Dynamic regulatory module networks for inference of cell type-specific transcriptional networks

Alireza Fotuhi Siahpirani[1,2,+], Sara Knaack[1+], Deborah Chasman[1,8,+], Morten Seirup[3,4], Rupa Sridharan[1,5], Ron Stewart[3], James Thomson[3,5,6], and Sushmita Roy[1,2,7*]

[1]Wisconsin Institute for Discovery, University of Wisconsin-Madison

[2]Department of Computer Sciences, University of Wisconsin-Madison

[3]Morgridge Institute for Research

[4]Molecular and Environmental Toxicology Program, University of Wisconsin-Madison

[5]Department of Cell and Regenerative Biology, University of Wisconsin-Madison

[6]Department of Molecular, Cellular, & Developmental Biology, University of California Santa Barbara

[7]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison

[8]Present address: Division of Reproductive Sciences, Department of Obstetrics and Gynecology, University of Wisconsin-Madison

[+]These authors contributed equally.

[*]To whom correspondence should be addressed.

Abstract

Changes in transcriptional regulatory networks can significantly alter cell fate. To gain insight into transcriptional dynamics, several studies have profiled bulk multi-omic datasets with parallel transcriptomic and epigenomic measurements at different stages of a developmental process. However, integrating these data to infer cell type-specific regulatory networks is a major challenge. We present Dynamic Regulatory Module Networks (DRMNs), a novel approach to infer cell type-specific *cis*-regulatory networks and their dynamics. DRMN integrates expression, chromatin state and accessibility to predict *cis*-regulators of context-specific expression, where context can be cell type, developmental stage or time point, and uses multi-task learning to capture network dynamics across linearly and hierarchically related contexts. We applied DRMNs to study regulatory network dynamics in three developmental processes, each exhibiting different temporal relationships and measuring a different combination of regulatory genomic datasets: cellular reprogramming, liver dedifferentiation and forward differentiation. DRMN identified known and novel regulators driving cell type-specific expression patterns demonstrating its broad applicability to examine dynamics of gene regulatory networks from linearly and hierarchically related multi-omic datasets.

Introduction

Transcriptional regulatory networks connect regulators such as transcription factors to target genes, and specify the context-specific patterns of gene expression, where context can be different cell types, cell states, developmental stages or time points. In developmental systems, changes in regulatory networks can significantly alter the type or state of a cell, which can affect both normal and disease processes. The regulatory interaction between a transcription factor (TF) and a target gene's promoter is dependent upon TF binding activity, histone modifications and open chromatin, that have all been associated with cell type-specific expression (Gonzalez et al. 2015; Lee and Young 2013; Osmanbeyoglu et al. 2019; Young 2011). To probe the dynamic and cell type-specific nature of mammalian regulatory networks, several research groups have generated matched transcriptomic and epigenomic data from short time courses or for cell types related by a branching lineage (Chronis et al. 2017; Lara-Astiaso et al. 2014; Wamstad et al. 2012). However, integrating these datasets to infer cell type-specific regulatory networks is an open challenge.

Existing computational methods to infer cell type-specific networks while integrating different types of measurements can be grouped into two main categories: (i) Regression-based methods, (ii) Statistical models for networks. Regression-based methods use linear and non-linear regression to predict mRNA levels as a function of chromatin marks (Dong et al. 2012; Rego et al. 2012) and/or transcription factor occupancies (Rego et al. 2012) and can infer a predictive model of mRNA for a single condition (time point or cell type). These regression approaches are applied to each context, such as a cell type, individually and have not been extended to model multiple related time points or cell types, which is important to study networks transition between different time points and cell states. Statistical models of networks constitute a large family of methods including probabilistic graphical models (Ernst et al. 2007; Gong et al. 2015; Jojic et al. 2013; Koch et al. 2017; Parikh et al. 2011; Pierson et al. 2015; Roy et al. 2011), correlation-based methods (Langfelder and Horvath 2008) and Boolean networks (Schwab et al. 2020), used to represent molecular networks. The vast majority of these methods (Gong et al. 2015; Jojic et al. 2013; Koch et al. 2017; Parikh et al. 2011; Pierson et al. 2015; Roy et al. 2011), are based on mRNA levels and require a sufficiently large number of mRNA samples for each time point or cell type to reliably estimate the statistical dependency structure. A few methods based on dynamic Bayesian networks (DBNs), including input-output Hidden Markov Models (Ernst et al. 2007) and time-varying DBNs (Gong et al. 2015) have been developed

to examine gene expression dynamics with static ChIP-seq datasets. However, both of these approaches are suited for time courses only, and do not accommodate branching structure of cell lineages.

To systematically integrate parallel transcriptomic and epigenomic datasets to predict cell type-specific regulatory networks, we have developed a novel dynamic network reconstruction method, Dynamic Regulatory Module Networks (DRMNs). DRMNs model regulatory networks in a cell type-specific manner by leveraging their relatedness, for example, by time or a lineage. DRMNs represent the cell type-specific regulatory network by a concise set of gene expression modules, defined by groups of genes with similar expression levels, and their associated regulatory programs (regulators predicted for that module and parameters capturing the strength and type of regulation). We applied DRMNs to four datasets measuring transcriptomic and epigenomic profiles from different cellular reprogramming and differentiation studies (Chronis et al. 2017; Roy and Sridharan 2017; Seirup et al. 2022; Xie et al. 2013) to link upstream regulatory programs to gene expression states associated with changes in cell state.

Results

Dynamic Regulatory Module Network (DRMN) for integrating diverse regulatory genomic measurements to infer regulatory networks on cell lineages

We developed DRMN to represent and learn context-specific regulatory networks, where contexts can be cell types, time points or cell states. We describe DRMNs with cell types as contexts, however our description applies to other types of contexts as well. We assume that each context has a small number of samples (e.g, one or two) but several types of measurements, such as RNA-seq, ChIP-seq and ATAC-seq. DRMN is based on a module-based representation of a regulatory network, where genes are grouped into modules and regulators are learned for each gene module. Each module, in turn, represents a discrete state of expression of the genes in the module, e.g. high or low expression (**Figure 1**). The module-based representation of the regulatory network of a cell type is implemented with a mixture model, each mixture component defining a gene module and a set of expression predictive features. The predictive features define the "regulatory program" for that expression state specifying the probability of a gene to belong to that expression state based on regulatory signals on a gene's promoter, such as sequence motifs, TF binding and epigenomic

signals. The module-based representation (Kundaje et al. 2007; Lee et al. 2009; Segal et al. 2003), enables DRMN to pool information from multiple genes to learn a predictive regulatory program and is appropriate when the number of samples per condition are too few to perform conventional gene regulatory network inference of estimating the regulators of individual genes. DRMN uses a multi-task learning framework of mixture models, one mixture model per task, which in turn corresponds to one cell type. The multi-task learning framework of DRMN leverages the relationship between cell types, e.g, on a lineage, to encourage similarity between the models learned for related contexts. DRMN produces two outputs: (1) a set of modules and regulators (e.g., TFs and histone marks) for each cell type, where the regulators are predictive of the expression levels of genes in a module, (2) "transitioning gene sets" comprising genes that change their module assignment across cell types and their regulators that predict their expression levels across the cell types (**Figure 1**). Using either output, we can infer a coarse regulatory network by adding edges between a regulator and genes in a module or transitioning gene set to which the regulator is connected (**Methods**).

DRMN offers a flexible framework to integrate diverse regulatory genomic measurements as predictive features, such as sequence-based motif strength, accessibility, and histone marks to examine their relative contribution for explaining variation in expression. In its simplest form, DRMN uses sequence motifs and expression only. Addition of accessibility and histone marks can help capture more variation in the expression levels (**Supplemental Methods**, **Supplemental Figure S1**). We found accessibility to be most informative when it was used together with sequence motifs (**Supplemental Figure S2**).

Multi-task learning is beneficial for learning cell type-specific expression patterns

We assessed the utility of DRMN's multi-task learning framework by comparing against a baseline model, Regulatory Module Network (RMN), which learned a mixture model for each context independently. We implemented two ways of multi-task learning: regularized regression (DRMN-FUSED) and graph structure prior (DRMN-ST) (See **Methods**). DRMN-ST is based on greedy structure learning of networks, while DRMN-FUSED is non-greedy, and uses regularized regression. We compared DRMN-ST to RMN-ST and DRMN-FUSED to RMN-FUSED. We used an array (Roy and Sridharan 2017) and a sequencing dataset (Chronis et al. 2017), both studying mouse cellular reprogramming, each measuring multiple histone marks, in addition to expression and accessibility (sequencing dataset only).

We learned DRMN and RMN models using different regulatory feature sets (**Figure 2**, **Supplemental Figure S1**): motif alone (Motif), histone marks alone (Histone) and combining motif and histone marks (Histone+Motif). We assessed their ability to model expression using the average Pearson's correlation of true and the generated expression of test genes based on their module, using three-fold cross-validation (**Figure 2A**, **Methods**). Note that the generated expression from the model requires the actual data to determine the mixture component and therefore should be interpreted as a model fit of the observed expression (See **Methods**). When comparing DRMNs to RMNs on array data, both versions of DRMN outperform corresponding RMN versions on Histone and Histone+motif features, (**Figure 2B,C, E, F**, blue for RMN and red for DRMN). On motifs, the difference between the models was dependent upon the cell line and the number of modules, k (**Figure 2D, G**). In particular, both DRMN-Fused and DRMN-ST outperformed RMN on the iPSC/ESC state, and was similar for pre-iPSC and MEF when comparing across different k. On sequencing data (**Figure 2H-M**), both DRMN-Fused and DRMN-ST were better than RMNs for most k in ESC and pre-IPSCs when considering Histone and Histone+Motifs. When using motifs, the performance depended upon the DRMN implementation and k (**Figure 2J, M**). In particular, DRMN-Fused was better than RMN for ESC, on par for pre-iPSC and MEF-48, and worse for MEF. DRMN-ST was on par or better than RMN-ST for $k = 3, 5$, however we observed decrease in performance in the MEF and MEF48 cell lines for higher k ($k = 7, 9, 11$). Across all cell states, DRMN-ST and DRMN-FUSED are better or comparable than the RMN versions, with greater benefits for the array data (**Supplemental Table S1**), suggesting that multi-task learning helps to improve the predictive power of these models. Between DRMN-FUSED and DRMN-ST, DRMN-FUSED was able to generally outperform DRMN-ST on different feature types and datasets (**Supplemental Figure S3**) likely because it learns a sparser model (**Supplemental Figure S4**). Hence, we report results of DRMN-FUSED application on different datasets.

We next compared DRMNs and RMNs to three baseline expression-based clustering methods: (i) GMM-Merged which applies a single Gaussian mixture model (GMM) to a merged matrix of expression values, (ii) GMM-Indep, which applies Gaussian mixture modeling per timepoint, (iii) ESCAROLE (Chasman et al. 2019), a non-stationary probabilistic model for expression clustering suitable for time-series, which was shown to better capture expression dynamics compared to standard clustering methods (Chasman et al. 2019). We used two metrics for comparison: (i) overall correlation computed using the predicted and

true expression of all genes, (ii) per module average correlation computed from the correlation of true and predicted expression per module. When using overall correlation, DRMNs, RMNs, GMM-Indep and ESCAROLE vastly outperform GMM-Merged (**Supplemental Figure S5A-F**). This suggests that the gene partitions are likely different between the different cell types and imposing a single structure for all genes as done in GMM-Merged, misses out on the cell type specific aspects of the data. DRMN models performed on par with RMN and GMM-Indep for most cases when using overall correlation (**Supplemental Figure S5A-F**). Based on per-module correlation, DRMN and RMN models outperform the expression-based clustering methods, which is expected as these methods do not model variation in expression within each module. These results demonstrate the advantage of modeling expression as a function of upstream features in DRMN/RMN over an expression-alone approach as the learned models capture a more fine-tuned model of expression variation as a function of *cis*-regulatory features (**Supplemental Figure S5G-L**).

DRMN accurately ranks regulators and regulatory network components across diverse developmental processes

We next compared DRMN's ability to predict regulators and coarse regulatory connections against multiple methods. For regulator prediction, we compared DRMN to DREM (Schulz et al. 2012) and regulators predicted based on motif enrichment of DRMN modules, and ESCAROLE modules (**Figure 3A**). DREM (Schulz et al. 2012) is an Input/Output HMM model which also models temporal expression profile as a function of regulatory signals such as histone marks and TF binding site. We considered three dynamic process for regulator prediction: cellular reprogramming (Chronis), hepatocyte dedifferentiation (Seirup) and early lineage specification from the embryonic stem cell state (Xie) (**Figure 3**, **Methods**). As gold standards, we used a literature curated set of regulators of embryonic stem cell state (ESC) for the Chronis and Xie datasets, and a literature curated set of hepatocyte regulators for the Seirup dataset (**Methods**). Note for DREM we did not have the cell stage-specific regulators from the command line interface, though this functionality may be available in the graphical user interface. For each method we computed the F-score comparing the predicted regulators to the gold standards. DRMN had the highest F-score (**Figure 3A**) compared to other methods. We also used Area Under the Precision Recall (AUPR) curve on the ranking of the regulators from each method. DRMN is better than enrichment across all datasets. DRMN and DREM are

comparable with DRMN outperforming DREM on the Chronis dataset and DREM outperforming DRMN on the Seirup dataset (**Supplemental Figure S6A**).

We next compared the regulatory relationships between transcription factors (TFs) and target genes from DRMN and several base line methods that inferred regulator-target relationships based on the enriched regulators in each module (Exp+Enr, DRMN+Enr, **Methods**). DREM's output was not amenable for this comparison. For DRMN, regulator-target relationships were weighted by the product of the regression weight of a regulator and the feature value (**Methods**). For Chronis and Xie, we used the ESC regulatory network downloaded from the ESCAPE database (Xu et al. 2013), which has both perturbation (regulator knock outs or knock downs) and ChIP-based regulator-target relationships. For Seirup, we additionally used ChIP-seq datasets in HepG2 cell line from the ENCODE Project Consortium to derive regulatory interactions more relevant to hepatocytes (**Figure 3B-D**, **Supplemental Figure S6B-D**). Based on F-score, computed on the top 50k edges, enrichment based networks were generally worse for the perturbation and perturbation+ChIP gold standards for Chronis (**Figure 3B**). The enrichment-based networks were better on the ChIP-based gold standards. A similar trend held for the Xie dataset although the difference on the perturbation networks was greater (**Figure 3C**). We observed similar trends on AUPR as well (**Supplemental Figure S6B, C**). For the Seirup dataset, DRMN again performed better than other methods when examining F-score on perturbation alone or perturbation+ChIP networks. Additionally DRMN networks exhibited greater variation across time which is indicative of context-specificity. On HepG2 the results were similar to the ESC ChIP-based network (**Figure 3D**). When using AUPR, we observed similar trends however the difference across methods was smaller (**Supplemental Figure S6D**). These results suggest DRMN generates meaningful TF-target relationships that more accurately recapitulate perturbation-based regulatory relationships and often capture greater context-specificity than simpler enrichment-based analyses.

DRMN predicts key regulatory network components indicative of inefficiency in cellular reprogramming

Cellular reprogramming is the process of converting a somatic cell type into a pluripotent cell type (Sridharan and Plath 2008). However, this process is inefficient as a small proportion of cells reprogram into the final pluripotency state. One hypothesis for low efficiency is the incomplete suppression of the somatic reg-

ulatory program (Chronis et al. 2017). To gain mechanistic insight into the regulatory program dynamics, a number of histone marks, select transcription factor (TF) binding, accessibility and RNA-seq measurements were profiled at four stages of reprogramming from mouse embryonic fibroblasts (MEFs) to induced pluripotent stem cells (iPSCs). We applied DRMN to examine how TFs interact with chromatin state to drive different expression states during this process.

We first described the global transcriptome state using seven DRMN modules (1-7), each corresponding to a distinct level of expression in each of the four reprogramming stages (**Figure 4A**). The modules were on average 30-90% similar in gene content, exhibiting the lowest similarity between MEF48 and pre-iPSC and the highest between MEF and MEF48 (**Supplemental Figure S7**). This agrees with the pre-iPSC state exhibiting a major change in transcriptional status during reprogramming. The repressed modules 1, 2 and 3 were less conserved across all cell types compared to the more highly expressed modules. Modules with high expression also exhibited more conserved enrichment of Gene Ontology (GO) processes compared to the repressed modules (**Supplemental Table S2**) and included house-keeping functions such as ribosome biogenesis and general metabolic processes. Processes that were specifically up-regulated in the iPSC and pre-iPSC stages included cell-cycle, while muscle development and cell adhesion processes were associated with MEF and MEF48. The enrichment of up-regulated cell cycle processes in the ESC/pre-iPSC and muscle related processes in MEF/MEF48 supports the biological relevance of our expression states.

We examined the regulators for the three modules associated with highest expression (5, 6 and 7, **Figure 4B**). Histone marks (H3K79me2 and H3K4me1) and accessibility were selected as predictive features for all four cellular stages. In contrast, the transcription factors (TFs) were selected in a more stage-specific manner with a few exceptions (e.g., INSM1 and BHLHE40) suggesting that TFs are important for the specificity of transcriptional programs. Several TFs that were predicted significant in a specific stage have known roles in that stage, e.g., NFE2L2 in ESC and pre-IPSC module 6, which is known to play an important role in the embryonic stem cell state (Dai et al. 2020), and MYCN and ESRRB in the pre-IPSC module 6, which are both known to be important for early embryonic stages. ESRRB has also been shown to be important for suppressing the MEF program during reprogramming (Chronis et al. 2017). SOX17 association with ESC was surprising because SOX17 is lowly expressed in ESC, however, SOX17 shares motif affinity with SOX2 which could explain this association (**Supplemental Figure S8**). We also found several muscle and

mesodermal factors associated with MEF (FOXL1 (Miyashita et al. 2020), MTF1 (Tavera-Montañez et al. 2019) in module 6), MEF48 (OSR1 (Vallecillo-García et al. 2017), LEFTY1 (Phan et al. 2020) in module 5), and both MEF and MEF48 (SOX5 (Ikeda et al. 2002) in module 5). The regulators were expressed in the cell states they were selected for prediction with the exception of FOXI1 (**Supplemental Figure S9A**). For FOXI1, we found other members of the Forkhead box family, to be expressed (**Supplemental Figure S9B)**.

Finally, we leveraged the DRMN modules to identify transitioning gene sets and their regulators, which can inform us about specific pluripotency and somatic expression programs that change during reprogramming (**Methods**). In total, we identified 85 gene sets with their predicted regulators (see **Methods**, **Supplemental Table S3**, **Supplemental File S1**). We focused on 42 gene sets with transitions into the high expression modules, 5, 6 or 7 as these indicate the strongest upregulated expression states during reprogramming (**Figure 4C**). These gene sets exhibited two main types of dynamics (**Figure 4C**, **Supplemental Figure S10**): (i) repressed in MEF/MEF48 and induced in ESC/iPSC (C216, **Figure 4D**), and (ii) induced in MEF/MEF48 and repressed in ESC/iPSC (e.g. C101, C220, **Supplemental Figure S10A,B**). The majority of these gene sets were predicted to be regulated by a combination of TFs and histone marks. Notably, nine gene sets (C186, C191, C200, C210, C182, C209, C203, C214, C212, (**Figure 4C**, red arrows) showed repressed expression in all but ESC (e.g., C191, C200, C203 **Supplemental Figure S10C, D, G**). In contrast, there were only two gene sets, C225 and C186 (**Supplemental Figure S10E, F**), that showed an opposite expression signature of repressed in ESC and induced in the other cell stages. Although these gene sets exhibited a similar trend of expression they were associated with different sets of regulators, which include a combination of elongation, promoter and repressive marks and distinct TFs (e.g., PLAG1 in C191, ZFP281 in C200, BCL6B in C203). Several of the regulators have known regulatory roles in pluripotency in development, e.g., ZFP281 (Fidalgo et al. 2011), PLAG1, which is involved in cancer and growth processes and was shown to effect embryonic development (Madissoon et al. 2019) and BCL6B which reduces reprogramming efficiency (Zhang et al. 2018). These gene sets can also provide clues into factors governing reprogramming efficiency as they retain a somatic MEF-like state in the pre-iPSC stage.

Using DRMN to gain insight into regulatory program dynamics across a long time course

The reprogramming study demonstrated DRMN application to a short time course (3-4 time points). We next tested the ability of DRMN to analyze temporal dynamics of a longer 16 time-point dataset measuring transcriptomic and accessibility profiles during hepatocyte dedifferentiation (Seirup et al. 2022). This dataset also showed DRMN application when only accessibility and gene expression are measured. Dedifferentiation of primary cells such as liver hepatocytes is a major challenge in studying normal liver function as well as for liver-related diseases (Elaut et al. 2006). Dedifferentiation could be due to the changes in the regulatory program over time, however, little is known about the transcriptional and epigenetic changes during this process. We applied DRMN to this dataset to characterize the transcriptional dynamics and predict regulators associated with major changes in expression.

Using DRMN we partitioned the genes into five levels of expression (1-5) at all time point (**Figure 5A**). Comparison of module assignments across time showed that module 1, associated with lowest expression was most conserved across time points, while modules 2, 3 and 4, exhibited transitions happening between 4 and 6 hrs and 14 and 16 hrs (**Supplemental Figure S11**). The relatively high conservation of the repressed module 1 was in contrast to DRMN modules from the reprogramming study, where it was least conserved. The repressed state was enriched for developmental processes (**Supplemental Table S2**) while the other modules were enriched for diverse metabolic processes. Modules 4 and 5, which are associated with induced expression are enriched for more liver-specific metabolic function such as co-enzyme metabolism, acetyl CoA metabolism and modules 2 and 3 were enriched for general housekeeping function such as nucleic acid metabolism. We next examined the regulators associated with each module predicted by DRMN (**Figure 5B**), which we previously showed to be more relevant to liver state than simpler enrichment based methods (**Figure 3A**). Focusing on the highly expressed modules (3, 4), we found several liver factors, e.g., NFKB (Luedde and Schwabe 2011), FOXK1, FOXK2 (Le Lay and Kaestner 2010; Sakaguchi et al. 2019), ONECUT2 (Clotman et al. 2005; Laudadio et al. 2012) and LHX9 (Rétaux et al. 1999) that are likely important for maintenance of the hepatocyte state (**Figure 5B**). In addition, we found several regulators involved in cell fate decision making e.g., TCF3 (Cole et al. 2008) and SMAD2 (Uemura et al. 2005). The cell fate regulators are enriched in the later part of the time course indicating their potential roles in dedifferentiation. As before we verified that the regulators were expressed in the time point they were selected; in several case

one or more members of the same family for the motif was expressed (**Supplemental Figure S12**).

To gain insight into fine-grained transition dynamics and the regulators driving these dynamics, we again used the DRMN module assignments to identify genes that transition from one module to another as a function of time. We identified 84 gene sets with predicted regulators at $\geq$60% confidence (**Methods**, **Supplemental File S1**). We focused on those gene sets with a transition into modules 3, 4 and 5 and identified a total of 25 gene sets with varying types of transitions (**Figure 5C**). Many of the transitions were between modules that are adjacent to each other based on expression levels, suggesting that the majority of the dynamic transitions are subtle (e.g., module 4 and 5). Several of these gene sets exhibited a gradual upregulation of expression, e.g., C444 (**Figure 5D**), C380, C539, C403, C460 (**Supplemental Figure S11B-E**), including transitions at 4 and 6 hrs which coincided with the largest changes in module assignment. Regulators associated with these gene sets included liver-specific (JUN (C444), STAT5A and STAT5B (C380)) or developmental regulators (IRX1 (Yu et al. 2017) (C380, C444), SIX6 (Diacou et al. 2018) (C444) and KLF (C380)). We also identified gene sets with down-regulation of expression, e.g., C437 (**Figure 5E**) and C383 (**Supplemental Figure S11F**). Key regulators included Hepatocyte nuclear factors, HBF4G, HNF4A, CEBPA (C437), the EGR family as well as several components of Wnt NF-kappa-B signaling pathways (C383). Early growth response (EGR) factors have been shown to play important roles in different liver-related functions including repair and injury (Magee and Zhang 2017) that have been implicated in liver-specific function (Rudraiah et al. 2016). To summarize, liver-related regulators were associated with both up and down-regulated genes, however, up-regulated genes were additionally associated with developmental regulators. These predictions implicate relevant regulators in liver cell fate maintenance and offer avenues for functional validation studies to understand dedifferentiation.

DRMN application to find regulators of lineage-specific expression on hierarchical lineages

To demonstrate the utility of DRMN on hierarchically related cell lineages, we considered a dataset profiling early differentiation of human embryonic stem cells (H1) into four lineages: mesendoderm, mesenchymal, neural progenitors, trophoblast (Xie et al. 2013). In addition to accessibility, this dataset measured eight different histone marks, H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K9ac, H3K79me2, H3K36me3 and H3K27me3. We applied DRMN to this dataset and identified genes at five major levels of expression, 1-5,

with 1 representing the lowest expression and 5 the highest (**Figure 6A**). The extent of gene conservation depended on the expression level, with states 1 and 2 exhibiting low conservation, while modules 3 and 4 exhibiting high conservation (**Figure 6B**). The low conservation of states 1 and 2 is consistent with our observations in the reprogramming study. Gene Ontology (GO) process enrichment showed that the repressed module is enriched for developmental and lineage specific functions, while the induced modules tended to be enriched for cell cycle and translation related processes (**Supplemental Table S2**). The most repressed genes exhibited the largest extent of cell type specific enrichments, while the induced state was enriched for similar processes.

We next examined the regulators selected by DRMN, focusing on the two highly expressed modules, 3 and 4 (5 was associated only with histone marks). We verified that the regulators were expressed in the corresponding lineage they were selected; in several case one or more members of the same family was expressed (**Supplemental Figure S13**). Similar to the reprogramming study, histone mark associations were more conserved across cell types than TFs, with the elongation mark, H3K36me3 and repressive mark, H3K27me3 being among the most conserved marks across cell types. Among the TFs associated with each module, several have known lineage-specific roles (**Figure 6C**). For example, we found several neuronal lineage regulators including FOX (Ferri et al. 2007) and MYB (Malaterre et al. 2008) proteins in module 3 of the Neural Progenitor cell type, and VSX1 (Francius et al. 2016) and SHOX2 (Scott et al. 2011) in Mesendoderm module 4. Similarly, TEAD4, a regulator important for trophoblast self renewal (Saha et al. 2020) was associated with Trophoblast module 4.

We used the DRMN results to examine fine-grained regulatory dynamics of early lineage specification. We focused on 72 gene sets transitioning into the high expression modules, 3, 4 and 5 for which we could predict regulators with high confidence (**Methods**, **Figure 6D, Supplemental Figure S14**). These gene sets exhibited up regulation in a single lineage (C74, C131 **Supplemental Figure S15A,B**) as well as in multiple lineages (C112 and C138, **Supplemental Figure S15C**, **D**). Xie et al independently identified lineage-specific genes based on mRNA levels. Majority of our gene sets with lineage-specific expression pattern have a significant overlap with the Xie et al gene sets. For example, gene sets C74, C218 (**Figure S15A, E**) and C235 (**Figure S15F**) included genes that were repressed in all but the Mesenchymal state and were enriched for Xie et al lineage specific genes for the Mesenchymal state (**Supplemental Figure S14**). We

see similar results for gene sets with a trophoblast (C226, **Figure S16A**) and neural progenitor upregulation (C131, C116 **Figure S15B**, **S16B**). The enrichment in the lineage-specific gene sets from Xie et al was a useful corroboration of our study, in addition we predicted TFs for several of these gene sets. In particular gene sets associated with neural progenitor cell state, were predicted to be regulated by relevant regulators such as INSM1 (C116), ZIC5 (C131), NFIX (Heng et al. 2014), HOXA2, HOXB2 (Davenne et al. 1999) and TCF4 (Wittmann and Häberle 2018) (all associated with C252) (**Figure 6D**). Similarly for the trophoblast associated gene sets, we predicted TEAD1 (C118), GMEB1 (C217), SNAI2 (C119) and NR3C2 (C126), as among the regulators (**Figure 6D**). TEAD1 which is a member of the Hippo signaling pathway is expressed in placental cells (Soncin and Parast 2020). GMEB1 and NR3C2 are involved in glucocorticoid signaling, which is implicated in the development of the placenta (Kisanga et al. 2018). SNAI2 is a regulator of the epithelial-mesenchymal transition (EMT), which is a key process in trophoblast development. Finally, we examined regulators of gene sets exhibiting up-regulation in multiple lineages (e.g., C120, C140, **Figure S16C,D**), and found these regulators are involved in multiple developmental lineages (e.g., GLI1) (Hui and Angers 2011). Taken together, DRMN predicted key transcription factors associated with lineage-specific patterns of expression, expanding on the Xie et al study which focused primarily on histone modifications. Several of these regulators are known to be involved in these specific lineages, while others are novel predictions participating in relevant pathways that can be followed with future functional studies.

Discussion

Cell type-specific gene expression patterns are established by a complex interplay of multiple regulatory levels including transcription factor binding, genome accessibility and histone modifications. To gain insight into regulatory network dynamics associated with cell type-specific expression patterns, time course and lineage-aware datasets measuring transcriptomes and epigenomes of a dynamic process are becoming increasingly available (Bunina et al. 2020; Chronis et al. 2017; Lara-Astiaso et al. 2014; Wamstad et al. 2012; Xie et al. 2013). Analyzing these datasets to identify the underlying gene regulatory network dynamics that drive context-specific expression changes is a major challenge. This is because of the large number of variables measured in each context (time point or cell type), but low sample size for each context. In this work, we developed DRMN, that simplifies genome scale regulatory networks from individual genes to gene modules and infers regulatory program for each module in all the input conditions. DRMN offers a flexible framework to integrate a variety of regulatory genomic signals. In its simplest form, DRMN can be applied to expression datasets with sequence-specific motifs to learn a predictive regulatory model. In its more general form, DRMN can integrate a variety of regulatory signals such as genome-wide chromatin accessibility, histone modification and transcription factor profiles measured using sequencing and array technologies. Furthermore DRMN is applicable to datasets of different experimental designs such as short time series (e.g., the reprogramming study), long time series (e.g., the hepatocyte dedifferentiation study) and hierarchically related cell types on lineages (e.g. in the cellular differentiation study).

Central to DRMN's modeling framework is joint learning of predictive regulatory programs for each cell type or time point by using multi-task learning. Using two different approaches to multi-task learning, we show that joint learning of regulatory programs is advantageous compared to an approach of learning regulatory programs independently per condition. While a few models handle multiple regulatory signals over time they are based entirely on clustering of multiple signals in a give loci (Ernst et al. 2011; Roy and Sridharan 2017). Predictive modeling of expression that also clusters genes into expression groups is more powerful than clustering expression alone, including approaches that explicitly model temporal and hierarchical relationships across datasets. Such models are able to capture fine-grained expression variation as a function of the upstream regulatory state of a gene.

We applied DRMNs to distinct types of dynamic processes which involved cell fate transitions: mouse

reprogramming from a differentiated fibroblast cell state to a pluripotent state (3-4 cell states), hepatocyte dedifferentiation (16 time points), and forward differentiation of human embryonic stem cells to different lineages (5 cell types). When comparing across processes, the repressed modules were least conserved in the reprogramming and differentiation study, while it was most conserved in the dedifferentiation study. Accordingly, biological processes were repressed in a cell state-specific manner in the reprogramming and forward differentiation study, while we saw a common down-regulation of developmental processes in the hepatocyte dedifferentiation dataset. Furthermore, there were greater changes in expression in the reprogramming time course compared to dedifferentiation indicating the difference in magnitude of change in the two processes. Using DRMN transitioning gene sets, we asked specific questions tailored to the process under study. In the reprogramming study we investigated transitioning genes that could inform us about the incomplete repression of the somatic program (Chronis et al. 2017; Sridharan and Plath 2008), which is considered a major barrier in reprogramming efficiency and identified previously implicated as well as novel TFs that can be followed up with perturbation studies. In the dedifferentiation study we predicted a combination of developmental and liver-specific regulators associated with genes induced or repressed as a function of time. Finally, in the forward differentiation study we identified lineage-specific genes many of which had significant overlap with the original study (Xie et al. 2013), and predicted several TFs in addition to histone modifications as important regulators for the neural and trophoblast lineages. By identifying TFs associated with different trends of expression we expanded on these previous studies which largely focused on the epigenome dynamics. These predictions offer testable hypothesis of regulators driving expression dynamics in cell fate transitions.

The DRMN model can be extended in several directions. One direction of future work would be to incorporate ATAC-seq signal from more distal regions by using genome-wide chromosome conformation capture assays (Fraser et al. 2015; Gorkin et al. 2014; Wit and Laat 2012) as well as computational long-range predictions (Whalen et al. 2016; Zhang et al. 2019). Another direction would be to use more generic sequence features, such as *k*-mers (Setty and Leslie 2015) or deep learning models (Avsec et al. 2021) to enable the discovery of novel regulatory elements and offer great flexibility in capturing sequence specificity and its role in predictive models of expression. DRMN's predictive model often selected elongation marks such H3K36me3 as good predictors of expression. However, these could represent downstream transcriptional

processes rather than upstream regulatory processes. Explicit modeling of the dependency structure among the regulatory features could assist with discriminating regulatory signals that are upstream versus downstream of expression. DRMN relies solely on the sequence specificity of a TF when linking to a module, which might not be sufficient to distinguish between paralogous TFs. Incorporation of the TF's expression level might be beneficial to distinguish between multiple TFs with similar sequence specificities.

In conclusion, DRMN offers a powerful and flexible framework to define context-specific gene regulatory programs from time series and hierarchically related regulatory genomic datasets. As additional datasets that profile epigenomic and transcriptomic dynamics of specific processes become available, methods like DRMN will become increasingly useful to examine regulatory network dynamics underlying context-specific expression.

Materials and methods

DRMN model description

We define a Regulatory Module Network (RMN) as the set of modules and their regulators of any one condition (e.g., cell types, time points). DRMN is a collection of RMNs, one for each condition related by time or a lineage tree. In our description below, we assume we have a set of related cell types, however the same description applies to multiple time points or related conditions.

RMN. An RMN $\mathbf{R}_c$ for cell type c is denoted by the tuple $< G_c, \Theta_c >$. G_c is a bipartite graph specifying the edges between F features and K modules, and Θ_c, are the parameters of the regulator programs for each module that relate the selected regulatory features to the expression of genes in the module. In RMN each cell type's data is modeled independently and we optimize the posterior probability of each cell type c's model defined as: $P(\mathbf{R}_c|\mathbf{X}_c, \mathbf{Y}_c)$, which is proportional to $P(\mathbf{X}_c|\mathbf{Y}_c, \mathbf{R}_c)P(\mathbf{R}_c)$. Here $\mathbf{X}_c$ is the expression data for cell type c, $\mathbf{Y}_c$ is an $N \times F$ matrix, with the g^{th} row, $\mathbf{Y}_c(g,:)$ specifying the values of F features for gene g. RMN models each gene g's expression in c, X_{cg}, as a mixture model as follows:

$$P(X_{cg}|M_{gc}, \mathbf{R}_c, \mathbf{Y}_c^{(i)}(g,:)) \sim \mathcal{N}\left(\sum_{f\in\{0\cdots F_{ci}\}} \theta_{ci}(f)\mathbf{Y}_c^{(i)}(g,f), \Sigma_{x|\mathbf{Y}_{cg}^{(i)}}\right) \quad (1)$$

Here, $M_{cg} = i$ denotes the module/mixture component, $f = 0$ corresponds to the bias term, $\mathbf{Y}_c^{(i)}$ are the features associated with component i, and F_{ci} denotes the number of columns of $\mathbf{Y}_c^{(i)}$. The data likelihood for each cell type, c is $\prod_g \sum_{M_{cg}} P(X_{cg}|M_{cg}, \mathbf{R}_c, \mathbf{Y}_c(g,:))P(M_{cg})|\mathbf{R}_c)$. See **Supplemental Methods** for details.

DRMN. The DRMN model is defined by a set of RMNs, $\mathbf{R} = \{\mathbf{R}_1, \cdots, \mathbf{R}_C\}$ linked via the lineage tree τ over C cell types. The posterior probability of the model given the data is:
$P(\mathbf{R}_1, \cdots, \mathbf{R}_C|\mathbf{X}_1, \cdots, \mathbf{X}_C, \mathbf{Y_1}, \cdots, \mathbf{Y}_C, \tau)$. Based on Bayes rule, this can be rewritten as
$P(\mathbf{X}_1, \cdots, \mathbf{X}_C|\mathbf{R}_1, \cdots, \mathbf{R}_C, \mathbf{Y}_1, \cdots, \mathbf{Y}_C)P(\mathbf{R}_1, \cdots, \mathbf{R}_C)$. In DRMN, we model the expression of a

gene g across all cell types together. The data likelihood is defined as:

$$\prod_g P(X_{1g}, \cdots, X_{Cg} | M_{1g}, \cdots, M_{Cg}, \mathbf{R}, \mathbf{Y}_1, \cdots, \mathbf{Y}_C) P(M_{1g}, \cdots, M_{Cg} | \mathbf{R}, \tau) P(\mathbf{R}|\tau) \quad (2)$$

Given the module assignments, M_{cg} the gene expression across cell types are independent of each other. Hence the probability of g's expression across all cell types is:

$$\left(\prod_c P(X_{cg} | M_{cg}, \mathbf{R}_c, \mathbf{Y}_c(g, :)) \right) P(M_{1g}, \cdots, M_{Cg} | \tau) P(\mathbf{R}|\tau)$$

Using the tree structure we assume that the module assignment in a cell type c, M_{cg} is independent of everything else, given its parent $\text{pa}(c) = c'$ in τ, which allows efficient computation of $P(M_{1g}, \cdots, M_{Cg}|\tau)$ using pairwise conditional distributions, $P(M_{cg}|M_{c'g})$. $P(X_{cg}|M_{cg}, \mathbf{R}_c, \mathbf{Y}_c(g, :))$ is computed in the same way as **Eqn 1**. $P(M_{cg}|M_{c'g})$ is obtained from the transition probabilities. $P(\mathbf{R}|\tau)$, specifies a prior over the structure/feature sets in each of the cell type-specific RMN models, $\mathbf{R}_c$. The prior controls what sets of features get associated with each module i, and how they change over time. We used two formulations for the prior to enable sharing information: DRMN-Structure Prior (DRMN-ST) defines a structure prior over the graph structures $P(G_1, ..., G_C)$ while DRMN-FUSED uses a regularized regression framework and implicitly defines priors on $P(\Theta_1, \cdots, \Theta_C)$. In both frameworks, we share information between the cell types/conditions to learn the regulatory programs of each cell type/condition. See **Supplemental Methods** for details.

DRMN learning

DRMNs are learned by optimizing the overall DRMN likelihood (**Eqn** 2), using an Expectation Maximization (EM) style algorithm that searches over the space of possible graphs for a local optimum (**Supplemental Methods**, **Algorithm 1**). In the Maximization (M) step, we estimate transition parameters (M1 step) and the regulatory program structure (M2 step). In the Expectation (E) step, we compute the expected probability of a gene's expression profile to be generated by one of the regulatory programs. The M2 step uses multi-task learning to jointly learn the regulatory programs for all cell types using either the framework of DRMN-ST or DRMN-FUSED. DRMN typically converges between the first 15-20 iterations (**Supplemental Figure S17**).

Identification of transitioning gene sets and their regulators

A transitioning gene is a gene whose module assignment changes in at least one time point/cell type. We grouped the genes into transitioning gene sets using a hierarchical clustering approach with city block as a distance metric and 0.05 as a distance threshold. We developed two strategies for selecting regulators for transitioning gene sets: Simple linear regression for short time courses (with 5 or fewer time points or conditions) and Multi-Task Group LASSO (MTG-LASSO) that is suitable for longer time courses. Both approaches take as input a list of genes from a transitioning set and predict the regulators that explain the variation in expression over time. The simple linear regression approach used the MERLIN algorithm (Roy et al. 2013) to find regulators that can explain the overall variation in the expression of genes in the set. The MTG-LASSO approach used a multi-task regression framework, which performs multiple regression tasks, one for each gene in the set to select regulatory features as predictors for each gene's expression levels. Once regulators were selected using either approach, we filtered the predicted regulators per gene set by assessing the correlation of the regulator/feature with the gene expression of a gene across the conditions and included a regulator if was correlated to at least 5 genes with a Pearson's correlation of 0.6 or higher. See **Supplemental Methods**, section "Predicting regulators for transitioning gene sets" for more details.

Evaluating the ability to model expression

We assessed the ability of DRMN to model expression based on the ability to capture variation in expression compared to previous expression clustering methods: simple expression clustering (GMM-Indep, GMM-merged), time point aware expression clustering (ESCAROLE (Chasman et al. 2019)), and single task learning methods (RMN-ST and RMN-FUSED). We compared the methods using 3-fold cross validation (CV). Briefly, we split the dataset along the gene dimension into three parts, each part comprising the expression data and features across conditions for genes in that part. We used each part as a test set and trained a model on the remaining two parts. We used the trained model to generate predictions on the test set as follows. Let h denote the index of a test gene in cell type c with measured expression value X_{hc} and let $\mathbf{Y}_c(h,:)$ be its feature values. As the DRMN, RMN, and GMM models are all based on a mixture model, we do not know which component of the mixture generated the expression. Therefore, we instead

first obtain the mixture component with the highest likelihood of generating the expression:

$$j = \arg\max_i P(X_{hc}|\mathbf{R}_c^{(i)}, \mathbf{Y}_c(h,:))$$

Next we predict the expression, $\widehat{X}_{hc}$ from the corresponding conditional mean of the selected component based on the feature values as:

$$\widehat{X}_{hc} = \theta_{ci}\mathbf{Y}_c(h,:),$$

where $\mathbf{Y}_c(h,:)$ is augmented with the first column of all 1s. To assess a model's ability to capture variation we computed prediction error between X_{hc} and $\widehat{X}_{hc}$. Note, as in mixture models used for clustering, the prediction models need the knowledge of the observed expression to determine the model component that should generate the data and should be interpreted in a way similar to test data likelihood. In the case of DRMN-FUSED, which requires hyperparameter tuning, we performed an internal 3-fold CV on the training set to select the best hyperparameter settings from a grid search on $\rho_1 \in \{30, 50, 70, 90, 110, 130, 150\}$, $\rho_2 \in \{0, 10, 30, 50, 70\}$ and $\rho_3 = \{0, 10, 30, 50, 70\}$). This way, we select the hyperparameter setting with the best performance on that training set and used the model with that hyperparameter setting to assess the predictive power of the model on the corresponding test set. As a result different hyperparameter settings can be selected for each test set in the outer loop of cross validation. For selecting the hyperparameters for running the algorithm on the full dataset, we used a simple 3 fold CV without internal CV (see Section **Application of DRMN to different datasets** for details of how parameters were selected).

Evaluation of predicted regulators and TF-target gene networks

We evaluated DRMN outputs of regulators and TF-target relationships against multiple baseline methods (**Figure 3**, **Supplemental Figure S6**). For regulator prediction comparison, we examined two enrichment-based approaches (regulator enrichment on ESCAROLE modules and DRMN modules), and DREM 2.0 (v2.0.3) (Schulz et al. 2012), which infers regulatory models of gene regulation from time-series data using an Input/Output HMM approach. To rank DRMN regulators, we computed change in prediction error after removing a regulator, and used this as a score for the importance of the regulators. For each module, we performed 5-fold cross validation to predict the expression of the module as a linear function of regulators

selected for that module. Next, we recalculated the prediction error after removing a regulator. The total change in prediction error of each regulator across all modules was used for ranking the regulators. For the enrichment baseline approaches, we tested modules from DRMN or ESCAROLE for enrichment of motif instances and used -log of FDR corrected p-values of a hyper-geometric test (summed over modules) as a score for ranking regulators. These approaches were applied to the Chronis et al reprogramming, Xie et al forward differentiation and Seirup et al hepatocyte dedifferentiation datasets.

To rank DREM regulators, we first applied DREM to the Chronis et al. reprogramming and Seirup et al. hepatocyte dedifferentiation datasets because of their temporal nature. For both data sets we gave DREM as input 1) normalized and log-transformed gene expression matrices, and 2) binarized Q-motif feature data in the required input format for DREM. To binarize the features, values above the genome-wide average were set to 1 and values below set to 0. DREM was applied to these data sets with recommended default settings from the command line. To obtain a ranking of features we used the absolute value of the regression weights of the features provided in the output regulatory state model of DREM; if a feature was selected more than once, we used the sum of the absolute value. The ranking was used for computing the Area Under the Precision-Recall (AUPR) curve. In addition, we also selected a subset of regulatory features with a coefficient of magnitude > 0.001 in at least one inferred regulatory state model and used this set for F-score (**Figure 3** and **Supplemental Figure S6**). Note, we did not have the information of which regulatory state model is associated with which cell stage-specific regulators in v2.0.3, though this functionality may be available in the graphical user interface in other versions of DREM.

We asked how well the predicted regulators recover known regulators associated with a specific cell type based on F-score and AUPR. For the reprogramming and the H1ESC differentiation dataset, we used a list of genes associated with the ESC state as a gold standard (**Supplemental Table S4**, **Supplemental File S2**). This list was created using genes assigned to GO terms "regulation of embryonic development" and "embryonic organ development" and gene lists from (Müller et al. 2008) and (Wong et al. 2008) listed in MSigDB (Subramanian et al. 2005). For the hepatocyte dedifferentiation dataset, we used a list of important regulators collected from literature (Kyrmizi et al. 2006; Odom et al. 2006; Sheaffer and Kaestner 2012; Velazquez et al. 2021; Wangensteen et al. 2015). Note, in the Sierup et al. and Xie et al. datasets Q-motif features were named with Cis-BP (Weirauch et al. 2014) motif IDs. Prior to comparison with the gold

standard regulators these IDs were converted to common names of the TFs. The Q-motif features for the Chronis et al. data were already mapped to common regulator names. We used AUCCalculator Java package to calculate AUPR (Davis and Goadrich 2006).

For regulator-target network comparison, we compared DRMN to two baseline approaches: enrichment of regulators in ESCAROLE modules, and enrichment of regulators in DRMN modules (**Supplemental Table S5**). For the enrichment methods, we created a network by mapping regulators associated with a module (by enrichment) to all genes in the module and scored edges by the feature values. For DRMN, we scored edges by the input Q-motif value multiplied by estimated regression coefficients. As gold standards, we used ChIP and regulator perturbation-based (knockdown and knockout of regulators) interactions from the ESCAPE database (Xu et al. 2013) that provides these interactions in human and mouse embryonic stem cells. We also added ChIP-based interactions identified in mESC and hESC from the ENCODE Project Consortium (The ENCODE Project Consortium 2012), and additional regulator-perturbation interactions identified in (Correa-Cerro et al. 2011; Nishiyama et al. 2013, 2009). For mouse hepatocyte dedifferentiation we additionally used ChIP-based interactions from the HepG2 cell line from the ENCODE Project Consortium. We compared the networks based on AUPR and also F-score on the top 50K edges from each approach. The gold standard regulators and networks are available as **Supplemental File S2**.

Application of DRMN to different datasets

We applied DRMN to four different datasets profiling three dynamic processes: (a) A microarray time course dataset of cellular reprogramming from mouse embryonic fibroblasts (MEFs) to induced pluripotent stem cells (iPSCs), (b) A sequencing time course dataset using the same system as (a), (c) A sequencing time course dataset profiling dedifferentation of hepatocyte cells, (d) differentiation of H1ESC cells to different lineages. Below we describe the dataset processing, feature generation, hyper-parameter selection and analysis of transitioning gene sets.

Reprogramming array data.

This dataset measures gene expression and eight chromatin marks in three cellular stages: MEFs, partially reprogrammed induced pluripotent stem cells (pre-iPSCs), and iPSCs, collected from multiple publications

(Maherali et al. 2007; Roy and Sridharan 2017; Sridharan et al. 2013, 2009). The expression values of 15,982 genes was measured by microarray. Eight chromatin marks were measured by ChIP-on-chip (chromatin immunoprecipitation followed by promoter microarray). For each gene promoter, each mark's value was averaged across a 8000-bp region associated with the promoter. The chromatin marks included those associated with active transcription (H3K4me3, H3K9ac, H3K14ac, and H3K18ac), repression (H3K9me2, H3K9me3, H3K27me3), and transcription elongation (H3K79me2). We considered the following set of features: Motif, Histone, Histone+Motif. We used the motif collection available with the PIQ software (Sherwood et al. 2014) from http://piq.csail.mit.edu/, which were sourced from multiple databases (Berger et al. 2006; Matys et al. 2003; Sandelin et al. 2004). From the full motif list, we used only those annotated with TF proteins (Ravasi et al. 2010) resulting in a total of 353 TFs. We used FIMO to find motif instances using the mm9 mouse genome (Grant et al. 2011). Motif instances for the same TF were further aggregated into a single feature per gene by selecting the most significant motif instance based on p-value. This resulted in a total of 353 dimensions for the Motif feature.

To determine suitable range of the number of modules we used visual inspection of modules, the Silhouette index and overall predictive power. We computed the Silhouette index on expression alone modules from ESCAROLE runs and visualized the expression heatmaps to assess the overall expression states and found $k = 5$ or $k = 7$ to be most appropriate (**Supplemental Figure S18**). To determine the hyper-parameter settings, we performed a grid search with three-fold cross validation on a range of values for ρ_1 (sparsity in each task), ρ_2 (selection of similar features for closely related cell types) and ρ_3 (selection of similar features for all cell types): $\rho_1 \in \{0.5, 1, 2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130\}$, $\rho_2 \in \{0, 10, 20, 30, 40, 50\}$ and $\rho_3 = \{0, 10, 20, 30, 40, 50\}$ (**Supplemental Figure S19**, **Supplemental Methods**). We focused on $k = 7$ as it gave visually distinct patterns of expression (**Supplemental Figure S20**) and had comparable predictive power as $k = 5$. The top 5 hyper-parameter configurations were: $\{(50, 50, 30), (50, 10, 30), (90, 10, 0), (40, 40, 50), (30, 50, 50)\}$. Inspection of the heat maps of DRMN modules showed that the results were similar across these configurations, hence we selected the first configuration $(50, 50, 30)$ for DRMN application on the full dataset. DRMN modules were interpreted using Gene Ontology enrichment using a Hyper-geometric test with FDR to control for multiple hypothesis correction and transitioning gene set analysis (**Supplemental Table S2**, **Supplemental Table S3**).

Reprogramming sequencing data.

This dataset generated by Chronis et al. (Chronis et al. 2017), assayed gene expression with RNA-seq, nine chromatin marks with ChIP-seq, and chromatin accessibility with ATAC-seq in different stages of reprogramming (MEF, MEF48 (48 hours after start of the reprogramming process), pre-iPSC, and embryonic stem cells (ESC)). The chromatin marks included H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, and H3K9me3. We aligned all sequencing reads to the mouse mm9 reference genome using Bowtie 2 (Langmead and Salzberg 2012). For RNA-seq data, we quantified expression to TPMs using RSEM (Li and Dewey 2011) and applied a log transform. After removing unexpressed genes (TPM$<$1), we had 17,358 genes. For the ChIP-seq and ATAC-seq datasets, we obtained per-base pair read coverage using BEDTools (Quinlan and Hall 2010), aggregated counts within $\pm 2,500$ bps of a gene's transcription start site and applied log transformation. Q-motif features were generated using the motif collection available with the PIQ software (Sherwood et al. 2014) (see **Supplemental Methods**).

To determine the number of modules we used Silhouette index of ESCAROLE clusters (**Supplemental Figure S18**) and found both $k = 5$ and 7 to be appropriate. We used $k = 7$ as this gave visually distinct patterns of expression (**Figure 4**). For hyper-parameter determination we used three-fold cross-validation mode with grid search (similar to the array dataset, **Supplemental Figure S21, S22, S23**) and determined the top 5 configurations of hyperparameters ρ_1, ρ_2, ρ_3, as $\{$(10,50,40), (20,50,40), (10,50,10), (40,50,0), (10,50,0)$\}$. As the DRMN modules were largely similar we present results for one of the configurations:$\rho_1 = 10, \rho_2 = 50, \rho_3 = 40$. Final results for this dataset are obtained by applying DRMN-FUSED to the Histone+Accessibility+Q-Motif feature set with $k = 7$ modules. We tested the modules for GO enrichment and defined transitioning gene sets based on genes that change their module assignments (**Supplemental Table S2**, **Supplemental Table S3**). Of the 17,358 genes, 11,152 genes changed their module assignments. We clustered them into sets of 5 or more genes and defined 111 gene sets consisting of 10,194 genes (the remaining were singleton or genes with similarity to less than 4 genes, **Supplemental Table S3**). We next used a simple regression model to predict regulators for each gene set (See **Section Identification of transitioning gene sets and their regulators**).

Hepatocyte dedifferentiation time course data.

The dedifferentiation time course consisted of samples extracted from adult mouse liver followed by RNA-seq and ATAC-seq analysis at 0 hrs, 0.5 hrs, 1 hrs, 2 hrs, 4 hrs, 6 hrs, 8 hrs, 10 hrs, 12 hrs, 14 hrs, 16 hrs, 18 hrs, 20 hrs, 22 hrs, 24 hours, and 36 hrs (16 time points in total, (Seirup et al. 2022)). Sequencing reads were aligned to mouse mm10 reference genome using Bowtie 2 (Langmead and Salzberg 2012), and gene expression was quantified using RSEM (Li and Dewey 2011). Any gene with TPM$= 0$ in all time points was removed, resulting in 14,794 genes with measurement in at least one time point. Per-base pair read coverage for ATAC-seq was obtained using BEDTools (Quinlan and Hall 2010), and counts were aggregated within $\pm 2,500$ bps of a gene's transcription start site. Both gene expression and accessibility data were quantile normalized across 16 time points and then log transformed.

We examined the Silhouette index of the ESCAROLE expression modules and found $k = 5$ to be appropriate (**Supplemental Figure S18**). Furthermore, as both $k = 5$ and $k = 7$ were reasonable for the reprogramming sequencing dataset, we used $k = 5$ as it was computationally faster. We applied DRMN-FUSED with Accessibility and Q-Motif as the feature set with $k = 5$ modules. Q-Motif features were generated similar to the Chronis et al dataset using PWMs from Cis-BP database (Weirauch et al. 2014). Motif instances were scored by the ATAC-seq signal resulting in a total of 2,856 features. The Q-Motif Features were quantile normalized and log transformed across the 16 time points. To determine the appropriate settings for the hyper-parameters for DRMN-FUSED, we scanned the following range of values: $\rho_1 \in \{30, 50, 70, 90, 110, 130, 150\}$, $\rho_2 \in \{0, 10, 30, 50, 70\}$ and $\rho_3 = \{0, 10, 30, 50, 70\}$ within a three-fold cross-validation setting (**Supplemental Figure S24**) and picked the setting with the lowest overall prediction error. The top 5 parameter configurations were $\{$(70,10,70), (70,0,70), (90,0,50), (70,30,50), (70,50,30)$\}$. We finally did a full run of DRMN-FUSED with $(\rho_1 = 70, \rho_2 = 10, \rho_3 = 70)$, which had the best prediction error. The DRMN modules were interpreted with GO enrichment analysis. We defined transitioning gene sets from the DRMN modules. Of the 14,793 genes, there were 5,762 genes that change their module assignments, which were grouped into a total of 150 transitioning gene sets of at least 5 genes spanning a total of 5,636 genes (**Supplemental Table S3**). We next predicted regulators for these transitioning gene sets using a regularized Group LASSO model, Multi-Task Group LASSO (See **Section**, **Identification of transitioning gene sets and their regulators** for more details).

H1ESC differentiation into different lineages.

The differentiation dataset from Xie et al. (Xie et al. 2013) profiled gene expression, accessibility and histone modifications in human embryonic stem cells (hESCs) and four lineages derived from hESCs, mesendoderm, neural progenitor, trophoblast-like, and mesenchymal stem cells. The dataset included eight histone modification marks: H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me1, and H3K9ac. Gene expression was measured with RNA-seq, histone modifications were measured with ChIP-seq and chromatin accessibility with DNase-seq. We aligned all sequencing reads to the human hg19 reference genome using Bowtie 2 (Langmead and Salzberg 2012). We do not expect our results to change if we used GRCh38 since most of the analysis is for protein coding genes and promoter proximal regions. We additionally realigned the RNA-seq data to GRCh38 and observe a correlation of 0.93-0.99 across the cell lines. RNA-seq data was quantified to TPMs using RSEM (Li and Dewey 2011), quantile normalized across cell lines and log transformed, resulting in 17,899 genes with expression across all cell lines. For both the ChIP-seq (chromatin marks) and DNase-seq data, we obtained per-base pair read coverage using BEDTools (Quinlan and Hall 2010) and aggregated counts within $\pm 2,500$ bps of a gene transcription start site (TSS).

DRMN was applied using the full set of features, Accessibility, Q-Motif, Histone with $k = 5$ modules. The number of modules was set following a similar approach as the hepatocyte dedifferentiation dataset. Q-Motif features were obtained using Cis-BP human motif PWM collection by applying utility tools from the PIQ software package (Sherwood et al. 2014) on ± 2500bp of a gene TSS. Each motif instance was scored with the DNase-seq signal resulting in a total of 2,998 features. These features were quantile normalized and log transformed across the cell lines. To determine the appropriate settings of the hyper-parameters, we performed three fold cross-validation with hyper parameter values in the following range: $\rho_1 \in \{30, 50, 70, 90, 110, 130, 150\}$, $\rho_2 \in \{0, 10, 30, 50, 70\}$ and $\rho_3 \in \{0, 10, 30, 50, 70\}$ (**Supplemental Figure S25**). The top 5 configurations were $\{$(150,0,30), (30,0,0), (150,0,10), (30,0,10), (130,0,30)$\}$ with similar results. The results reported here were generated by applying DRMN on the full dataset with $\rho_1 = 150, \rho_2 = 0, \rho_3 = 30$. We note that for this dataset, as the relationships between the cell types is captured by a two-level tree, the ρ_2 parameter likely does not have a large effect and most of the task sharing is sufficiently captured by the global ρ_3 parameter. Once modules were defined we analyzed them as above and generated transitioning gene sets. We identified a total of 6988 (of 17904) transitioning genes that changed their expression state.

We grouped these into 122 gene sets of at least 5 genes and included 6730 genes (**Supplemental Table S3**). We identified regulators for the transitioning gene sets using our simple regression-based approach (See **Section Identification of transitioning gene sets and their regulators** for more details).

Data Access and Software availability

The DRMN code is available at `https://github.com/Roy-lab/drmn` along with usage instructions and also available as **Supplemental File S3**. Data pre-processing and feature generation scripts are available at `https://github.com/Roy-lab/drmn_utils` and also available as **Supplemental File S4**. DRMN input files have been uploaded to zenodo and are available at `https://zenodo.org/record/6461721` and `https://zenodo.org/record/6587927`, and also available as **Supplemental File S5**. DRMN outputs have been provided as **Supplemental File S6**. The necessary mapping between motif names and TF names are available in **Supplemental File S7**. The following accession numbers correspond to the different datasets used: GSE90895 (reprogramming sequencing data), SRP000941 (H1ESC differentiation), GSE97222 (reprogramming array data).

Competing interest statement

The authors do not have any competing interests.

Acknowledgment

This work was made possible in part by NIH NIGMS grant 1R01GM117339 to S.R. The authors thank the Center for High Throughput Computing (CHTC) for computing resources.

Figures

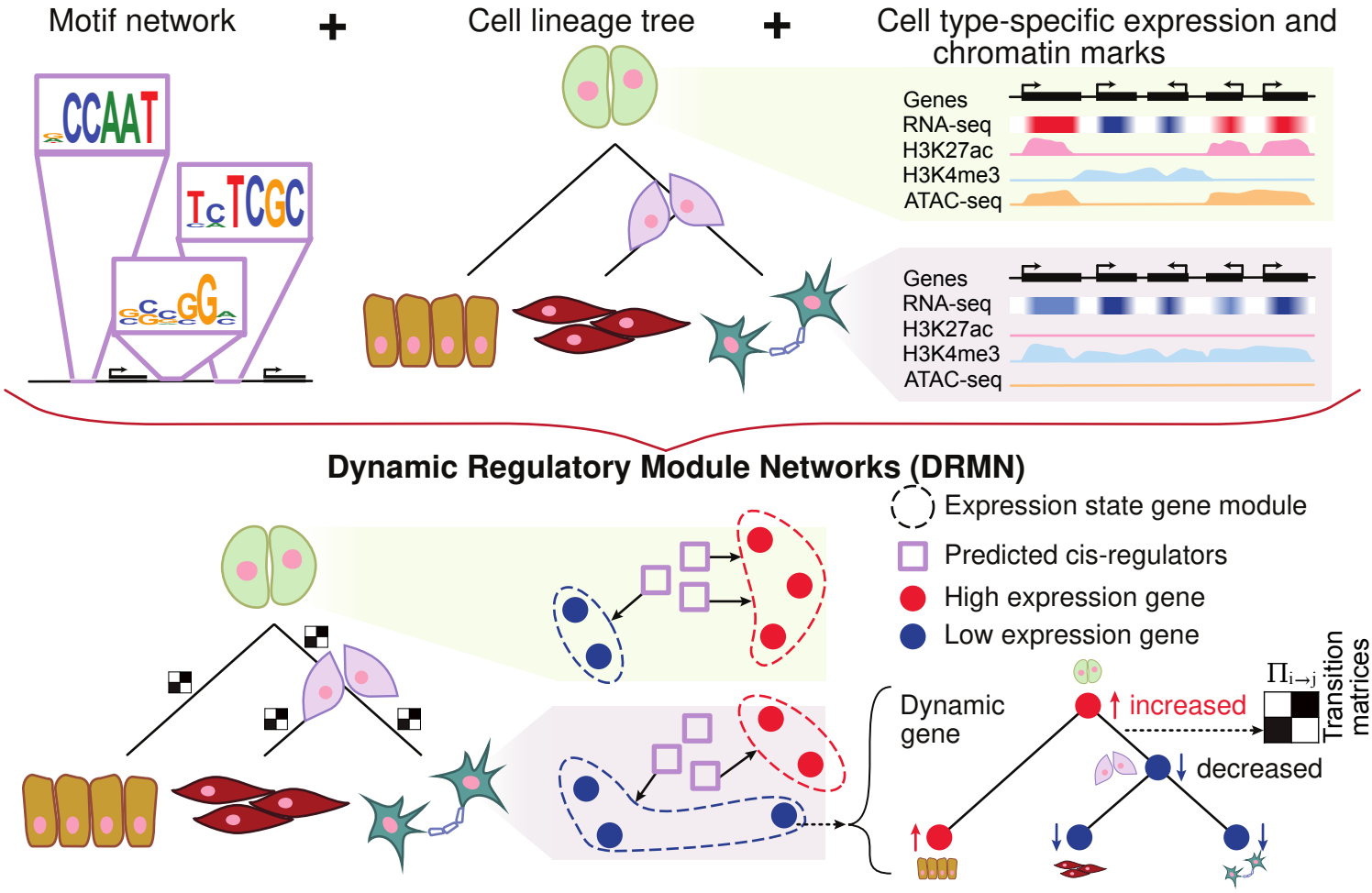
Motif network
+
Cell lineage tree
+
Cell type-specific expression and chromatin marks
Genes
RNA-seq
H3K27ac
H3K4me3
ATAC-seq
Genes
RNA-seq
H3K27ac
H3K4me3
ATAC-seq
Dynamic Regulatory Module Networks (DRMN)
Expression state gene module
Predicted cis-regulators
High expression gene
Low expression gene
Dynamic gene
increased
decreased
Transition matrices

Figure 1. Overview of Dynamic Regulatory Module Networks (DRMNs). Inputs are a lineage tree over the cell types, cell type-specific expression levels, a motif-based network, and optionally cell type-specific features such as histone modification marks or chromatin accessibility signal. The output is a learned DRMN, which consists of cell type-specific expression state modules, their regulatory programs, and transition matrices (white-black matrices) describing dynamics between the cell types. Expression states of individual genes can be traced on the tree to identify "dynamic" or "transitioning" genes.

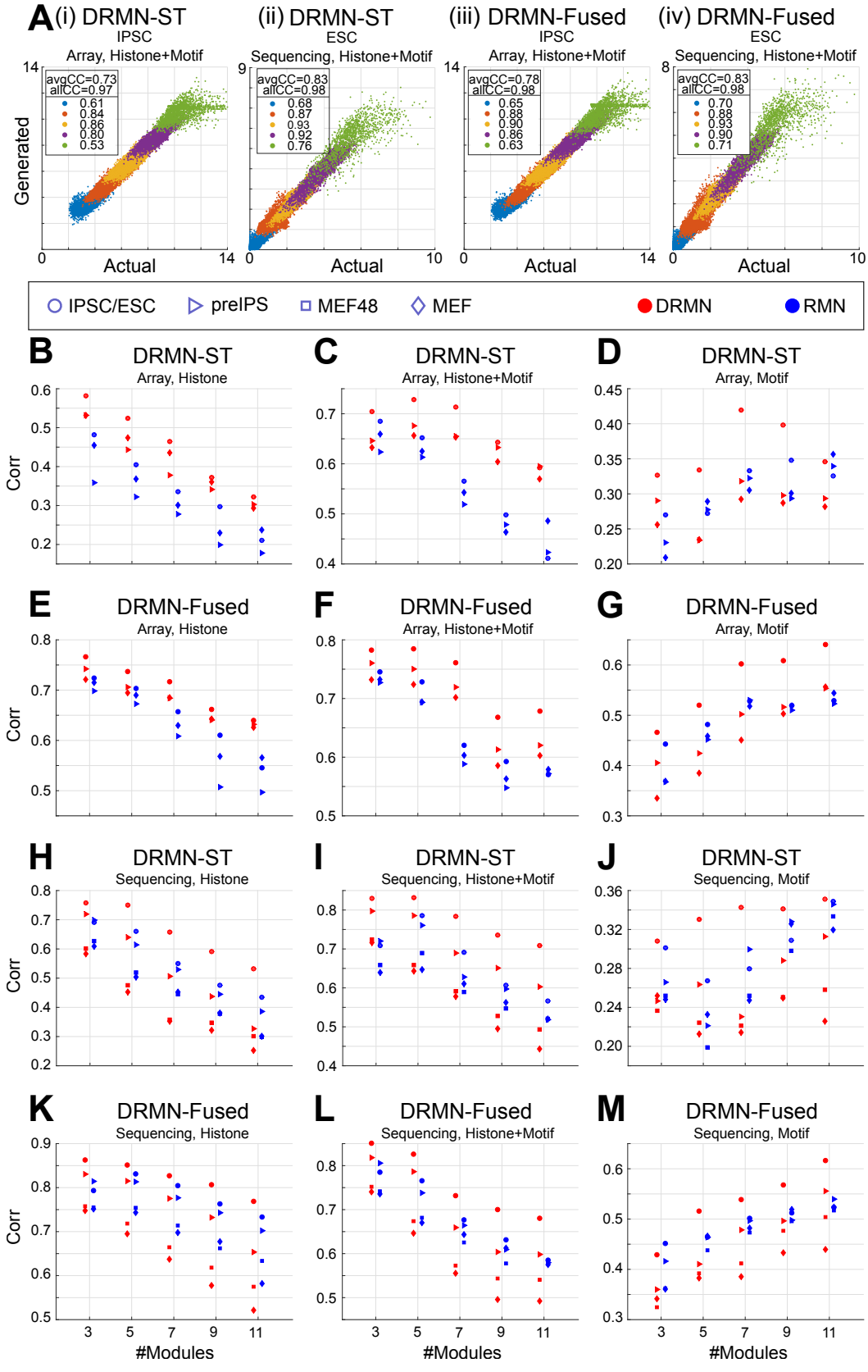
A
(i) DRMN-ST
IPSC
Array, Histone+Motif
avgCC=0.73
allCC=0.97
0.61
0.84
0.86
0.80
0.53
(ii) DRMN-ST
ESC
Sequencing, Histone+Motif
avgCC=0.83
allCC=0.98
0.68
0.87
0.93
0.92
0.76
(iii) DRMN-Fused
IPSC
Array, Histone+Motif
avgCC=0.78
allCC=0.98
0.65
0.88
0.90
0.86
0.63
(iv) DRMN-Fused
ESC
Sequencing, Histone+Motif
avgCC=0.83
allCC=0.98
0.70
0.88
0.93
0.90
0.71
Generated
Actual
IPSC/ESC
preIPS
MEF48
MEF
DRMN
RMN
B
DRMN-ST
Array, Histone
C
DRMN-ST
Array, Histone+Motif
D
DRMN-ST
Array, Motif
E
DRMN-Fused
Array, Histone
F
DRMN-Fused
Array, Histone+Motif
G
DRMN-Fused
Array, Motif
H
DRMN-ST
Sequencing, Histone
I
DRMN-ST
Sequencing, Histone+Motif
J
DRMN-ST
Sequencing, Motif
K
DRMN-Fused
Sequencing, Histone
L
DRMN-Fused
Sequencing, Histone+Motif
M
DRMN-Fused
Sequencing, Motif
Corr
#Modules

Figure 2. Comparing DRMN vs RMN models. **A**. Predicted (generated) expression from a module *vs.* measured (actual) expression, for iPSC/ESC, for **(i)** DRMN-ST on array dataset, **(ii)** DRMN-ST on sequencing dataset, **(iii)** DRMN-FUSED on array dataset, and **(iv)** DRMN-FUSED on sequencing dataset. Colors correspond to different modules. The values reported in the legend correspond to per module correlation. **B-M**. Average per-module correlation for individual cell lines as a function of different number of modules for single task and multi task versions of the method, for DRMN-ST on array dataset (**B-D**), DRMN-FUSED on array dataset (**E-G**), DRMN-ST on sequencing dataset (**H-J**), and DRMN-FUSED on sequencing dataset (**K-M**). Each shape correspond to a cell state and each color corresponds to a different method. Note, for expression prediction the generative models need the information about the observed expression.

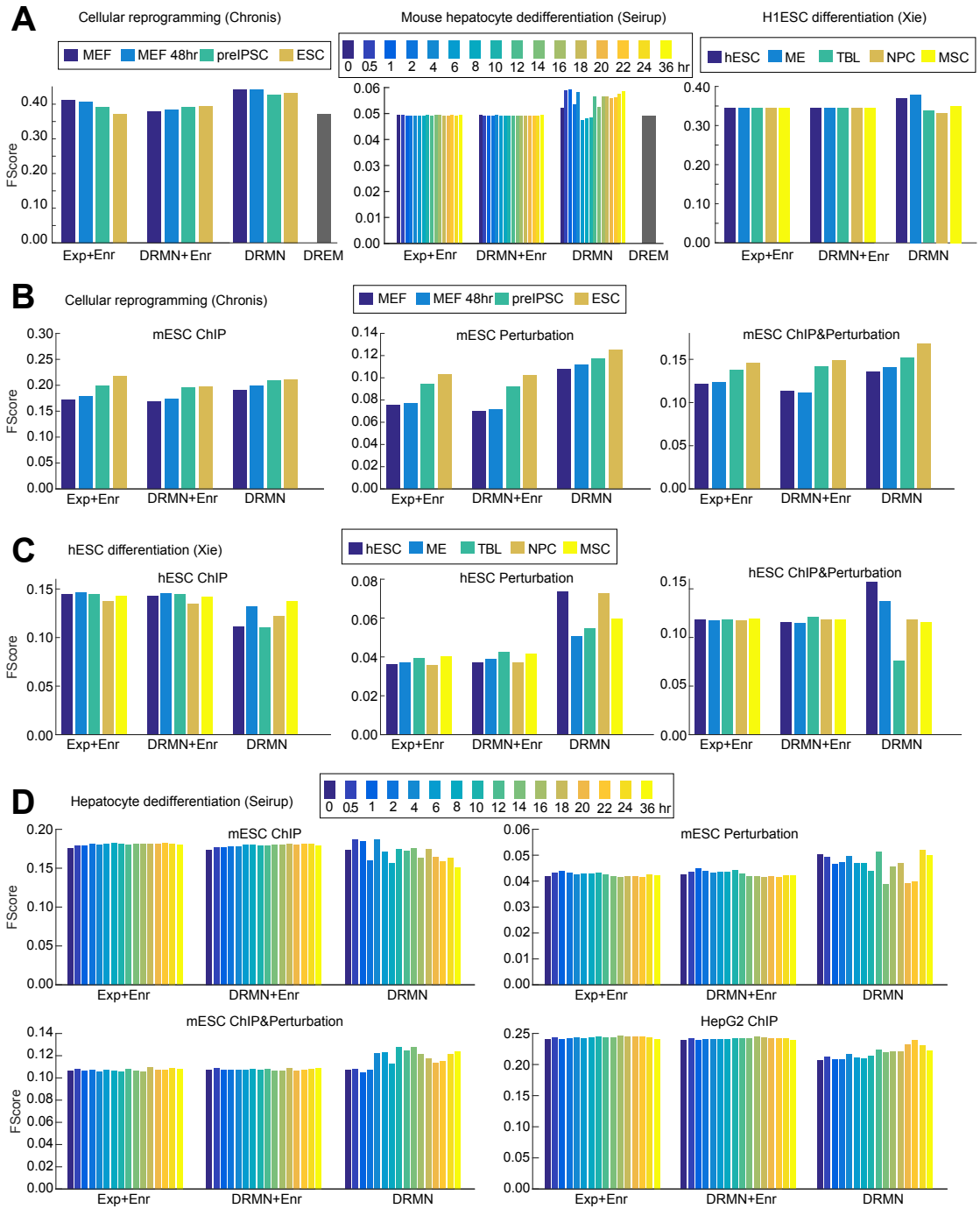
A
Cellular reprogramming (Chronis)
MEF
MEF 48hr
preIPSC
ESC
Mouse hepatocyte dedifferentiation (Seirup)
0 0.5 1 2 4 6 8 10 12 14 16 18 20 22 24 36 hr
H1ESC differentiation (Xie)
hESC
ME
TBL
NPC
MSC
FScore
Exp+Enr
DRMN+Enr
DRMN
DREM
B
Cellular reprogramming (Chronis)
mESC ChIP
mESC Perturbation
mESC ChIP&Perturbation
C
hESC differentiation (Xie)
hESC ChIP
hESC Perturbation
hESC ChIP&Perturbation
D
Hepatocyte dedifferentiation (Seirup)
mESC ChIP
mESC Perturbation
mESC ChIP&Perturbation
HepG2 ChIP

Figure 3. Performance of regulator ranking and inferred networks of DRMN. **A**. The F-score of regulator ranking with different strategies for cellular reprogramming, mouse hepatocyte dedifferentiation, and H1ESC differentiation. Each plot shows regulator enrichment of expression based modules (Exp+Enr), regulator enrichment of DRMN modules (DRMN+Enr), regulator ranking based on inferred DRMN networks (DRMN), and regulator ranking from DREM (when applicable). **B**. F-score based performance of inferred networks and baseline networks for cellular reprogramming, when compared to ChIP-based, perturbation-based, and intersection of the ChIP and perturbation-based gold standards in mESC. **C**. F-score based performance of inferred networks and baseline networks for H1ESC differentiation, when compared to ChIP based, perturbation-based, and intersection of the ChIP and perturbation-based gold standards in hESC. **D**. F-score based performance of inferred networks and baseline networks for mouse hepatocyte dedifferentiation, when compared to ChIP-based, perturbation-based, and intersection of the ChIP and perturbation-based gold standards for mESC, and ChIP-based gold standard from HepG2 cell line. In each panel, we show the baseline networks from regulator enrichment of expression based modules (Exp+Enr), regulator enrichment of DRMN modules (DRMN+Enr), and inferred networks with edges scored by input feature multiplied by estimated regression coefficients (DRMN).

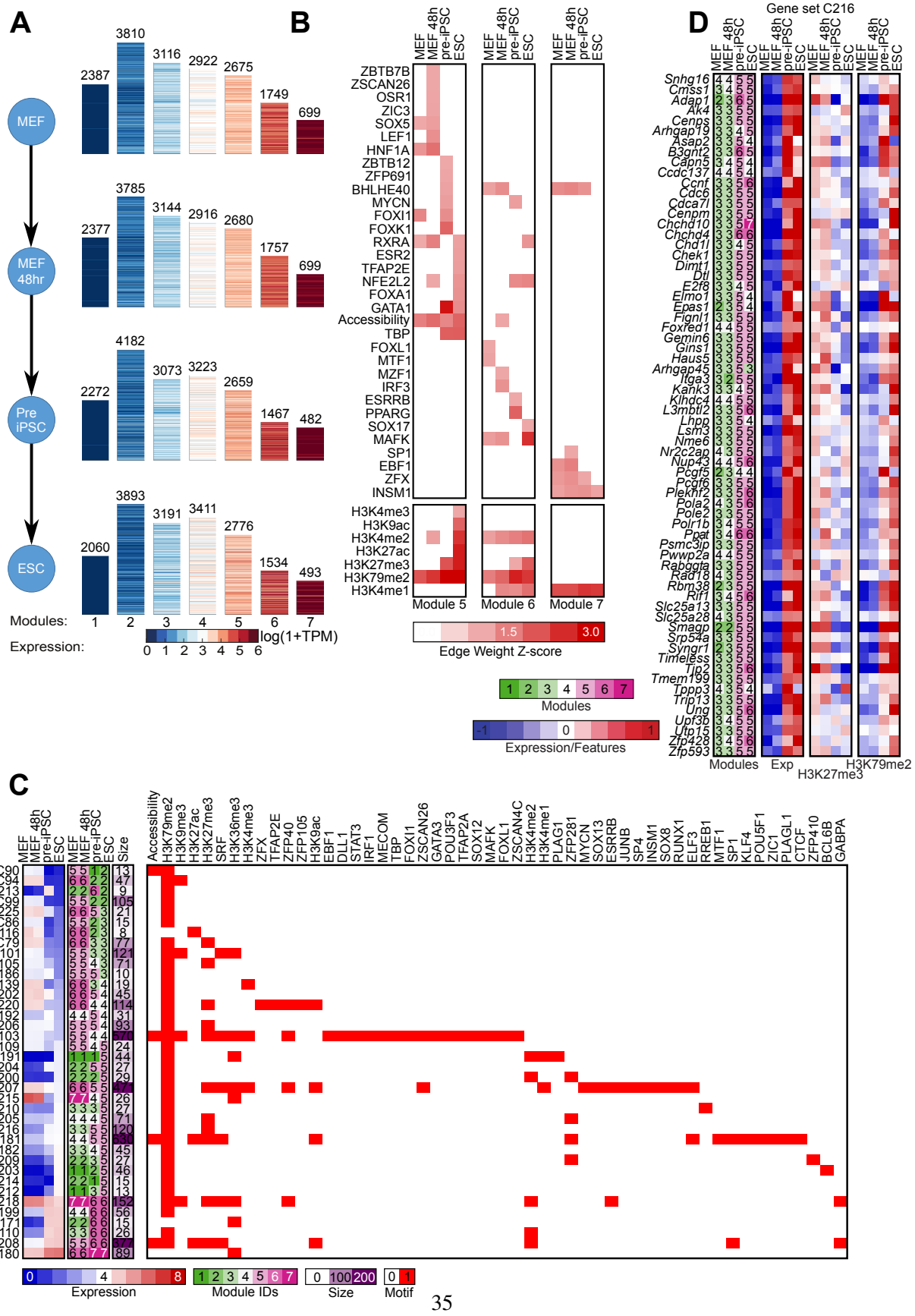
A
MEF
MEF 48hr
Pre iPSC
ESC
2387
3810
3116
2922
2675
1749
699
2377
3785
3144
2916
2680
1757
699
2272
4182
3073
3223
2659
1467
482
2060
3893
3191
3411
2776
1534
493
Modules: 1 2 3 4 5 6 7
Expression:
0 1 2 3 4 5 6 log(1+TPM)
B
MEF
MEF 48h
pre-iPSC
ESC
ZBTB7B
ZSCAN26
OSR1
ZIC3
SOX5
LEF1
HNF1A
ZBTB12
ZFP691
BHLHE40
MYCN
FOXI1
FOXK1
RXRA
ESR2
TFAP2E
NFE2L2
FOXA1
GATA1
Accessibility
TBP
FOXL1
MTF1
MZF1
IRF3
ESRRB
PPARG
SOX17
MAFK
SP1
EBF1
ZFX
INSM1
H3K4me3
H3K9ac
H3K4me2
H3K27ac
H3K27me3
H3K79me2
H3K4me1
Module 5
Module 6
Module 7
1.5
3.0
Edge Weight Z-score
1 2 3 4 5 6 7
Modules
-1 0 1
Expression/Features
D
Gene set C216
Snhg16
Cmss1
Adap1
Ak4
Cenps
Arhgap19
Asap2
B3gnt2
Capn5
Ccdc137
Ccnf
Cdc6
Cdca7l
Cenpm
Chchd10
Chchd4
Chd1l
Chek1
Dimt1
Dtl
E2f8
Elmo1
Epas1
Fignl1
Foxred1
Gemin6
Gins1
Haus5
Arhgap45
Itga3
Kank3
Klhdc4
L3mbtl2
Lhpp
Lsm3
Nme6
Nr2c2ap
Nup43
Pcgf5
Pcgf6
Plekhf2
Pola2
Pole2
Polr1b
Ppat
Psmc3ip
Pwwp2a
Rabggta
Rad18
Rbm39
Rif1
Slc25a13
Slc25a28
Smagp
Srp54a
Syngr1
Timeless
Tip2
Tmem199
Tppp3
Trip13
Ung
Upf3b
Utp15
Zfp428
Zfp593
Modules
Exp
H3K79me2
H3K27me3
C
MEF
MEF 48h
pre-iPSC
ESC
Size
Accessibility
H3K79me2
H3K9me3
H3K27ac
H3K27me3
SRF
H3K36me3
H3K4me3
ZFX
TFAP2E
ZFP40
ZFP105
H3K9ac
EBF1
DLL1
STAT3
IRF1
MECOM
TBP
FOXI1
ZSCAN26
GATA3
POU3F3
TFAP2A
SOX12
MAFK
FOXL1
ZSCAN4C
H3K4me2
H3K4me1
PLAG1
ZFP281
MYCN
SOX13
ESRRB
JUNB
SP4
INSM1
SOX8
RUNX1
ELF3
RREB1
MTF1
SP1
KLF4
POU5F1
ZIC1
PLAGL1
CTCF
ZFP410
BCL6B
GABPA
C90 C94 C213 C99 C225 C86 C116 C79 C101 C105 C186 C139 C202 C220 C192 C206 C103 C109 C191 C204 C200 C207 C215 C210 C205 C216 C181 C182 C209 C203 C214 C212 C218 C199 C171 C110 C208 C180
0 4 8
Expression
1 2 3 4 5 6 7
Module IDs
0 100 200
Size
0 1
Motif

Figure 4. Application of DRMNs to the cellular reprogramming sequencing dataset using histone marks, accessibility and Q-motifs. **A**. Shown are gene expression (log 1+TPM) patterns of the $k = 7$ modules for each cell state (major row). The number above the heatmap is the number of genes in that module. The color corresponds to the level of expression, with more red denoting high expression and more blue denoting low expression. **B**. Inferred regulators for each module across time. Only modules with highest expression and that had TFs as regulators (5, 6, and 7) are shown. The red intensity is proportional to the z-score significance of the regression weight of a regulator. See **Supplemental Table S6** for all values. **C**. Transitioning gene sets exhibiting changes into the high expression modules (5, 6, 7). Shown are the mean expression levels of genes in the gene set (left, red-blue heat map), the module assignment (second), the number of genes in each module (third) and the set of regulators for each gene set (white-red heat map). Red arrows depict gene sets discussed in the text. **D**. Selected transitioning gene set C216 in the cellular reprogramming dataset. The panel shows the member genes of the transitioning gene set. The columns show the module assignment of each gene, followed by its expression level in each cell state (Expression). The subsequent groups of columns are the levels of the regulator on the gene promoters. The name of the regulator is specified at the bottom of the heat map.

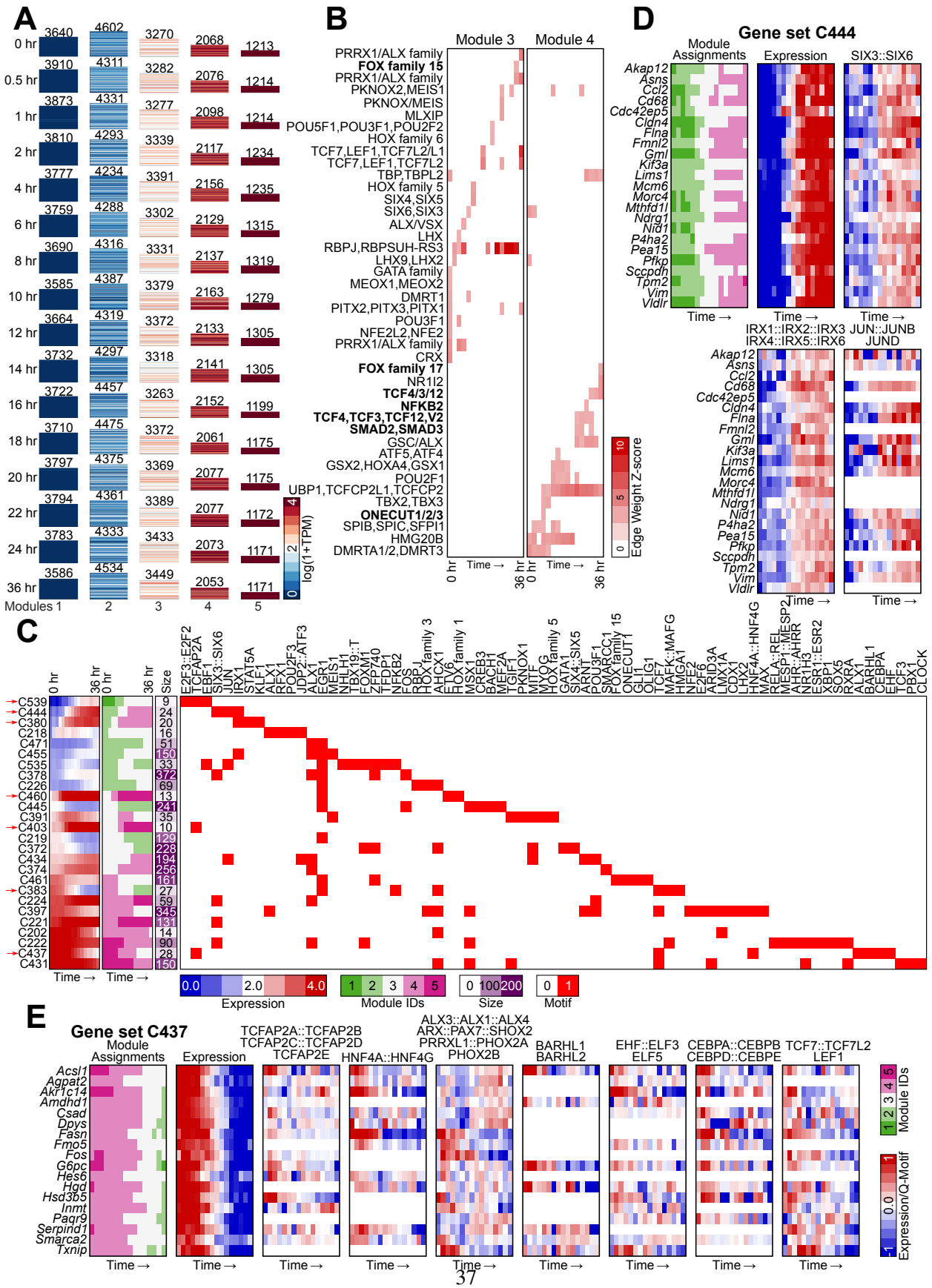
A
B
C
D
E
Gene set C444
Gene set C437
Module Assignments
Expression
SIX3::SIX6
IRX1::IRX2::IRX3
IRX4::IRX5::IRX6
JUN::JUNB
JUND
Module 3
Module 4
Time →
Modules 1 2 3 4 5
log(1+TPM)
Edge Weight Z-score
Expression
Module IDs
Size
Motif
Expression/Q-Motif

Figure 5. Application of DRMNs to the hepatocyte dedifferentiation dataset using accessibility and Q-motifs. **A**. Shown are gene expression (log 1+TPM) patterns of the $k = 5$ modules for each time point (major row). The number above the heatmap is the number of genes in that module. The color corresponds to the level of expression, with more red denoting high expression and more blue denoting low expression. **B**. Inferred regulators for each module across time. Only modules with highest expression and that had TFs as regulators (3 and 4) are shown. The red intensity is proportional the z-score significance of the regression weight of a regulator. See **Supplemental Table S6** for all values. **C**. Transitioning gene sets exhibiting changes into the high expression modules (3, 4, 5). Shown are the mean expression levels of genes in the gene set (left, red-blue heat map), the module assignment (second), the number of genes in each module (third) and the set of regulators for each gene set (white-red heat map). Red arrows depict gene sets discussed in the text. Some of the regulator names were shortened for space, the full names are available in **Supplemental File S7**. **D, E**. Selected transitioning gene sets in the hepatocyte dedifferentiation dataset. Each panel shows the member genes of a transitioning gene set (label on top). The columns show the module assignment of each gene, followed by its expression level in each cell type (Expression). The subsequent groups of columns are the levels of the regulator on the gene promoters. The name of the regulator is specified at the top of the heat map. All significant regulators are shown.

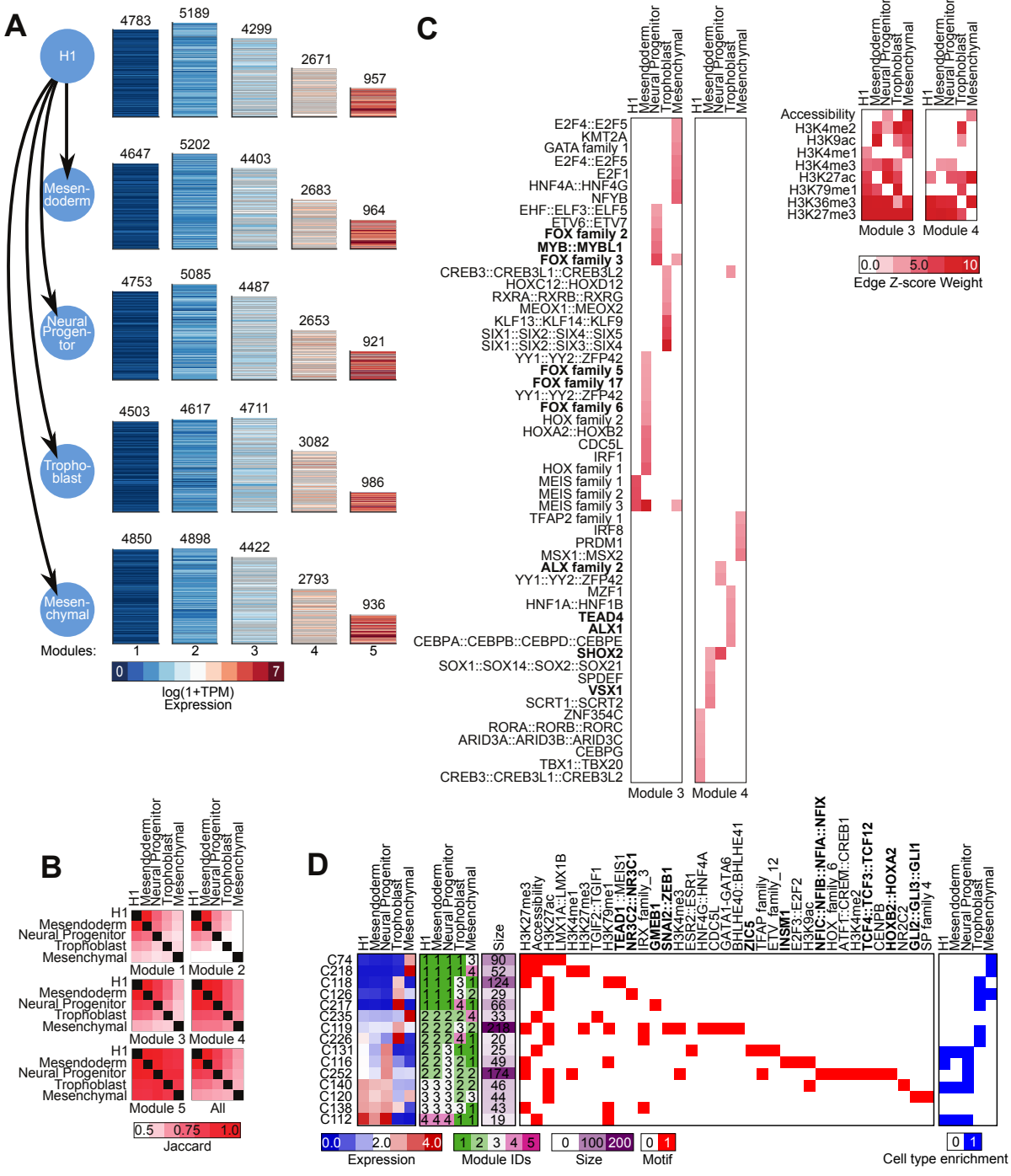
A
H1
Mesen-doderm
Neural Progen-tor
Tropho-blast
Mesen-chymal
4783
5189
4299
2671
957
4647
5202
4403
2683
964
4753
5085
4487
2653
921
4503
4617
4711
3082
986
4850
4898
4422
2793
936
Modules: 1 2 3 4 5
0 7
log(1+TPM) Expression
B
H1
Mesendoderm
Neural Progenitor
Trophoblast
Mesenchymal
Module 1
Module 2
Module 3
Module 4
Module 5
All
0.5 0.75 1.0
Jaccard
C
E2F4::E2F5
KMT2A
GATA family 1
E2F4::E2F5
E2F1
HNF4A::HNF4G
NFYB
EHF::ELF5::ELF5
ETV6::ETV7
FOX family 2
MYB::MYBL1
FOX family 3
CREB3::CREB3L1::CREB3L2
HOXC12::HOXD12
RXRA::RXRB::RXRG
MEOX1::MEOX2
KLF13::KLF14::KLF9
SIX1::SIX2::SIX4::SIX5
SIX1::SIX2::SIX3::SIX4
YY1::YY2::ZFP42
FOX family 5
FOX family 17
YY1::YY2::ZFP42
FOX family 6
HOX family 2
HOXA2::HOXB2
CDC5L
IRF1
HOX family 1
MEIS family 1
MEIS family 2
MEIS family 3
TFAP2 family 1
IRF8
PRDM1
MSX1::MSX2
ALX family 2
YY1::YY2::ZFP42
MZF1
HNF1A::HNF1B
TEAD4
ALX1
CEBPA::CEBPB::CEBPD::CEBPE
SHOX2
SOX1::SOX14::SOX2::SOX21
SPDEF
VSX1
SCRT1::SCRT2
ZNF354C
RORA::RORB::RORC
ARID3A::ARID3B::ARID3C
CEBPG
TBX1::TBX20
CREB3::CREB3L1::CREB3L2
Module 3
Module 4
Accessibility
H3K4me2
H3K9ac
H3K4me1
H3K4me3
H3K27ac
H3K79me1
H3K36me3
H3K27me3
Module 3
Module 4
0.0 5.0 10
Edge Z-score Weight
D
Size
C74
C218
C118
C126
C217
C235
C119
C226
C131
C116
C252
C140
C120
C138
C112
H3K27me3
Accessibility
H3K27ac
LMX1A::LMX1B
H3K4me1
H3K27me3
TGIF2::TGIF1
H3K79me1
TEAD1::MEIS1
NR3C2::NR3C1
IRX family_3
GMEB1
SNAI2::ZEB1
H3K4me3
ESR2::ESR1
HNF4G::HNF4A
CDC5L
GATA1-GATA6
BHLHE40::BHLHE41
ZIC5
TFAP family
ETV family_12
INSM1
E2F3::E2F2
H3K9ac
NFIC::NFIB::NFIA::NFIX
HOX family 6
ATF1::CREM::CREB1
H3K4me2
TCF4::TCF3::TCF12
CENPB
HOXB2::HOXA2
NR2C2
GLI2::GLI3::GLI1
SP family 4
0.0 2.0 4.0
Expression
1 2 3 4 5
Module IDs
0 100 200
Size
0 1
Motif
0 1
Cell type enrichment

Figure 6. Application of DRMNs to ESC differentiation dataset using histone marks, accessibility and Q-motifs. **A**. Shown are gene expression (log 1+TPM) patterns of the $k = 5$ modules for each lineage (major row). The number above the heatmap is the number of genes in that module. The color corresponds to the level of expression, with more red denoting high expression and more blue denoting low expression. **B**. Similarity of modules across lineages as measured by Jaccard index. The color intensity is proportional to the module match. **C**. Inferred regulators for each modules across time. Only modules with highest expression (3 and 4) and that had TFs as regulators are shown. The red intensity is proportional the z-score significance of the regression weight of a regulator. See **Supplemental Table S6** for all values. **D**. Selected transitioning gene set profiles exhibiting changes into the high expression modules (3, 4, 5). Shown are the mean expression levels of genes in the gene set (left, red-blue heat map), the module assignment (second), the number of genes in each module (third), the set of regulators for each gene set (white-red heat map) and enrichment in lineage-specific gene sets from Xie et al. The complete transitioning gene set is shown in **Supplemental Figure S14**. Some of the regulator names were shortened for space, the full names are available in **Supplemental File S7**.

References

Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, Fropf R, McAnany C, Gagneur J, Kundaje A, et al. 2021. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*. **53**: 354–366.

Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, and Bulyk ML. 2006. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology*. **24**: 1429–1435.

Bunina D, Abazova N, Diaz N, Noh KM, Krijgsveld J, and Zaugg JB. 2020. Genomic Rewiring of SOX2 Chromatin Interaction Network during Differentiation of ESCs to Postmitotic Neurons. *Cell Systems*. **10**: 480–494.e8.

Chasman D, Iyer N, Fotuhi Siahpirani A, Estevez Silva M, Lippmann E, McIntosh B, Probasco MD, Jiang P, Stewart R, Thomson JA, et al. 2019. Inferring Regulatory Programs Governing Region Specificity of Neuroepithelial Stem Cells during Early Hindbrain and Spinal Cord Development. *Cell Systems*. **9**: 167–186.e12.

Chronis C, Fiziev P, Papp B, Butz S, Bonora G, Sabri S, Ernst J, and Plath K. 2017. Cooperative Binding of Transcription Factors Orchestrates Reprogramming. *Cell*. **168**: 442–459.e20.

Clotman F, Jacquemin P, Plumb-Rudewiez N, Pierreux CE, Van der Smissen P, Dietz HC, Courtoy PJ, Rousseau GG, and Lemaigre FP. 2005. Control of liver cell fate decision by a gradient of TGF beta signaling modulated by Onecut transcription factors. *Genes Dev*. **19**: 1849–1854.

Cole MF, Johnstone SE, Newman JJ, Kagey MH, and Young RA. 2008. Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells. *Genes & development*. **22**: 746–755.

Correa-Cerro LS, Piao Y, Sharov AA, Nishiyama A, Cadet JS, Yu H, Sharova LV, Xin L, Hoang HG, Thomas M, et al. 2011. Generation of mouse ES cell lines engineered for the forced induction of transcription factors. *Scientific Reports*. **1**:

Dai X, Yan X, Wintergerst KA, Cai L, Keller BB, and Tan Y. 2020. Nrf2: Redox and Metabolic Regulator of Stem Cell State and Function. *Trends in Molecular Medicine*. **26**: 185–200.

699 Davenne M, Maconochie MK, Neun R, Pattyn A, Chambon P, Krumlauf R, and Rijli FM. 1999. Hoxa2
700 and Hoxb2 control dorsoventral patterns of neuronal development in the rostral hindbrain. *Neuron*. **22**:
701 677–691.
702 Davis J and Goadrich M 2006. The Relationship Between Precision-Recall and ROC Curves. In: *Proceed-*
703 *ings of the 23rd International Conference on Machine Learning (ICML 2006)*. ICML '06. Pittsburgh,
704 Pennsylvania, USA: ACM, pp. 233–240.
705 Diacou R, Zhao Y, Zheng D, Cvekl A, and Liu W. 2018. Six3 and Six6 Are Jointly Required for the
706 Maintenance of Multipotent Retinal Progenitors through Both Positive and Negative Regulation. *Cell*
707 *Reports*. **25**: 2510–2523.e4.
708 Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, Gingeras TR, Gerstein M, Guigó R,
709 Birney E, et al. 2012. Modeling gene expression using chromatin features in various cellular contexts.
710 *Genome biology*. **13**: R53.
711 Elaut G, Henkens T, Papeleu P, Snykers S, Vinken M, Vanhaecke T, and Rogiers V. 2006. Molecular Mech-
712 anisms Underlying the Dedifferentiation Process of Isolated Hepatocytes and Their Cultures. *Current*
713 *Drug Metabolism*. **7**: 629–660.
714 Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne
715 M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*. **473**:
716 43–49.
717 Ernst J, Vainas O, Harbison CT, Simon I, and Bar-Joseph Z. 2007. Reconstructing dynamic regulatory maps.
718 *Molecular systems biology*. **3**: 74.
719 Ferri ALM, Lin W, Mavromatakis YE, Wang JC, Sasaki H, Whitsett JA, and Ang SL. 2007. Foxa1 and
720 Foxa2 regulate multiple phases of midbrain dopaminergic neuron development in a dosage-dependent
721 manner. *Development*. **134**: 2761–2769.
722 Fidalgo M, Shekar PC, Ang YS, Fujiwara Y, Orkin SH, and Wang J. 2011. Zfp281 functions as a transcrip-
723 tional repressor for pluripotency of mouse embryonic stem cells. *Stem Cells*. **29**: 1705–1716.
724 Francius C, Hidalgo-Figueroa M, Debrulle S, Pelosi B, Rucchin V, Ronellenfitch K, Panayiotou E, Makrides
725 N, Misra K, Harris A, et al. 2016. Vsx1 Transiently Defines an Early Intermediate V2 Interneuron

Precursor Compartment in the Mouse Developing Spinal Cord. *Frontiers in molecular neuroscience*. **9**: 145–145.

Fraser J, Williamson I, Bickmore WA, and Dostie J. 2015. An Overview of Genome Organization and How We Got There: from FISH to Hi-C. *Microbiology and Molecular Biology Reviews*. **79**: 347–372.

Gong W, Koyano-Nakagawa N, Li T, and Garry DJ. 2015. Inferring dynamic gene regulatory networks in cardiac differentiation through the integration of multi-dimensional data. *BMC bioinformatics*. **16**: 74.

Gonzalez AJ, Setty M, and Leslie CS. 2015. Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. *Nature genetics*. **47**: 1249–1259.

Gorkin DU, Leung D, and Ren B. 2014. The 3D Genome in Transcriptional Regulation and Pluripotency. *Cell Stem Cell*. **14**: 762–775.

Grant CE, Bailey TL, and Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics*. **27**: 1017–1018.

Heng YHE, McLeay RC, Harvey TJ, Smith AG, Barry G, Cato K, Plachez C, Little E, Mason S, Dixon C, et al. 2014. NFIX regulates neural progenitor cell differentiation during hippocampal morphogenesis. *Cerebral cortex (New York, N.Y. : 1991)*. **24**: 261–279.

Hui Cc and Angers S. 2011. Gli Proteins in Development and Disease. *Annual Review of Cell and Developmental Biology*. **27**: 513–537.

Ikeda T, Zhang J, Chano T, Mabuchi A, Fukuda A, Kawaguchi H, Nakamura K, and Ikegawa S. 2002. Identification and characterization of the human long form of Sox5 (L-SOX5) gene. *Gene*. **298**: 59–68.

Jojic V, Shay T, Sylvia K, Zuk O, Sun X, Kang J, Regev A, Koller D, Immunological Genome Project Consortium, Best AJ, et al. 2013. Identification of transcriptional regulators in the mouse immune system. *Nat Immunol*. **14**: 633–643.

Kisanga EP, Tang Z, Guller S, and Whirledge S. 2018. Glucocorticoid signaling regulates cell invasion and migration in the human first-trimester trophoblast cell line Sw.71. *Am J Reprod Immunol*. **80**: e12974.

Koch C, Konieczka J, Delorey T, Lyons A, Socha A, Davis K, Knaack SA, Thompson D, O'Shea EK, Regev A, et al. 2017. Inference and Evolutionary Analysis of Genome-Scale Regulatory Networks in Large Phylogenies. *Cell Systems*. **4**: 543–558.e8.

Kundaje A, Lianoglou S, Li X, Quigley D, Arias M, Wiggins CH, Zhang L, and Leslie C. 2007. Learning Regulatory Programs That Accurately Predict Differential Expression with MEDUSA. *Annals of the New York Academy of Sciences*. **1115**: 178–202.

Kyrmizi I, Hatzis P, Katrakili N, Tronche F, Gonzalez FJ, and Talianidis I. 2006. Plasticity and expanding complexity of the hepatic transcription factor network during liver development. *Genes & Development*. **20**: 2293–2305.

Langfelder P and Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. **9**: 559.

Langmead B and Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*. **9**: 357–359.

Lara-Astiaso D, Weiner A, Lorenzo-Vivas E, Zaretsky I, Jaitin DA, David E, Keren-Shaul H, Mildner A, Winter D, Jung S, et al. 2014. Chromatin state dynamics during blood formation. *Science (New York, N.Y.)* **345**: 943–949.

Laudadio I, Manfroid I, Achouri Y, Schmidt D, Wilson MD, Cordi S, Thorrez L, Knoops L, Jacquemin P, Schuit F, et al. 2012. A Feedback Loop Between the Liver-Enriched Transcription Factor Network and Mir-122 Controls Hepatocyte Differentiation. *Gastroenterology*. **142**: 119–129.

Le Lay J and Kaestner KH. 2010. The Fox genes in the liver: from organogenesis to functional integration. *Physiol Rev*. **90**: 1–22.

Lee SI, Dudley AM, Drubin D, Silver PA, Krogan NJ, Pe'er D, and Koller D. 2009. Learning a Prior on Regulatory Potential from eQTL Data. *PLoS Genet*. **5**: e1000358+.

Lee TI and Young RA. 2013. Transcriptional regulation and its misregulation in disease. *Cell*. **152**: 1237–1251.

Li B and Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*. **12**: 323.

Luedde T and Schwabe RF. 2011. NF-κB in the liver–linking injury, fibrosis and hepatocellular carcinoma. *Nature reviews. Gastroenterology & hepatology*. **8**: 108–118.

779 Madissoon E, Damdimopoulos A, Katayama S, Krjutškov K, Einarsdottir E, Mamia K, De Groef B, Hovatta
780 O, Kere J, and Damdimopoulou P. 2019. Pleomorphic Adenoma Gene 1 Is Needed For Timely Zygotic
781 Genome Activation and Early Embryo Development. *Scientific Reports*. **9**: 8411.
782 Magee N and Zhang Y. 2017. Role of early growth response 1 in liver metabolism and liver cancer. *Hep-*
783 *atoma research*. **3**: 268–277.
784 Maherali N, Sridharan R, Xie W, Utikal J, Eminli S, Arnold K, Stadtfeld M, Yachechko R, Tchieu J, Jaenisch
785 R, et al. 2007. Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread
786 tissue contribution. *Cell stem cell*. **1**: 55–70.
787 Malaterre J, Mantamadiotis T, Dworkin S, Lightowler S, Yang Q, Ransome MI, Turnley AM, Nichols NR,
788 Emambokus NR, Frampton J, et al. 2008. c-Myb is required for neural progenitor cell proliferation and
789 maintenance of the neural stem cell niche in adult brain. *Stem Cells*. **26**: 173–181.
790 Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-
791 Margoulis OV, et al. 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic*
792 *acids research*. **31**: 374–378.
793 Miyashita N, Horie M, Suzuki HI, Saito M, Mikami Y, Okuda K, Boucher RC, Suzukawa M, Hebisawa A,
794 Saito A, et al. 2020. FOXL1 Regulates Lung Fibroblast Function via Multiple Mechanisms. *Am J Respir*
795 *Cell Mol Biol*. **63**: 831–842.
796 Müller FJ, Laurent LC, Kostka D, Ulitsky I, Williams R, Lu C, Park IH, Rao MS, Shamir R, Schwartz PH,
797 et al. 2008. Regulatory networks define phenotypic classes of human stem cell lines. *Nature*. **455**:
798 Nishiyama A, Sharov AA, Piao Y, Amano M, Amano T, Hoang HG, Binder BY, Tapnio R, Bassey U,
799 Malinou JN, et al. 2013. Systematic repression of transcription factors reveals limited patterns of gene
800 expression changes in ES cells. *Scientific Reports*. **3**:
801 Nishiyama A, Xin L, Sharov AA, Thomas M, Mowrer G, Meyers E, Piao Y, Mehta S, Yee S, Nakatake Y,
802 et al. 2009. Uncovering Early Response of Gene Regulatory Networks in ESCs by Systematic Induction
803 of Transcription Factors. *Cell Stem Cell*. **5**: 420–433.
804 Odom DT, Dowell RD, Jacobsen ES, Nekludova L, Rolfe PA, Danford TW, Gifford DK, Fraenkel E, Bell
805 GI, and Young RA. 2006. Core transcriptional regulatory circuitry in human hepatocytes. *Molecular*
806 *Systems Biology*. **2**: 2006.0017.

807 Osmanbeyoglu HU, Shimizu F, Rynne-Vidal A, Alonso-Curbelo D, Chen HA, Wen HY, Yeung TL, Jelinic
808 P, Razavi P, Lowe SW, et al. 2019. Chromatin-informed inference of transcriptional programs in gyne-
809 cologic and basal breast cancers. *Nature Communications*. **10**: 4369.
810 Parikh AP, Wu W, Curtis RE, and Xing EP. 2011. TREEGL: reverse engineering tree-evolving gene net-
811 works underlying developing biological lineages. *Bioinformatics (Oxford, England)*. **27**: i196–i204.
812 Phan QM, Fine GM, Salz L, Herrera GG, Wildman B, Driskell IM, and Driskell RR. 2020. Lef1 expression
813 in fibroblasts maintains developmental potential in adult skin to regenerate wounds. *eLife*. **9**: e60066.
814 Pierson E, GTEx Consortium, Koller D, Battle A, Mostafavi S, Ardlie KG, Getz G, Wright FA, Kellis M,
815 Volpi S, et al. 2015. Sharing and Specificity of Co-expression Networks across 35 Human Tissues. *PLoS*
816 *Comput Biol*. **11**: e1004220.
817 Quinlan AR and Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features.
818 *Bioinformatics*. **26**: 841–842.
819 Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, Akalin A, Schmeier S, Kanamori-
820 Katayama M, Bertin N, et al. 2010. An Atlas of Combinatorial Transcriptional Regulation in Mouse and
821 Man. *Cell*. **140**: 744–752.
822 Rego TG do, Roider HG, Carvalho FAT de, and Costa IG. 2012. Inferring epigenetic and transcriptional reg-
823 ulation during blood cell development with a mixture of sparse linear models. *Bioinformatics (Oxford,*
824 *England)*. **28**: 2297–2303.
825 Rétaux S, Rogard M, Bach I, Failli V, and Besson MJ. 1999. Lhx9: A Novel LIM-Homeodomain Gene
826 Expressed in the Developing Forebrain. *Journal of Neuroscience*. **19**: 783–793.
827 Roy S, Lagree S, Hou Z, Thomson JA, Stewart R, and Gasch AP. 2013. Integrated Module and Gene-Specific
828 Regulatory Inference Implicates Upstream Signaling Networks. *PLoS Comput. Biol.* **9**: e1003252+.
829 Roy S and Sridharan R. 2017. Chromatin module inference on cellular trajectories identifies key transition
830 points and poised epigenetic states in diverse developmental processes. *Genome research*. **27**: 1250–
831 1262.
832 Roy S, Werner-Washburne M, and Lane T. 2011. A multiple network learning approach to capture system-
833 wide condition-specific responses. *Bioinformatics*. **27**: 1832–1838.

Rudraiah S, Zhang X, and Wang L. 2016. Nuclear Receptors as Therapeutic Targets in Liver Disease: Are We There Yet? *Annual Review of Pharmacology and Toxicology*. **56**: 605–626.

Saha B, Ganguly A, Home P, Bhattacharya B, Ray S, Ghosh A, Rumi MAK, Marsh C, French VA, Gunewardena S, et al. 2020. TEAD4 ensures postimplantation development by promoting trophoblast self-renewal: An implication in early human pregnancy loss. *Proceedings of the National Academy of Sciences*. **117**: 17864–17875.

Sakaguchi M, Cai W, Wang CH, Cederquist CT, Damasio M, Homan EP, Batista T, Ramirez AK, Gupta MK, Steger M, et al. 2019. FoxK1 and FoxK2 in insulin regulation of cellular and mitochondrial metabolism. *Nature Communications*. **10**: 1582.

Sandelin A, Alkema W, Engström P, Wasserman WW, and Lenhard B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*. **32**: D91–D94.

Schulz MH, Devanny WE, Gitter A, Zhong S, Ernst J, and Bar-Joseph Z. 2012. DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. *BMC Systems Biology*. **6**: 104.

Schwab JD, Kühlwein SD, Ikonomi N, Kühl M, and Kestler HA. 2020. Concepts in Boolean network modeling: What do they all mean? *Computational and structural biotechnology journal*. **18**: 571–582.

Scott A, Hasegawa H, Sakurai K, Yaron A, Cobb J, and Wang F. 2011. Transcription factor short stature homeobox 2 is required for proper development of tropomyosin-related kinase B-expressing mechanosensory neurons. *J Neurosci*. **31**: 6741–6749.

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, and Friedman N. 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**: 166–176.

Seirup M, Sengupta S, Swanson S, McIntosh BE, Collins M, Chu LF, Cheng Z, Gorkin DU, Duffin B, Bolin JM, et al. 2022. Rapid changes in chromatin structure during dedifferentiation of primary hepatocytes in vitro. *Genomics*. **114**: 110330.

Setty M and Leslie CS. 2015. SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. *PLOS Computational Biology*. **11**: 1–21.

Sheaffer KL and Kaestner KH. 2012. Transcriptional Networks in Liver and Intestinal Development. *Cold Spring Harbor Perspectives in Biology*. **4**:

Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, and Gifford DK. 2014. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol*. **32**: 171–178.

Soncin F and Parast MM. 2020. Role of Hippo signaling pathway in early placental development. *Proceedings of the National Academy of Sciences*. **117**: 20354–20356.

Sridharan R, Gonzales-Cope M, Chronis C, Bonora G, McKee R, Huang C, Patel S, Lopez D, Mishra N, Pellegrini M, et al. 2013. Proteomic and genomic approaches reveal critical functions of H3K9 methylation and heterochromatin protein-1γ in reprogramming to pluripotency. *Nature cell biology*. **15**: 872–882.

Sridharan R and Plath K. 2008. Illuminating the black box of reprogramming. *Cell Stem Cell*. **2**: 295–297.

Sridharan R, Tchieu J, Mason MJ, Yachechko R, Kuoy E, Horvath S, Zhou Q, and Plath K. 2009. Role of the murine reprogramming factors in the induction of pluripotency. *Cell*. **136**: 364–377.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*. **102**: 15545–15550.

Tavera-Montañez C, Hainer SJ, Cangussu D, Gordon SJV, Xiao Y, Reyes-Gutierrez P, Imbalzano AN, Navea JG, Fazzio TG, and Padilla-Benavides T. 2019. The classic metal-sensing transcription factor MTF1 promotes myogenesis in response to copper. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*. **33**: 14556–14574.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*. **489**: 57–74.

Uemura M, Swenson ES, Gaça MDA, Giordano FJ, Reiss M, and Wells RG. 2005. Smad2 and Smad3 play different roles in rat hepatic stellate cell function and alpha-smooth muscle actin organization. *Molecular biology of the cell*. **16**: 4214–4224.

Vallecillo-García P, Orgeur M, Hofe-Schneider S vom, Stumm J, Kappert V, Ibrahim DM, Börno ST, Hayashi S, Relaix F, Hildebrandt K, et al. 2017. Odd skipped-related 1 identifies a population of em-

bryonic fibro-adipogenic progenitors regulating myogenesis during limb development. *Nature Communications*. **8**: 1218.

Velazquez JJ, LeGraw R, Moghadam F, Tan Y, Kilbourne J, Maggiore JC, Hislop J, Liu S, Cats D, Chuva de Sousa Lopes SM, et al. 2021. Gene Regulatory Network Analysis and Engineering Directs Development and Vascularization of Multilineage Human Liver Organoids. *Cell Systems*. **12**: 41–55.e11.

Wamstad JA, Alexander JM, Truty RM, Shrikumar A, Li F, Eilertson KE, Ding H, Wylie JN, Pico AR, Capra JA, et al. 2012. Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell*. **151**: 206–220.

Wangensteen KJ, Zhang S, Greenbaum LE, and Kaestner KH. 2015. A genetic screen reveals Foxa3 and TNFR1 as key regulators of liver repopulation. *Genes & Development*. **29**: 904–909.

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*. **158**: 1431–1443.

Whalen S, Truty RM, and Pollard KS. 2016. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics*. **48**: 488–496.

Wit E de and Laat W de. 2012. A decade of 3C technologies: insights into nuclear organization. *Genes & Development*. **26**: 11–24.

Wittmann MT and Häberle BM. 2018. Linking the Neuropsychiatric Disease Gene TCF4 to Neuronal Activity-Dependent Regulatory Networks. *Journal of Neuroscience*. **38**: 2653–2655.

Wong DJ, Liu H, Ridky TW, Cassarino D, Segal E, and Chang HY. 2008. Module Map of Stem Cell Genes Guides Creation of Epithelial Cancer Stem Cells. *Cell Stem Cell*. **2**: 333–344.

Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, Whitaker JW, Tian S, Hawkins RD, Leung D, et al. 2013. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*. **153**: 1134–1148.

Xu H, Baroukh C, Dannenfelser R, Chen EY, Tan CM, Kou Y, Kim YE, Lemischka IR, and Ma'ayan A. 2013. ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. *Database (Oxford)*. **2013**: bat045.

Young RA. 2011. Control of the Embryonic Stem Cell State. *Cell*. **144**: 940–954.

Yu W, Li X, Eliason S, Romero-Bustillos M, Ries RJ, Cao H, and Amendt BA. 2017. Irx1 regulates dental outer enamel epithelial and lung alveolar type II epithelial differentiation. *Developmental biology*. **429**: 44–55.

Zhang J, Cao H, Xie J, Fan C, Xie Y, He X, Liao M, Zhang S, and Wang H. 2018. The oncogene Etv5 promotes MET in somatic reprogramming and orchestrates epiblast/primitive endoderm specification during mESCs differentiation. *Cell Death & Disease*. **9**: 224.

Zhang S, Chasman D, Knaack S, and Roy S. 2019. In silico prediction of high-resolution Hi-C interaction matrices. *Nature Communications*. **10**: 5449.