are available through WiCell Research Institute (www.wicell.org; NHLBI Next Gen Collection). FASTQ sequencing data for bulk RNA-seq is available through dbGaP (phs001325). FASTQ files for bulk ATAC-seq have been deposited into GSE_XXX, along with the reference narrow peak file, TMM peak accessibility matrix and metadata. WGS data for iPSCORE subjects is available as a VCF file and as FASTQ files from dbGaP (phs001325). Tables including output for the LMM co-accessibility, and TOBIAS TF binding predictions have been deposited on Figshare (https://figshare.com/account/home#/projects/XXX).

CODE AVAILABILITY

All scripts developed to perform this study are available in "GitHub [https://github.com/XXX]".

AUTHOR INFORMATION

iPSCORE Consortium

Angelo D. Arias[4], Timothy D. Arthur[1,2], Paola Benaglio[4], W. Travis Berggren[9], Juan Carlos Izpisua Belmonte[10], Victor Borja[5], Megan Cook[5], Matteo D'Antonio[2.5], Agnieszka D'Antonio-Chronowska[4], Christopher DeBoever[3], Kenneth E. Diffenderfer[9], Margaret K.R. Donovan[2,3], KathyJean Farnam[5], Kelly A. Frazer[4,5], Kyohei Fujita[4], Melvin Garcia[5], Olivier Harismendy[2], David Jakubosky[1,2], Kristen Jepsen[5], He Li[4], Hiroko Matsui[5], Naoki Nariai[4], Jennifer P. Nguyen[2,3], Daniel T. O'Connor[11], Jonathan Okubo[5], Athanasia D. Panopoulos[10], Fengwen Rao[11], Joaquin Reyna[5], Bianca M. Salgado[5], Erin N. Smith[4], Josh Sohmer[5], Shawn Yost[3], William W. Young Greenwald[3]

[9]Stem Cell Core, Salk Institute for Biological Studies, La Jolla, CA 92037, USA

[10]Gene Expression Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037, USA

[11]Department of Medicine, University of California, San Diego, La Jolla, CA 92093, USA

AUTHOR CONTRIBUTIONS

TDA, WWYG, KAF, and iPSCORE consortium members conceived the study. TDA, JPN, MD, HM, and iPSCORE consortium members performed the computational analyses. ADC, NS, and iPSCORE consortium members generated molecular data. KAF, AL, and iPSCORE consortium members oversaw the study. TDA, MFP, and KAF prepared the manuscript.

ACKNOWLEDGEMENTS

This work was supported by a California Institute for Regenerative Medicine grant GC1R-06673-B, NIH grants HG008118 (KAF), HL107442 (KAF) and HG011558 (KAF), F31DK131867 (JPN), F31HL158198 (TDA) and T15LM011271 (JPN&TDA), and by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

DECLARATION OF INTERESTS

WWYG and KAF are co-founders of Synthalogy Therapeutics and WWYG is an employee of Guardant Health.

FIGURE TITLES AND LEGENDS

Figure 1. Community detection algorithm identifies hiPSC gene networks

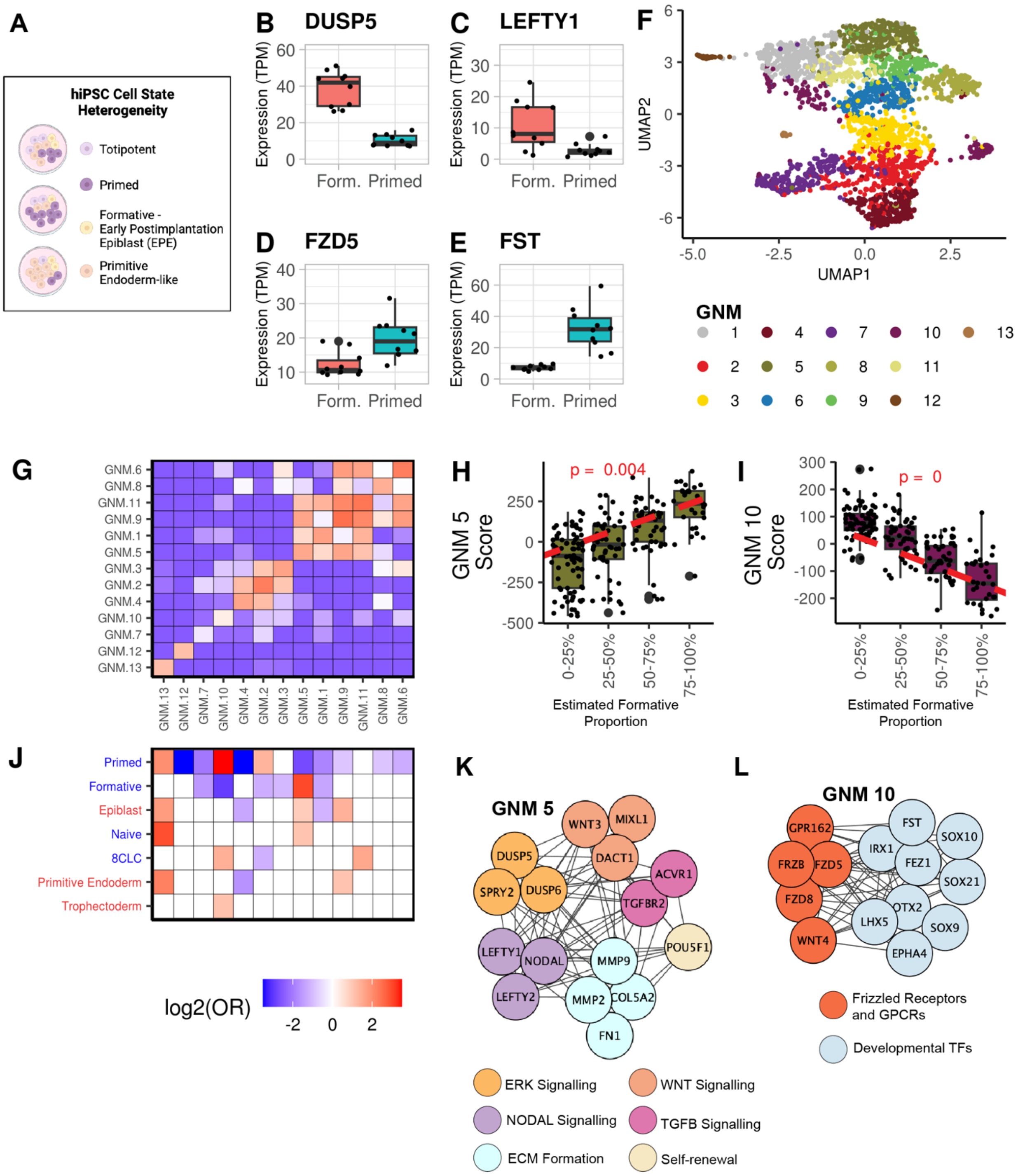

Figure 1. Community detection algorithm identifies hiPSC gene networks

(A) Cartoon depicting our hypothesis that hiPSCs are composed of varying proportions of pluripotent cell state subpopulations.

(**B-E**) Boxplots of gene expression in the 10 hiPSCs with the highest estimated proportions versus the 10 hiPSCs with the lowest estimated proportions of formative-state cells. (B) DUSP5, (C) LEFTY1, (D) FZD5, and (E) FST.

(F) UMAP plot displaying the expression of the 2,946 Pareto genes (colored by GNM membership) across 213 hiPSC RNA-seq samples.

(G) Heatmap showing that genes within a GNM are enriched for co-expression. Pairwise Fisher's Exact tests were performed to validate that genes within the same GNM were more co-expressed with each other than with genes in other GNMs. Each cell is filled with the $\log_2$(Odds Ratio). For plot legibility, the enrichment range was set to -3.5 to 3.5.

(H) Boxplots showing the association between GNM 5 and the estimated proportion of formative-state cells across the 213 RNA-seq samples.

(I) Boxplots showing the association between GNM 10 and the estimated proportion of formative-state cells across the 213 RNA-seq samples.

(J) Heatmap showing the enrichment of pluripotency cell state-associated genes in the 13 GNMs. A Fisher's Exact Test was performed on seven published gene sets associated with hiPSC cell states on each GNM to calculate the Odds Ratio. Published gene set labels on the y-axis were colored to indicate whether they were curated from *in vivo* (red) or *in vitro* (blue) experiments. Each cell is filled with the $\log_2$(Odds Ratio) for significant GNM cell state associations (red = enrichment, blue = depletion, white = non-significant). For plot legibility, the enrichment range was set to -3.5 to 3.5 and the "Primitive Endoderm" label represents the primitive endoderm-primed founder cells.

(K) Network graph showing a subset of the co-expressed genes in GNM 5.

(L) Network graph showing a subset of the co-expressed genes in GNM 10.

Figure 2. Community detection algorithm identifies hiPSC regulatory networks

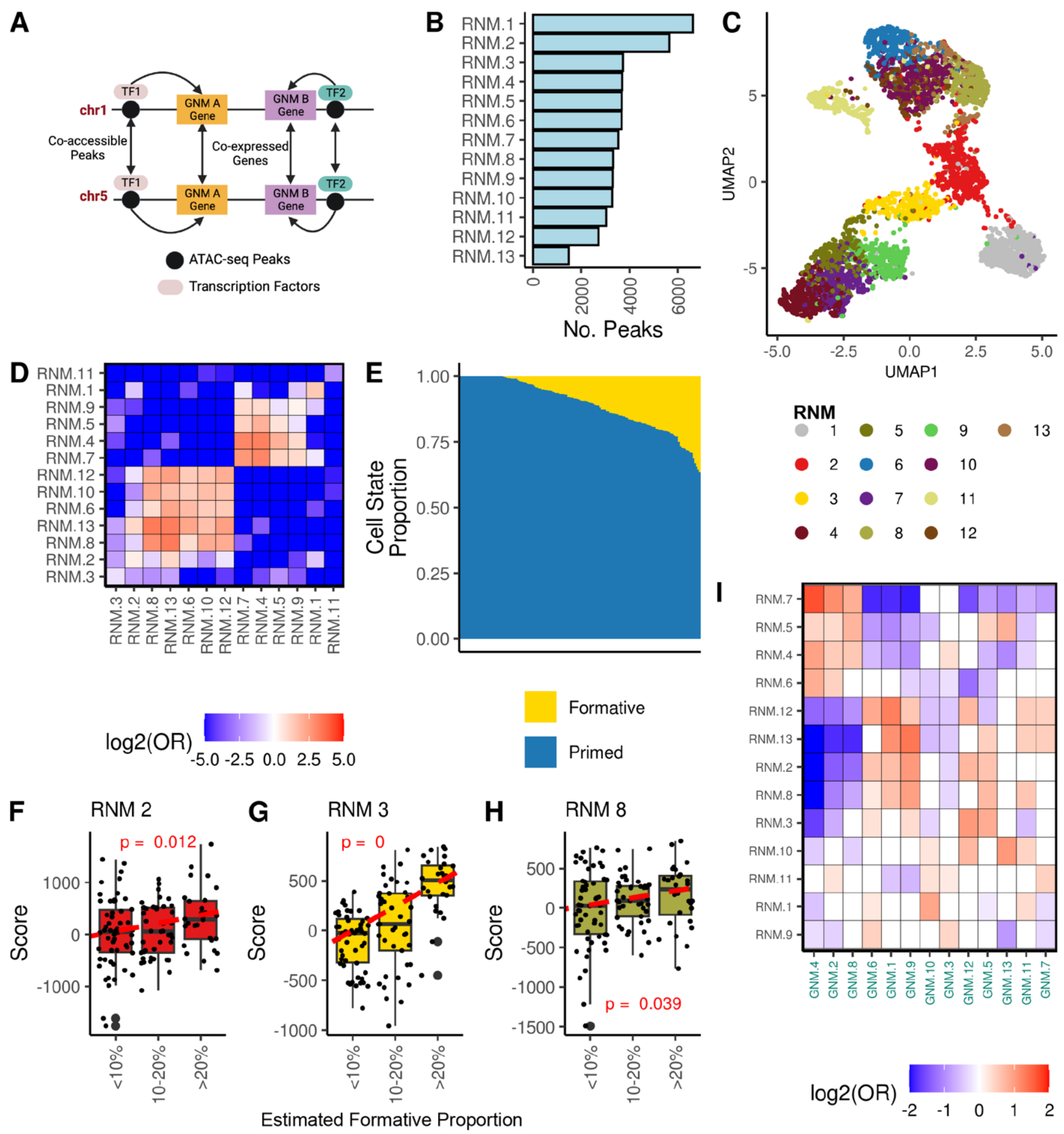

Figure 2. Community detection algorithm identifies hiPSC regulatory networks

(**A**) Diagram of proposed molecular mechanisms underlying gene co-expression networks. Co-expression of genes located on different chromosomes but within the same GNM (A or B) is mediated via regulatory elements that are co-accessible because they bind the same transcription factors (TF1 or TF2).

(B) Histogram displaying the number of peaks in each of the 13 major RNMs.

(C) UMAP plot of 4,772 ATAC-seq peaks in the 13 major RNMs. We plotted ATAC-seq peaks with the top 10% intramodular co-accessibility in each RNM, as opposed to the Pareto peaks (n = 9,545) for plot legibility. Each point represents a peak colored by its corresponding RNM.

(D) Heatmap showing the pairwise associations between 13 RNMs based on the co-accessibility enrichment. Each cell is filled with the $\log_2$(Odds Ratio) for the corresponding RNM pair.

(E) Barplot showing the estimated cell state proportions across 150 ATAC-seq samples. Previously published ATAC-seq peaks from FACs sorted cells representing formative and primed pluripotency states[1] were used to perform cellular deconvolution on the 150 iPSCORE hiPSC ATAC-seq samples. Each stacked barplot corresponds to an ATAC-seq sample and the colors correspond to the estimated formative and primed cell states

(F-H) Boxplots demonstrating RNM 2 (F), 3 (G), and 8 (H) associations with the estimated proportion of cells in the formative state. A linear model was used to calculate the association between RNM scores (summed inversed normal transformed accessibility of RNM Pareto peaks) and the estimated formative proportion of 123 ATAC-seq samples (< 100% estimated formative high proportion). Each point represents an ATAC-seq sample. Samples were binned by estimated proportion for boxplot legibility.

(I) Heatmap showing GNM-RNM associations. 32,327 ATAC-seq peaks in the 13 major RNMs were annotated with a putative target gene (see Methods). We performed pairwise Fisher's Exact test to calculate enrichments ($\log_2$(Odds Ratio)) of RNM peaks in GNMs. Non-significant associations are filled in white.

Figure 3: RNM Functional Characterization

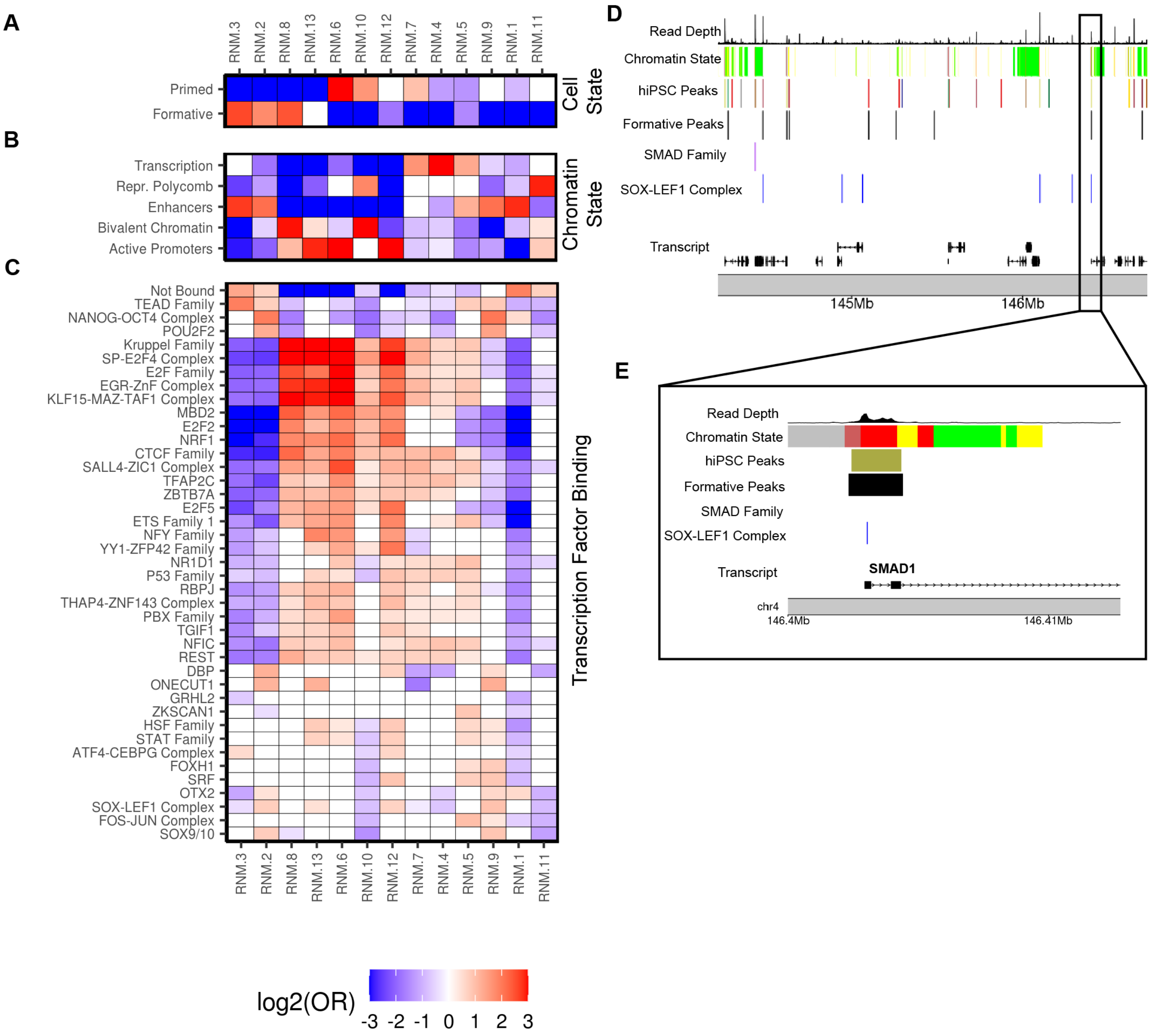

Figure 3: RNM Functional Characterization

(A) Heatmap displaying the enrichment (Odds Ratio) of formative and primed-state associated peaks in each RNM calculated using Fisher's Exact tests. Note: For Figure 3A-C, the following features are consistent; 1) the RNM order on the x-axis, 2) each cell is filled with the log$_2$(Odds Ratio) for the corresponding RNM (red = enrichment, blue = depletion, white = non-significant), and 3) the enrichment range was set from -3 to 3 for plot legibility.

(B) Heatmap showing enrichment (Odds Ratio) of the annotations for 5 collapsed iPSC chromatin states in each RNM calculated using Fisher's Exact tests (Figure S7B).

(C) Heatmap displaying enrichment (Odds Ratio) of 41 selected TF groups in each RNM calculated using Fisher's Exact tests. A heatmap reporting the RNM enrichment for all 92 TF groups and "Not Bound" is shown in Figure S10.

(D) Genome browser plot exhibiting the distribution of the three epigenetic annotations in the *chr4:144,200,000:146,700,000* locus. The top track is the read depth of the merged BAM file of the 34 reference samples, the track below is the five collapsed chromatin states (colored by chromatin state membership) (Figure S7B), the following four tracks respectively show co-accessibility of hiPSC ATAC-seq peaks (colored by RNM membership), the formative-state ATAC-seq peaks reanalyzed from Lau et al[1], predicted binding sites for SMAD Family and SOX-LEF1 Complex, and transcript hg19 coordinates.

(E) Detailed genome browser view highlighting the *SMAD1* region of the *chr4:146,400,000:146,415,000* locus. As indicated by the black box, the plot is a subset of the locus displayed in Figure 3D, and has concordant track ordering.

Figure 4. RNM enrichments for fetal cell type-specific ATAC-seq peaks

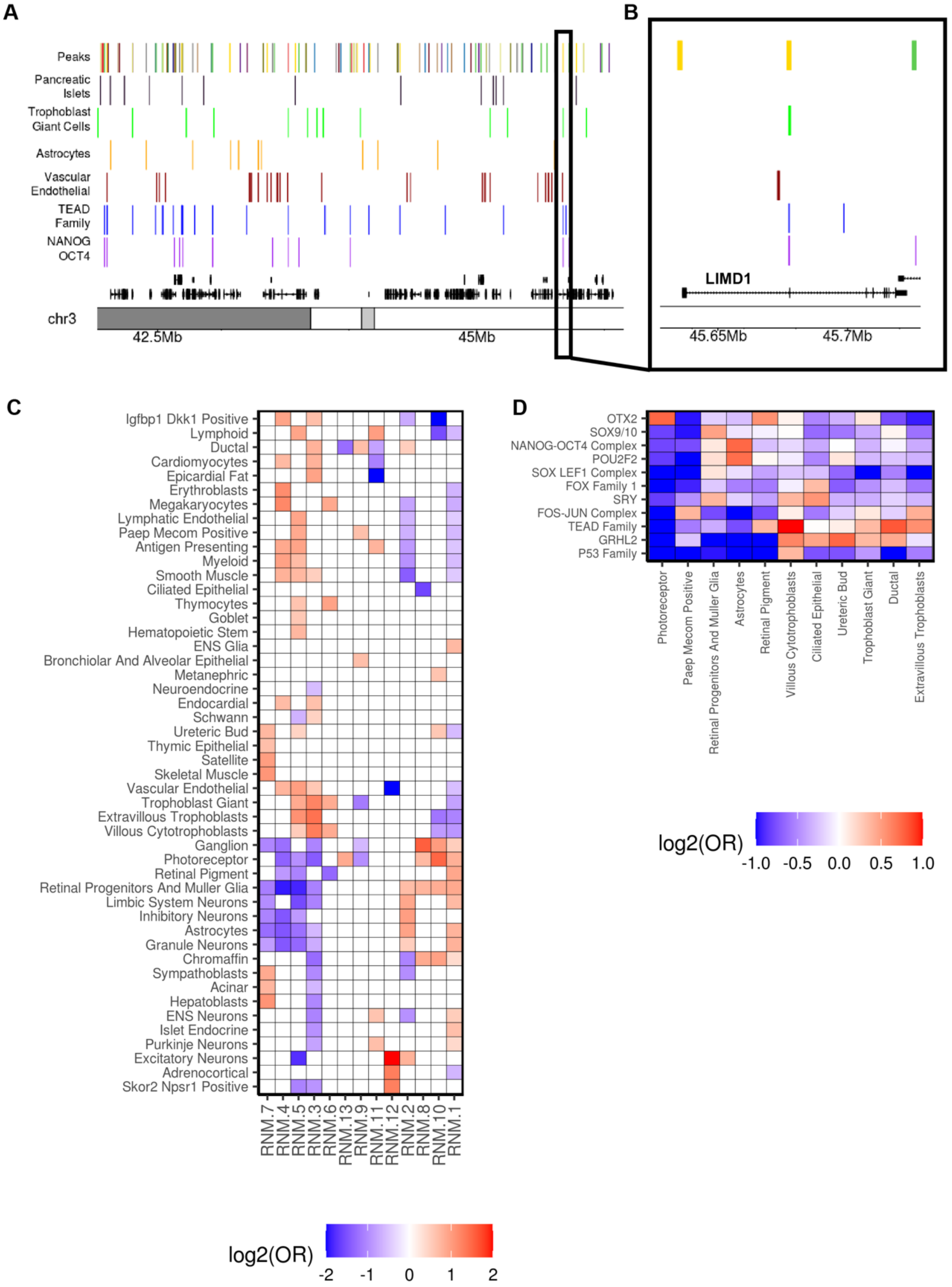

Figure 4. RNM enrichments for fetal cell type-specific ATAC-seq peaks

(A) Genome browser visualization showing (from the top): 1) hiPSC ATAC-seq peaks (colored by RNM); fetal cell type-specific peaks from the Descartes Atlas for 2) pancreatic islets, 3) trophoblast giant cells, 4) astrocytes, 5) vascular endothelial cells; TF footprints for 6) TEAD Family, and 7) NANOG-OCT4 Complex; and 8) transcript hg19 coordinates.

(B) As indicated by the black box, the view is a section of Figure 4A at higher resolution and demonstrates that an RNM 3 hiPSC ATAC-seq peak bound by NANOG-OCT4 and TEAD Family TFs overlaps a trophoblast giant cell-specific peak near an *LIMD1* exon.

(C) Heatmap displaying enrichment of the 13 RNMs in fetal cell type-specific peaks. Fisher's Exact tests were used to calculate the enrichment (Odds Ratio) of the cell type-specific peaks for 54 fetal cell types in each RNM. 5 fetal cell type-specific peaks were omitted from the plot because they were not enriched an RNM. Each cell is filled with the $\log_2$(Odds Ratio) for the corresponding fetal cell types and RNM (red = enrichment, blue = depletion, white = non-significant).

(D) Heatmap showing TF enrichment in the ATAC-seq peaks underlying enrichments in RNM 10. Fisher's Exact tests were used to calculate the enrichment (Odds Ratio) of the 92 TF groups in the hiPSC peak underlying the fetal cell type associations. Each cell is filled with the $\log_2$(Odds Ratio) for the indicated fetal cell type and TF group (red = enrichment, blue = depletion, white = non-significant).

Figure 5. RNMs and TFBSs exhibit differential allelic imbalance

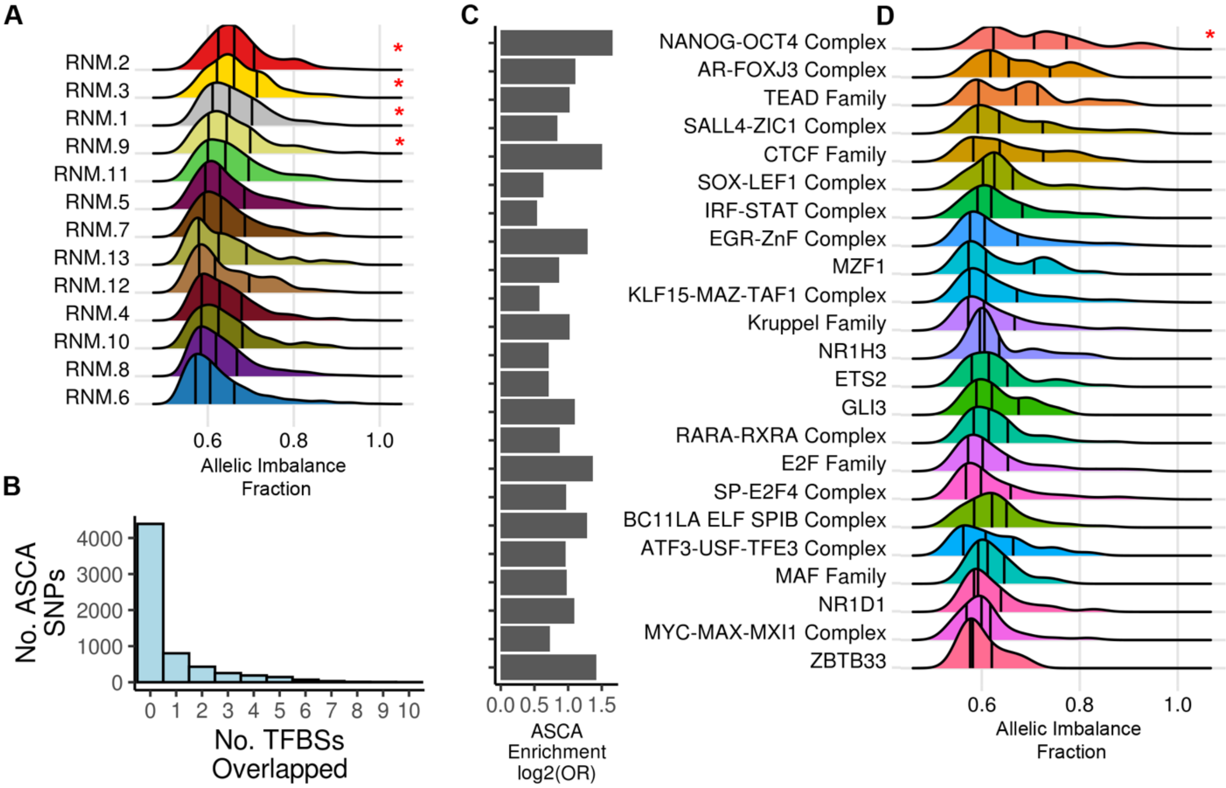

Figure 5. RNMs and TFBSs exhibit differential allelic imbalance

(A) Density plot showing the allelic imbalance fraction (AIF) of ASCA SNPs in the 13 RNMs. The AIF densities (x-axis) for the 13 RNMs (y-axis) demonstrate that 4 RNMs (indicated by red asterisks) contain ASCA SNPs that have significantly greater allelic imbalance compared to ASCA SNPs in other RNMs (Mann-Whitney U Test, adjusted P-value < 0.05).

(B) Histogram showing the number of ASCA SNPs (y-axis) that overlapped predicted TFBSs (x-axis). 6,323 SNPs in 4,241 ATAC-seq peaks exhibited allele-specific chromatin accessibility (ASCA). ASCA SNPs were intersected with predicted TBFSs for all 92 TF groups. 4,299 ASCA SNPs did not overlap a TFBS and 1,933 overlapped one or more TFBSs.

(C) Barplot demonstrating enrichment for ASCA SNPs in 23 TF groups. A Fisher's Exact test was used to calculate enrichment (Odds Ratio) for ASCA SNPs in the 92 TF groups (Table S12). The barplot displays the enrichment ($\log_2$(Odds Ratio)) for the 23 TF groups that were significant (adjusted P-value < 0.05). The TF groups are ordered based on their AIF densities in Figure 5D.

(D) Density plot showing the allelic imbalance fraction (AIF) of ASCA SNPs in the 23 TF groups. The AIF densities (x-axis) for the 23 TF groups (y-axis) demonstrate that NANOG-OCT4 Complex binding sites (indicated by red asterisk) contain ASCA SNPs that have greater allelic imbalance compared to ASCA SNPs in the other 22 TF groups (Mann-Whitney U Test, adjusted P-value < 0.05).

SUPPLEMENTAL NOTE 1

Weighted gene co-expression network analysis (WGCNA)[6] is the most commonly used gene module detection method, however, it cannot account for kinship (genetically related individuals). In this study, we used hiPSC lines from 219 individuals (Table S1) recruited as part of the iPSCORE resource, of which 140 belonged to families composed of two or more subjects (range: 2–14 subjects). To address this confounding factor, we first applied an LMM to calculate gene co-expression and ATAC-seq peak co-accessibility, using kinship as the random effects term (See Methods). We loaded the edges of the significantly co-expressed genes and co-accessibility ATAC-seq peaks into a network and applied the Leiden community detection algorithm to detect modules. To determine if our approach or WGCNA is more suitable for module detection using iPSCORE resource samples**,** we applied WGCNA to calculate gene co-expression in the 213 hiPSC RNA-seq samples. Downstream analyses showed that the WGCNA modules were correlated with biological and technical covariates, as well as family structure (Figure S3). We also determined that the LMM-Leiden module detection approach was more precise at identifying modules associated with formative-state-specific gene expression than WGCNA (Figure S4). These results show that the conventional WGCNA module detection approach can be affected by donor relatedness and that accounting for kinship leads to more accurate module membership.

Figure S1. Pluripotency cell states in human induced pluripotent stem cell lines.

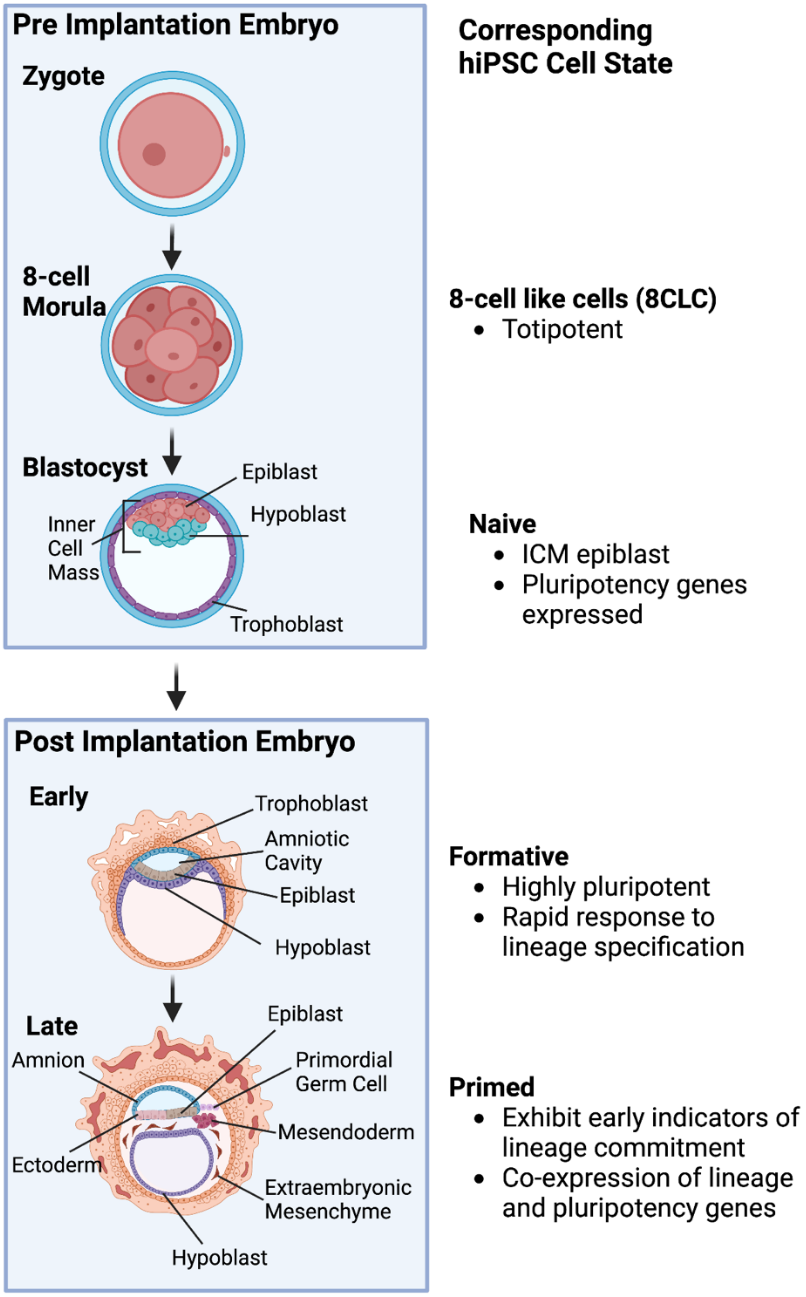

Figure S1. Pluripotency cell states in human induced pluripotent stem cell lines.

Recent studies have shown that human induced pluripotent stem cell (hiPSC) lines are mosaics, composed of varying proportions of cells in different, interconvertible pluripotent cell states with distinct transcriptional circuitry and epigenomic profiles[1]. Live cell staining[1] and single-cell analyses[2] in hiPSCs have revealed that the interconvertible pluripotent cell states have morphological, transcriptional, and epigenetic profiles that correspond to different embryonic and extraembryonic lineage fates[1,2]. Subpopulation heterogeneity complicates the molecular characterization of these interconvertible stem cell stages. Most cells within a conventional hiPSC culture resemble the late post-implantation epiblast stem cells[1], which resembles the primitive streak and are referred to as “primed”. Primed stem cells co-express lineage-specific and pluripotency genes. “Naïve” stem cells represent the cellular state of the inner cellular mass (ICM) preimplantation epiblast which gives rise to the embryo proper. At this same pre-implantation stage are the extra-embryonic primitive endoderm cells (PrE) that give rise to the primary yok sac. A pluripotent state called “formative” has been identified as developmentally between the naïve and primed states, which represents the early post-implantation epiblast (EPE). The formative pluripotent stage has been shown to be comprised of cells enriched for high self-renewal[1]. Typically, a small proportion of cells within an hiPSC line are totipotent and resemble the 8-cell morula (8 cell-like cells; 8CLC)[3]. The 8CLC subpopulation is enriched in catabolic processes, such as protein synthesis and RNA metabolism[3].

Figure S2. Formative cell state deconvolution of 213 hiPSC lines

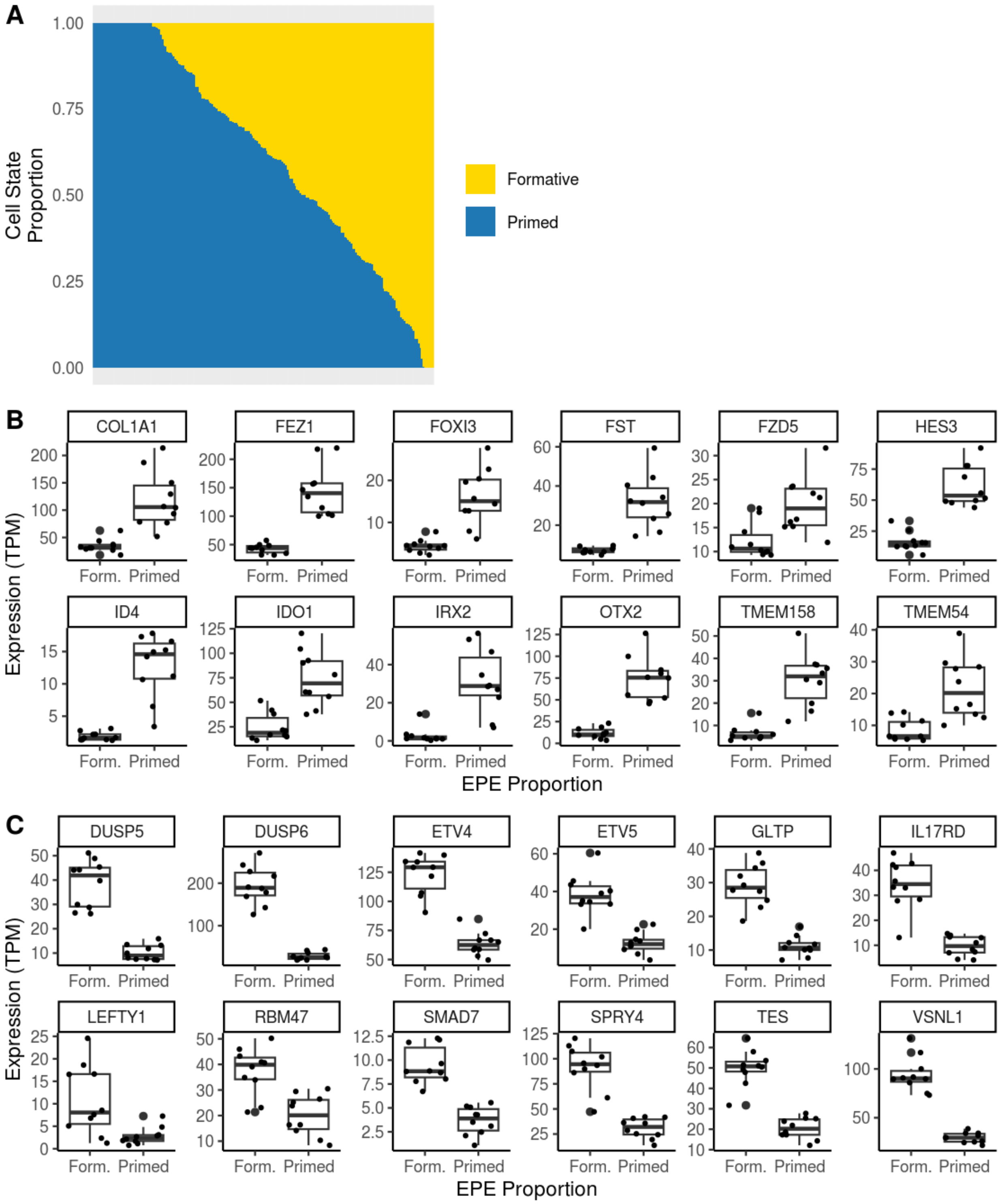

Figure S2. Formative cell state deconvolution of 213 hiPSC lines

(A) Stacked bar plot showing estimated proportions of cells in formative and primed states across 213 hiPSC lines. Each bar on the x-axis represents a hiPSC line with the corresponding estimated proportions of each cell state on the y-axis. We generated gene signatures using bulk RNA-seq data for FACS-sorted formative and paired unsorted (e.g. primed) cells[1]. While there are multiple pluripotency states present in hiPSCs, the analysis was limited to estimating the fraction of cells that were formative-like (e.g. formative, naïve, totipotent) and the fraction that were primed-like (e.g. general population). We applied the CIBERSORT deconvolution algorithm with a signature matrix containing the 100 most differentially expressed genes between the two populations (Table S2), and observed that the estimated fraction of cells in the formative state exhibited a wide range (0-100%) across the 213 hiPSC lines. Of note, while the RNA-seq and ATAC-seq data were generated from the same hiPSC lines, there were batch effects because of differences in culture conditions (i.e., the ATAC-seq data were generated from hiPSCs that had been cultured with ROCK inhibitor; see Methods). The estimated ratios of formative:primed cells differ between the deconvolution analysis using RNA-seq data and the deconvolution analysis using ATAC-seq data shown in Figure 2E. We feel that technical differences (the aforementioned batch effects and the FACs sorted cell populations used to generate the gene and peak signatures were slightly different[1]) underlie most of the variance in the estimated ratios of formative:primed cells but biological factors may also contribute.

(B) Boxplots showing the differential expression of 12 primed-specific signature genes between ten hiPSC lines with the lowest and highest estimated formative proportion.

(C) Boxplots showing the differential expression of 12 formative-specific signature genes between ten hiPSC lines with the lowest and highest estimated formative proportion.

Figure S3. Heatmap demonstrating the associations between WGCNA module correlations with covariates.

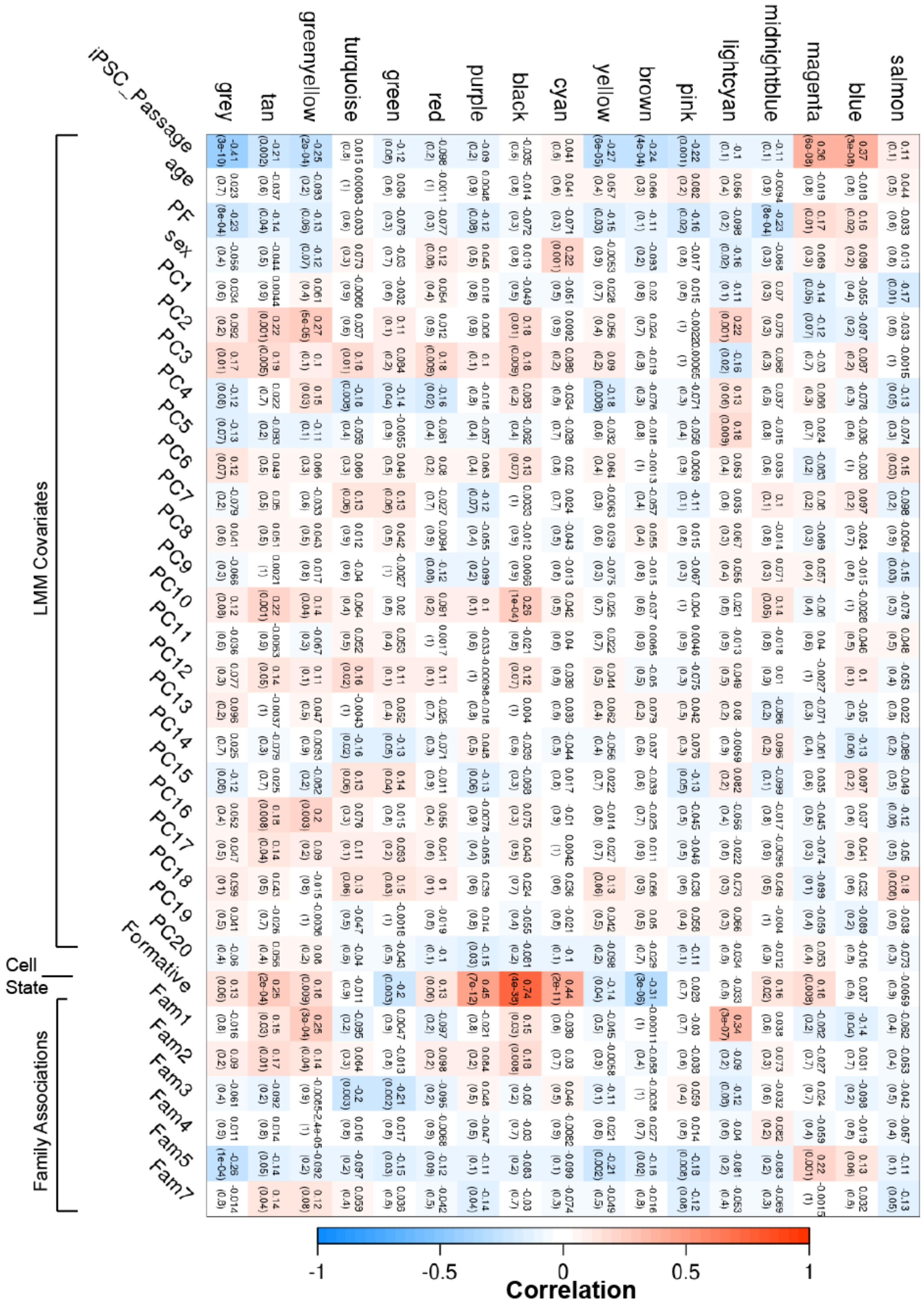

Figure S3. Heatmap demonstrating the associations between WGCNA module correlations with covariates. WGCNA[6] is the most commonly used R package for gene module identification. However, WGCNA cannot account for data that includes samples from related individuals or correct for covariates. To evaluate whether WGCNA gene module

identification was affected by the presence of related individuals, we applied it to the 213 hiPSCs RNA-seq data using the standard workflow outlined in the tutorial and identified 17 modules (top referred to by colors). We used WGCNA to calculate associations between the 17 modules and the covariates used in our LMM (age, sex, hiPSC passage, number of reads passing filters, and 20 ancestry PCs), as well as estimated formative (EPE) proportion, and binary annotations for large families (Fam 1, 2, 3, 4, 5 and 7). Each cell is filled with correlation (top value) and the p-value of association (bottom value). Several covariates exhibited significant associations with several modules. For example, the “black” module is significantly associated with global genotype PCs 2, 3, 10, and samples from families 1 and 2, and the formative-state cell proportion. In total, 11 modules were significantly associated with global genotype PCs, and 13 modules were associated with specific families. This indicates that genetic factors influence WGCNA detection of biologically relevant modules.

Figure S4: WGCNA Cell State Enrichment

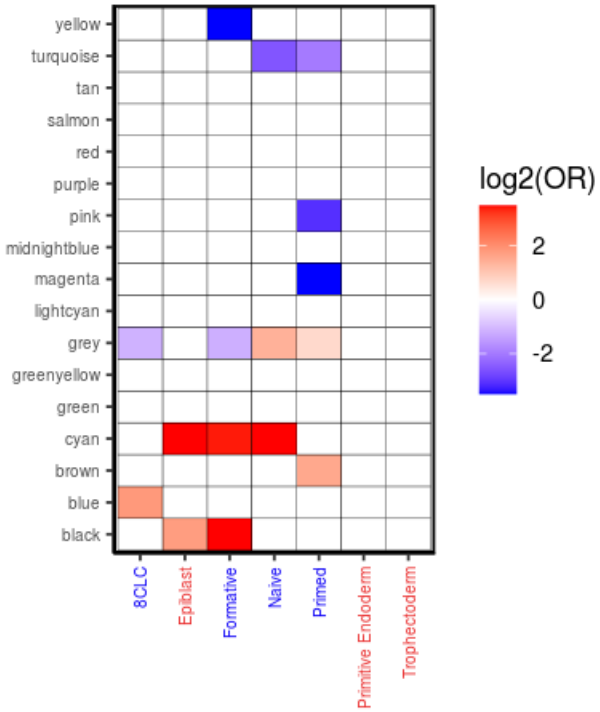

Figure S4. Cell state enrichments in WGCNA gene modules.

There is no consensus on gene module validation^{6} , however, assessing the biological enrichments is commonly used to determine which modules capture genes involved in the same biological processes. We demonstrated that our LMM approach using kinship as a random effects term to identify significantly co-expressed genes, followed by loading the edges of the co-expressed genes into a network and applying the Leiden community detection algorithm to detect modules identified a single GNM (5) enriched with formative-specific genes (Figure 1J-K). To evaluate WGCNA precision, we re-performed the stem cell state enrichment analysis (Figure 1J) using the WGCNA modules (Supplemental Note 1). Published gene set labels on the y-axis were colored to indicate whether they were curated from *in vivo* (red) or *in vitro* (blue) experiments. We observed that the formative-state gene set is strongly enriched in the “black” and “cyan” WGCNA modules (Figure S3). This suggests that in the iPSCORE cohort, which includes samples from related individuals, our LMM model more precisely identifies gene module(s) associated with the formative state than WGCNA.

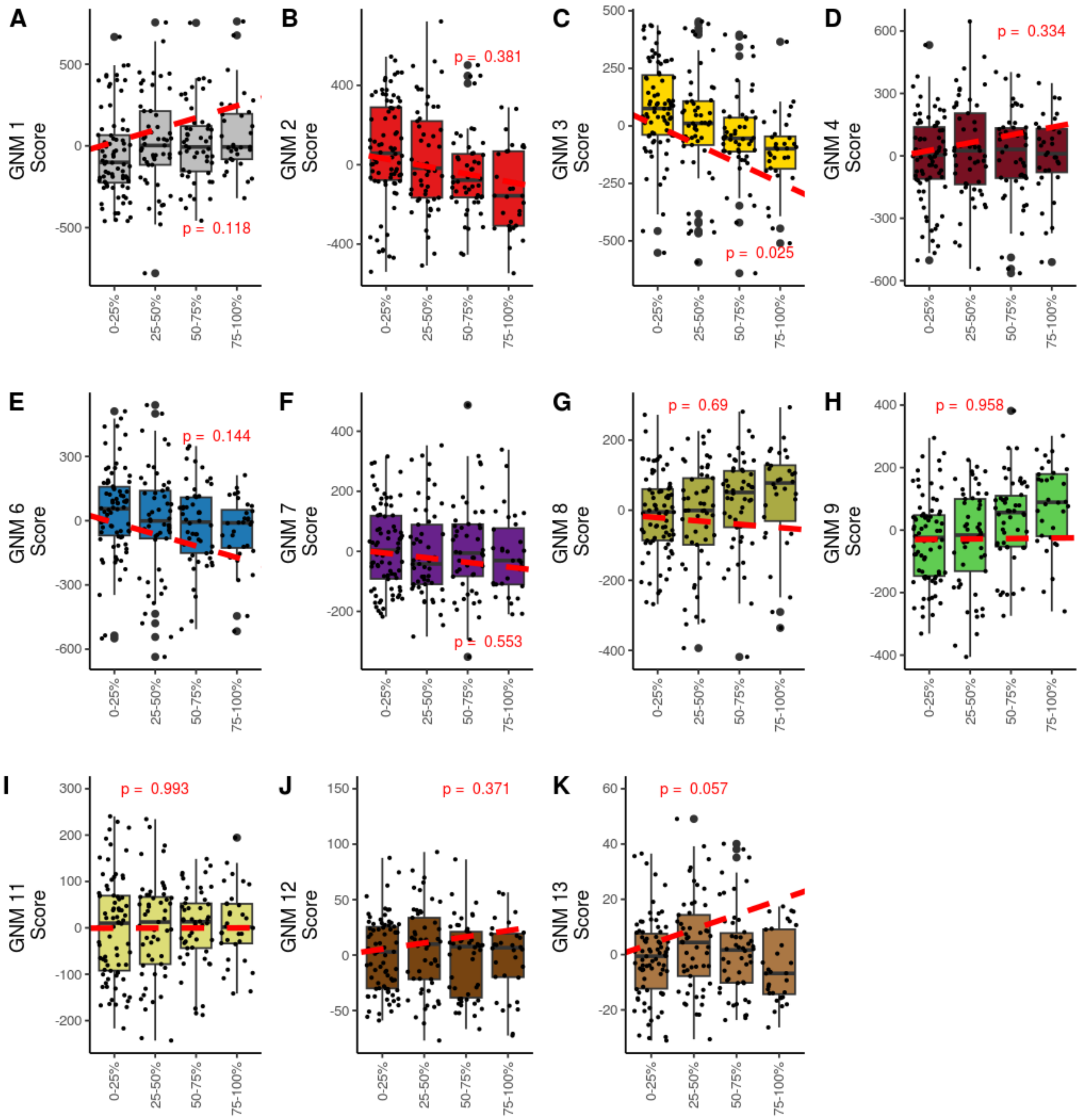

Figure S5. Box plots displaying the association between estimated proportions of formative-state cells and GNMs. Scores for hiPSC for each GNM were calculated by summing the inverse normal transformed expression of the Pareto genes. We used a linear model to calculate the association between scores for each GNM and the estimated formative-state proportion (see Methods). For plot legibility, samples were binned into quartiles based on estimated formative state and plotted against GNM scores. Each point represents an RNA-seq sample.

Figure S6. ATAC-seq peak calling strategy and quality control.

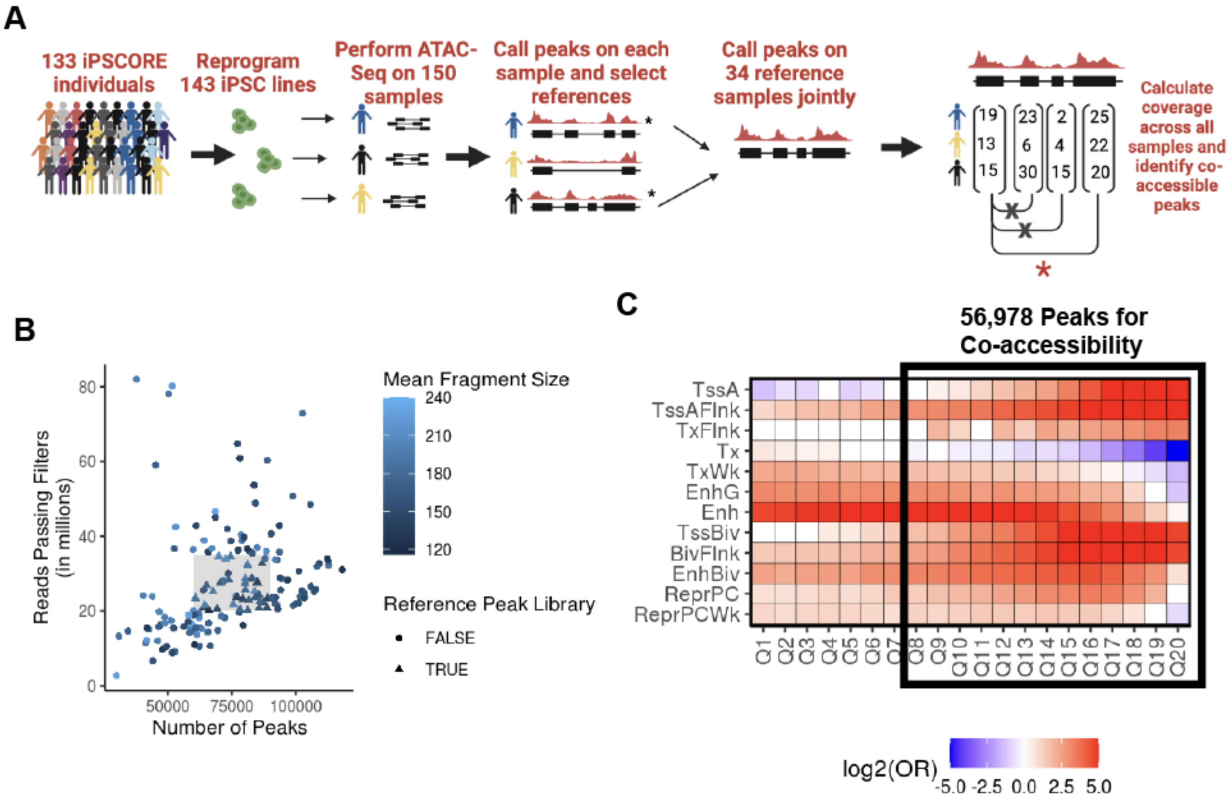

Figure S6. ATAC-seq peak calling strategy and quality control

(A) Diagram of ATAC-seq peak calling strategy. We set out to identify a reference set of high quality ATAC-seq peaks to use the co-accessibility analysis (see Methods for a detailed description). Briefly, we first used MACS2 to call broad peaks on each of the 150 ATAC-seq samples individually. We then evaluated the quality of each sample by examining the number of reads passing filters, the mean fragment size, and the number of broad peaks. Using these quality metrics, we selected 34 reference samples from unrelated iPSCORE individuals and used MACS2 to call narrow peaks on them jointly to establish a set of 136,333 reference peaks (including 132,225 autosomal peaks and 4,108 peaks on sex chromosomes, MACS2 score > 100). We then used *featureCounts* to count the number of reads in each of the 136,333 reference peaks for each of the 150 ATAC-seq samples. We selected 56,978 autosomal peaks (see details in Methods) and calculated co-accessibility.

(B) Scatter plot showing quality metrics of the 150 ATAC-seq samples. As described above, we evaluated the quality of samples based on three technical variables; number of reads passing filters, mean fragment size, and number of broad peaks. In the plot, each point represents an ATAC-seq sample, the box represents the window where we selected the 34 reference samples (triangles) from unrelated iPSCORE individuals.

(C) Heatmap showing the enrichment of ATAC-seq peaks in hiPSC-18 ChromHMM chromatin states by MACS2 score quantile. To use chromatin state enrichments in order to prioritize ATAC-seq peaks for downstream co-accessibility analysis, we binned the peaks into 20 quantiles by their MACS2 score and created bed files containing the coordinates for peaks in each quantile. We then calculated their enrichment (Odds Ratio) in each of the 12 active chromatin states for hiPSC-18, using *bedtools fisher*. It has been extensively shown that ATAC-seq identifies accessible chromatin that is enriched with active regulatory elements (TssA, Enh), bivalent chromatin (TssBiv, EnhBiv), and repressed polycomb regions (ReprPC, ReprPCWk) in iPSCs, therefore we evaluated the enrichment of these chromatin states across the 20 quantiles. We noticed that TssA are enriched in quantiles 8-20 and Enh are enriched across all 20 quantiles. Additionally, TssBiv and ReprPC started to exhibit diminishing enrichment around quantile 8. These observations supported our decision to use peaks in quantiles 8-20 (annotated by the black box) for downstream analyses.

Figure S7. Epigenetic Characterization of Reference ATAC-seq Peaks

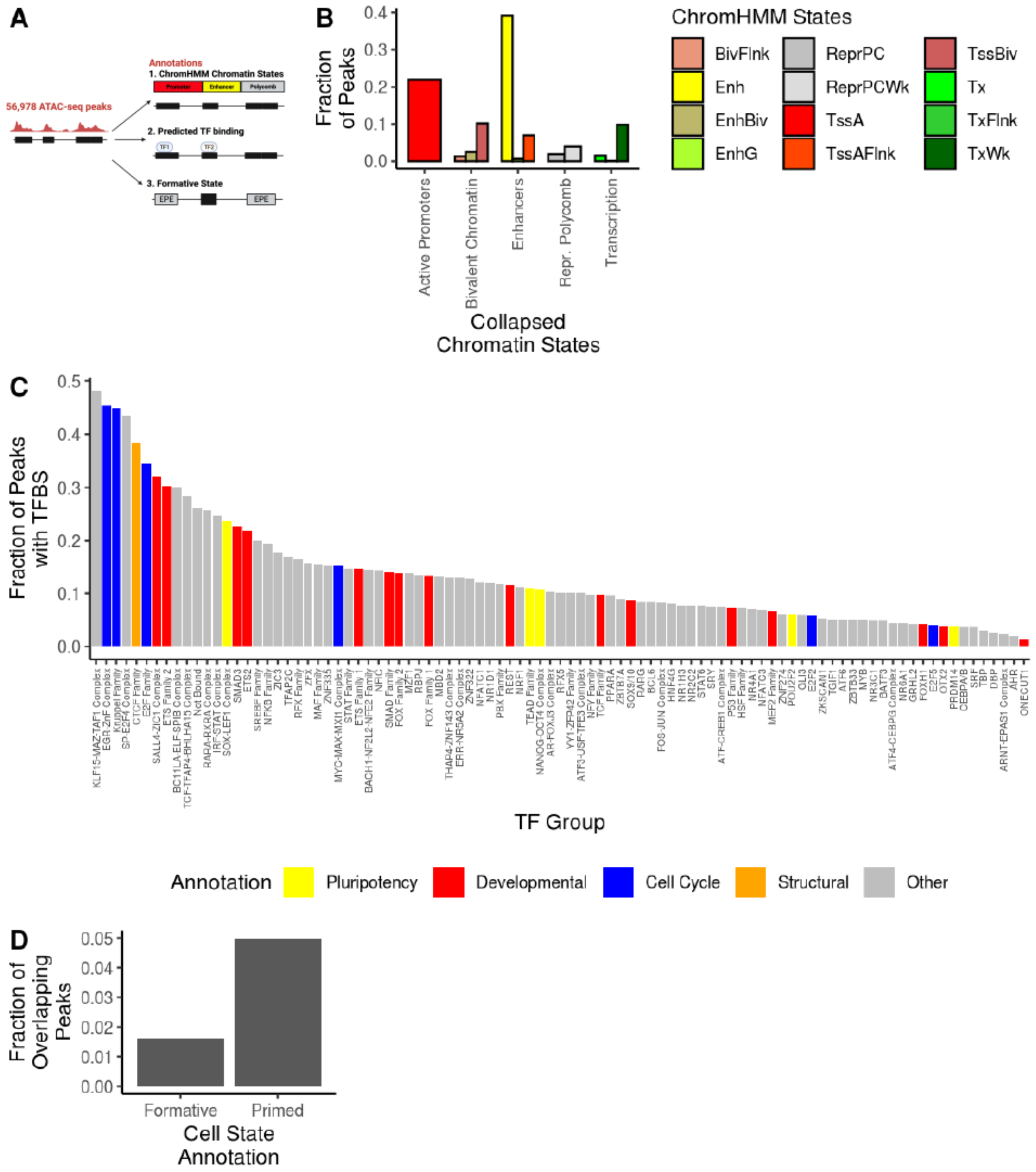

Figure S7. Epigenetic Characterization of Reference ATAC-seq Peaks

(A) Diagram of Epigenome Annotation Strategy. The 56,978 ATAC-seq peaks were characterized with three epigenetic annotations; 1) ChromHMM iPSC-18 chromatin states, 2) TF binding predicted by TOBIAS (v0.15.1) and 3) formative-state associated peaks.

(B) Barplot showing the fraction of ATAC-seq peaks annotated by the iPSC-18 ChromHMM chromatin states and the five collapsed chromatin states. We filtered ATAC-seq peaks in 3 inactive chromatin states (ZNF genes & repeats, heterochromatin, and quiescent chromatin). We binned the remaining 12 chromatin states into 5 collapsed states by molecular similarities. "Active promoters" (TssA) were not collapsed, the collapsed "Enhancer" annotation consisted of peaks in enhancers (Enh), genic enhancers (EnhG), and flanking active promoters (TssAFlnk), “Bivalent Chromatin” consisted of bivalent promoters (TssBiv), bivalent enhancers (EnhBiv), and regions flanking bivalent chromatin (BivFlnk), “Transcription” consists of strong (Tx) and weak (TxWk) transcription, and flanking transcription (TxFlnk), and “Repressed Polycomb” consists of both polycomb states (ReprPC, ReprPCWk). We then calculated co-accessibility on ATAC-seq peaks in these 5 collapsed chromatin states.

(C) Barplot showing the fraction of ATAC-seq peaks bound by a selected subset of 93 TF groups. As indicated by the legend, bars are colored by an associated biological process curated by a review of the literature.

(D) Barplot showing the fraction of ATAC-seq peaks overlapping formative or primed-associated peaks. Formative peaks overlapped GCTM-2^{high}CD9^{high}EPCAM^{high} specific peaks and primed peaks overlapped GCTM-2^{mid}-CD9^{mid} specific peaks obtained from reanalyzing the Lau et al. dataset[1].

Figure S8. Validation of TOBIAS TFBS predictions

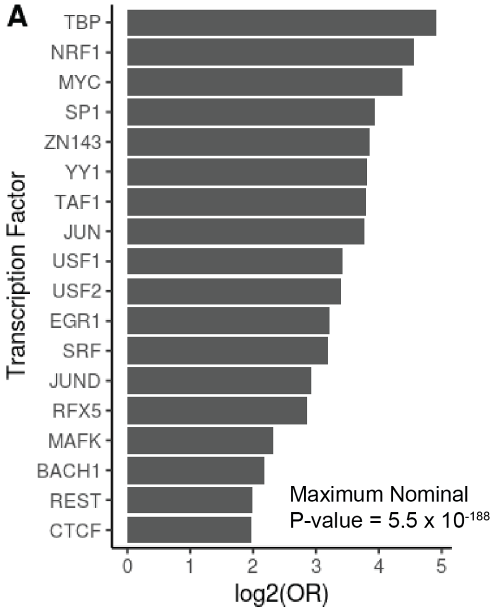

Figure S8. Validation of TOBIAS TFBS predictions

(A) Barplot demonstrating the accuracy of TOBIAS TFBS predictions for 18 TFs. To validate the TOBIAS TFBS binding sites, we obtained IDR-corrected TF ChIP-seq peaks from H1 embryonic stem cells for 18 transcription factors from ENCODE. We performed Fisher's Exact tests on the predicted TFBSs for the corresponding TF based on whether there is overlap with the experimentally validated TF ChIP-seq peaks. This reveals that TOBIAS predicted TFBSs were highly enriched in experimentally validated ChIP-seq peaks (maximum p-value = $5.5\text{x}10^{-188}$, Table S9).

Figure S9. Defining TF groups by binding site similarity

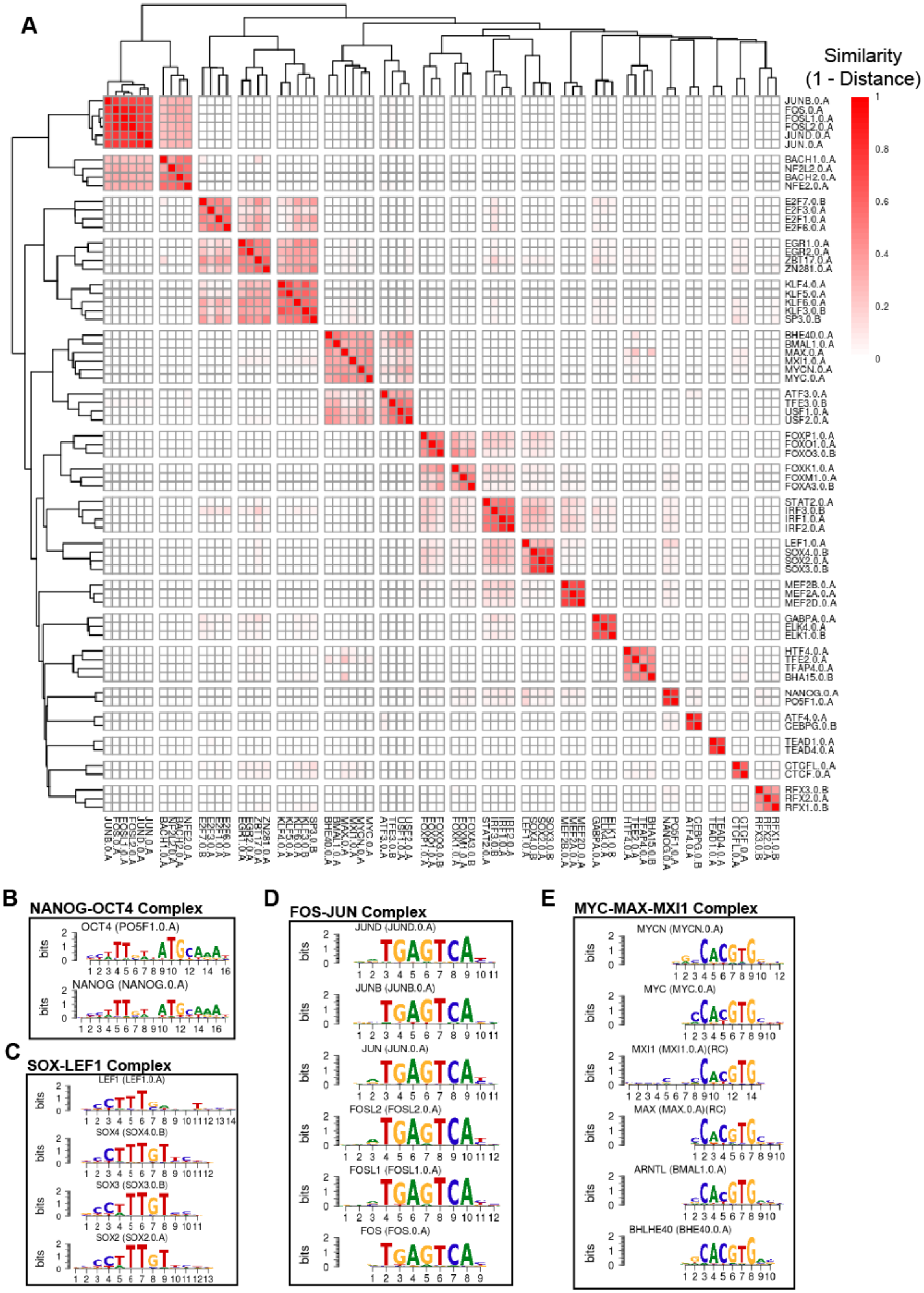

Figure S9. Defining TF groups by binding site similarity

(A) Heatmap displaying TFBS similarities for collapsed TF groups. To account for motif similarity of closely-related TFs and co-binding TFs that form complexes, we used the motif distance matrix from the TOBIAS output to binding site similarities (1 – distance) across 187 motifs. We used *cutree* (h = 0.75) to collapse the 187 motifs into 92 TF groups (Table S10). For plot legibility, only 19 collapsed groups consisting of 68 motifs are shown.
(B-E) Consensus sequences of motifs collapsed into TF groups. (B) NANOG.0.A and PO5F1.0.A (OCT4) have similar, long motifs that capture the distinct binding sequences for both TFs and form the NANOG-OCT4 Complex. (C) SOX TFs (SOX2, SOX3, and SOX4) have similar motifs to LEF1, thus form a complex, (D) TFs from the FOS and JUN families have highly similar motifs and form the FOS-JUN Complex. (E) Others have demonstrated that MYC (MYC.0.A) and MYCN (MYCN.0.A) form complexes with MAX, MXI1, ARTNL (BMAL.0.A) and BHLHE40 (BHE40.0.A), and collectively they form the MYC-MAX-MXI1 Complex. The plot titles include the standard gene name, the motif name, and whether the aligned sequence is the reverse complement (RC).

Figure S10. RNM Enrichments with Three Classes of Epigenetic Markers

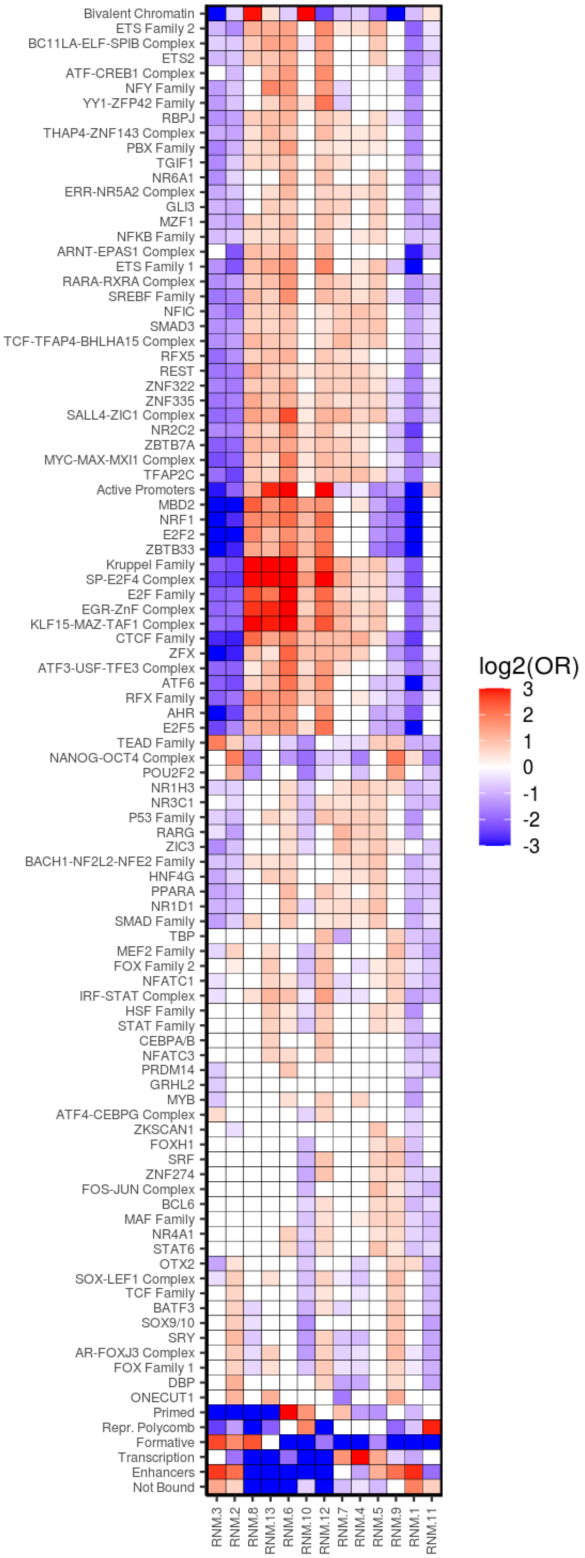

Figure S10. RNM Enrichments with Three Classes of Epigenetic Markers

Heatmap displaying epigenetic annotations of the 13 major RNMs. This figure is a comprehensive version of Figures 3A-C and consists of RNM enrichments of 1) all 92 TF groups and "Not Bound" peaks; 2) 5 collapsed hiPSC chromatin states

(Figure S7B), and 3) two self-renewal cell states (Formative-state and primed-state associated peaks (Figure S7D). Only significant enrichments are shown (non-significant enrichments are filled in white). The results of this analysis are reported in Table S12.

SUPPLEMENTAL TABLE LEGENDS

Table S1. iPSCORE Subject Information.

Information about each of the 219 iPSCORE individuals that were included in this study, including subject iPSCORE ID (Column A), subject universally unique identifier (UUID, Column B), whole genome sequencing sample UUID (Column C), sex (Column D), age at enrollment (Column E), if there is corresponding RNA-seq data (Column F), ATAC-seq data (Column G), and the top 20 genotype PCs for global ancestry (Columns H-AA) for the individual. The kinship matrix describing the relatedness of the samples was deposited on GEO (GSE_XXX).

Table S2: RNA-seq Sample Information

Information about the 213 RNA-seq samples used in this study, including subject iPSCORE ID (Column A), subject universally unique identifier (UUID, Column B), RNA-seq sample UUID (Column C), the iPSC iPSCORE ID which consists of the iPSCORE ID, iPSC clone number and the iPSC passage number (Column D), the iPSC clone (Column E) and passage (Column F) numbers, the number of properly paired reads (Column G), and the estimated formative proportion (Column H).

Table S3: Signature Gene Matrix for RNA-seq Cellular Deconvolution

Signature gene expression matrix for CIBERSORT cellular deconvolution of 213 RNA-seq samples. The table includes the Gencode v34 gene ID and gene name (Columns A and B), the expression (TPM) of 100 differentially expressed genes between the FACS-sorted formative (GCTM-2^{high}CD9^{high}EPCAM^{high}, Column C) and primed (unsorted, column D) populations from Lau et al. 2020.

Table S4: Gene Network Module Memberships

Information about gene network module (GNM) analysis annotations for the 16,110 expressed genes, including the Gencode v34 gene ID (Column A) and gene name (Column B), the corresponding GNM (Column C), whether the corresponding GNM was used for downstream analyses (Major GNM, Column D), whether the gene was considered Pareto (Column E), the gene's co-expression degree connectivity across all 16,110 expressed genes (genome-wide degree, Column F), and the gene's co-expression degree connectivity (Column G) within its corresponding module (intramodular degree).

Table S5: GNM Enrichment Results

Results from the Fisher's Exact test to calculate pluripotency cell state gene set enrichments. Information includes the annotation tested (Column A), the GNM tested (Column B), the odds ratio (Column C), the p-value (Column D), and the Benjamini-Hochberg corrected p-value (Column E).

Table S6: ATAC-seq Sample Information

Information about the 263 individual ATAC-seq samples from iPSCORE individuals that were merged by library into 150 ATAC-seq samples. The table includes subject universally unique identifier (UUID, Column A), the iPSCORE subject ID (Column B), the iPSC iPSCORE ID which consists of the iPSCORE subject ID, clone number and passage number (Column C), sample UUIDs for the 150 merged ATAC-seq libraries (Column D), sample UUIDs for the 263 ATAC-seq samples before merging (Column E), the iPSC clone number (Column F), the iPSC passage (Column G) and number of reads passing filters (Column H), the mean fragment size (Column I), the number of broad peaks used for QC (Column J), the ratio of 100bp reads to 150bp reads in the merged sample (1=100% 100bp reads and 0=100% 150bp reads; see Methods; Column K), the estimated formative proportion (Column L), and whether the merged sample was used as a reference for establishing a set of reference narrow peaks (Column M). **Note:** To obtain non-redundant data for the 150 merged ATAC-seq samples used in downstream analyses, use the unique rows from columns A-D and F-M.

Table S7: Signature Peak Matrix for ATAC-seq Cellular Deconvolution

Signature gene expression matrix for CIBERSORT cellular deconvolution of 150 ATAC-seq samples. The table includes the narrow peak ID (Column A), the accessibility (TMM) of 200 differentially accessible peaks between the FACS-sorted formative (GCTM-2^{high}CD9^{high}EPCAM^{high}, Column B) and primed (GCTM-2^{mid}-CD9^{mid}, column C) populations from Lau et al. 2020.

Table S8: Regulatory Network Module (RNM) Memberships and Annotations

Information about regulatory network module (RNM) analysis annotations for the 56,978 accessible peaks, including the peak ID (Column A), peak chromosome, start and end positions (Column B-D), the corresponding RNM (Column E), whether the RNM was one of the 13 used for downstream analyses (Major RNM, Column F), whether the peak was considered a Pareto peak (Column G), the peak's co-accessibility degree connectivity (Column H) across all 56,978 accessible peaks (genome-wide degree), and the peak's co-accessibility degree connectivity (Column I) within its corresponding module (intramodular degree), the iPSC-18 ChromHMM chromatin state annotation (Column J), the collapsed chromatin states (Column K), and the Gencode v34 gene ID (Column L), gene name (Column M), the distance in base pairs (Column N) of the closest expressed gene after ROCK kinase inhibitor stimulation (see Methods), and whether the peak overlaps a Formative (Column O) or Primed (Column P) peak.

Table S9: TOBIAS Predicted Binding Site Validation

This table includes information about the TOBIAS prediction validation analysis, including the ENCODE ID of the TF ChIP-seq data for H1 ESCs (Column A), the corresponding transcription factor (Column B), the p-value and the odds ratio for the Fisher's Exact tests (Columns C-D).

Table S10: Transcription Factor Group Motif Memberships

Information about TF groups determined by TOBIAS predicted binding similarities for 187 motifs, including the HOCOMOCO motif ID (Column A), the Gencode v34 gene ID (Column B), gene name (Column C) and the name of the collapsed TF group to which the motif belongs (Column D). **Note:** The TOBIAS motif distance matrix and predicted binding sites for all 187 motifs were uploaded to GEO (GSE_XXX) and FigShare (XXX).

Table S11: Annotations of 56,978 peaks for binding of 92 TF groups

This table includes the TOBIAS-predicted transcription factor binding sites for all 56,978 ATAC-seq peaks. Information includes; the peak ID (Column A), and binary annotations for the 92 collapsed TF groups and "Not Bound" peaks (Columns B-CP), where 1 indicates that there is a bound TF group on the corresponding peak.

Table S12: RNM Annotation Enrichment Results

This table contains the results for enrichments in the RNM Pareto peaks, including the annotation type (transcription factor, chromatin state, and cell state, Column A), the corresponding annotation (Column B), the tested RNM (Column C), the odds ratio, p-value, and corrected p-value (Columns D-F).

Table S13: RNM Fetal Tissue ATAC-seq Enrichment Results

This table contains the results for the single-cell fetal tissue-specific peak RNM enrichments, including the tested fetal cell type (Column A), the tested RNM (Column B), the odds ratio, p-value, and corrected p-value (Columns C-E).

Table S14: Fetal Tissue TFBS Enrichment Results

This table contains the results for the single-cell fetal tissue-specific peak TFBS enrichments, including the tested fetal cell type (Column A), the tested TFBS (Column B), the odds ratio, p-value, and corrected p-value (Columns C-E).

Table S15: Allele-Specific Chromatin Accessibility (ASCA) Results

This table contains information on SNPs tested for ASCA, including SNP ID and gnomad RSID (Columns A-B), Peak ID (Column C), number of reads mapping to the reference and alternative alleles (Columns D-E), the allelic imbalance fraction (Column F), the number of heterozygous individuals tested (Column G), the minor allele frequency (Column H), the p-value and Benjamini-Hochberg corrected p-value (Columns I-J), whether the SNP has ASCA (adjusted P-value < 0.05, Column K).

REFERENCES

1. Lau, K.X., Mason, E.A., Kie, J., De Souza, D.P., Kloehn, J., Tull, D., McConville, M.J., Keniry, A., Beck, T., Blewitt, M.E., et al. (2020). Unique properties of a subset of human pluripotent stem cells with high capacity for self-renewal. Nat. Commun. *11*, 2420. 10.1038/s41467-020-16214-8.

2. Hough, S.R., Thornton, M., Mason, E., Mar, J.C., Wells, C.A., and Pera, M.F. (2014). Single-cell gene expression profiles define self-renewing, pluripotent, and lineage primed states of human pluripotent stem cells. Stem Cell Rep. *2*, 881–895. 10.1016/j.stemcr.2014.04.014.

3. Mazid, M.A., Ward, C., Luo, Z., Liu, C., Li, Y., Lai, Y., Wu, L., Li, J., Jia, W., Jiang, Y., et al. (2022). Rolling back human pluripotent stem cells to an eight-cell embryo-like stage. Nature *605*, 315–324. 10.1038/s41586-022-04625-0.

4. Cornacchia, D., Zhang, C., Zimmer, B., Chung, S.Y., Fan, Y., Soliman, M.A., Tchieu, J., Chambers, S.M., Shah, H., Paull, D., et al. (2019). Lipid Deprivation Induces a Stable, Naive-to-Primed Intermediate State of Pluripotency in Human PSCs. Cell Stem Cell *25*, 120-136.e10. 10.1016/j.stem.2019.05.001.

5. Theunissen, T.W., Powell, B.E., Wang, H., Mitalipova, M., Faddah, D.A., Reddy, J., Fan, Z.P., Maetzel, D., Ganz, K., Shi, L., et al. (2014). Systematic Identification of Culture Conditions for Induction and Maintenance of Naive Human Pluripotency. Cell Stem Cell *15*, 524–526. 10.1016/j.stem.2014.09.003.

6. Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559. 10.1186/1471-2105-9-559.

7. Lemoine, G.G., Scott-Boyer, M.-P., Ambroise, B., Périn, O., and Droit, A. (2021). GWENA: gene co-expression networks analysis and extended modules characterization in a single Bioconductor package. BMC Bioinformatics *22*, 267. 10.1186/s12859-021-04179-4.

8. Yin, W., Mendoza, L., Monzon-Sandoval, J., Urrutia, A.O., and Gutierrez, H. (2021). Emergence of co-expression in gene regulatory networks. PloS One *16*, e0247671. 10.1371/journal.pone.0247671.

9. Xin, J., Zhang, H., He, Y., Duren, Z., Bai, C., Chen, L., Luo, X., Yan, D.-S., Zhang, C., Zhu, X., et al. (2020). Chromatin accessibility landscape and regulatory network of high-altitude hypoxia adaptation. Nat. Commun. *11*, 4928. 10.1038/s41467-020-18638-8.

10. Liu, Q., Jiang, C., Xu, J., Zhao, M.-T., Van Bortle, K., Cheng, X., Wang, G., Chang, H.Y., Wu, J.C., and Snyder, M.P. (2017). Genome-Wide Temporal Profiling of Transcriptome and Open Chromatin of Early Cardiomyocyte Differentiation Derived From hiPSCs and hESCs. Circ. Res. *121*, 376–391. 10.1161/CIRCRESAHA.116.310456.

11. Pliner, H.A., Packer, J.S., McFaline-Figueroa, J.L., Cusanovich, D.A., Daza, R.M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., et al. (2018). Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol. Cell *71*, 858-871.e8. 10.1016/j.molcel.2018.06.044.

12. Dong, K., and Zhang, S. (2021). Joint reconstruction of cis-regulatory interaction networks across multiple tissues using single-cell chromatin accessibility data. Brief. Bioinform. *22*, bbaa120. 10.1093/bib/bbaa120.

13. Kartha, V.K., Duarte, F.M., Hu, Y., Ma, S., Chew, J.G., Lareau, C.A., Earl, A., Burkett, Z.D., Kohlway, A.S., Lebofsky, R., et al. (2022). Functional inference of gene regulation using single-cell multi-omics. Cell Genomics *2*, 100166. 10.1016/j.xgen.2022.100166.

14. Kashyap, V., Rezende, N.C., Scotland, K.B., Shaffer, S.M., Persson, J.L., Gudas, L.J., and Mongan, N.P. (2009). Regulation of stem cell pluripotency and differentiation involves a mutual regulatory circuit of the

NANOG, OCT4, and SOX2 pluripotency transcription factors with polycomb repressive complexes and stem cell microRNAs. Stem Cells Dev. *18*, 1093–1108. 10.1089/scd.2009.0113.

15. Coronado, D., Godet, M., Bourillot, P.-Y., Tapponnier, Y., Bernat, A., Petit, M., Afanassieff, M., Markossian, S., Malashicheva, A., Iacone, R., et al. (2013). A short G1 phase is an intrinsic determinant of naïve embryonic stem cell pluripotency. Stem Cell Res. *10*, 118–131. 10.1016/j.scr.2012.10.004.

16. Li, M., and Belmonte, J.C.I. (2017). Ground rules of the pluripotency gene regulatory network. Nat. Rev. Genet. *18*, 180–191. 10.1038/nrg.2016.156.

17. Costa, Y., Ding, J., Theunissen, T.W., Faiola, F., Hore, T.A., Shliaha, P.V., Fidalgo, M., Saunders, A., Lawrence, M., Dietmann, S., et al. (2013). NANOG-dependent function of TET1 and TET2 in establishment of pluripotency. Nature *495*, 370–374. 10.1038/nature11925.

18. D'Antonio, M., Nguyen, J.P., Arthur, T.D., iPSCORE Consortium, Matsui, H., D'Antonio-Chronowska, A., and Frazer, K.A. (2023). Fine mapping spatiotemporal mechanisms of genetic variants underlying cardiac traits and disease. Nat. Commun. *14*, 1132. 10.1038/s41467-023-36638-2.

19. DeBoever, C., Li, H., Jakubosky, D., Benaglio, P., Reyna, J., Olson, K.M., Huang, H., Biggs, W., Sandoval, E., D'Antonio, M., et al. (2017). Large-Scale Profiling Reveals the Influence of Genetic Variation on Gene Expression in Human Induced Pluripotent Stem Cells. Cell Stem Cell *20*, 533-546.e7. 10.1016/j.stem.2017.03.009.

20. Chen, B., Khodadoust, M.S., Liu, C.L., Newman, A.M., and Alizadeh, A.A. (2018). Profiling Tumor Infiltrating Immune Cells with CIBERSORT. Methods Mol. Biol. Clifton NJ *1711*, 243–259. 10.1007/978-1-4939-7493-1_12.

21. Pera, M.F., and Rossant, J. (2021). The exploration of pluripotency space: Charting cell state transitions in peri-implantation development. Cell Stem Cell *28*, 1896–1906. 10.1016/j.stem.2021.10.001.

22. Li, M., and Izpisua Belmonte, J.C. (2018). Deconstructing the pluripotency gene regulatory network. Nat. Cell Biol. *20*, 382–392. 10.1038/s41556-018-0067-6.

23. Panopoulos, A.D., D'Antonio, M., Benaglio, P., Williams, R., Hashem, S.I., Schuldt, B.M., DeBoever, C., Arias, A.D., Garcia, M., Nelson, B.C., et al. (2017). iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types. Stem Cell Rep. *8*, 1086–1100. 10.1016/j.stemcr.2017.03.012.

24. Therneau, T. (2012). coxme: mixed effects Cox models.

25. Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. *9*, 5233. 10.1038/s41598-019-41695-z.

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

27. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabási, A.-L. (2000). The large-scale organization of metabolic networks. Nature *407*, 651–654. 10.1038/35036627.

28. Pareto, V. (1964). Cours d'économie politique. (Librairie Droz).

29. Stirparo, G.G., Boroviak, T., Guo, G., Nichols, J., Smith, A., and Bertone, P. (2018). Integrated analysis of single-cell embryo data yields a unified transcriptome signature for the human pre-implantation epiblast. Dev. Camb. Engl. *145*, dev158501. 10.1242/dev.158501.

30. Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216. 10.1038/nmeth.1906.

31. Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. Nat. Protoc. *12*, 2478–2492. 10.1038/nprot.2017.124.

32. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330. 10.1038/nature14248.

33. Bentsen, M., Goymann, P., Schultheis, H., Klee, K., Petrova, A., Wiegandt, R., Fust, A., Preussner, J., Kuenne, C., Braun, T., et al. (2020). ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. Nat. Commun. *11*, 4267. 10.1038/s41467-020-18035-1.

34. Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., et al. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. *46*, D252–D259. 10.1093/nar/gkx1106.

35. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. 10.1038/nature11247.

36. Cruz-Molina, S., Respuela, P., Tebartz, C., Kolovos, P., Nikolic, M., Fueyo, R., van Ijcken, W.F.J., Grosveld, F., Frommolt, P., Bazzi, H., et al. (2017). PRC2 Facilitates the Regulatory Topology Required for Poised Enhancer Function during Pluripotent Stem Cell Differentiation. Cell Stem Cell *20*, 689-705.e9. 10.1016/j.stem.2017.02.004.

37. Asenjo, H.G., Gallardo, A., López-Onieva, L., Tejada, I., Martorell-Marugán, J., Carmona-Sáez, P., and Landeira, D. (2020). Polycomb regulation is coupled to cell cycle transition in pluripotent stem cells. Sci. Adv. *6*, eaay4768. 10.1126/sciadv.aay4768.

38. Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell *125*, 315–326. 10.1016/j.cell.2006.02.041.

39. Lou, S., Li, T., Kong, X., Zhang, J., Liu, J., Lee, D., and Gerstein, M. (2020). TopicNet: a framework for measuring transcriptional regulatory network change. Bioinforma. Oxf. Engl. *36*, i474–i481. 10.1093/bioinformatics/btaa403.

40. Sethi, A., Gu, M., Gumusgoz, E., Chan, L., Yan, K.-K., Rozowsky, J., Barozzi, I., Afzal, V., Akiyama, J.A., Plajzer-Frick, I., et al. (2020). Supervised enhancer prediction with epigenetic pattern recognition and targeted validation. Nat. Methods *17*, 807–814. 10.1038/s41592-020-0907-8.

41. Karvas, R.M., David, L., and Theunissen, T.W. (2022). Accessing the human trophoblast stem cell state from pluripotent and somatic cells. Cell. Mol. Life Sci. *79*, 604. 10.1007/s00018-022-04549-y.

42. Dattani, A., Huang, T., Liddle, C., Smith, A., and Guo, G. (2022). Suppression of YAP safeguards human naïve pluripotency. Development *149*, dev200988. 10.1242/dev.200988.

43. Viukov, S., Shani, T., Bayerl, J., Aguilera-Castrejon, A., Oldak, B., Sheban, D., Tarazi, S., Stelzer, Y., Hanna, J.H., and Novershtern, N. (2022). Human primed and naïve PSCs are both able to differentiate into trophoblast stem cells. Stem Cell Rep. *17*, 2484–2500. 10.1016/j.stemcr.2022.09.008.

44. Wei, Y., Wang, T., Ma, L., Zhang, Y., Zhao, Y., Lye, K., Xiao, L., Chen, C., Wang, Z., Ma, Y., et al. (2021). Efficient derivation of human trophoblast stem cells from primed pluripotent stem cells. Sci. Adv. *7*, eabf4416. 10.1126/sciadv.abf4416.

45. Boward, B., Wu, T., and Dalton, S. (2016). Concise Review: Control of Cell Fate Through Cell Cycle and Pluripotency Networks. Stem Cells *34*, 1427–1436. 10.1002/stem.2345.

46. Deb-Rinker, P., Ly, D., Jezierski, A., Sikorska, M., and Walker, P.R. (2005). Sequential DNA methylation of the Nanog and Oct-4 upstream regions in human NT2 cells during neuronal differentiation. J. Biol. Chem. *280*, 6257–6260. 10.1074/jbc.C400479200.

47. Po, A., Ferretti, E., Miele, E., De Smaele, E., Paganelli, A., Canettieri, G., Coni, S., Di Marcotullio, L., Biffoni, M., Massimi, L., et al. (2010). Hedgehog controls neural stem cells through p53-independent regulation of Nanog. EMBO J. *29*, 2646–2658. 10.1038/emboj.2010.131.

48. Chong, J.A., Tapia-Ramírez, J., Kim, S., Toledo-Aral, J.J., Zheng, Y., Boutros, M.C., Altshuller, Y.M., Frohman, M.A., Kraner, S.D., and Mandel, G. (1995). REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. Cell *80*, 949–957. 10.1016/0092-8674(95)90298-8.

49. Chen, D., Zhao, M., and Mundy, G.R. (2004). Bone morphogenetic proteins. Growth Factors Chur Switz. *22*, 233–241. 10.1080/08977190412331279890.

50. Morgani, S.M., and Hadjantonakis, A.-K. (2020). Signaling regulation during gastrulation: Insights from mouse embryos and in vitro systems. Curr. Top. Dev. Biol. *137*, 391–431. 10.1016/bs.ctdb.2019.11.011.

51. Domcke, S., Hill, A.J., Daza, R.M., Cao, J., O'Day, D.R., Pliner, H.A., Aldinger, K.A., Pokholok, D., Zhang, F., Milbank, J.H., et al. (2020). A human cell atlas of fetal chromatin accessibility. Science *370*, eaba7612. 10.1126/science.aba7612.

52. Su, Z., Zhang, Y., Liao, B., Zhong, X., Chen, X., Wang, H., Guo, Y., Shan, Y., Wang, L., and Pan, G. (2018). Antagonism between the transcription factors NANOG and OTX2 specifies rostral or caudal cell fate during neural patterning transition. J. Biol. Chem. *293*, 4445–4455. 10.1074/jbc.M117.815449.

53. Naama, M., Rahamim, M., Zayat, V., Sebban, S., Radwan, A., Orzech, D., Lasry, R., Ifrah, A., Jaber, M., Sabag, O., et al. (2023). Pluripotency-independent induction of human trophoblast stem cells from fibroblasts. Nat. Commun. *14*, 3359. 10.1038/s41467-023-39104-1.

54. Walentin, K., Hinze, C., Werth, M., Haase, N., Varma, S., Morell, R., Aue, A., Pötschke, E., Warburton, D., Qiu, A., et al. (2015). A Grhl2-dependent gene network controls trophoblast branching morphogenesis. Dev. Camb. Engl. *142*, 1125–1136. 10.1242/dev.113829.

55. Avery, S., Hirst, A.J., Baker, D., Lim, C.Y., Alagaratnam, S., Skotheim, R.I., Lothe, R.A., Pera, M.F., Colman, A., Robson, P., et al. (2013). BCL-XL mediates the strong selective advantage of a 20q11.21 amplification commonly found in human embryonic stem cell cultures. Stem Cell Rep. *1*, 379–386. 10.1016/j.stemcr.2013.10.005.

56. Merkle, F.T., Ghosh, S., Genovese, G., Handsaker, R.E., Kashin, S., Meyer, D., Karczewski, K.J., O'Dushlaine, C., Pato, C., Pato, M., et al. (2022). Whole-genome analysis of human embryonic stem cells enables rational line selection based on genetic variation. Cell Stem Cell *29*, 472-486.e7. 10.1016/j.stem.2022.01.011.

57. Andrews, P.W., Barbaric, I., Benvenisty, N., Draper, J.S., Ludwig, T., Merkle, F.T., Sato, Y., Spits, C., Stacey, G.N., Wang, H., et al. (2022). The consequences of recurrent genetic and epigenetic variants in human pluripotent stem cells. Cell Stem Cell *29*, 1624–1636. 10.1016/j.stem.2022.11.006.

58. Degtyareva, A.O., Antontseva, E.V., and Merkulova, T.I. (2021). Regulatory SNPs: Altered Transcription Factor Binding Sites Implicated in Complex Traits and Diseases. Int. J. Mol. Sci. *22*, 6454. 10.3390/ijms22126454.

59. Benaglio, P., D'Antonio-Chronowska, A., Ma, W., Yang, F., Young Greenwald, W.W., Donovan, M.K.R., DeBoever, C., Li, H., Drees, F., Singhal, S., et al. (2019). Allele-specific NKX2-5 binding underlies multiple genetic associations with human electrocardiographic traits. Nat. Genet. *51*, 1506–1517. 10.1038/s41588-019-0499-3.

60. Johnston, A.D., Simões-Pires, C.A., Thompson, T.V., Suzuki, M., and Greally, J.M. (2019). Functional genetic variants can mediate their regulatory effects through alteration of transcription factor binding. Nat. Commun. *10*, 3472. 10.1038/s41467-019-11412-5.

61. Abramov, S., Boytsov, A., Bykova, D., Penzar, D.D., Yevshin, I., Kolmykov, S.K., Fridman, M.V., Favorov, A.V., Vorontsov, I.E., Baulin, E., et al. (2021). Landscape of allele-specific transcription factor binding in the human genome. Nat. Commun. *12*, 2751. 10.1038/s41467-021-23007-0.

62. Endo, Y., Kamei, K.-I., and Inoue-Murayama, M. (2020). Genetic Signatures of Evolution of the Pluripotency Gene Regulating Network across Mammals. Genome Biol. Evol. *12*, 1806–1818. 10.1093/gbe/evaa169.

63. Yagi, R., Kohn, M.J., Karavanova, I., Kaneko, K.J., Vullhorst, D., DePamphilis, M.L., and Buonanno, A. (2007). Transcription factor TEAD4 specifies the trophectoderm lineage at the beginning of mammalian development. Development *134*, 3827–3836. 10.1242/dev.010223.

64. Kumar, B., Navarro, C., Winblad, N., Schell, J.P., Zhao, C., Weltner, J., Baqué-Vidal, L., Salazar Mantero, A., Petropoulos, S., Lanner, F., et al. (2022). Polycomb repressive complex 2 shields naïve human pluripotent cells from trophectoderm differentiation. Nat. Cell Biol. *24*, 845–857. 10.1038/s41556-022-00916-w.

65. Nguyen, J.P., Arthur, T.D., Fujita, K., Salgado, B.M., Donovan, M.K.R., iPSCORE Consortium, Matsui, H., D'Antonio-Chronowska, A., D'Antonio, M., and Frazer, K.A. (2021). Disease-associated regulatory variation often displays plasticity or temporal-specificity in fetal pancreas (Genomics) 10.1101/2021.03.17.435846.

66. Jakubosky, D., D'Antonio, M., Bonder, M.J., Smail, C., Donovan, M.K.R., Young Greenwald, W.W., Matsui, H., i2QTL Consortium, D'Antonio-Chronowska, A., Stegle, O., et al. (2020). Properties of structural variants and short tandem repeats associated with gene expression and complex traits. Nat. Commun. *11*, 2927. 10.1038/s41467-020-16482-4.

67. Smith, E.N., Jepsen, K., Arias, A.D., Shepard, P.J., Chambers, C.D., and Frazer, K.A. (2014). Genetic ancestry of participants in the National Children's Study. Genome Biol. *15*, R22. 10.1186/gb-2014-15-2-r22.

68. Luo, Y., Hitz, B.C., Gabdank, I., Hilton, J.A., Kagda, M.S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., et al. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. *48*, D882–D889. 10.1093/nar/gkz1062.

69. Ruiz, S., Brennand, K., Panopoulos, A.D., Herrerías, A., Gage, F.H., and Izpisua-Belmonte, J.C. (2010). High-Efficient Generation of Induced Pluripotent Stem Cells from Human Astrocytes. PLoS ONE *5*, e15526. 10.1371/journal.pone.0015526.

70. Panopoulos, A.D., D'Antonio, M., Benaglio, P., Williams, R., Hashem, S.I., Schuldt, B.M., DeBoever, C., Arias, A.D., Garcia, M., Nelson, B.C., et al. (2017). iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types. Stem Cell Rep. *8*, 1086–1100. 10.1016/j.stemcr.2017.03.012.

71. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinforma. Oxf. Engl. *29*, 15–21. 10.1093/bioinformatics/bts635.

72. Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773. 10.1093/nar/gky955.

73. Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. Bioinformatics *31*, 2032–2034. 10.1093/bioinformatics/btv098.

74. Tischler, G., and Leonard, S. (2014). biobambam: tools for read pair collation based algorithms on BAM files. Source Code Biol. Med. *9*, 13. 10.1186/1751-0473-9-13.

75. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323. 10.1186/1471-2105-12-323.

76. Nariai, N., Greenwald, W.W., DeBoever, C., Li, H., and Frazer, K.A. (2017). Efficient Prioritization of Multiple Causal eQTL Variants via Sparse Polygenic Modeling. Genetics *207*, 1301–1312. 10.1534/genetics.117.300435.

77. D'Antonio-Chronowska, A., Donovan, M.K.R., Young Greenwald, W.W., Nguyen, J.P., Fujita, K., Hashem, S., Matsui, H., Soncin, F., Parast, M., Ward, M.C., et al. (2019). Association of Human iPSC Gene Signatures and X Chromosome Dosage with Two Distinct Cardiac Differentiation Trajectories. Stem Cell Rep. *13*, 924–938. 10.1016/j.stemcr.2019.09.011.

78. Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol. *109*, 21.29.1-21.29.9. 10.1002/0471142727.mb2129s109.

79. Gögelein, H., and Hüby, A. (1984). Interaction of saponin and digitonin with black lipid membranes and lipid monolayers. Biochim. Biophys. Acta *773*, 32–38. 10.1016/0005-2736(84)90547-9.

80. Garcia-Ruiz, C., Mari, M., Colell, A., Morales, A., Caballero, F., Montero, J., Terrones, O., Basañez, G., and Fernández-Checa, J.C. (2009). Mitochondrial cholesterol in health and disease. Histol. Histopathol. *24*, 117–132. 10.14670/HH-24.117.

81. van Meer, G., Voelker, D.R., and Feigenson, G.W. (2008). Membrane lipids: where they are and how they behave. Nat. Rev. Mol. Cell Biol. *9*, 112–124. 10.1038/nrm2330.

82. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. 10.1093/bioinformatics/btp352.

83. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137. 10.1186/gb-2008-9-9-r137.

84. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinforma. Oxf. Engl. *30*, 923–930. 10.1093/bioinformatics/btt656.

85. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140. 10.1093/bioinformatics/btp616.

86. Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. *9*, 9354. 10.1038/s41598-019-45839-z.

87. Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. Nat. Protoc. *12*, 2478–2492. 10.1038/nprot.2017.124.

88. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. 10.1186/s13059-014-0550-8.

89. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. 10.1086/519795.

90. Lancichinetti, A., and Fortunato, S. (2011). Limits of modularity maximization in community detection. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *84*, 066122. 10.1103/PhysRevE.84.066122.

91. van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J.K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat. Methods *12*, 1061–1063. 10.1038/nmeth.3582.

92. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples (Genomics) 10.1101/201178.

93. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. 10.1038/s41586-020-2308-7.