TITLE

Unsupervised Machine Learning Identifies Chromatin Accessibility Regulatory Networks that Define Cell State Transitions in Pluripotency

AUTHORS

Timothy D. Arthur[1,2], Jennifer P. Nguyen[2,3], Agnieszka D'Antonio-Chronowska[4], Hiroko Matsui[5], Nayara Silva[4,5], Isaac N. Joshua[5], iPSCORE Consortium[*], William W. Young Greenwald[3], Matteo D'Antonio[2, 5], Martin F. Pera[6], 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]The Jackson Laboratory, Bar Harbor, ME 04609, USA

*Correspondence: kafrazer@health.ucsd.edu

SUMMARY

Stem cells exist *in vitro* in a spectrum of interconvertible pluripotent cell states. Understanding the genetic and epigenetic regulatory processes underlying cell state transitions between these pluripotency states will have broad applications. Here, we applied a machine learning algorithm to analyze RNA-seq and ATAC-seq data derived from hundreds of human induced pluripotent stem cells (hiPSCs) resulting in the discovery of 24 gene network modules (GNMs) and 20 regulatory network modules (RNMs). Characterization of the network modules revealed that the GNMs and RNMs were highly correlated with each other and enabled us to decipher the roles that individual modules play in pluripotency and self-renewal. Genetic analyses identified regulatory variants that disrupted transcription factor binding and were associated with both decreased co-accessibility of regulatory elements within an RNM and increased stability of a particular pluripotency state. Our findings uncover novel pluripotency regulatory mechanisms and provide a rich resource for future stem cell research.

KEYWORDS

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

INTRODUCTION

Characterizing the molecular mechanisms underlying self-renewal, pluripotency and cell state transitions 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, with the fractions of these subpopulations varying across lines[1,3]. Biological networks are constructed by looking for pairs of molecular elements that exhibit similar expression patterns across hundreds of samples, and community detection algorithms identify modules of elements involved in common biological processes. Therefore, examining the co-expression of genes and genome-wide co-accessibility of regulatory elements across hundreds of hiPSC lines using network and unsupervised community detection algorithms could provide novel insights into self-renewal and pluripotency cell state-specific network modules, and how genetic background contributes to variability in network activity across cell lines.

The increasing use of hiPSC to model early human embryonic development has wide-ranging implications for our understanding of human embryology and the factors required to establish and maintain a healthy pregnancy, and for the production of tissues and organs from stem cell sources[4]. The fidelity of these models depends on a clear definition of the composition of the cell population used to initiate the experiment. Recent studies have shown that hiPSCs lines are mosaics, composed of varying proportions of cells in different, interconvertible pluripotent cell states[5] that are stabilized under different culturing conditions. Live cell staining[1] and single-cell analyses[2,6–11] in hiPSCs have revealed that the interconvertible pluripotent cell states have morphological, transcriptional, and epigenetic profiles that correspond to different distinct stages of embryonic development. Most cells within a conventional hiPSC culture resemble the late post-implantation epiblast stem cells[1] and their pluripotency cell state is referred to as 'primed' because lineage-specific genes are poised to rapidly transition from a bivalent to active state. Under different culture conditions, hiPSCs can be maintained in a state that resembles the preimplantation epiblast, referred to as naive pluripotency. Finally, an intermediate cellular state in hiPSCs, referrred to as 'formative' pluripotency, is found as a subpopulation in conventional hiPSC cultures, and can also be stabilized by modifications to the culture system. 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]. Effective use of hiPSC as embryo models requires a clear definition of the substates of pluripotency represented in culture and the molecular machinery that governs state transitions, and how both vary between cell lines derived from different genetic backgrounds. This is best approached through the study of hundreds of hiPSC lines cultured under indentical conditions.

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[12–14]. However, network-based algorithms have not yet been used to analyze ATAC-seq[15] data to identify co-accessible regulatory elements distributed across the genome. While co-expression has been conjectured to result from genes sharing common regulators (i.e., regulatory

elements that bind similar transcription factors), this hypothesis has not yet been experimentally tested due to the lack of suitable datasets. Therefore, it is important to discover chromatin co-accessibile regulatory networks with an unsupervised approach, in order to examine their correlation with gene co-expression networks and underlying elements for shared epigenomic features. Expression networks represent genes as nodes and edges convey information about transcriptome-wide co-expression of the genes. A regulatory element network could represent ATAC-seq peaks as nodes with edges representing genome-wide co-accessibility of the peaks. The ability to bin the precise locations and epigenetic profiles of all elements into descrete regulatory networks in hiPSCs would provide novel insights into important cellular processes including self-renewal, pluripotency and cell cycle transitions.

In this study, 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 RNA-seq dataset[16,17] for 213 hiPSC lines also in iPSCORE. 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 regulatory elements in 20 high resolution regulatory networks associated with distinct pluripotency states in hiPSCs; and examine associations between regulatory variants that disrupt TF binding sites and the stability of different pluripotency cell states.

RESULTS

Study Overview

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 iPSCORE individuals (Figure 1A). We hypothesized that that these regulatory network modules could be identified due to heterogenity arising from variation in the proportions of pluripotent subpopulations (Figure S1) across hiPSC lines, and would underlie coordinated expression of gene network modules. We generated 150 ATAC-seq samples, calculated coverage at 59,499 reference peaks and then determined genome-wide co-accessibility between all pairwise combinations of the reference peaks. Likewise for 213 RNA-seq samples, we determined co-expression between 16,110 genes. We then applied an unsupervised machine learning algorithm independently to each dataset which detected 20 regulatory network modules (RNMs) and 24 gene network modules (GNMs). We demonstrate that the both the RNMs and GNMs are 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 cellular states across the hiPSC samples.

To functionally characterize the RNMs we annotated the ATAC-seq peaks with chromatin states[18,19] and transcription factor (TF) binding sites[20–22] (Figure 1B). We then annotated the peaks for association with the formative pluripotency subpopulation by leveraging a previously published dataset of early postimplantation epiblast-like (EPE)- associated ATAC-seq peaks[1]. 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 20 discrete functionally characterized regulatory networks whose elements tend to share similar TF binding, chromatin state profiles and pluripotency state enrichments.

We further show that the 20 RNMs largely cluster into two groups likely representing the genome-wide networks in two pluripotency states (primed and formative) that most likely control pluripotency, self-renewal, and other cellular processes (Figure 1C). 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). Finally, we investigate genetic regulatory variation that disrupts local chromatin accessibility and show that SNPs associated with allele specific chromatin accessibility (ASCA-SNPs) can also impact genome-wide co-accessibility of regulatory elements and be associated with increased stability of defined pluripotency states.

Unsupervised discovery of gene co-expression network modules

To identify high resolution modules of co-expressed genes, we analyzed RNA-seq generated from 213 hiPSCs (Figure 1, Table S1) and calculated the pairwise correlation between 16,110 expressed autosomal genes (TPM ≥ 1 in at least 20% samples) using a linear mixed model (LMM)[17,23,24]. We identified 3,533,609 co-expressed gene pairs (adjusted P-value < 0.05) with positive associations and created a global gene network (GN) using these gene pairs as edges and the effect size of the association as weights. We then applied the unsupervised Leiden community detection algorithm[25,26] to the GN which resulted in the identification of 33 high resolution gene co-expression network modules (GNMs) consisting of between 10 to 1,204 genes (Figure S2A, Table S2, Table S3) and 1,199 genes that were not co-expressed with any other gene. GNM memberships consist of genes in molecular pathways that are characteristic of different stem cell states; for example, GNM 14 consists of frizzled receptors, GTPase Coupled Receptors and genes associated with early neural specification (Figure 2A), all characteristic of the primed pluripotent stem cell state[1,10]; and GNM 32 consists of genes in Nodal signaling, ERK signaling, self-renewal processes, extracellular (ECM) matrix formation and Wnt signaling pathways (Figure 2B), characteristic of formative and naïve pluripotent stem cell states[1].

We examined if the GNMs were differentially expressed across the hiPSC lines due to varying proportions of subpopulations with different pluripotency cell states (Figure S1). As pluripotency cell states are defined by complex biological processes that involve many genes, we focused our downstream analyses on 24 GNMs with a membership of at least 100 genes (mean = 599.9±265.0 genes, Figure S2A). 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 analyzed the expression of 2,862 GNM-specific Pareto genes (top 20 percentile of intra-GNM degree connectivity) and observed that the network properties correspond to GNM-specific clustering in UMAP space (Figure 2C, Figure S2B), indicating GNM-specific Pareto genes capture highly co-expressed genes with similar expression profiles across the 213 hiPSC lines. To further assess differential expression of GNMs across the hiPSCs, we calculated GNM scores by ranking the sum of GNM-specific Pareto gene expression (quantile-normalized TPM) in each of the 213 RNA-seq samples, which clearly showed that some GNMs were positively correlated while others were negatively correlated with each other (Figure S3). For example, GNM 32 and GNM 14 expression were negatively correlated (Figure 2D-F) indicating that they had opposite expression patterns across the 213 hiPSCs.

After determining that the GNMs were differentially expressed across the hiPSC lines, we sought to better characterize the nature of their enrichments for different pluripotency cell states (Figure 2G-J, Table S4). Utilizing marker genes for hiPSC cell states defined in previous studies, we discovered that four GNMs (19, 27, 36, and 37) were associated with the 8-cell like cells (8CLC) totipotent cell state[3] (Figure 2G); the naïve[6], epiblast[6] and formative (EPE)[1] cell states were all enriched for GNM 32; the primitive endoderm (PrE)[6] and conventional hiPSCs[6] were enriched for both GNM 32 and GNM 14; while the general population[1] was enriched for GNM 14. Examining correlations of the 24 GNMs revealed two major groupings (Figure 2H, Table S5), one of which contained GNM 32 and the other which contained GNM 14. We also performed functional enrichment analysis of the GNMs using the MSigDB C5 biological process gene sets[29,30] (Figure 2I, Table S4) and observed enrichments concordant with previous characterizations of pluripotency cell states[1,6], including, the associations of both GNMs 14 and 32 with non-specific developmental terms, and the strong enrichment of GNM 32 for genes involved in primary germ layer formation and cell fate commitment. Altogether, these findings show that the profiled hiPSCs were composed of subpopulations of naïve to primed pluripotency states (Figure S1).

Previous studies have suggested that self-renewal might be a function of the formation of a critical mass of cells expressing survival or growth factors at a sufficiently high local level[1]. ROCK inhibitor, which is commonly used in hiPSC maintenance and single cell plating during differentiation protocols, increases hiPSC cell survival while sustaining their stemness characteristics including self-renewal and differentiation capacity[31]. To characterize ROCK inhibitor's effect on gene expression in hiPSCs, we compared RNA-seq data from 3 iPSCORE hiPSC lines cultured with or without ROCK inhibitor (Figure 2J, Table S6). We identified 49 genes upregulated and 116 genes downregulated (FDR < 0.05 and $-1 > \log_2(FC) > 1$) after culturing with ROCK inhibitor and observed that many of the upregulated genes, such as *NODAL, LEFTY1*, *LEFTY2* and *DACT1*, were previously reported to be associated with the formative (EPE) cell state[1]. Furthermore, the upregulated genes were enriched for GNM 32 while the downregulated genes were enriched for GNM 14 (Figure 2G). These data indicate that culturing with a ROCK inhibitor shifts the composition of pluripotency cell states decreasing the number of cells in a primed-like state and increasing the number in the formative-like state.

Altogether, we identified 24 modules of co-expressed genes associated with the differential expression of marker genes defining hiPSC interconvertible pluripotency cell states and showed that their discovery was due to differences in the proportion of these pluripotency cell states across the 213 hiPSC lines.

Unsupervised discovery of genome-wide regulatory element network 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 3A). We generated 150 independent ATAC-seq libraries from 143 hiPSC lines (expanded with ROCK inhibitor, see Methods) derived from 133 iPSCORE individuals (7 lines were cultured independently twice and 5 individuals each had two or three independent clones) (Table S7).

To identify high resolution modules of co-accessible regulatory elements, we calculated the accessibility of 59,499 peaks across the 150 ATAC-seq samples (Figure 1, Figure S4A-D, Table S8) and applied an LMM to identify pairs of co-accessible ATAC-seq peaks. We created a weighted genome-wide regulatory network (RN) using 8,661,299 co-accessible

autosomal ATAC-seq peak pairs (P-value < $5x10^{-8}$, Effect Size > 0) as edges and the effect size as weights. We applied the unsupervised Leiden community detection algorithm[24,25] to the weighted RN and identified 20 regulatory network modules (RNMs) each consisting of at least 500 ATAC-seq peaks (mean = 2,246.95 ± 1,271.45 peaks) (Figure 3B). We analyzed the expression of 8,972 Pareto peaks (top 20% intra-RNM degree connectivity) and observed RNM-specific clustering in the UMAP space (Figure 3C) indicating that the RNMs capture ATAC-seq peaks with similar varying accessibility across the hiPSC lines. To further assess differential accessibility of RNMs across the hiPSCs, we calculated RNM scores by ranking the sum of quantile normalized Pareto peak accessibility for each RNM independently in each of the 150 ATAC-seq samples (Figure 3D-E, Figure S5), which clearly showed that accessibility in RNMs can be negatively or positively correlated with each other across the hiPSCs.

To examine whether genome-wide ATAC-seq peak co-accessibility underlies coordinated gene expression, we annotated 22,936 ATAC-seq peaks in the 20 RNMs with 11,121 neighboring candidate target genes (See Methods) and then performed a permutation test (n = 1000) to calculate enrichments of RNMs in GNMs. We found that 85% of the RNMs had an association with at least one GNM (Figure 3F). For example, five RNMs (10, 22, 24, 28 and 30) were positively associated with GNM 32 (Figure 2B, Figure 2G-I) while three RNMs (14, 18 and 21) were associated GNM 14 (Figure 2A, Figure 2G-I) suggesting that these regulatory networks are active in different pluripotency states.

Altogether, these data show that we discovered 20 genome-wide regulatory network modules by examining chromatin accessibility across 143 hiPSCs because the lines contained varying proportions of pluripotency subpopulations; and that these regulatory modules were strongly associated with, and likely mechanistically underlie, the coordinated expression within genes 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 regards to their expression profiles and biological functions, regulatory elements have not been well characterized. Therefore, in order to functionally characterize the 20 RNMs, we annotated the ATAC-seq peaks with three epigenetic annotations (Figure 4A): 1) hiPSC-specific chromatin states[18,19,32], 2) TF binding sites[20–22], and 3) formative (EPE)-associated ATAC-seq peaks[1].

We initially annotated each of the 59,499 reference ATAC-seq peaks with chromatin states (See Methods, Table S8) and observed that 48.1% of the ATAC-seq peaks were in active regulatory elements (active promoters;TssA, enhancers;Enh), 12.4% were in bivalent or poised chromatin (bivalent promoters;TssBiv, regions flanking bivalent marks;BivFlnk and bivalent ehancers; EnhBiv), 4.4% were in repressed polycomb regions (represssed polycomb; ReprPC and weak repressed polycomb; ReprPCWk) and 7.5% were in transcribed regions (Tx and TxWk) (Figure 4B). Our annotations were consistent with previous characterizations of hiPSC regulatory elements[1,33–35], specifically; 1) the relatively large fraction of the peaks in bivalent chromatin indicating open but inactive regulatory elements, 2) the presence of peaks in polycomb regions, and 3) the fact that while the vast majority of the genome is quiescent chromatin[32], only 20.4% of peaks were in these regions indicating a strong depletion in inactive chromatin.

We identified 187 TFs that were expressed in the hiPSCs and predicted their binding across the 59,499 reference ATAC-seq peaks using TOBIAS, a digital footprinting method[20] (Table S9). To validate the TOBIAS predictions, we used ENCODE ChIP-seq data for 18 TFs generated using H1 embryonic stem cells (ESCs)[22,36,37] (See Methods, Figure S6, Table S10). We observed that TOBIAS predicted binding sites were strongly enriched in corresponding TF ChIP-seq peaks (Fisher's Exact test; maximum nominal $p = 5.5x10^{-188}$), 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 S7, Figure S8) into 92 TF groups (49 single motif, 24 same TF family, 19 contained TFs from different families and are referred to as "Complexes") (Table S9, Table S11). Among the 19 TF Complexes several were composed of TFs with well-known interactions or competitive binding including: 1) NANOG and POU5F1/OCT4 (NANOG-OCT4 Complex, Figure S7B)[38,39], 2) SOX2, SOX3, SOX4 and LEF1 (SOX-LEF1 Complex, Figure S7C)[40], 3) ARNTL, BHLHE40, MYC, MXI1, MAX, MYCN (MYC-MAX-MXI1 Complex, Figure S7D)[41], and 4) FOS, FOSL1, FOSL2, JUN, JUNB, JUND (FOS-JUN Complex. Figure S7E)[42]. We annotated the TF groups with one of six biological functions curated through literature review and observed that functionally distinct TFs are bound in different proportions of the 59,499 reference ATAC-seq peaks (Figure 4C, Figure S8); for instance, TFs involved in transcriptional regulation (EGR-ZnF Complex, Kruppel Family), chromatin organization (CTCF Family) and cell cycle (SP-E2F4, E2F Family) were bound in a high fraction of ATAC-seq peaks (38-45%), while those involved in pluripotency (NANOG-OCT4 Complex, SOX-LEF1 Complex and POU2F2/OCT2) were bound in a lower fraction (6.5-25.5%). Approximately 22% of the 59,499 reference ATAC-seq peaks were predicted to be not bound by TFs.

To assess whether certain TF groups were more likely to bind in the same ATAC-seq peaks, we performed Fisher's Exact tests on all possible pairs of the 92 TF groups. We observed that distinct TF groups clustered together: 1) pluripotency-associated TFs (NANOG-OCT4 Complex, POU2F2/OCT2, and TEAD Family); 2) cell cycle regulatory TFs (SP-E2F4 complex, Kruppel Family and E2F Family), and 3) developmental TFs (SMAD Family, REST, ETS Family 1 and 2) (Figure 4D, Figure S9). To assess whether certain TFs preferentially bind chromatin in different states, we calculated enrichment with Fisher's Exact test (Figure 4E, Figure S10). We observed that cell cycle-associated TFs tended to be more strongly enriched in either active promoters (TssA: ETS Family 1, E2F2, and E2F5) or bivalent promoters (TssBiv: SP-E2F4 complex, Kruppel Family and E2F Family), CTCF Family was enriched in repressed chromatin (ReprPC, ReprPCWk and Quies) and regions with active transcription (Tx and TxWk), while pluripotency-associated TFs were enriched in active enhancers (Enh: NANOG-OCT4 Complex, POU2F2, TEAD Family), bivalent enhancers (EhnBiv, specifically NANOG-OCT4 Complex) and regions flanking active promoters (TssAFlnk). Altogether these data indicate that across the genome cell cycle-associated TFs co-bind with each other at active and bivalent promoters, while pluripotency-associated TFs show a stronger affinity for co-binding at upstream regulatory elements.

We leveraged previous work characterizing the formative pluripotency state[1] to annotate 1,665 ATAC-seq peaks as formative-associated and 4,029 peaks as non-formative (GCTM-2^{mid}-$CD9^{mid}$), because they overlapped peaks specific to a hiPSC subpopulations that exhibited either high self renewal capabilities and other properties of the early post-implantation epiblast (Table S8) or a subpopulation which resembled the primed state of pluripotency. We performed a UMAP analysis

on the formative/non-formative annotated peaks that were also Pareto peaks and observed two distinct clusters (Figure 4F), further validating that our dataset capture pluripotency state specific regulatory elements.

Integration of hiPSC ATAC-seq peak functional annotations

To demonstrate the utility of integrating all three annotations (chromatin state, predicted TFBSs and pluripotency state), we plotted a 6MB interval on chromosome 4 surrounding *BMPR1B* (bone morphogenetic protein receptor type 1B), which is a receptor for BMP ligands that regulate ealy cell fate decisions during specification of trophoblast and later, gastrulation[43–45]. We observed four formative (EPE)-associated peaks and multiple peaks binding SMAD Family (SMAD2 and SMAD4) and SOX-LEF1 Complex (SOX2, SOX3, SOX4, LEF1) TFs (Figure 5A). Focusing on the ~100kb interval harboring the 5' end of *BMPR1B*, we observed three ATAC-seq peaks, of which one is in RNM 10, formative (EPE)-associated and upstream of the promoter; one is in RNM 22, bound by a SMAD Family TF and overlaps the bivalent promoter; and one is in RNM 4, bound by a SOX-LEF1 Complex TF and in an intronic region (Figure 5B). RNM 10 was associated with GNM 32, which was active in formative-like states, suggesting that ATAC-seq peaks in RNM 10 contain regulatory elements associated with pluripotency processes (Figure 2G-I, Figure 3F). The RNM 22 ATAC-seq peak overlapping the bivalent promoter was shown to be bound by SMAD Family TFs which are key regulators of embryogenesis through modulation of BMPs and other TGFβ signalling pathways[46]. Altogether these observations demonstrate the utility of annotating ATAC-seq peaks with epigenomic and cell state-specific features.

Regulatory network modules form two clusters associated with distinct pluripotency cell states

To further characterize the 20 RNMs we calculated their enrichments for formative-state ATAC-seq peaks (Figure 5C), the 15 chromatin states (Figure 5E), and the 92 TF groups including "Not Bound" peaks (Figure 5F; Figure S11, Table S12). We also calculated the pairwise correlations between the 20 RNMs using the mean accessibility of the 8,972 Pareto peaks (Figure 5D, Table S13). These analyses revealed two major RNM clusters corresponding to the regulatory networks underlying different pluripotency states and two RNMs that did not cluster in either of these two pluripotency states (Figure 5D).

We initially examined the larger cluster (Cluster A) composed of eleven RNMs including all four RNMs (10, 22, 28, 30) enriched for formative-state peaks. RNM 30 was enriched for pluripotency TFs (NANOG-OCT4 complex, POU2F2/OCT2, SOX-LEF1 complex) in active and bivalent upstream regulatory elements (Enh, TssAFlnk, EnhBiv), suggesting that it represents the hallmark hiPSC pluripotency regulatory network active in the formative state (Figure 5E-G). RNM 10 was highly enriched for active upstream regulatory elements (Enh, TssAFlnk) and the TEAD Family (TEAD1 and TEAD4) (Figure 5E-F, H, Figure S7B). TEAD signaling is strongly implicated in the differentiation of hPSC to the trophoblast lineage[44], and recently Dattani et al.[47] showed that suppression of YAP/TEAD signaling was critical to insulation of naive hPSC from trophoblast differentiation. Though the capacity for conventional hPSC to undergo trophoblast differentiation has been the subject of considerable controversy[44,48], it is certainly possible that EPE cells, closer to the naive state, might be capable to enter this differentiation pathway, and that sites in RMN 10 are in some way primed for activation in naive and EPE cell types [49]. The formative-state has a highly proliferative phenotype and an abbreviated G_1 phase[1,50], which is

consistent with the five RNMs (3, 7, 8, 22, 28) in Cluster A enriched for cell cycle-associated TFs (E2F Family, E2F2, E2F5, SP-E2F Complex) that co-bind with active (TssA) and bivalent promoters (TssBiv, Figure 4E, D, Figure 5F). RNMs 7 and 8, which are enriched in TssA, but not TssBiv (Figure 5E), likely underlie the mechanism utilized by highly pluripotent stem cells to rapidly traverse G_1 phase[50]. RNMs 3, 18 and 22 are enriched with TssBiv (Figure 5E, I), which is consistent the vital role of developmental gene repression in highly pluripotent stem cells[51]. RNM 18 was also enriched for ATAC-seq peaks in repressed polycomb (ReprPC, Figure 4E) which is a well-documented mechanism for maintaining naïve pluripotency by restricting differentiation[52]. Of note, there were three RNMs (4, 28, 20) in Cluster A enriched for unbound regulatory elements or CTCF Family in quiescent regions and depleted for other chromatin marks (Quies, Figure 4E). Overall, these analyses show that the RNMs in Cluster A encompass well known epigenomic and cell state-specific features present in formative to naïve cell states and importantly captured the coordinated activity of the majority of regulatory elements underpinning the spectrum of pluripotency traits in these cells.

The smaller cluster (Cluster B, Figure 5D) contained seven RNMs several of which had similar epigenomic features to RNMs in Cluster A. For example, RNM 11 was enriched for pluripotency TFs (NANOG-OCT4 complex, POU2F2/OCT2, SOX-LEF1 complex) in active upstream regulatory elements (Enh, TssAFlnk) similar to RNM 30 in Cluster A, suggesting that there are distinct mechanisms for maintaining self renewal in primed and formative states (Figure 5E-F). Likewise, RNM 24 was enriched for active upstream regulatory elements (Enh, TssAFlnk) and TEAD TF family binding similar to RNM 10 in Cluster A. Despite the similarities of RNM 24 to its counterparts in Cluster A there were important distinctions including its enrichment for methylated enhancers (EnhG). Five RNMs (1, 6, 21, 24, 37) in Cluster B were enriched for transcription (Tx, TxWk) and RNM 24 was also enriched for FOS-JUN transcriptional activators that regulate the expression of genes during cell proliferation. Six of the RNMs showed enrichment for ZIC3, which was recently shown to direct OCT4 to primed state-specific enhancer sites[53]. In total, the seven RNMs in Cluster B capture many of the previously identified epigenomic and cell state-specific features of primed cells and most likely represent the coordinated activities of the regulatory networks in this pluripotent state.

RNM 14 is enriched with NANOG-OCT4 and OTX2, and not associated with any other pluripotency factors (Figure 5E). The fact that the loss of NANOG-OCT4 mediated self-renewal initiates neuronal differentiation[54,55], and that REST (RE1-Silencing Transcription factor), which is involved in the repression of neural genes in non-neuronal cells, was depleted[56], suggests RNM 14 could represent a regulatory network underlying neural specification.

In summary, these findings present novel insights into 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. We show that regulatory elements with similar epigenetic features and TFBS profiles in primed and formative pluripotency states (e.g., respectively, RNM 11 and RNM 30) are commonly present in different regulatory pathways. The corollary to this is that regulatory pathways with closely related functions in different pluripotency states were composed of different elements (i.e., each reference ATAC-seq peak is present in only one RNM).

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[57]. Of the 44,939 hiPSC reference ATAC-seq peaks in the 20 major RNMs we identified 10,700 that overlapped cell type-specific peaks for at least one of the 54 fetal cell types (Figure 6A-B, Table S14). For each RNM, we then examined if the overlapping peaks were enriched in one or more of the 54 fetal cell types (Figure 6C).

Based on the functional annotations and pairwise correlations of the RNMs (Figure 5), we had hypothesized that RNM 10, and RNM 14 could represent regulatory networks respectively underlying the emergence of placental trophoblast and neural lineages. We observed that RNM 10 was enriched for cell type specific peaks in villous cytotrophoblasts, extravillious trophoblasts, trophoblast giant and three other cell types (Figure 6C), while RNM 14 was enriched in multiple neural cell types including astrocytes, granule neurons, limbic system neurons, ENS glia, Purkinje neurons and ENS neurons. These findings suggested that RNMs present in specific hiPSC subpopulations could represent networks important for lineage specification and be conserved in the derived cell types.

We reasoned that RNM conservation during development would involve binding of the same TF group to the same regulatory elements in both hiPSCs and the derived cell type. Because of the role TEAD4 plays in specification of the trophectoderm lineage[58] and the fact that RNM 10 was enriched for binding the TEAD TF family (TEAD1 and TEAD4) (Figure 5F, Figure S8B), we examined the TF occupancy of the RNM 10 and the trophoblast-specific overlapping peaks. Of note, we annotated the occupancy based on the TFs that bound to the peaks in the hiPSCs. The RNM 10 peaks that overlapped the three "trophoblast" cell type-specific peaks were enriched for binding the TEAD Family TFs (Figure 6D). These data suggest that RNM 10 regulatory elements that bind the TEAD Family in hiPSCs, also bind the TEAD Family in "trophoblasts". Furthermore, because we analyzed cell type-specific peaks these data suggest that the shared RNM 10 regulatory elements that bind the TEAD Family show co-accessibility in "trophoblast" cell types.

Altogether, these results suggest that the some hiPSC regulatory networks are conserved in fetal cell types, and that the molecular mechanism underlying conservation is binding of the same TF Family to the same regulatory elements in both stages of development.

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

Previous studies have shown that both genetic and epigenetic variation affect hiPSC phenotypes[59–61], however the underlying mechanisms are poorly understood. One likely mechanism through which genetic variation influences chromatin accessibility is by affecting the binding specifity of TF motifs in regulatory elements[62–65]. We sought to detemine if specific TF groups, RNMs and chromatin states in our study were enriched for peaks with SNPs impacting allele specific chromatin accessibility (ASCA-SNPs) (Figure 7A). We used the coordinates of 1,861,239 TFBSs in 35,039 ATAC-seq peaks in the 20 major RNMs to identify variants that potentially disrupt TF binding. We found 17,805 SNPs (MAF $\geq$ 0.05, HWE p-value > $1x10^{-6}$) on 11,309 accessible ATAC-seq peaks to test for ASCA. In total 2,234 (12.5%) SNPs, which overlapped

3,941 TF group motifs in 1,821 peaks, exhibited ASCA (FDR < 0.05) (See Methods, Table S15). We evaluated the distribution of the TFBS-disrupting ASCA-SNPs by TF group and observed that only a small fraction of TFBSs have a ASCA-SNP (range per TF group: 0-0.76%, Figure 7B). NANOG-OCT4 Complex had the highest fraction of disrupted TFBSs in the 20 major RNMs (0.76%), followed by promoter-binding zinc-finger associated TFBSs (SP-E2F4 Complex, Kruppel Family, KLF15-MAZ-TAF1 Complex and EGR-ZnF Complex) (range 0.69-0.74%) (Figure 7B). To further characterize the peaks with TFBS-disrupting ASCA-SNPs, we performed Fisher's Exact tests to calculate their enrichment in RNMs and chromatin states and observed enrichments in RNMs 7, 21, 37, active promoters; and depletions in RNMs 4, 10, 14, 28, 30 and enhancers, regions flanking bivalent promoters/enhancers, and quiescient chromatin (Figure 7C-D). The strong enrichment of RNM 7 for ASCA-SNPs was congruent with the fact that it is also enriched for active promoters and promoter-binding zinc-finger associated TFs (Figure 5C-F). However, despite NANOG-OCT4 Complex having the highest fraction of TFBSs with ASCA-SNPs many formative-asociated RNMs (10, 28, 30) exhibited strong depletion for ASCA-SNPs (Figure 7C), suggesting that regualtory elements in these RNMs bound by factors other than NANOG or OCT4 are under evolutionary constraint.

Taken together, these data show that specific TF groups, chromatin states and the RNMs which they comprise show relative enrichment or depletion for ASCA-SNPs.

Genetic variants associated with proportion of pluripotency cell states in hiPSCs

We hypothesized that the effects of disrupting the binding of a transcription factor by an ASCA-SNP could potentially be propagated through a regulatory network by impacting the co-accessibility of the corresponding ATAC-seq peak with partner peaks. To identify ASCA-SNPs that influence the co-accessibility of their corresponding ATAC-seq peak with other peaks in an RNM, we used the ASCA-SNP genotypes as a covariate in the original linear mixed model (LMM) employed to identify co-accessibility (Figure 7A, Table S15). We then identified intramodular peak partners that were co-accessible in original LMM that were no longer accessible (P-value > $5x10^{-8}$, and Δ-$\log_{10}$(P-value) > 2) after accounting for the ASCA-SNP genotype. We determined that 342 (15.3%) ASCA-SNPs disrupted the co-accessibility of the corresponding ATAC-seq peak with at least one intramodular partner peak (Table S15) while 1,892 (84.7%) had no effect on co-accessibility. 101 (29.5%) of the co-acessibility disrupting ASCA-SNPs in 87 unique ATAC-seq peaks exerted large effects (disrupted co-accessibility with five or more intramodular partner peaks which in total comprised at least 1% of all partner peaks, Figure 7E). Using the 101 large effect ASCA-SNPs, we ran a linear regression to evaluate whether the fraction of partner peaks with impacted co-accessibility was correlated with the allelic imbalance fraction (AIF) of the corresponding ASCA-SNP (Figure 7F). We observed a modest but significant association (p-value = 9.2×10^{-5}, $\beta = 1.03$, Figure 7F), which suggests that ASCA-SNPs that had the largest impact on TF binding were more likely to also have a larger impact on the intramodular co-accessibility of the corresponding ATAC-seq peak.

We hypothesized that the large effect ASCA-SNPs could impact RNM integrity (i.e., overall accessibility of the Pareto peaks) and hence be associated with the fraction of different pluripotency cell states across the hiPSC lines. We ran a LMM (see Methods) to calculate the association between the 101 large effect ASCA-SNP genotypes and the RNM scores for each of the 150 ATAC-seq samples (i.e., the summed accessibility of the corresponding RNM-specific Pareto peaks; Figure 3D-

E) for each of the 20 major RNMs independently (Table S16). Fourteen (13.8%) of the large effect ASCA-SNPs on 11 unique ATAC-seq peaks (2 peaks contained 2 ASCA-SNPs) were significantly associated with the integrity of at least one major RNM (FDR < 0.05, Figure 7G). Five ASCA-SNPs (rs9350250, rs3760638, rs2774290, rs188702962, and rs111261603) had opposite correlations in the formative- and primed-associated RNMs indicating that they were associated with the proportions of different pluripotency states (Figure 7G). Moreover, six of the 11 ATAC-seq peaks with an RNM integrity disrupting ASCA-SNP were Pareto peaks suggesting that the peak centrality is associated with the ability of regulatory variants to exert large effects on RNM integrity (Figure 7G).

Rs9350250 was strongly associated with the accessibility of the corresponding peak (a narrow peak in RNM 24 numbered 183865) (Figure 7H) and the RNM scores for RNMs 2, 3, 8, 24 and 37 (Figure 7G,I-L, RNM 2 not shown). Although narrow peak 183865 is not a Pareto peak, rs9350250 exerted large effects because it disrupted 100% of the peak's intramodular co-accessibility (n=17, Figure 7M). We examined the TFBSs disrupted by rs9350250 and determined that it overlapped a core nucleotide within a MAFF motif from the MAF Family TF group (Figure 7N). Narrow peak 183865 was located in an intron of *E2F3* (Figure 7O), which is an E2F Family TF that plays an important role in the induction of S-phase[66]. As mentioned above, formative state stem cells have an abbreviated G_1 phase[1,50], indicating that the E2F Family TFs play an integral role in maintaining pluripotency by influencing the G_1 phase duration. Here, we show that rs9350250 is associated with decreased accessibility of three primed-associated RNMs (21, 24, and 37), and increased accessibility of formative-associated RNMs 3 and 8 (Figure 7G) suggesting that regulatory variants can influence the stability of different pluripotency cell states in hiPSC lines.

Altogether, these data show that 0.66% (14 out of 2,234) of regulatory variants that disrupted TF binding were associated with the overall accessibility of the Pareto peaks (i.e. the RNM scores for each of the 150 ATAC-seq samples) in one or more RNMs. Furthermore, these observations suggest that future studies maybe able to identifiy genetic signatures associated with increased stability of the formative pluripotency state in hiPSC lines under defined media conditions.

DISCUSSION

In recent years, gene co-expression networks have emerged as a powerful approach to identify co-regulated genes that have roles in shared biological processes and cell types[14]. Co-expression has been conjectured to result from genes driven by regulatory elements with similar epigenomic features (i.e. TF binding, histone acetylation), however this hypothesis has not been rigorously tested, partly due to the lack of suitable datasets. While single cell RNA-seq and single nuclei ATAC-seq can identify co-accessible cis regulatory elements (CREs) within 1.5Mb of their putative target gene they cannot directly resolve genome-wide regulatory networks[67]. One approach for inferring cell type-specific regulatory networks from single cell data is to collapse CREs within the same cluster of cells[57], with the assumption that these elements are co-accessible because they are present in a given cell type. When applied to pluripotent stem cells, single cell technologies have not yielded highly distinct cell clusters[68] because hiPSC heterogeneity is due to varying proportions of pluripotency cell states and is much more subtle than what is observed in mature tissues. Additionally, collapsing CREs within the same cluster of cells does not identify functionally discrete regulatory networks each of which underlies different cellular processes.

Furthermore, single cell technologies are typically applied to limited sample sizes and hence the role of genetics on regulatory network stability cannot be examined.

Bulk sequencing approaches have been criticized because they cannot identify signals that arise from different cell states or type present in the sample, as they report the average phenotype (i.e gene expression, chromatin accessibility) across cells in a sample. This phenomenon, known as the Simpson's paradox, is an artifact of sample size, and can be addressed by generating data for more samples. In this study, we generated bulk ATAC-seq data for 143 hiPSC lines from 133 genetically diverse individuals to better characterize pluripotent cell state heterogeneity. We calculated the pairwise co-accessibility of 59,499 reference ATAC-seq peaks and employed an unsupervised machine learning algorithm to identify 20 high resolution regulatory networks with distinct epigenetic features. These 20 regulatory networks formed 2 large clusters that correspond to primed and formative pluripotent cell states (Figure 3, Figure 5), and most likely mechanistically underlie important cellular processes including pluripotency, self-renewal and cell cycle transitions.

The formative cell state is enriched for self renewal processes[1], compared to primed cells which can exhibit emergent features of lineage commitment[5]. Historically, epigenomic regulation in pluripotent cells has been characterized by a single regulatory network primarily mediated by NANOG, OCT4 and SOX2 binding (RNM 30)[69]. Our study reveals that there are several regulatory networks with distinct epigenetic profiles (chromatin state, TF binding), that mediate biological processes that are indispensible for maintaining the formative and naïve cell states (Figure 5). For example, the TEAD4 regulatory network (RNM 10, Figure 5F) works as a pioneering factor for trophoectoderm lineage commitment[49,70] in the naive state. Our analyses show that the TEAD4 regulatory network (RNM 10) is conserved during the development of the trophoectoderm-derived placenta whose regulatory elements share epigenomic similarities (Figure 6C-D). Additionally, we identified five formative-associated regulatory networks (RNMs 3, 7, 8, 12 and 28) that are enriched with active promoters and binding sites for E2F TFs (E2F Family, E2F2, E2F5, and SP-E2F4 Complex) (Figure 5D-F), that represent novel independent networks that regulate cell cycle transitions in pluripotent stem cells. These analyses revealed epigenomic 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 networks.

Another advantage of our approach is the enhanced capability to evaluate the influence of genetics on regulatory networks. One of the primary mechanisms of regulatory variation is thought to be disruption or alteration of TFBSs[62–65] and single nuclei ATAC-seq studies have identified allele specific chromatin accessibility (ASCA)[57], but are inherently underpowered to test genotypic affects on co-accessibility within networks. In our study, we identified 2,234 SNPs within TFBSs that resulted in ASCA and characterized how these variants perturbed peak co-accessibility. In most cases, ASCA-SNPs demonstrated little to no effect on peak co-accessibility, but in some cases, they disrupted co-accessibility with all peak partners in the regulatory network (Figure 7E,H-O). We identified 101 ASCA-SNPs that exerted large effects, of which 14 were associated with the integrity of at least one major RNM (i.e. the corresponding RNM score for each of the 150 ATAC-seq samples) (Figure 7G). Five of these 14 ASCA-SNPs exhibited opposite correlations with primed and formative-associated RNMs, indicating that these variants were associated with the fraction of pluripotency cell states across the hiPSC

lines and could potentially be leveraged to identify genetic backgrounds with propensity for differentiation into germline, soma, and extraembryonic tissue.

In summary, our study leveraged the large sample sizes in the iPSCORE resource to identify novel regulatory networks in primed pluripotent and formative/naïve cell states and regulatory variants that influenced pluripotency state stability[11]. Future applications of the approach we have developed to characterize genetic effects on regulatory networks in mature tissues could provide novel insights into molecular mechanisms underlying polygenic diseases.

ACKNOWLEDGEMENTS

This work was supported by a California Institute for Regenerative Medicine grant GC1R-06673-B, and NIH grants HG008118 (KAF), HL107442 (KAF) and HG011558 (KAF), F31DK131867 (JPN), F31HL158198 (TDA) and T15LM011271 (JPN&TDA).

AUTHOR CONTRIBUTIONS

TDA, WWYG and KAF conceived the study. TDA, JPN, MD and HM performed the analyses. ADC generated ATAC-seq data. HM performed quality check on RNA-seq and ATAC-seq samples. KAF oversaw the study. TDA, MFP and KAF prepared the manuscript.

iPSCORE CONSORTIUM

University of California, San Diego, La Jolla, CA 92093, USA

Angelo D. Arias, Timothy D. Arthur, Paola Benaglio, Victor Borja, Megan Cook, Matteo D'Antonio, Agnieszka D'Antonio-Chronowska, Christopher DeBoever, Margaret K.R. Donovan, KathyJean Farnam, Kelly A. Frazer, Kyohei Fujita, Melvin Garcia, Olivier Harismendy, David Jakubosky, Kristen Jepsen, Isaac Joshua, He Li, Hiroko Matsui, Naoki Nariai, Jennifer P. Nguyen, Daniel T. O'Connor, Jonathan Okubo, Fengwen Rao, Joaquin Reyna, Lana Ribeiro Aguiar, Bianca Salgado, Nayara Silva, Erin N. Smith, Josh Sohmer, William W. Young Greenwald, Shawn Yost

Salk Institute for Biological Studies, La Jolla, CA 92037, USA

Athanasia D. Panopoulos, W. Travis Berggren, Kenneth E. Diffenderfer

DECLARATION OF INTERESTS

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

METHODS

Subject Information

We used hiPSC lines from 219 individuals (Table S1) recruited as part of the iPSCORE project[17,23,71]. 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)[72]. 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[17,23] 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 S7;
- Previously published[17,23,71] RNA-seq data from 213 hiPSC lines (collected after culturing in mTeSR1 medium on Matrigel) from 213 individuals, Table S1.
- 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.

We used the following publicly available datasets:

- ChIP-seq data for 18 Transcription Factors from ENCODE[22,36,73] (Table S10);
- Seven hiPSC cell state gene sets obtained and curated from 3 published studies:
    - One gene set for 8-cell like cells[3]
    - Two gene sets: one for formative-state (early postimplantation epiblast-like; EPE) and the "general population" which consisted of non-formative-state cells[1]
    - Four gene sets for epiblast, conventional PSCs, naïve PSCs, primitive endoderm[6]
- Fifty-five ATAC-seq peak sets from 2 published studies:
    - Formative-associated (EPE) ATAC-seq peaks[1]
    - Fifty-four sets for fetal tissue specific ATAC-seq peaks from the Descartes; Human Chromatin Accessibility During Development Atlas[57]

- ChromHMM chromatin states in the ROADMAP Epigenomics hiPSC-18 line[19];

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[74]. 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 [75]. 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[17] 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: As previously described[75], 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[17]. Briefly, FASTQ files were aligned to the hg19 reference genome using STAR 2.5.0a[76] and Gencode V.34lift37[77] 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[78] and marked duplicates using biobambam2 (2.0.95) bammarkduplicates[79] and then re-indexed. To quantify gene expression (TPM), we used RSEM v1.2.20[80] 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 quantile normalization: To normalize gene expression and usage across samples, we performed quantile-normalization 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[16,81].

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). 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 2.5×10^4 cells for the ATAC-seq assay.

We performed ATAC-seq on the 150 hiPSC samples using a modified version of the Buenrostro et al.[15] 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[82], permeabilized the nuclear membranes (containing 20-35% cholesterol and cholesterol esters[82,83] without disrupting the mitochondrial membranes (0.5-3% cholesterol)[84]. 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 S7). The 263 FASTQ files were aligned to the hg19 reference genome with STAR[76] 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[85]. 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)[85]. 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 S7).

ATAC-seq Peak Calling and Quality Control: MACS2 v.2.2.7[86] 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 S4B, Table S7). 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). 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[87] . 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[88].

ATAC-seq peak quantile normalization: TMM values for all 136,333 ATAC-seq peaks were quantile-normalized 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.

***Identification of 59,499 autosomal reference peaks*:** To reduce the computational burden in downstream analyses, we selected a subset of 59,499 reference peaks based on their MACS2 peak score and enrichment in active chromatin and for TF binding sites. We binned the 132,225 ATAC-seq peaks on autosomes into 20 MACS2 score quantiles each containing 6,611 peaks, where quantile 20 consisted of the peaks with highest scores. We looked at the density of ATAC-seq peak length by the 20 quantiles and observed that the peaks in higher quantiles were longer than lower quantiles (Figure S4C). We then examined the enrichment of the 132,225 ATAC-seq peaks in iPSC-18 ChromHMM chromatin states by MACS2 score quantile (Figure S4D). We created 20 bed files containing the coordinates of the peaks in each quantile and calculated their enrichment (Odds Ratio) in each of the 15 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; but were depleted for transcriped (Tx, TxWk) and repressed (Het, Quies) chromatin. Finally, we examined the distribution of ATAC-seq peaks with at least one TOBIAS predicted TFBS across the 20 MACS2 score quantiles (see below: *Prediction of transcription binding with TOBIAS*) and observed that the 92.3% TOBIAS predicted TFBSs (n = 45,291) were on the 59,499 ATAC-seq peaks in quantiles 12-20 (data not shown). Based on these observations, we calculated pairwise co-accessibility for the 59,499 ATAC-seq peaks in MACS2 quantiles 12-20 (see *Identifying co-accessible peaks* section below), which alleviated the computational burden by eliminating $7x10^9$ pairwise tests between peaks with low accessibility.

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 S7.

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[16]. 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[16] 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.

Gene Co-expression Analyses

We leveraged the previously published RNA-seq dataset of 213 hiPSC lines from iPSCORE individuals[17,71,75] to identify gene networks that are active in stem cells. We integrated and curated gene sets of defined pluripotency cell states from external sources[1,3,6,75] 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*).

<u>Formula #1:</u>

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

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 mth genotype principal component for the individual associated with sample *i*, γ_m is the effect size of the mth genotype principal component, *M* is the number of genotype principal components used (M =20), C_{ip} is the covariate of the pth covariate for sample *i*, γ_p the effect size of the pth 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 and the effect size of each gene (β_k) as weights. 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 with following default parameters *cluster_leiden(objective_function = "modularity", beta = 0.01, n_iterations = 5)*. We evaluated the GNMs resulting from resolutions 1-5 and selected resolution 4 which yielded 24 GNMs. 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 intermodule edges (Table 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 quantile normalized TPM expression of the intramodular Pareto genes in the 24 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 24 GNMs by summing the quantile-normalized TPM expression of the corresponding GNM-specific Pareto genes (i.e., each RNA-seq sample had 24 GNM scores). We then performed a PC analysis on the transposed quantile-normalized Pareto gene TPM expression matrix to analyze GNM expression for samples, rather than genes. To visualize the differential expression of the GNM-specific Pareto genes across the 213 samples, we performed UMAP dimensionality reduction, ranked samples from 1-213 by their GNM scores, and then colored each sample by their GNM rank (Figure S3).

Functional Characterization: Gene Co-expression Network Modules

ROCK inhibition differential expression analysis: We examined nine RNA-seq samples (three iPSCORE hiPSC lines each with three technical replicates) that had been expanded in mTeSR1 medium with ROCK inhibitor on Matrigel (see above: *Generation of RNA-seq for 3 hiPSC lines expanded with ROCK inhibitor*). We performed differential expression analysis between these nine RNA-seq samples and the three paired (i.e., same hiPSC lines) ROCK inhibitor-naive hiPSC RNA-seq samples used in the co-expression analysis, using the DESeq2 (v.1.34) *R* package[90] (Table S6). We annotated 416 genes in the 24 major GNMs based on whether they were upregulated (Adjusted P-value < 0.05 and $\log_2$(FC) > 1) or downregulated (Adjusted P-value < 0.05 and $\log_2$(FC) < -1) by ROCK inhibitor.

***GNM Enrichment in Stem Cell States*:** We integrated differential gene expression data from the ROCK inhibition analysis and four published studies describing 7 hiPSC subpopulations representing different pluripotency cell states including: (1) 47 genes upregulated by ROCK inhibitor (Adjusted P-value < 0.05 & $\log_2$(FC) > 1) and 369 genes downregulated by ROCK inhibitor (Adjusted P-value < 0.05 & $\log_2$(FC) < -1); (2) 199 8-cell like cell (8CLC)-associated genes (Adjusted P-value < 0.05 & Samples = "8CLC") from Mazid et al[3]; (3) 176 primitive endoderm (PrE)-associated genes (Adjusted P-value < 0.05 & Z-score > 5), 109 epiblast-associated genes (Adjusted P-value < 0.05 & Z-score < -5); 775 Naïve-associated PSCs genes (Adjusted P-value < 0.05 & $\log_2$(FC) > 3), and 310 conventional PSC genes (Adjusted P-value < 0.05 & $\log_2$(FC) < -3) from Stirparo et al.[6]; and (4) 113 formative (EPE) cell-associated genes (Adjusted P-value < 0.05 & $\log_2$(FC) < -0.5) and 327 general population-associated genes (Adjusted P-value < 0.05 & $\log_2$(FC) > 0.5) from Lau et al.[1]. 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 23 major GNMs as background.

Calculation of GNM Correlations: To assess the relationship between gene networks, we calculated the pairwise Pearson correlation between the GNMs. We first calculated the mean quantile-normalized expression of the 2,862 Pareto genes in the 24 major GNMs. We then used the base *R* cor function to calculate the pairwise Pearson correlation of the mean quantile normalized GNM Pareto gene expression across the 213 RNA-seq samples. We calculated the corresponding p-values using the p.mat function from the corrplot package (v0.92)(Table S5).

Calculation of GNM Biological Process Functional Enrichment: To identify biological processes associated with each GNM, we performed a functional enrichment analysis using the MSigDB (v.7.5)[29] C5 database of gene ontologies for biological processes. For each gene set, we removed non-autosomal genes and then performed a Fisher's Exact test to calculate the enrichment in each GNM (Table S4).

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 59,499 reference peaks (see above: *Identification of 59,499 autosomal reference peaks*) which reduced the number of tests to 1.74 x 10^9. To identify pairwise correlations of accessibility between the 59,499 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 quantile-normalized TMMs across the 150 ATAC-seq samples for each of the 59,499 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 59,499 reference peaks, 44,939 were present in the 20 major RNMs. In total, we identified 8,661,299 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 59,499 accessible peaks as nodes, the 8,661,299 co-accessible peaks (P-value < $5x10^{-8}$, Effect Size > 0) as edges and the effect size

as weights. We extracted the genome-wide degree of each peak using the *degree* function on the RN. To identify peak co-accessibility regulatory network modules (RNMs), we applied the Leiden community detection algorithm (from the *igraph* R package[26] to the RN with following default parameters *cluster_leiden(objective_function = "modularity", beta = 0.01, n_iterations = 5)* and evaluated RNMs resulting from resolutions 1-5). We selected resolution 4 which yielded 20 major RNMs (membership ≥ 500 ATAC-seq peaks). 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 a 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 defined intramodular Pareto peaks as the top 20% interconnected ATAC-seq peaks within each RNM (intramodular degree, Table S8, n = 8,972 peaks). We performed a PC analysis on the quantile normalized TMM of the intramodular Pareto peaks from the 20 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 RNM Score: For each of the 150 ATAC-seq samples we calculated a RNM score for each of the 20 RNMs by summing the quantile-normalized TPM matrix of the corresponding RNM-specific Pareto peaks (i.e., each ATAC-seq sample had 20 RNM scores). We then performed a PC analysis on the transposed quantile-normalized Pareto peak TPM matrix to analyze RNM accessibility for samples, rather than peaks. To visualize the differential accessibility of the RNM-specific Pareto peaks across the 150 samples, we performed UMAP dimensionality reduction, ranked samples from 1-150 by their RNM scores, and then colored each sample by their RNM rank (Figure S5). The RNM score, not the rank, was used for downstream analyses (**See Section**: *Calculation of the effects of ASCA-SNPs on accessibility of RNM-specific Pareto peaks)*.

Correlation of GNMs and RNMs

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

Identifying Putative Gene Targets for ATAC-seq Peaks: To identify candidate target genes for the 59,499 reference ATAC-seq peaks, we generated a bed file of the TSSs for the 16,752 genes expressed (TPM > 1 in at least one of the nine samples, see above: *RNA-seq data processing*) after culturing with 5μM ROCK inhibitor and performed *bedtools closest* to identify the closest TSS within 50kb 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 stimulated hiPSC lines and annotated the ATAC-peak with the gene with the maximum expression. In total, we annotated 34,432 ATAC-seq peaks with 14,455 unique genes (Table S8).

Identifying Associations Between Gene and Regulatory Networks: The GNMs were identified using ROCK inhibitor-naïve hiPSCs but the ATAC-seq peaks were identified in hiPSCs after culturing with ROCK inhibitor. To annotate peaks with putative gene targets we: 1) identified genes expressd after culturing with ROCK inhibitor, and then 2) annotated each

peak with a single gene (distance < 50kb and highest expressed gene). To identify associations between the GNMs and RNMs we only used genes: 1) expressed in both ROCK inhibitor-naïve hiPSCs and ROCK inhibitor-exsposed hiPSCs, 2) in one of the 24 major GNMs, and 3) annotated as a putative target in one of the 20 major RNMs. In total 11,121 unique genes corresponding to 22,936 peaks were used for the association test. We calculated the observed number of occurrences of all GNM-RNM pairs. We permuted the GNM annotations 1000 times and calculated GNM-RNM pair occurrences after permutations as the expected number of occurences and calculated the Z-score and P-values.

Functional Annotation of hiPSC ATAC-seq peaks

To functionally characterize the hiPSC epigenome, we annotated the 59,499 reference ATAC-seq peaks with three epigenetic annotations: 1) hiPSC-specific chromatin states, 2) TF binding, 3) pluripotency cell state.

Annotating hiPSC ATAC-seq Peaks with ChromHMM-defined chromatin states: From the Roadmap Epigenomics Resource (http://www.roadmapepigenomics.org) we obtained the ChromHMM chromatin states for the iPSC-18 line. We calculated enrichment for the 59,499 ATAC-seq peaks in 15 chromatin states in the core ChromHMM model v1.10[19] using bedtools fisher and hg19 chromosome sizes (obtained from http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/). We resolved a single chromatin state for each peak by intersecting the maximum summits with the iPS-18 chromatin states (Table S8).

Prediction of transcription binding with TOBIAS: The TOBIAS algorithm[20] 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. First, we identified 187 transcription factors with experimentally validated high-confidence motifs (Quality A and B) included in the HOCOMOCO V 11 collection[21] 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. We examined the distribution of ATAC-seq peaks with at least one predicted TFBS across the 20 MACS2 score quantiles (see above: *Identification of 59,499 autosomal reference peaks*) and observed that the 92.3% predicted TFBSs (n = 45,291) were on the 59,499 reference ATAC-seq peaks in quantiles 12-20 (data not shown). 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[22,36,73] (Table S10). We obtained peaks from 19 sets of transcription factor ChIP-seq peak files for 18 TFs in H1 embryonic stem cells from ENCODE[22,36,73] (Table 10). 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.

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 S7B) and were predicted to be bound in similar proportions of ATAC-seq peaks (Figure S8B); 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 S8). 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 S7, Figure S8). The 187 TF motifs and their corresponding collapsed TF groups are in Table S10.

Annotation with Formative Cell State: To identify ATAC-seq peaks associated with the formative pluripotency state, we integrated ATAC-seq peaks from Lau et al 2020[1] which describe a hiPSC subpopulation (GCTM-2^{high}CD9^{high}EPCAM^{high}) that exhibit properties of the early post-implantation epiblast (EPE). We used *bedtools intersect -r 0.25* to identify 1,665 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 4,029 non-formative associated peaks using the same approach.

Functional characterization of RNMs

***Epigenetic Feature Enrichment within hiPSC Regulatory Networks*:** To molecularly characterize hiPSC regulatory networks, for each of the 20 RNMs we calculated enrichment of the formative (EPE) state, 15 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 19 remaining major RNMs as background (Table S12). We considered enrichments with a Bonferroni-corrected p-value < 0.05 significant.

Calculation of RNM Correlations: To assess the relationship between regulatory networks, we calculated the pairwise Pearson correlation between the RNMs. We first calculated the mean quantile-normalized accessibility of the 8,972 Pareto peaks in the 20 major RNMs. We then used the base *R* cor function to calculate the pairwise Pearson correlation of the mean quantile normalized RNM Pareto peak accessibility across the 150 ATAC-seq samples. We calculated the corresponding p-values using the p.mat function from the corrplot package (v0.92) (Table S13).

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[57]. 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 44,939 reference ATAC-seq peaks in the 20 major RNMs and identified 10,700 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 10,700 hiPSC ATAC-seq peaks, using the remaining 19 RNMs as background (Table S14).

TF enrichment in fetal cell type associated hiPSC peaks: To assess whether hiPSC regulatory networks are reused during fetal development, we assessed the TF occupancy of peaks present in both hiPSCs and fetal cell types. For each RNM enriched with fetal cell type specific peaks, we performed Fisher's Exact tests to calculate enrichment of the 92 TF groups and "Not Bound" annotation in peaks present in each RNM-fetal cell type pair, using the remaining 10,700 peaks without the tested TFBS as background.

Allele Specific Chromatin Accessibility (ASCA) analyses

To identify regulatory variants that affect transcription factor binding and network connectivity, 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[71] was obtained from dbGaP (phs001325). We used the coordinates of 1,861,239 TOBIAS predicted TFBSs to merge them into 329,455 sets of overlapping TF Group binding sites. For example, if two TF group binding sites overlapped, i.e. TF group 1 has a site chr1:100-125 and TF group 2 has a site chr:120-150, we merged from chr1:100-150. The 329,455 sets of overlapping TF Group binding sites were present in 35,039 accessible ATAC-seq peaks in the 20 major RNMs. We extracted SNPs in these 329,455 regions: 1) with minor allele frequency > 0.05 in all 273 iPSCORE individuals[71] 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. 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 haplotypes using snp2hd5 in WASP. We then filtered poorly mapped reads (MAPQ < 20%), duplicates and reads less than 38bp and greater than 2000bp with samtools view. We 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, 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 17,805 bi-allelic SNPs in 11,309 accessible ATAC-seq peaks used for ASCA analysis. We annotated 17,778 SNPs with their

corresponding rsid from the gnomAD database (v2)[93], and used the chromosome, position, reference and alternate allele to annotate the 27 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 17,805 SNPs with the RNMs and chromatin states associated with the ATAC-seq peak. To annotate a variant with TF groups, we merged the binding sites for all 187 TFs by their corresponding TF group and determined overlap with the 17,805 SNPs using *bedtools intersect*.

***Enrichment of ASCA-SNPs within hiPSC Regulatory Networks and Chromatin States*:** We examined whether specific hiPSC regulatory networks and chromatin states were enriched for peaks with ASCA-SNPs. For each of the 20 RNMs and 15 chromatin states, we performed a Fisher's exact test on the 44,939 ATAC-seq peaks in the 20 major RNMs (2,234 of which contained an ASCA-SNP), respectively using the 19 remaining RNMs or the 14 remaining chromatin states as background.

Calculation of the effects of ASCA on co-accessibility: To assess the effect of genetic variation on regulatory network co-accessibility, we modified the original LMM (see above: Formula #2) by adding a covariate to account for ASCA-SNP genotypes. For each ASCA-SNP, we ran the modified LMM on the corresponding RNM to calculate co-accessibility between peaks while adjusting for the ASCA-SNP genotypes. We defined an ASCA-disrupted co-accessible peak as one that was determined to be co-accessible (P-value < $5x10^{-8}$ and Effect Size > 0) in the original LMM model but was not co-accessible in the ASCA-modified LMM (P-value > $5x10^{-8}$ and Δ-$\log_{10}$(P-value) > 2). To calculate the fraction of peaks showing disrupted co-accessibility for each ASCA-SNP, we divided the number of disrupted co-accessible peaks by the total number of peaks in the original LMM with which the peak containing the ASCA-SNP was co-accessible (i.e., its intramodular degree; see above: *Construction of the Genome-wide Regulatory Co-accessibility Network*) (Tables S8, S15).

Calculation of the effects of ASCA-SNPs on accessibility of RNM-specific Pareto peaks: To assess the effect of ASCA-SNPs on RNM integrity (i.e., overall accessibility of RNM-specific Pareto peaks), we tested each of the 101 large effect ASCA-SNPs for association with RNM score (see above: *Calculation of RNM Score*) for the 150 ATAC-seq samples. We applied the LMM described above (see Formula #2) using the same covariates (see *Covariates for co-expression and co-accessibility analyses*) to calculate the association between the 101 ASCA-SNP genotypes and the RNM scores for each of the 20 major RNMs independently. We corrected the P-values for all 2,020 tests using Benjamini-Hochberg and considered ASCA-SNPs with Benjamini-Hochberg adjusted p-value < 0.05 significant (Table S16).

FIGURES

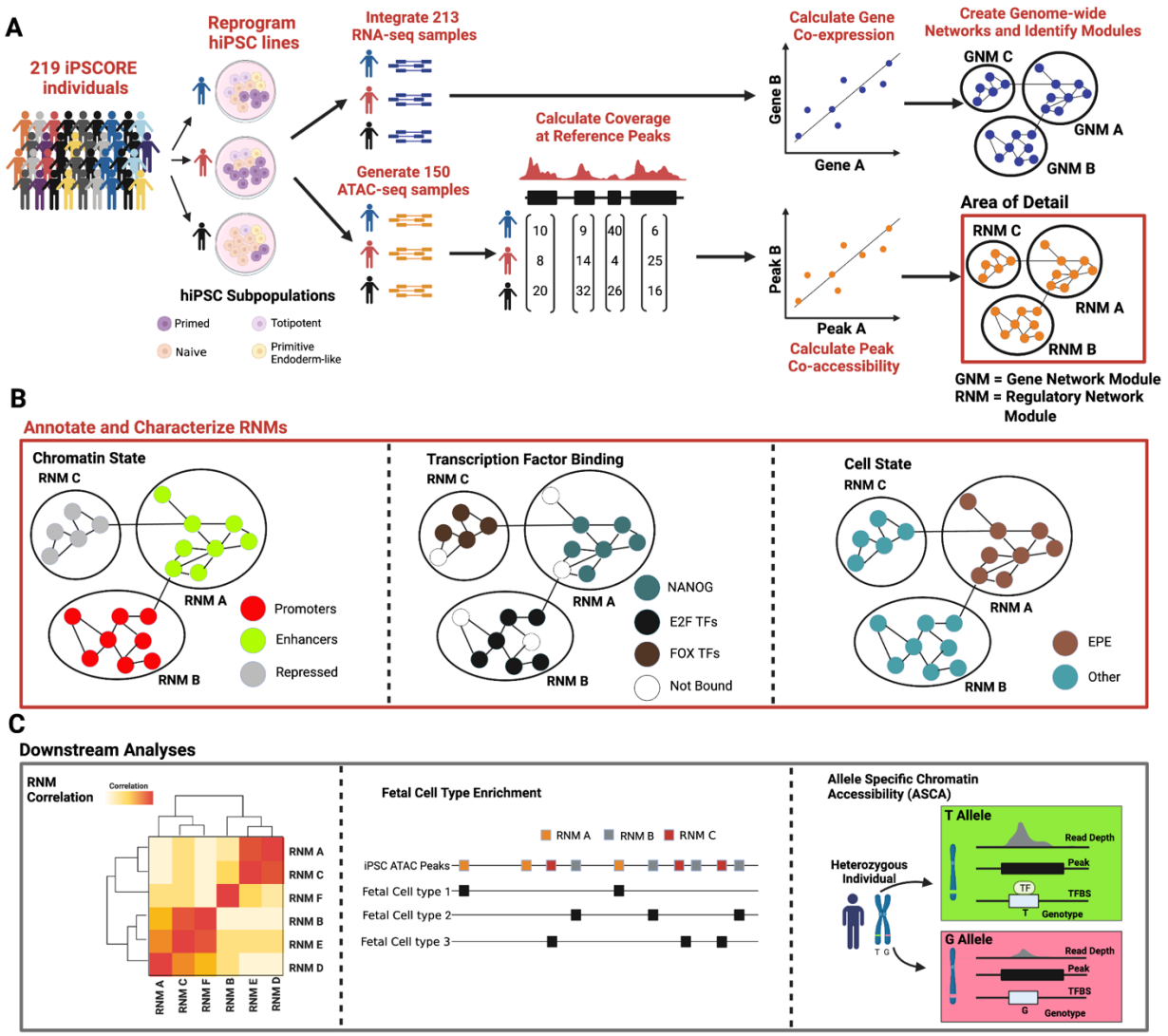

Figure 1. Overview of Study Design

Identification of genome-wide hiPSC regulatory networks. To identify and characterize regulatory networks underlying pluripotency and self-renewal, we analyzed hiPSC ATAC-seq samples and RNA-seq samples generated from hundreds of individuals in the iPSCORE resource. Figure was generated with Biorender (license: UB259BHRFI).

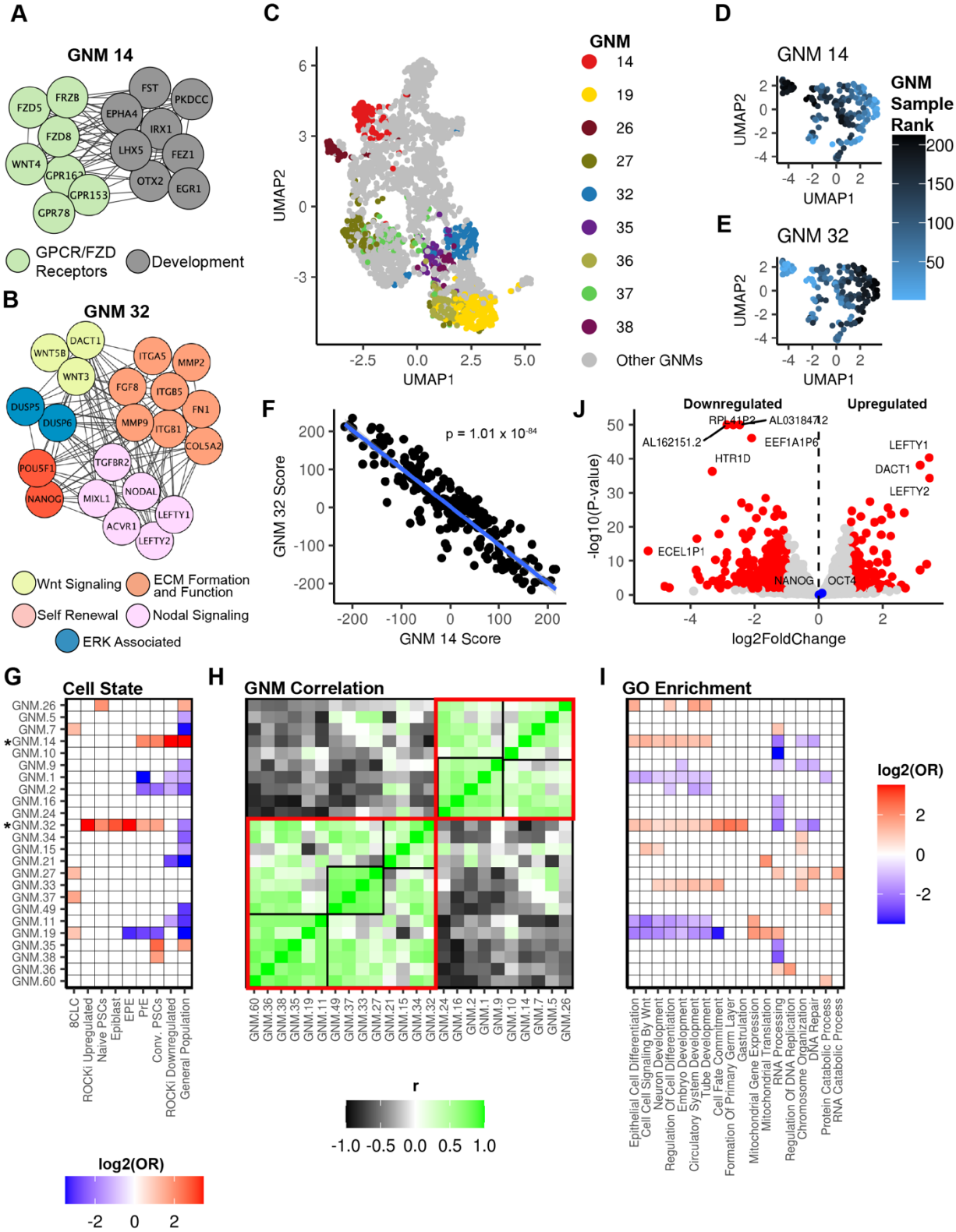

Figure 2. Unsupervised machine learning algorithm identifies hiPSC gene networks

(A) Network graph showing a subset of the co-expressed genes in GNM 14. Nodes are colored by their associated biological process; Frizzled and GPC receptor signaling (green) and embryonic development (gray).

(B) Network graph showing a subset of the co-expressed genes in GNM 32. Nodes are colored by their associated biological process; self renewal processes (light-orange), Wnt signaling (light-green), Nodal signaling (pink) and extracellular matrix formation (dark-orange).

(C) UMAP plot displaying the expression of the 973 Pareto genes (colored by GNM membership) of nine GNMs selected to represent a spectrum of pluripotent cell states (H). Pareto genes within the other 15 major GNMs shown as "Other GNMs" (gray). A corresponding UMAP plot displaying all 2,862 Pareto genes for the 24 major GNMs is shown in Figure S2B.

(D-E) UMAP plots exhibiting the separation of 213 RNA-seq samples based on proportion of (D) GNM 14 and (E) GNM 32. Each point represents a RNA-seq sample colored by their corresponding GNM rank.

(F) Scatter plot showing the negative linear relationship of the summed Pareto gene expression for GNM 14 versus GNM 32. Each point represents one of the 213 RNA-seq samples.

(G) Heatmap showing the enrichment of pluripotency cell state-associated genes in the 24 GNMs. A Fisher's Exact Test was performed on gene sets from seven hiPSC cell states and genes whose expression was affected by ROCK inhibitor (expression up and down) on each GNM to calculate the Odds Ratio (Table S4). 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.

(H) Heatmap showing the pairwise associations between 24 GNMs based on the mean expression of Pareto genes. Each cell is filled with the Pearson correlation (r), where -1 (black) represents an anti-correlation between two GNMs and 1 (green) represents a positive association between GNMs. There are two main clusters of correlated GNMs, subclusters are outlined by black boxes (Table S5).

(I) Heatmap showing the enrichment of biological processes in 24 GNMs. A Fisher's Exact test was performed on all MSigDB C5 Biological Process gene sets for each GNM to calculate an Odds Ratio (Table S4). Each cell is filled with the $\log_2$(Odds Ratio) for the corresponding gene set and GNM (red = enrichment, blue = depletion, white = non-significant). For plot legibility, the enrichment range was set to -3.5 to 3.5.

(J) Volcano plot displaying the differential gene expression in hiPSCs cultured with ROCK inhibitor. Genes that are significantly affected by ROCK inhibitor (FDR < 0.05 and -1 > logFC > 1) are colored red. For figure legibility, highly significant genes (p-value < $1x10^{-50}$) were assigned p-value equal to $1x10^{-50}$. Genes not differentially expressed are colored in light gray except for NANOG and OCT4 (blue).

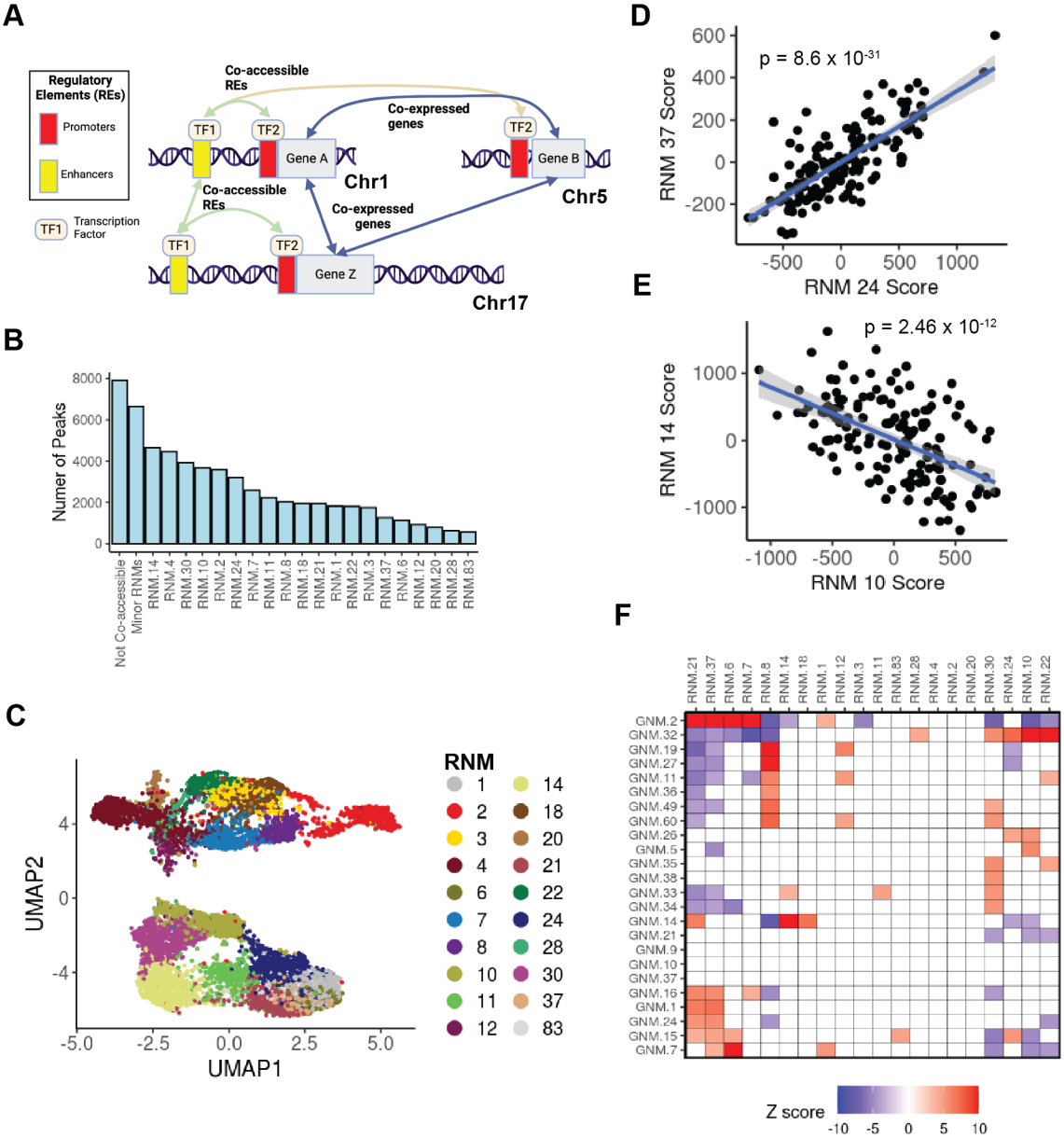

Figure 3. Unsupervised machine learning algorithm identifies hiPSC regulatory networks

(A) Diagram of proposed molecular mechanisms underlying gene co-expression networks. Co-expression of genes (gray boxes) located on different chromosomes is mediated via co-accessible regulatory elements (i.e, promoters (red boxes) and enhancers (yellow boxes) that bind to similar transcription factors.

(B) Histogram displaying the number of peaks in each of the 20 major RNMs, and the total number of non-coaccessible peaks and peaks in minor RNMs.

(C) UMAP plot of 8,972 RNM Pareto ATAC-seq peaks (top 10% intramodular degree connectivity). We performed Leiden clustering on the genome-wide ATAC-seq peak co-accessibility network and identified 20 major RNMs (membership > 500 ATAC-seq peaks). Each point represents a Pareto peak colored by its corresponding RNM.

(D-E) Scatter plots showing the negative linear relationship of the summed Pareto peak accessibility for (D) RNM 37 versus RNM 24 and (E) RNM 14 versus RNM 10. Each point represents one of the 150 ATAC-seq samples.

(F) Heatmap showing GNM-RNM associations. 22,936 ATAC-seq peaks in the 20 major RNMs were annotated with a putative target gene using the genes in the 24 major GNMs (see Methods). We performed a permutations test to identify associations between GNMs and RNMs. We identified 18 major RNMs associated with at least one GNM. The heatmap displays the associations (Z score) between RNM (x-axis) and GNM (y-axis). Non-significant associations are filled in white.

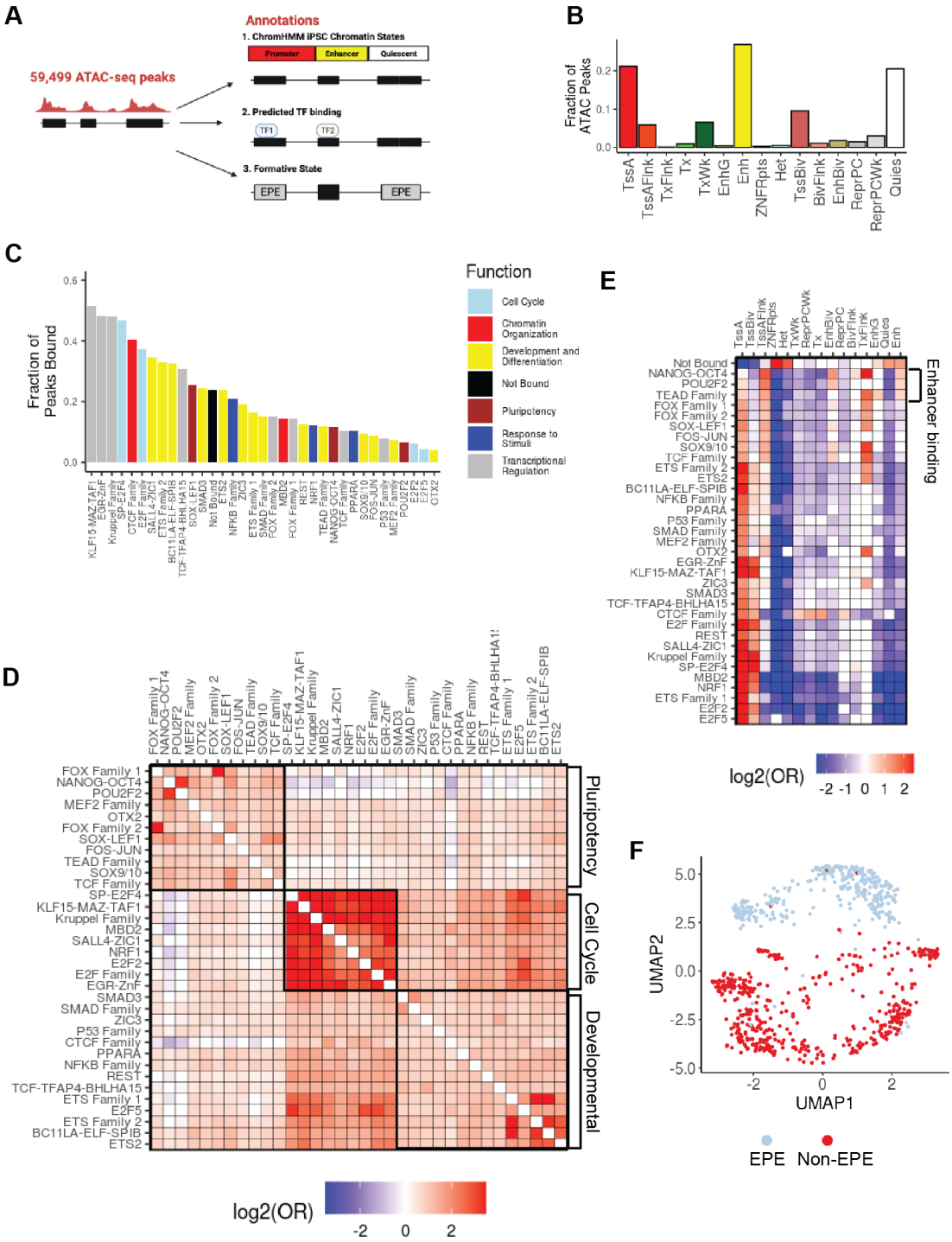

Figure 4. Epigenetic Characterization of Reference ATAC-seq Peaks

(A) Diagram of Epigenome Annotation Strategy. The 59,499 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 (EPE)-associated peaks.

(B) Barplot showing the fraction of ATAC-seq peaks annotated by each of the 15 iPSC-18 ChromHMM chromatin states.

(C) Barplot showing the fraction of ATAC-seq peaks bound by a selected subset of 35 TF groups. As indicated by the legend, bars are colored by an associated biological process curated by review of the literature. The fraction of peaks bound by all 92 TF groups (and "Not Bound" peaks) and 187 TF motifs before grouping is shown in Figure S8.

(D) Heatmap showing TF group co-binding enrichments fall into three distinct clusters. Pairwise Fisher's Exact tests were performed to calculate the enrichment of co-occurrence of the selected subset of 35 TF groups on ATAC-seq peaks. A heatmap displaying co-binding enrichments for all 92 TF groups are reported in Figure S9.

(E) Heatmap showing associations between the selected subset of 35 TF groups and 15 chromatin states. Each cell is filled with the enrichment ($\log_2$OR) for each TF group and chromatin state association. A heatmap showing the chromatin state associations for all 92 TF groups and "Not Bound" peaks is available in Figure S10.

(F) UMAP plot of formative-associated peak accessibility. A UMAP analysis was performed on the Pareto peaks that were annotated by formative (EPE) and non-formative (non-EPE; GCTM-2^{mid}-CD9^{mid}) peaks from the Lau et al. dataset[1].

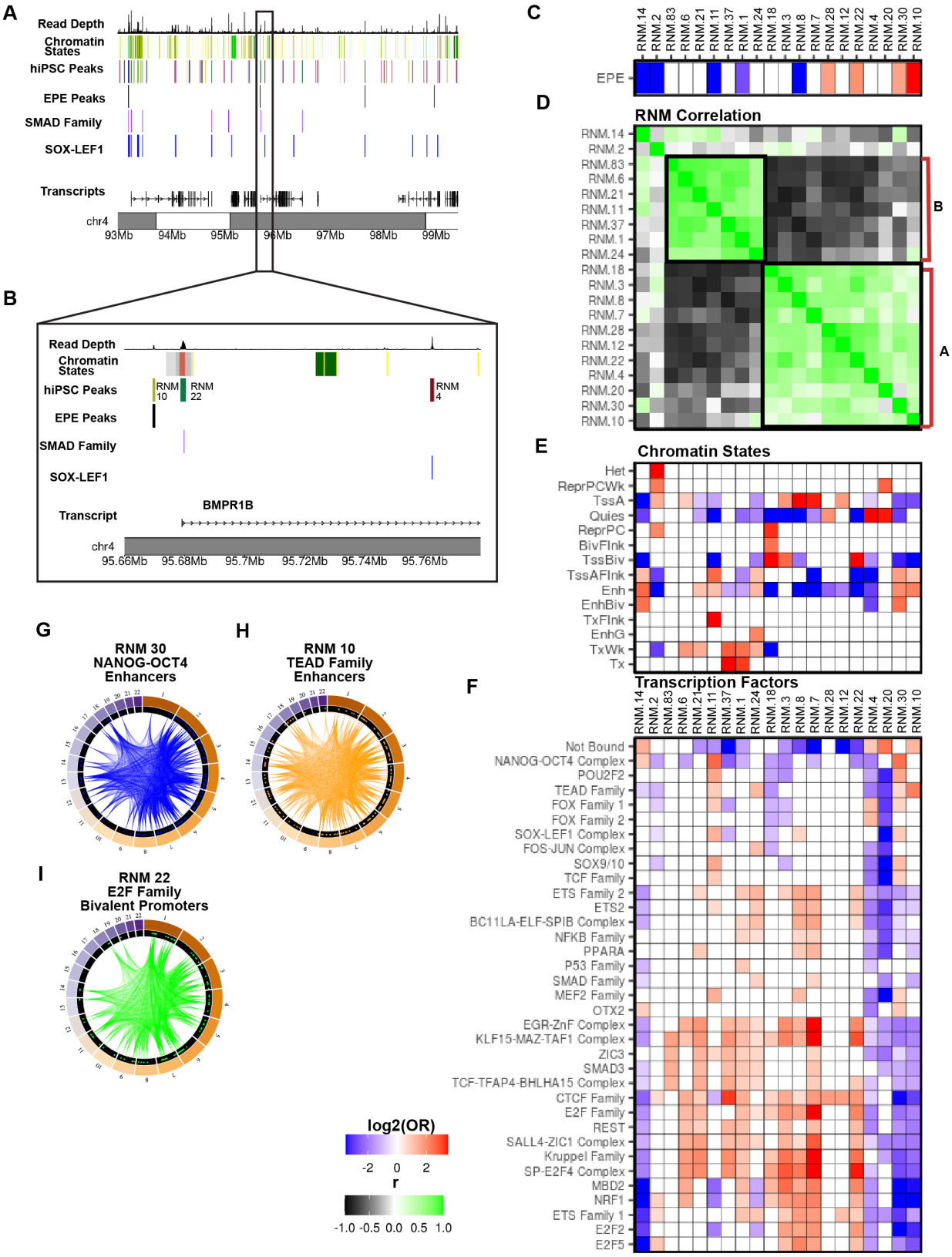

Figure 5: RNM Functional Chacterizations

(A) Genome browser plot exhibiting the distribution of the three epigenetic annotations in the *chr4:93,000,000:99,500,000* locus. The top track is the read depth of the merged BAM file of the 34 reference samples,

the track below is the iPSC-18 chromatin states colored by the their ChromHMM assigned colors (Figure 4B), the following tracks show co-accessibility of hiPSC ATAC-seq peaks (colored by RNM membership), the formative-state (EPE) ATAC-seq peaks from Lau et al[1], predicted binding sites for SMAD Family and SOX-LEF1 Complex, and transcript hg19 coordinates.

(B) Detailed genome browser view highlighting the *BMPR1B* region of the *chr4:95,660,000:95,770,000* locus. As indicated by the black box, the plot is a subset of the locus displayed in Figure 5A, and has concordant track ordering.

(C) Heatmap displaying the enrichment of formative-associate (EPE) peaks in RNMs. Fisher's Exact tests were used to calculate the enrichment (Odds Ratio) of EPE-associated peaks in each RNM. Each cell is filled with the log2(Odds Ratio) for EPE-associated peaks and the corresponding GNM (red = enrichment, blue = depletion, white = non-significant). For plot legibility, the enrichment range was set to -3.5 to 3.5. Note: the RNM order on the x-axis is retained in Figure 5C-F, to avoid redundant labeling.

(D) Heatmap showing the pairwise associations between 20 RNMs based on the mean expression of Pareto peaks. Each cell is filled with the Pearson correlation (r), where -1 (black) represents an anti-correlation between two RNMs and 1 (green) represents a positive association between RNMs. Clusters of correlated RNMs are outlined by black boxes. Cluster A and Cluster B were respectively composed of eleven and seven RNMs.

(E) Heatmap showing enrichment of iPSC-18 chromatin states in RNMs. Fisher's Exact tests were used to calculate the enrichment (Odds Ratio) of the annotations for 15 iPSC chromatin states in each RNM. Each cell is filled with the $\log_2$(Odds Ratio) for the corresponding chromatin state and RNM (red = enrichment, blue = depletion, white = non-significant). For plot legibility, the enrichment range was set to -3.5 to 3.5.

(F) Heatmap displaying enrichment of 35 selected TF groups in RNMs. Fisher's Exact tests were used to calculate the enrichment (Odds Ratio) of the TF group annotations in each RNM. Each cell is filled with the $\log_2$(Odds Ratio) for the corresponding TF group and RNM (red = enrichment, blue = depletion, white = non-significant). For plot legibility, the enrichment range was set to -3.5 to 3.5. A heatmap reporting the RNM enrichment for all 92 TF groups and "Not Bound" is shown in Figure S11.

(G-I) Biocircos plots displaying enhancer co-accessibility in RNMs. The Biocircos plots consist of 3 layers; 1) the outermost layers (brown-purple scale) represent the 22 autosomal chromosomes, 2) the middle circle (black) contains the points representing positions of co-accessible ATAC-seq on the corresponding chromosomes (layer #1), and 3) the links represent co-accessibility between ATAC-seq peaks (layer #2) across the chromosomes. (G) Shows the co-accessibility of enhancers in RNM 30 bound by NANOG-OCT4 Complex in enhancers (dark blue), (H) Shows the co-accesibility of enhancers in RNM 10 bound by TEAD Family in enhancers (orange), (I) Shows the co-accessibility of enhancers in RNM 22 bound by E2F Family in bivalent promoters (green).

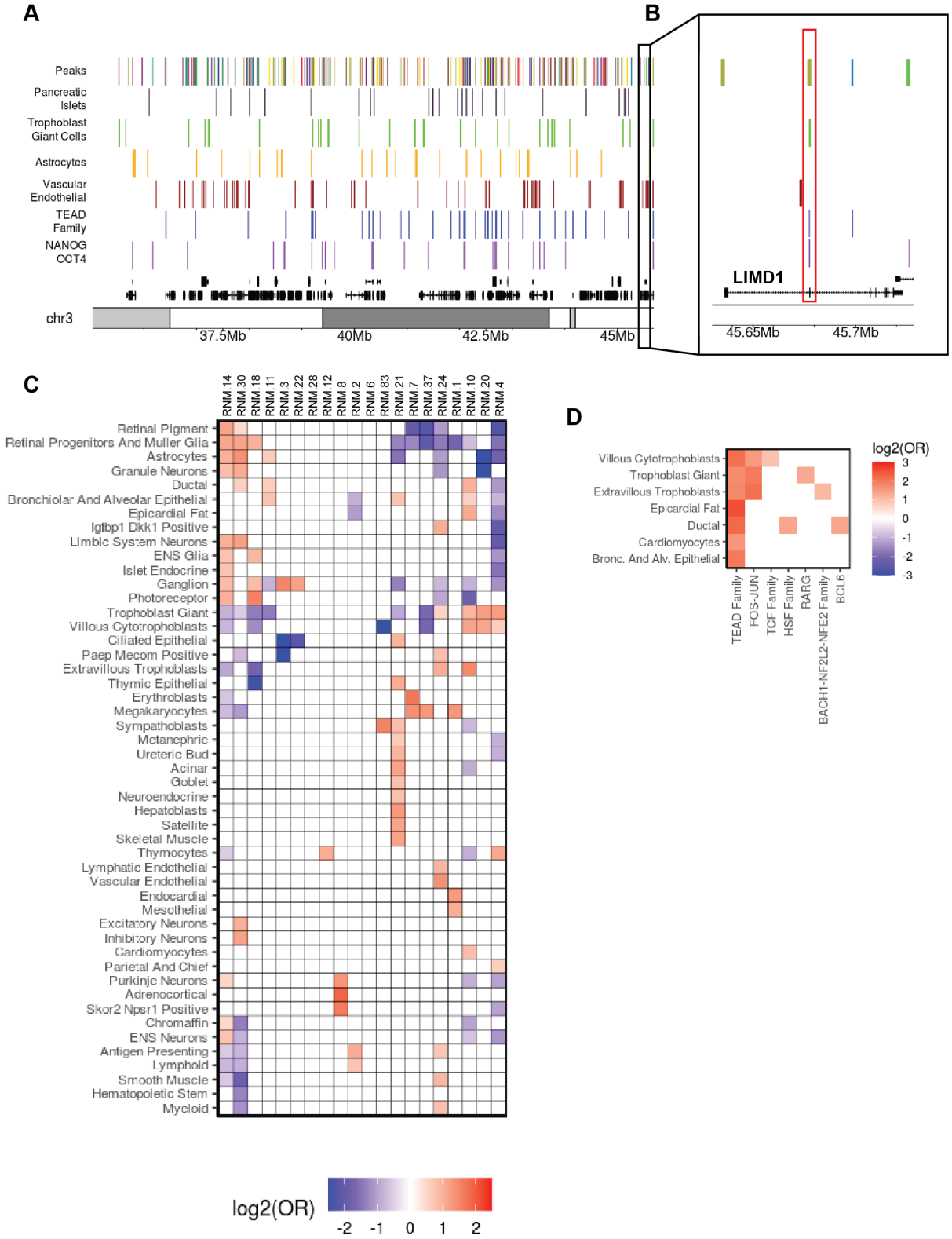

Figure 6. RNMs enrichments for fetal cell type-specific ATAC-seq peaks

(A) Genome browser visualization showing (from the top): 1) hiPSC ATAC-seq peaks (colored by RNM); fetal 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 6A at higher resolution, and demonstrates that an RNM 10 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 fetal cell type-specific peaks in the 20 RNMs. 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. 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 TFs bound in hiPSC ATAC-seq peaks underlying the cell type associations with RNM 10. Each cell is filled with the $\log_2$(Odds Ratio) for the corresponding fetal cell types and TFs bound in RNM 10 hiPSC peaks (red = enrichment, blue = depletion, white = non-significant).

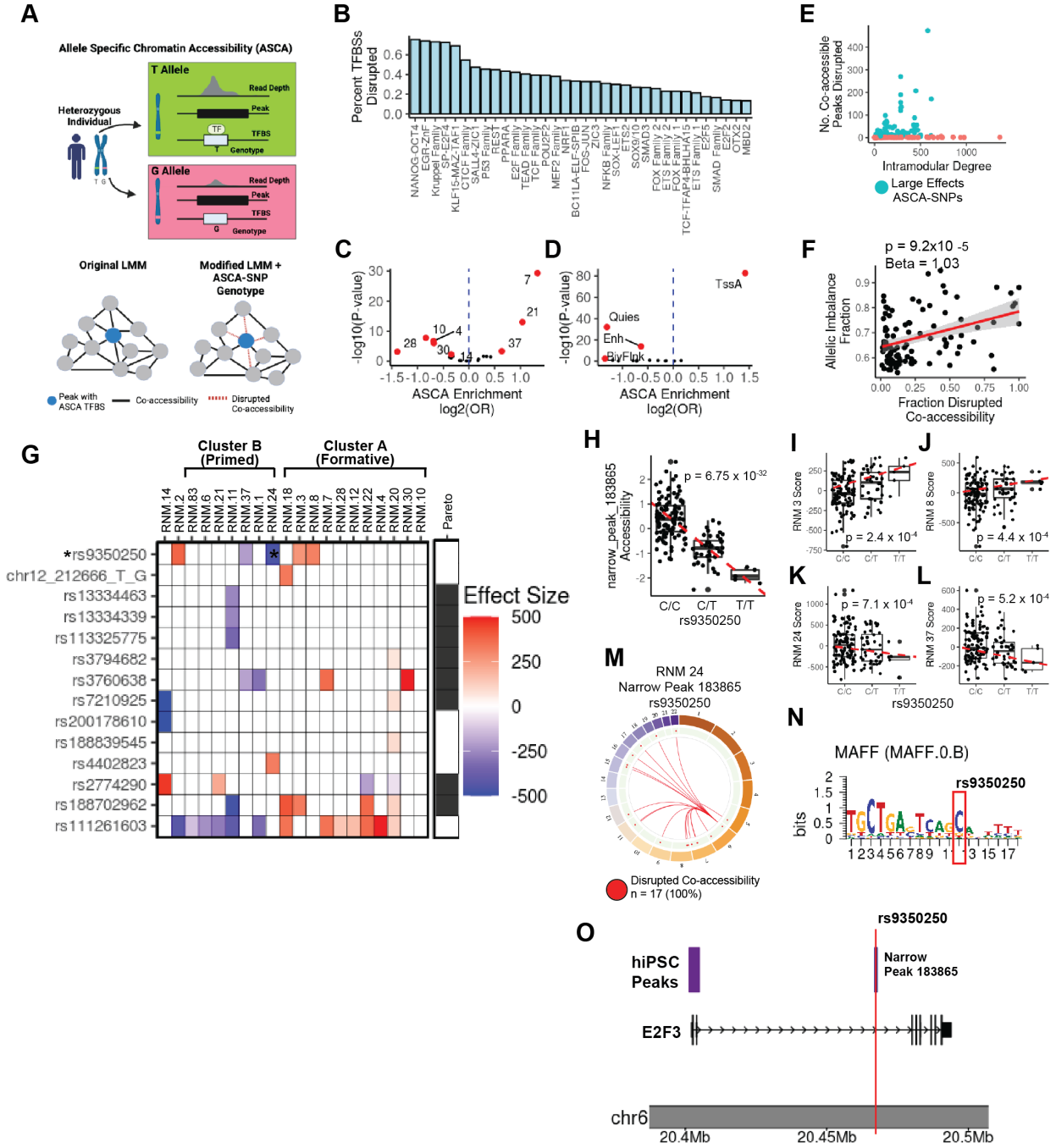

Figure 7. Genetic Perturbations of Regulatory Networks

(A) Diagram showing the functional impact of an Allele Specific Chromatin Accessibility SNP (ASCA-SNP) and genetic perturbation of regulatory networks

(B) Barplot showng the percent of hiPSC ATAC-seq peaks that contained the indicated TF group which showed ASCA

(C) Volcano plot showing enrichments of the 20 RNMs for ASCA-SNPs

(D) Volcano plot showing enrichments of the 15 chromatin states for ASCA-SNPs

(E) Scatter plot of the 343 ASCA-SNPs with disrupted co-accessibility with at least one peak. Shown are the number of peak pairs that were co-accessible (Intramodular Degree) with the peak containing the ASCA-SNP in the original LMM and the number of peak pairs whose co-accessibility was disrupted when accounting for ASCA-SNP genotypes (No. Co-accessible Peaks Disrupted). The 101 ASCA-SNPs with large effects are shown in turquoise.

(F) Scatter plot showing the association between AIF and disrupted co-accessibility. Each point represents an ASCA-SNP that disrupts the integrity of it's corresponding RNM. The regression line was generated from the linear model applied to calculate the association between the fraction of disrupted co-accessibility and AIF.

(G) Heatmap displaying associations between ASCA-SNPs and the integrity of the 20 major RNMs. The RNMs are in the same order as Figure 5D, where they clustered based on pluripotency state enrichments: A (formative) and B (primed). The LMM effect sizes (β) for the 14 ASCA-SNPs with at least one significant association are shown. Non-significant results are shown in white. To the right of the heatmap, is a binary annotation indicating whether the ASCA-SNP is in a Pareto peak (black). An asterisk demarcates ACSA-SNP rs9350250 which is located within RNM 24 narrow peak 183865.

(H) Boxplot showing association between rs9350250 and accessibility of narrow peak 183865 as determined by applying an LMM. The quantile normalized narrow peak 183865 is shown on the y-axis, the rs9350250 genotypes are on the x-axis and the LMM p-value is embedded in the plot.

(I-L) Boxplot showing association between rs9350250 and Pareto peak accessibility for RNMs 3 (I), 8 (J), 24 (K) and 37 (L). The RNM score is shown on the y-axis, the rs9350250 genotypes are on the x-axis and the LMM p-value is embedded in the plot.

(M) Biocircos plot displaying ACSA-SNP rs9350250 disruption of co-accessibility between the narrow peak 183865 and 17 partner peaks in RNM 24. The outermost layer represent the 22 autosomal chromosomes, the middle layer displays the location of peaks (represented by points) that are co-accessibile with narrow peak 183865 in the original LMM and the links in the innermost layer represent that all co-accessibility with narrow peak 183865 is disrupted (red) by ACSA-SNP rs9350250.

(N) Consensus motif sequence for MAFF (MAF Family TF) disrupted by rs9350250. The black retangle indicates the relative position of rs9350250 in the motif.

(O) Genome browser track showing the *E2F3* locus with ASCA-SNP rs9350250. The red line represents the rs9350250 position, the first track is the location of the peak (purple rectangle), the third and fourth tracks show the coordinates of *E2F3*.

SUPPLEMENTAL FIGURES

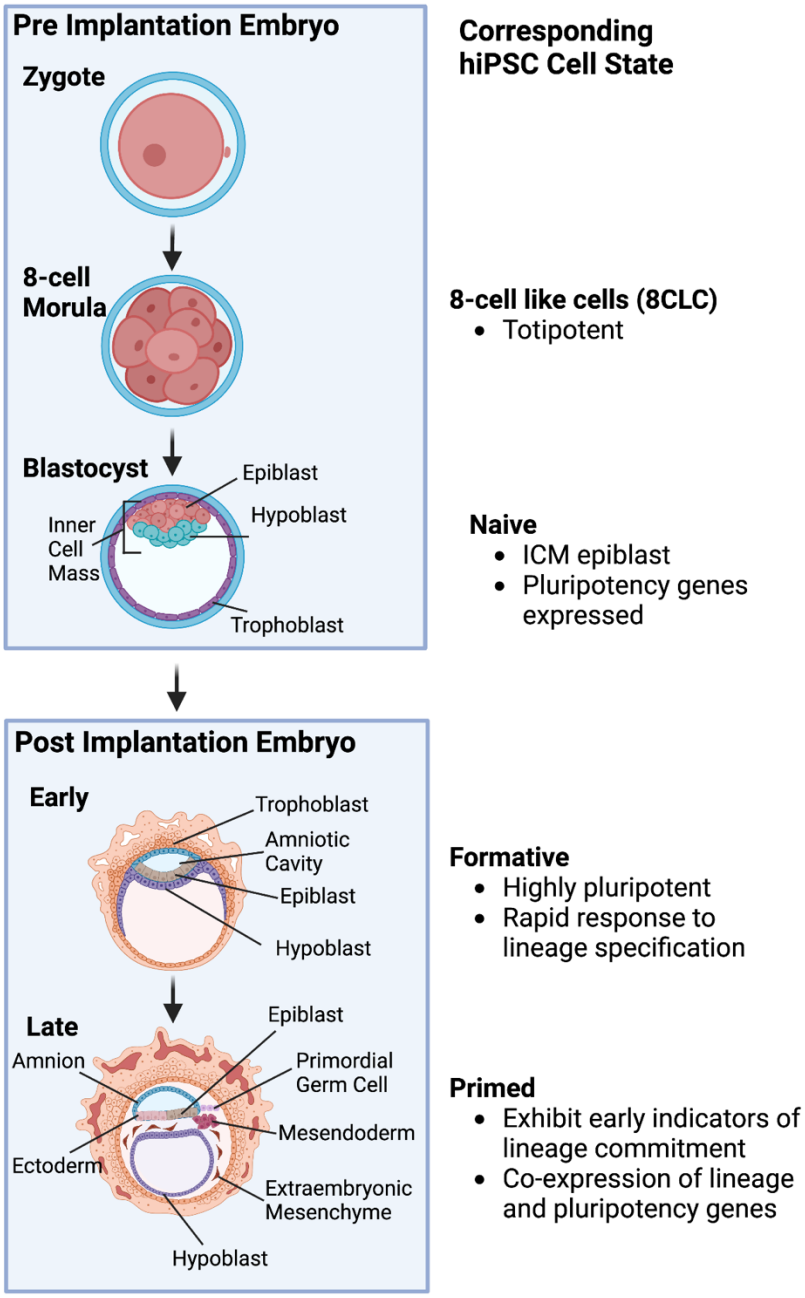

Figure S1. Pluripotency cell states in human induced pluripotent stem cell lines.

This figure was adapted from Pera and Rossant[5] and was generated to describe early embryonic cell states and their analogs in hiPSC lines. Recent studies have shown that 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[6,8] 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[1,6]. A very small proportion of stem cells in hiPSC lines are totipotent 8-cell like cells (8CLC) and resemble the 8-cell

morula. “Naïve” stem cells represent the cellular state of the inner cellular mass (ICM) preimplantation epiblast which gives rise to the embryo proper. 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 stem cells enriched for high self-renewal[1]. Most stem cells within a conventional hiPSC culture resemble the late post-implantation epiblast stem cells[1,5] that comprise the primitive streak, and are referred to as “primed”. Primed stem cells exhibit early signs of cell fate commitment by co-expressing lineage-specific and pluripotency genes. Figure was generated with Biorender (license: XY259BIAH1).

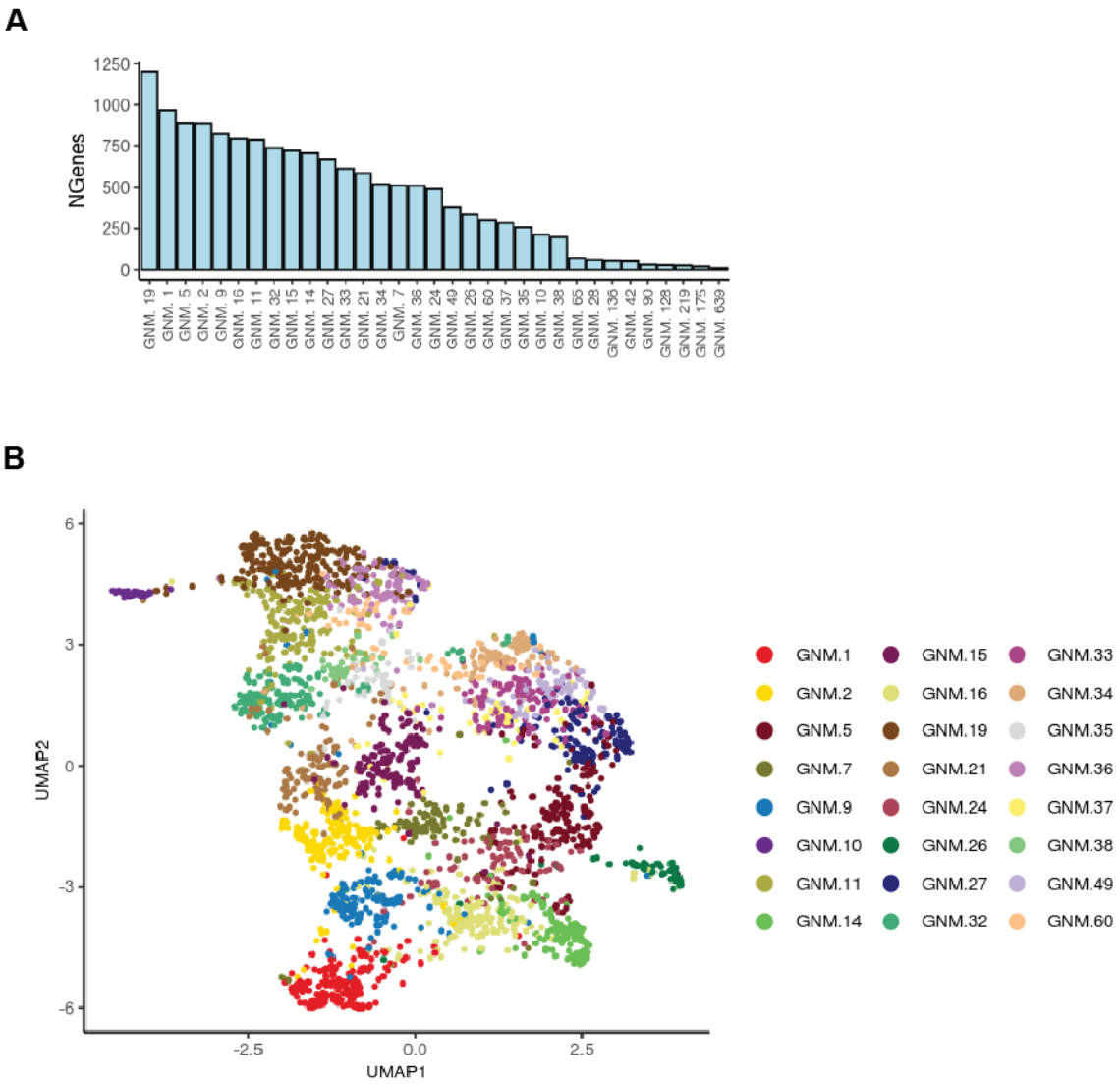

Figure S2: Properties of the gene co-expression network

(A) Barplot showing the number of genes in each of the 33 GNMs (membership > 10 genes).

(B) UMAP plot showing the expression of 2,862 Pareto genes for the 24 major GNMs (membership > 100 genes) across 213 RNA-seq samples

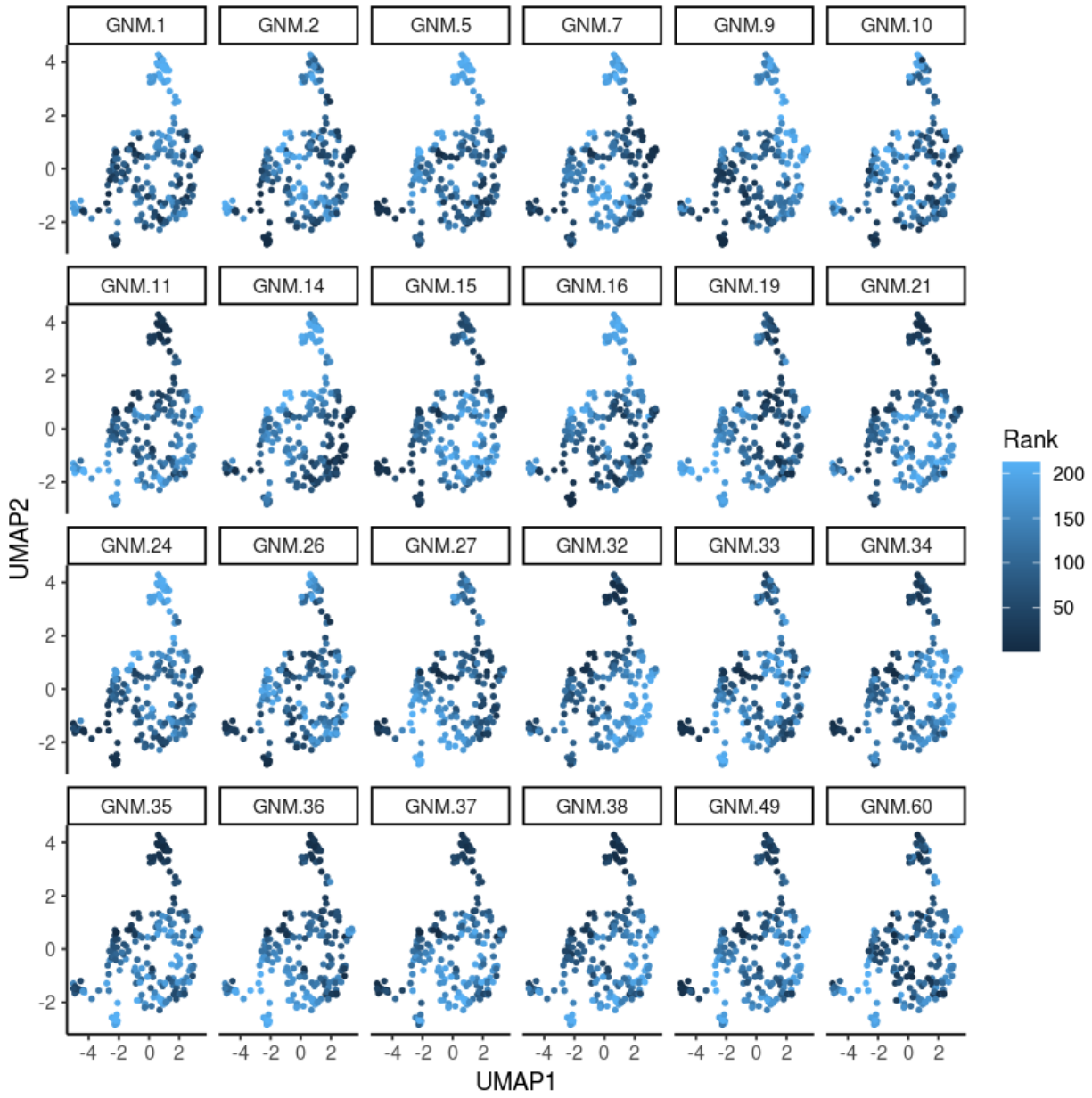

Figure S3. GNMs are differentially expressed across hiPSC lines

UMAP plots showing the GNM expression scores for all 24 major GNMs in the 213 hiPSC RNA-seq samples. A UMAP analysis was performed on all 213 RNA-seq based on the expression of the 2,862 Pareto genes in the 24 major GNMs. A GNM expression score was calculated for each sample (see Methods), then, for visualization purposes, samples were ranked from 1 to 213 and colored based on their GNM score. The plots demonstrate that varying proportions of pluripotency cell state subpopulations across the 213 hiPSC lines result in clustering based on the expression of cell state-specific gene networks.

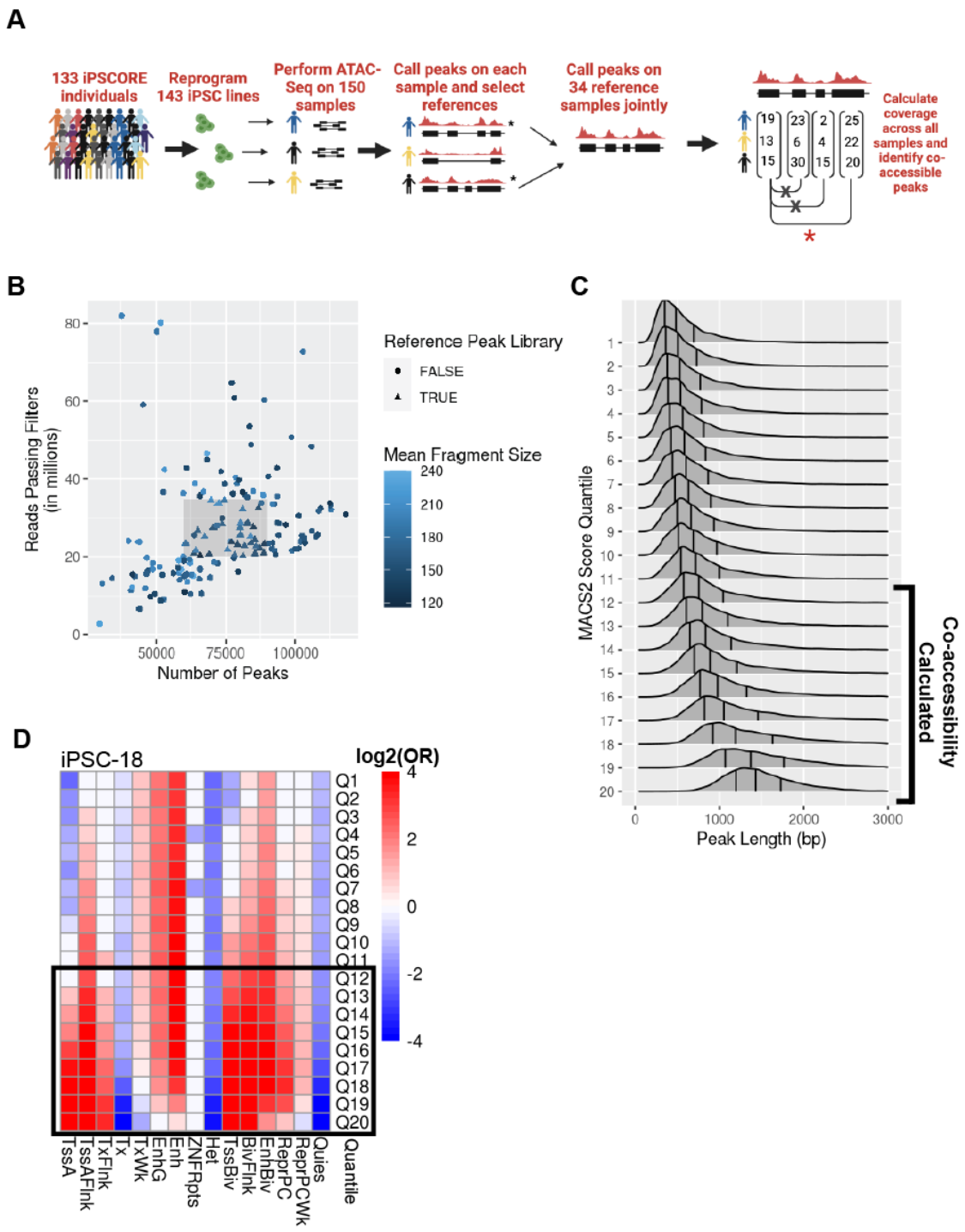

Figure S4. ATAC-seq peak calling strategy and quality control

(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 59,499 autosomal peaks (see details below) 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) Ridge plot demonstrating association of ATAC-seq peak length and MACS2 peak score. To reduce the computational burden of testing ~1.7B pairwise combinations of ATAC-seq peaks, we selected a subset of 59,499 autosomal peaks based on their MACS2 peak score. We binned the 132,225 ATAC-seq peaks on autosomes into 20 MACS2 score quantiles each containing 6,611 peaks, where quantile 20 consisted of the peaks with highest scores. We then looked at the density of ATAC-seq peak length by the 20 quantiles and observed that peaks in quantiles 12-20 were longer than lower quantiles. In conjunction with the chromatin state enrichments (D) and predicted transcription factor occupancy (Figure S8A), we selected the 59,499 ATAC-seq peaks in quantiles 12-20 to calculate pairwise co-accessibility, as indicated by the label.

(D) Heatmap showing the enrichment of ATAC-seq peaks in hiPSC-18 ChromHMM chromatin states by MACS2 score quantile. As described above (C), we used chromatin state enrichments to prioritize ATAC-seq peaks for downstream co-accessibility analysis. We created 20 bed files containing the coordinates for peaks in each quantile. We then calculated their enrichment (Odds Ratio) in each of the 15 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 13-20 and Enh are enriched across all 20 quantiles. Additionally, TssBiv and ReprPC started to exhibit diminishing enrichment around quantile 12. These observations supported our decision to use peaks in quantiles 12-20 (annotated by the black box) for downstream analyses.

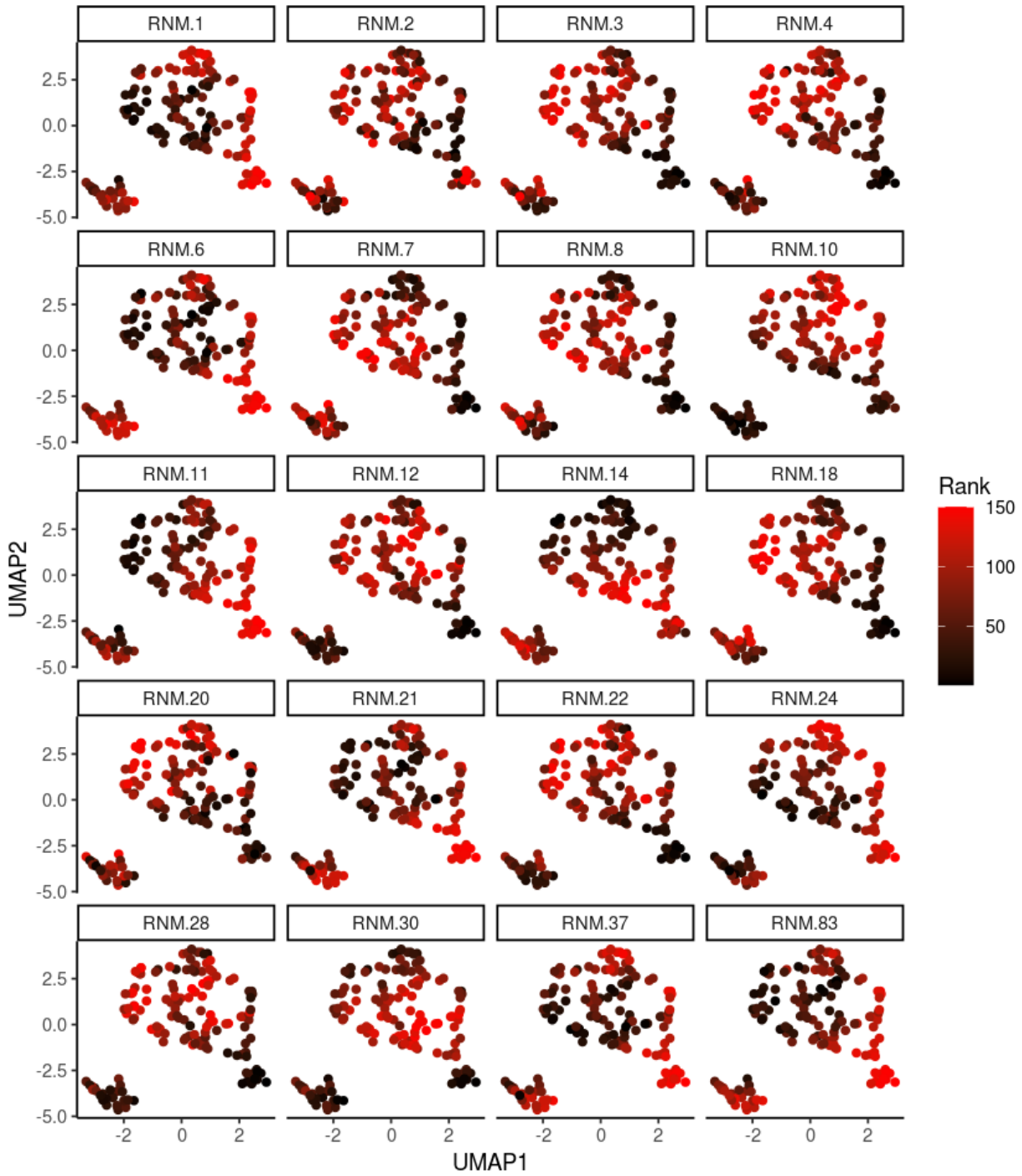

Figure S5. RNMs are differentially accessible across hiPSC lines

UMAP plots showing the RNM accessibility scores for all 20 major GNMs in the 150 hiPSC ATAC-seq samples. A UMAP analysis was performed on all 150 ATAC-seq based on the accessibility of the 8,972 Pareto peaks in the 20 major RNMs. A RNM accessibility score was calculated for each sample (see Methods), then, for visualization purposes, samples were ranked from 1 to 150 and colored based on their RNM score. The plots demonstrate that varying proportions of pluripotency cell state subpopulations across the 150 hiPSC lines result in clustering based on the accessibility of cell state-specific regulatory networks.

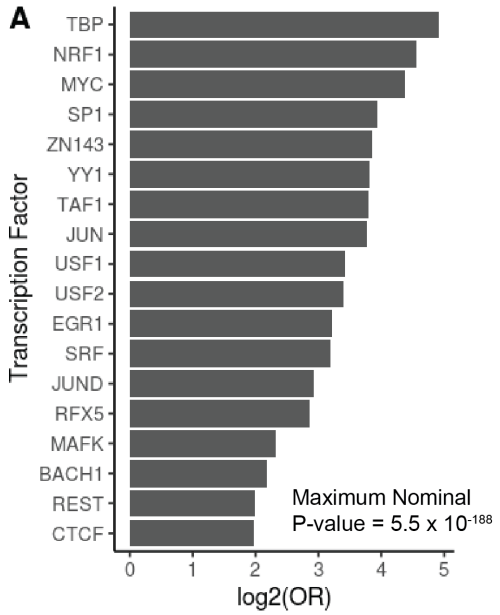

Figure S6. Validation of TOBIAS TFBS predictions

(A) Barplot demonstrating 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 = 5.5×10^{-188}, Table S10).

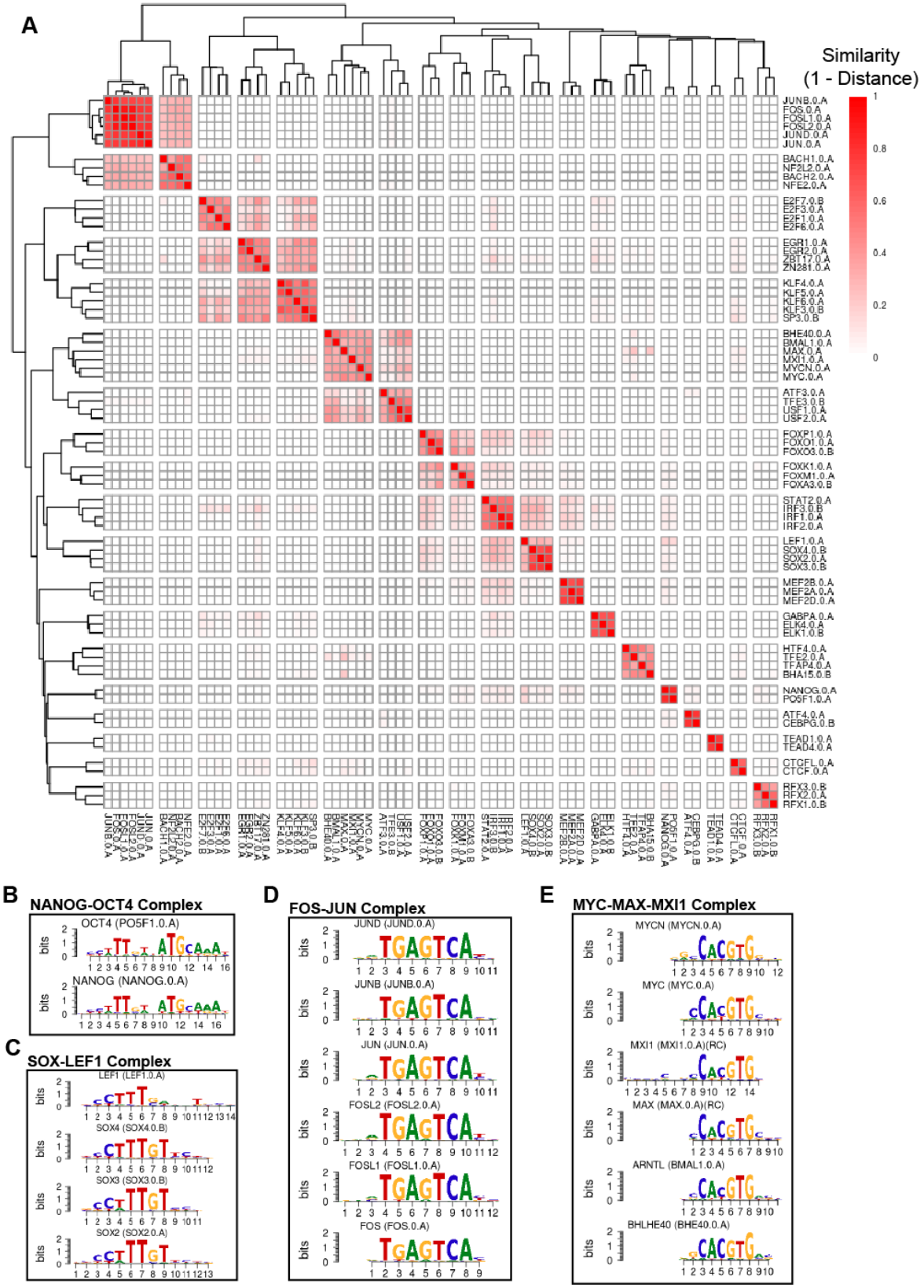

Figure S7. Defining TF groups by binding site similarity

(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 S9). 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).

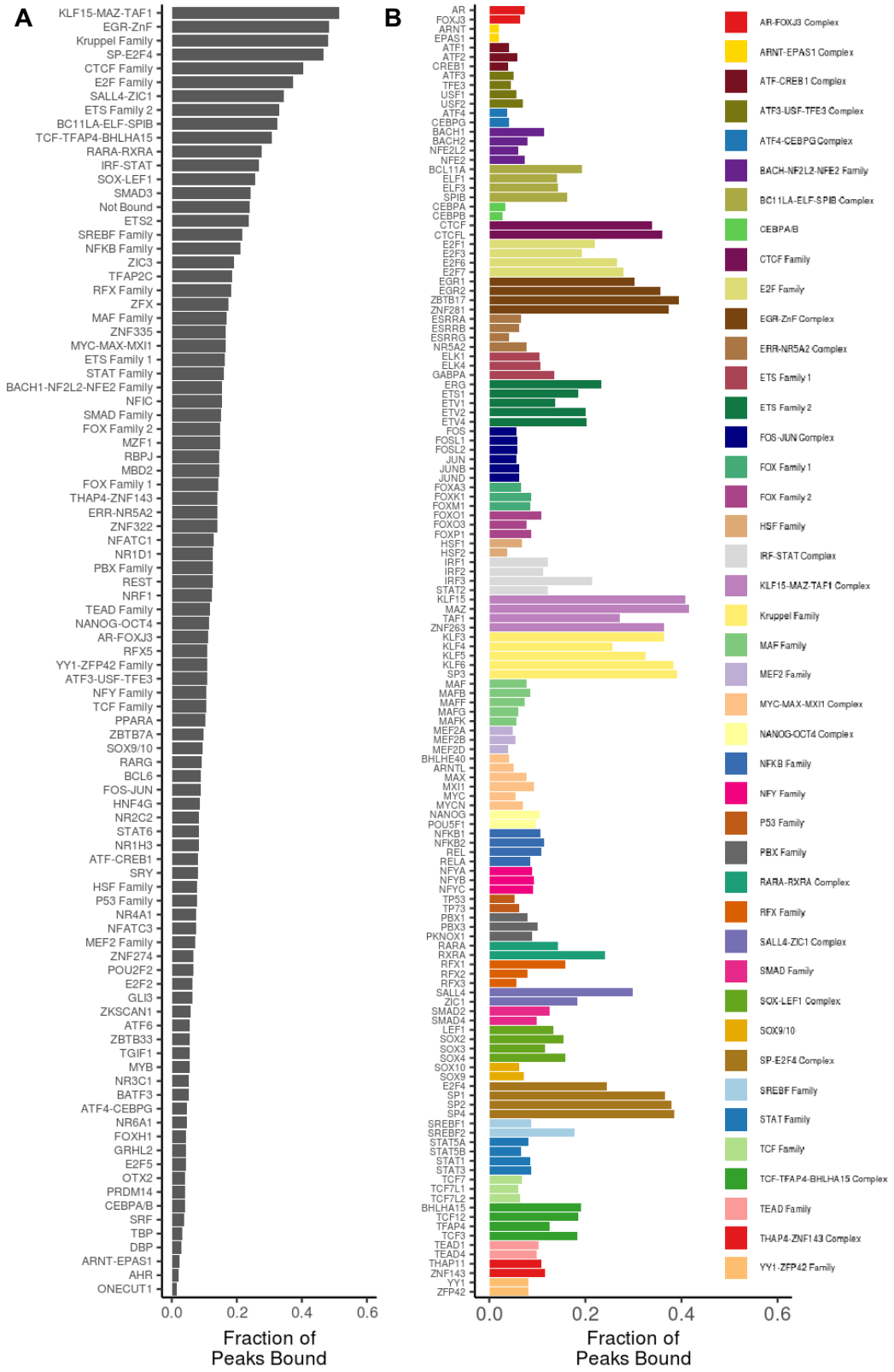

Figure S8. Profile of predicted TFBS across the hiPSC epigenome

(A) Barplot showing the fraction of ATAC-seq peaks bound with a predicted TFBSs. Quantiles 12-20 captured 92.3% of ATAC-seq peaks on autosomes with at least one predicted binding site (data not shown). We calculated the fraction of the 59,499 ATAC-seq peaks (quantiles 12-20) that are bound by the 187 individual TFs.

(B) 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 S6) into 92 TF groups (49 single motif, 24 same TF family, 19 contained TFs from different families and are referred to as Complexes). The individual TFs that were collapsed into one of the 43 TF groups are colored and sorted corresponding to their TF group. Many of the grouped TFs were often predicted to be bound at the same site and hence are also predicted to be bound in similar proportions of ATAC-seq peaks. For these reasons, the identity of the exact TF bound cannot be resolved, thus justifying the curation of TF groups. A subset of the TF groups is shown in Figure 4C.

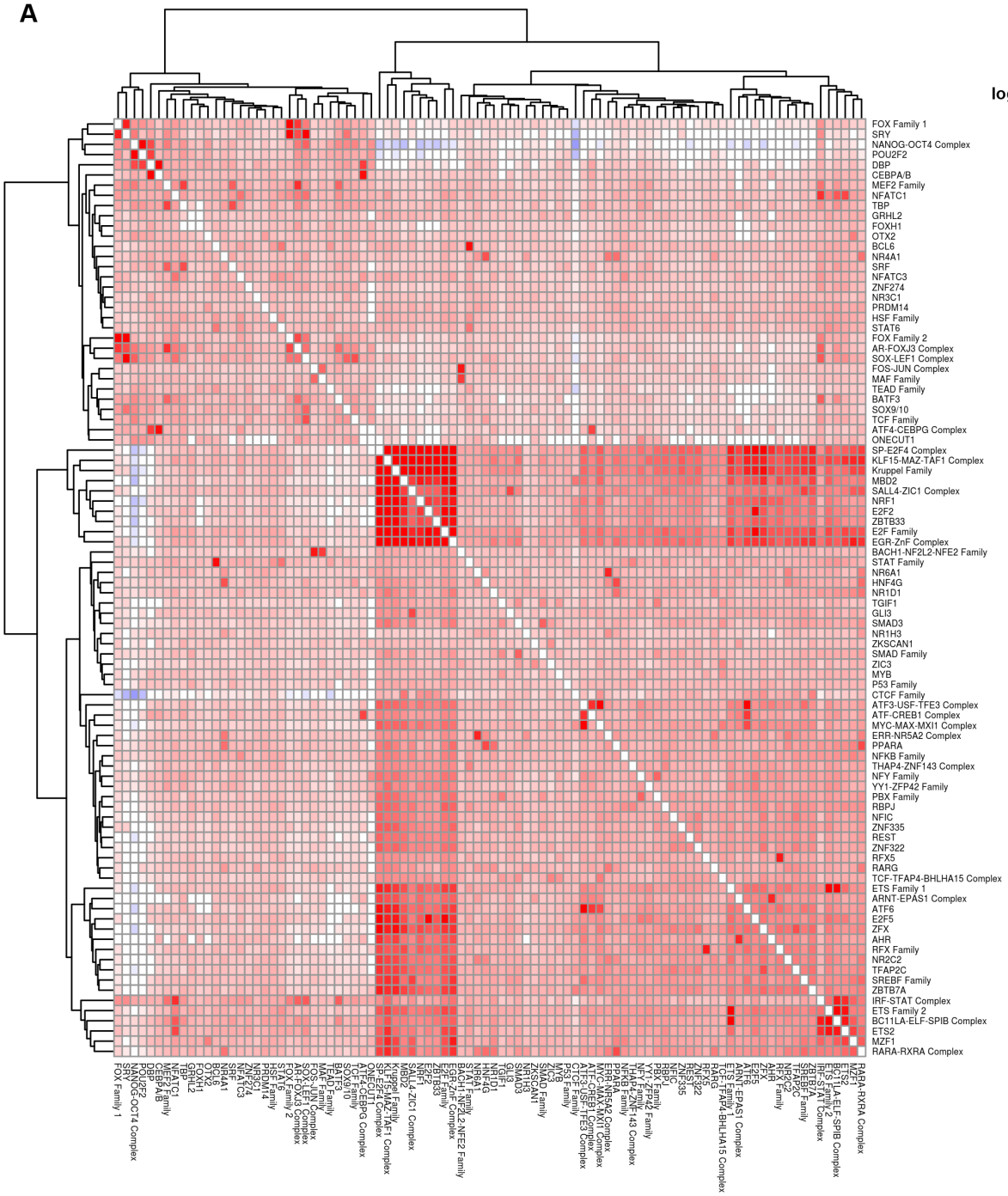

Figure S9. High degree of TF group co-binding in ATAC-seq peaks

(A) Heatmap demonstrating the co-occurrence of TF groups within a single ATAC-seq peak. Fisher's Exact tests were performed to calculate the enrichment of pairwise TF-TF co-binding. Most TFs are enriched for co-binding with other TFs, indicating that the ATAC-seq peaks are highly enriched with diverse regulatory elements. Notably, pluripotency-associated TF groups (NANOG-OCT4 Complex and POU2F2/OCT2) cluster with FOX Family 1 and SRY. ATAC-seq peaks bound NANOG-OCT4 Complex are depleted for CTCF Family, E2F TFs, zinc finger-associated TFs (Kruppel Family, EGR-ZnF Complex), and other TFs that are highly active during the G1 phase, which is consistent with previous characterizations of cells in the formative pluripotency state. A subset of these data is shown in Figure 4D.

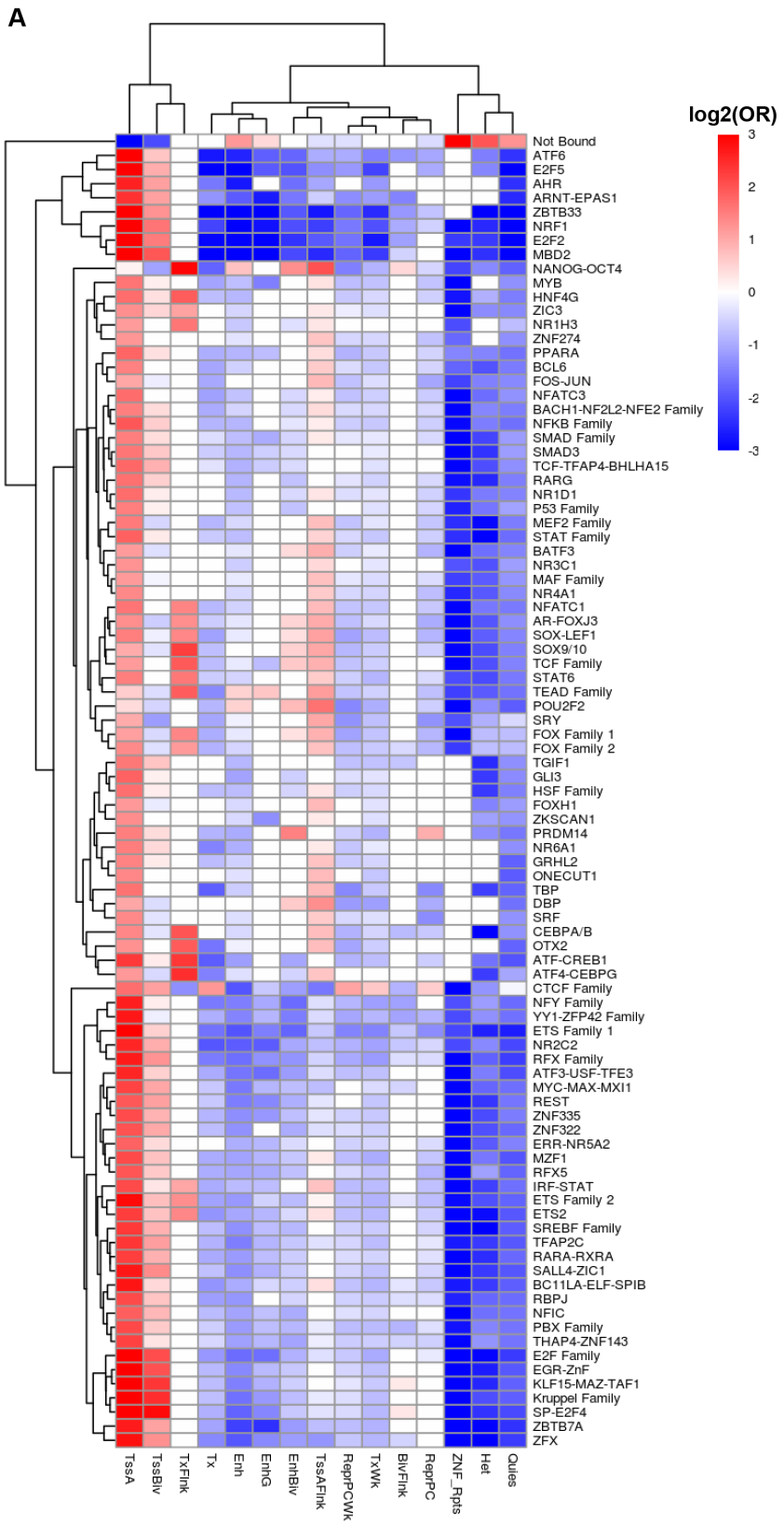

Figure S10. TF groups exhibit variable affinities to different chromatin states

(A) Heatmap showing TF group affinities for different chromatin states. We permuted the TF group and chromatin state annotations for the 59,499 ATAC-seq peaks 100 times to calculate TFBS affinities (Z scores) for 15 chromatin states. We

observe that certain TF groups are more likely to be bound in active or bivalent promoters (E2F TFs, zinc finger-associated TFs, ETS TFs, etc.), while others have stronger affinities for enhancers and regions flanking promoters (NANOG-OCT4, FOX TFs, TEAD Family, SOX TFs, and FOS-JUN Family) and CTCF Family is bound in quiescent, regions of active transcription, and repressed polycomb regions. A subset of these data is shown in Figure 4E).

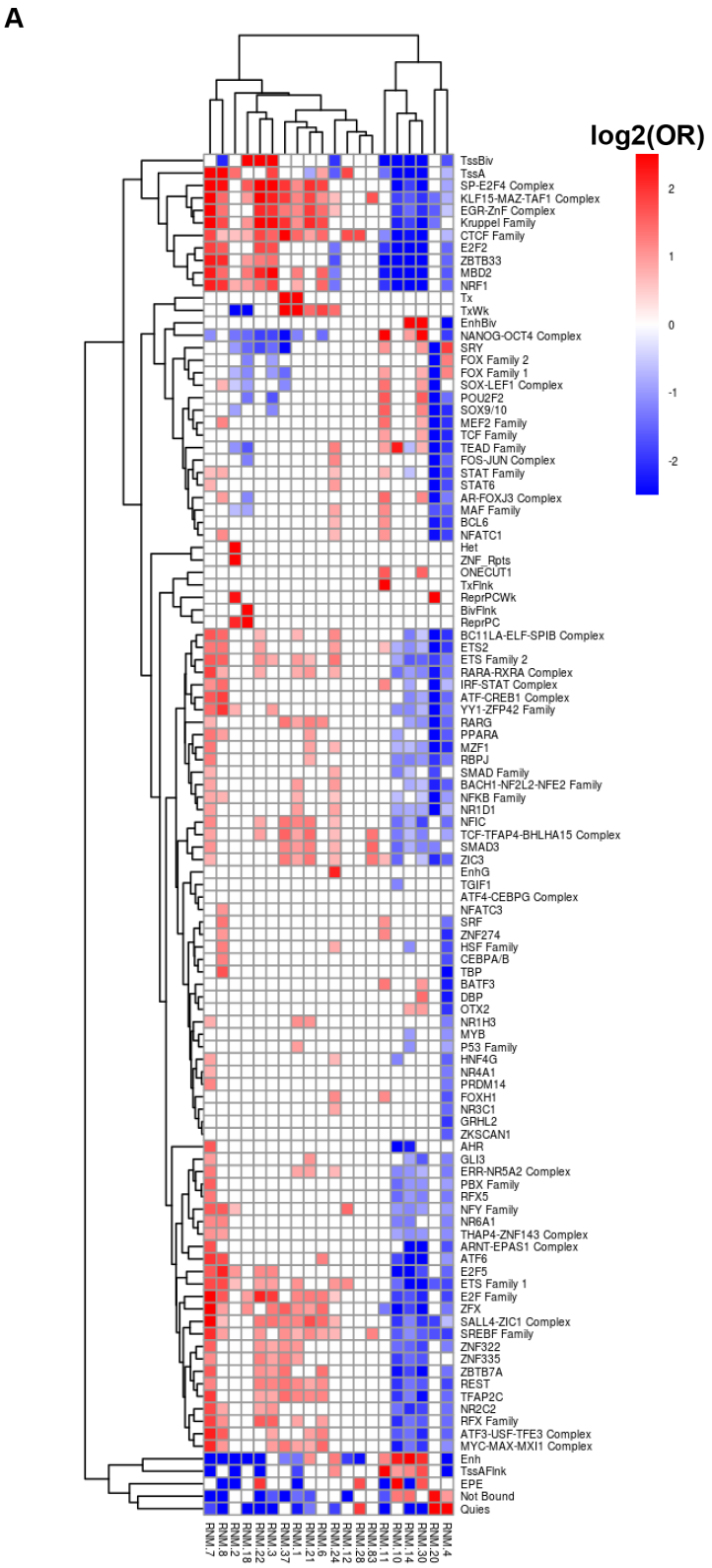

Figure S11. RNM Enrichments with Three Classes of Epigenetic Markers

(A) Heatmap displaying epigenetic annotations of the 20 major RNMs. This figure is a comprehensive version of Figure 5C, E, F and consists of RNM enrichments of all 92 TF groups, "Not Bound" peaks, 15 hiPSC chromatin states and 1 self-

renewal cell state. Only significant enrichments are shown (non-significant enrichments are filled in white). The results for 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), the 20 genotype PCs for global ancestry. If there is corresponding RNA-seq data (Column F), ATAC-seq data (Column G), and the top 20 genotype PCs (Columns H-AA) for the individual.

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 clone (Column D) and passage (Column E) numbers, and the number of properly paired reads (Column F).

Table S3: Gene Network Module Memberships

Information about gene network module (GNM) analysis annotations for the 16,110 expressed genes, including the Gencode (v24) 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 (Column F), and the gene's co-expression degree connectivity (Column H) within its corresponding module (intramodular degree).

Table S4: GNM Enrichment Results

Results from the Fisher's Exact test to calculate pluripotency cell state and gene ontology (GO) enrichments. Information includes the corresponding analysis (GO enrichment or cell state) (Column A), the annotation tested (Column B), the GNM tested (Column C), the odds ratio (Column D), the p-value (Column E) and the Benjamini-Hochberg corrected p-value (Column F). Only significant test results are shown.

Table S5: GNM Correlation Results

Resuls from the pairwise Spearman Correlation tests to calculate similarities between GNMs. Information includes the two GNMs being compared (Columns A and B), the Spearman's correlation (r, Column C), the corresponding p-value and the 95% confidence interval (Column E-F).

Table S6: ROCK inhibitor Stimulation Differential Expression

Differential expression results from three iPSCORE iPSC lines cultured with 5μM ROCK inhibitor. Information includes Gencode (v34) gene id (Column A), gene name (Column B), and the standard DESeq2 output (Columns C-H).

Table S7: ATAC-seq Sample Information

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

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

Information about regulatory network module (RNM) analysis annotations for the 59,499 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 20 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 59,499 accessible peaks, 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), and the Gencode v34 gene id (Column K), gene name (Column L), the distance in base pairs (Column M) of the closest expressed gene after ROCK kinase inhibitor stimulation, and whether the peak overlaps an EPE (Column N) or Non-EPE (Column O) peak.

Table S9: 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).

Table S10: 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 S11: Annotations of 59,499 peaks for binding of 92 TF groups

This table includes the TOBIAS-predicted transcription factor binding sites for all 59,499 ATAC-seq peaks. Information includes; the peak ID, peak chromosome, peak start coordinates, peak end coordinates (Column A-D), and binary annotations for the 92 collapsed TF groups and "Not Bound" peaks (Columns E-CS), 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 (EPE), 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 Correlation Results

Results from the pairwise Spearman Correlation tests to calculate similarities between RNMs. Information includes the two RNMs being compared (Columns A and B), the Spearman's correlation (r, Column C), the corresponding p-value and the 95% confidence interval (Column E-F).

Table S14: 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 S15: Allele Specific Chromatin Accessibility (ASCA) and Network Disruption Results

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

Table S16: Associations between ASCA-SNPs and accessibility of Pareto peaks

This table includes the results from the LMM used to calculate the association between ASCA-SNP genotypes and RNM scores, including the SNP ID (Column A), the RNM tested (Column B), the effect size, standard error, and P-value from the LMM (Columns C-E), and the Benjamini-Hochberg corrected p-value (Column F).

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. Sozen, B., Conkar, D., and Veenvliet, J.V. (2022). Carnegie in 4D? Stem-cell-based models of human embryo development. Semin. Cell Dev. Biol. *131*, 44–57. 10.1016/j.semcdb.2022.05.023.

5. 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.

6. 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.

7. Petropoulos, S., Edsgärd, D., Reinius, B., Deng, Q., Panula, S.P., Codeluppi, S., Plaza Reyes, A., Linnarsson, S., Sandberg, R., and Lanner, F. (2016). Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell *165*, 1012–1026. 10.1016/j.cell.2016.03.023.

8. Nakanishi, M., Mitchell, R.R., Benoit, Y.D., Orlando, L., Reid, J.C., Shimada, K., Davidson, K.C., Shapovalova, Z., Collins, T.J., Nagy, A., et al. (2019). Human Pluripotency Is Initiated and Preserved by a Unique Subset of Founder Cells. Cell *177*, 910-924.e22. 10.1016/j.cell.2019.03.013.

9. Gropp, M., Waldhorn, I., Gil, Y., Steiner, D., Turetsky, T.T., Smith, Y., Sabag, O., Falick-Michaeli, T., Even Ram, S., and Reubinoff, B.E. (2022). Laminin111-based defined culture promoting self-renewing human pluripotent stem cells with properties of the early post-implantation epiblast. Stem Cell Rep. *17*, 2643–2660. 10.1016/j.stemcr.2022.10.010.

10. 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.

11. Yu, L., Wei, Y., Sun, H.-X., Mahdi, A.K., Pinzon Arteaga, C.A., Sakurai, M., Schmitz, D.A., Zheng, C., Ballard, E.D., Li, J., et al. (2021). Derivation of Intermediate Pluripotent Stem Cells Amenable to Primordial Germ Cell Specification. Cell Stem Cell *28*, 550-567.e12. 10.1016/j.stem.2020.11.003.

12. 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.

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

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

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

20. 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.

21. 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.

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

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. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–15550. 10.1073/pnas.0506580102.

30. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740. 10.1093/bioinformatics/btr260.

31. Ohgushi, M., and Sasai, Y. (2011). Lonely death dance of human pluripotent stem cells: ROCKing between metastable cell states. Trends Cell Biol. *21*, 274–282. 10.1016/j.tcb.2011.02.004.

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. 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.

34. 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.

35. 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.

36. 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.

37. 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.

38. Liang, J., Wan, M., Zhang, Y., Gu, P., Xin, H., Jung, S.Y., Qin, J., Wong, J., Cooney, A.J., Liu, D., et al. (2008). Nanog and Oct4 associate with unique transcriptional repression complexes in embryonic stem cells. Nat. Cell Biol. *10*, 731–739. 10.1038/ncb1736.

39. Pan, G., and Thomson, J.A. (2007). Nanog and transcriptional networks in embryonic stem cell pluripotency. Cell Res. *17*, 42–49. 10.1038/sj.cr.7310125.

40. Hamilton, D.J., Hein, A.E., Holmes, Z.E., Wuttke, D.S., and Batey, R.T. (2022). The DNA-Binding High-Mobility Group Box Domain of Sox Family Proteins Directly Interacts with RNA *In Vitro*. Biochemistry *61*, 943–951. 10.1021/acs.biochem.2c00218.

41. Zervos, A.S., Gyuris, J., and Brent, R. (1993). Mxi1, a protein that specifically interacts with Max to bind Myc-Max recognition sites. Cell *72*, 223–232. 10.1016/0092-8674(93)90662-A.

42. Angel, P., and Karin, M. (1991). The role of Jun, Fos and the AP-1 complex in cell-proliferation and transformation. Biochim. Biophys. Acta BBA - Rev. Cancer *1072*, 129–157. 10.1016/0304-419X(91)90011-9.

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

44. 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.

45. 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.

46. Derynck, R., and Zhang, Y.E. (2003). Smad-dependent and Smad-independent pathways in TGF-beta family signalling. Nature *425*, 577–584. 10.1038/nature02006.

47. 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.

48. 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.

49. 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.

50. 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.

51. Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S.H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. Cell *132*, 1049–1061. 10.1016/j.cell.2008.02.039.

52. 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.

53. Buecker, C., Srinivasan, R., Wu, Z., Calo, E., Acampora, D., Faial, T., Simeone, A., Tan, M., Swigut, T., and Wysocka, J. (2014). Reorganization of enhancer patterns in transition from naive to primed pluripotency. Cell Stem Cell *14*, 838–853. 10.1016/j.stem.2014.04.003.

54. 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.

55. 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.

56. 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.

57. 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.

58. 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.

59. 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.

60. 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.

61. 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.

62. 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.

63. 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.

64. 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.

65. 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.

66. Leone, G., DeGregori, J., Yan, Z., Jakoi, L., Ishida, S., Williams, R.S., and Nevins, J.R. (1998). E2F3 activity is regulated during the cell cycle and is required for the induction of S phase. Genes Dev. *12*, 2120–2130. 10.1101/gad.12.14.2120.

67. Preissl, S., Gaulton, K.J., and Ren, B. (2023). Characterizing cis-regulatory elements using single-cell epigenomics. Nat. Rev. Genet. *24*, 21–43. 10.1038/s41576-022-00509-1.

68. 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.

69. 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.

70. Dong, C., Beltcheva, M., Gontarz, P., Zhang, B., Popli, P., Fischer, L.A., Khan, S.A., Park, K.-M., Yoon, E.-J., Xing, X., et al. (2020). Derivation of trophoblast stem cells from naïve human pluripotent stem cells. eLife *9*, e52504. 10.7554/eLife.52504.

71. 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.

72. 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.

73. 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.

74. 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.

75. 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.

76. 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.

77. 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.

78. 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.

79. 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.

80. 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.

81. 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.

82. 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.

83. 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.

84. 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.

85. 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.

86. 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.

87. 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.

88. 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.

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. 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.

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.