MICA: A multi-omics method to predict gene regulatory networks in early human embryos

Gregorio Alanis-Lobato[1,*], Thomas E. Bartlett[2,*], Qiulin Huang[1,3*], Claire Simon[1], Afshan McCarthy[1], Kay Elder[4], Phil Snell[4], Leila Christie[4] & Kathy Niakan[1,3,5,6]

1 Human Embryo and Stem Cell Laboratory, The Francis Crick Institute, London, UK.
2 Department of Statistical Science, University College, London WC1E 7HB, UK.
3 The Centre for Trophoblast Research, Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge CB2 3EG, UK.
4 Bourn Hall Clinic, Bourn, Cambridge, CB23 2TN.
5 Wellcome Trust – Medical Research Council Stem Cell Institute, University of Cambridge, Jeffrey Cheah Biomedical Centre, Puddicombe Way, Cambridge CB2 0AW, UK
6 Epigenetics Programme, Babraham Institute, Cambridge CB22 3AT, UK
*Equal contribution
Corresponding authors: kkn21@cam.ac.uk; g.alanis.lobato@gmail.com; thomas.bartlett.10@ucl.ac.uk

Abstract

Recent advances in single cell omics have been transformative in the characterisation of challenging to study biological contexts, including when the source material is precious, such as the early human embryo. Single cell datasets are technically challenging to infer transcription factor-gene regulatory interactions, due to low read-depth leading to zero inflated data. Here we have systematically assessed the application of four different machine learning linear or non-linear gene regulatory network prediction strategies to single cell simulated and human embryo transcriptome datasets. We have also compared how the method of gene expression normalisation impacts on regulatory network predictions. Integrating chromatin accessibility together with transcript expression datasets improved the reproducibility of the predicted gene regulatory networks. We found that the application of a non-linear network prediction method based on mutual information (MI) to single cell transcriptome datasets refined with chromatin accessibility (CA) (called MICA), exhibited higher reproducibility compared to the alternative network interference methods tested. Moreover, MICA was used to make predictions about GRNs in the preimplantation human embryo, which were supported by previous findings in other developmental and stem cell contexts. Based on the gene regulatory networks predicted by MICA, we discovered co-localisation of the AP-1 transcription factor subunit proto-oncogene JUND and the TFAP2C transcription factor AP-2γ in human preimplantation embryos. Overall, our comparative analysis of gene regulatory network prediction methods defines a pipeline that can be applied to single-cell multi-omics datasets to infer interactions between transcription factor expression and target gene regulation, which can then be functionally tested.

Introduction

Following the fusion of the oocyte and sperm, the zygote undergoes a series of cell divisions until it forms a blastocyst prior to implantation into the uterus. A human blastocyst is comprised of a fluid-filled cavity and approximately 200 cells that comprise three distinct cell types: the trophectoderm (TE), which gives rise to fetal components of the placenta; the primitive endoderm (PE), which forms the yolk sac; and the pluripotent epiblast (EPI), which gives rise to the embryo proper (Blakeley et al., 2015). The specification of these three lineages represents the earliest cell-fate decisions in humans. Understanding the molecular mechanisms that regulate these decisions is important for applications including stem cell biology, regenerative medicine, and reproductive technologies (Niakan et al., 2012).

We do not yet understand how cell fate specification is regulated in the human embryo. Transcription factor (TF) and target gene regulatory interactions associated with cell fate specification in this context would be informative and has not been elucidated. Defining the gene regulatory networks (GRNs) associated with a given cell type at a distinct time in development facilitates characterisation of cell type identity and prediction of the transcriptional regulators and cis-regulatory DNA sequences that may underlie cell fate specification (Davidson and Erwin, 2006; Materna and Davidson, 2007; Peter and Davidson, 2011). Challenges to determining GRNs in early embryos include the small number of cells that contribute to the distinct cell lineages of the human embryo at these early stages of development. Moreover, while single cell omics technologies facilitate the characterisation of genome-wide gene expression and chromatin accessibility changes in human blastocysts (Blakeley et al., 2015; Li et al., 2018; Liu et al., 2019; Petropoulos et al., 2016; Yan et al., 2013), it is unclear if resolution at this level would allow for accurate predictions of GRNs contributing to cell type specific identity.

The most common computational approach to infer TF-gene regulatory interactions is to model the expression of each target gene as either a linear or non-linear combination of the expression of a set of potential regulators (e.g., the TFs in the dataset of interest) (Dobra et al., 2004; Marbach et al., 2012). The 'target gene' method uses advanced regression methods, such as sparse penalised regression or random forests, to model the expression level of a particular 'target' gene (the response), conditional on the expression levels of a set of other genes (predictors, such as TFs) (Haury et al., 2012; Huynh-Thu et al., 2010). By combining these local network model fits genome-wide, with each gene taking a turn as the response or 'target gene' (Dobra et al., 2004), the GRN is constructed. The linear regression model has previously been used in combination with chromatin accessibility to infer GRNs in human brain organoids(Fleck et al., 2022). Bayesian networks also model the expression of each gene conditional on a set of regulating genes (i.e., parents in the network) by computing how likely it is that each TF in the dataset was responsible for the expression of a certain gene (Kamimoto et al., 2020; Marbach et al., 2012; Pe'er, 2005; Tipping, 2001). The Bayesian/Bagging Ridge model in combination with chromatin accessibility has been used effectively to predict the consequence of gene perturbations and to understand cell fate transitions in the context of development and cellular reprogramming (Argelaguet et al., 2022; Kamimoto et al., 2020). Of note, when fitting Bayesian models it is usually necessary to assume global compatibility (of the local network models), which in turn typically requires that the graph contains no loops (acyclicity). This an assumption which is challenging for biological systems, where feedback loops are a persistent feature (Brandman and Meyer, 2008). In another important group of methods, TF-gene (or more generally gene-gene) interactions are ranked based on a pairwise similarity measure of their expression across samples, such as a variant of correlation or mutual information (Daub et al., 2004; Krishnaswamy et al., 2014). Again, alternatives are available that consider conditional statistics (e.g., partial correlation or partial information decomposition) (Chan et al., 2017). Finally, a range of 'black-box' machine learning methods are available that model the expression of each gene in hard to visualise ways using for

example sophisticated non-linear combinations of TF expression values (Shu et al., 2021; Wang et al., 2020).

Although GRN inference methods have been comprehensively compared and developed for whole-tissue or bulk transcriptome analysis (Cahan et al., 2014; Chai et al., 2014; Marbach et al., 2012; Morris et al., 2014; Thompson et al., 2015), the interest in cell type-specific regulatory networks has prompted the application of these methods to single cell RNA-seq (scRNA-seq) datasets (Aibar et al., 2017; Chen and Mar, 2018; Fiers et al., 2018; Iacono et al., 2019; Kang et al., 2021; Mochida et al., 2018; Nguyen et al., 2021; Pratapa et al., 2020; Stone et al., 2021). These previous studies showed that the approaches with the best overall performance belong to the regression and mutual information categories and highlighted the challenges of performing GRN inference on single cell datasets. This is attributed mainly to the heterogeneous and stochastic nature of gene expression in individual cells (Nguyen et al., 2021). Of note is that even state-of-the-art methods that have been evaluated on well-defined benchmark datasets predict a considerable number of false positive interactions (Kang et al., 2021; Marbach et al., 2012; Nguyen et al., 2021; Pratapa et al., 2020; Stone et al., 2021). Most of these false predictions originate from indirect associations, for example a path $a \rightarrow b \rightarrow c$ can result in the prediction of $a \rightarrow c$ even if there is no direct link between those nodes (i.e., a 'transitive edge' in the network). Several methods to eliminate such effects have been proposed (Barzel and Barabási, 2013; Chan et al., 2017; Feizi et al., 2013; Margolin et al., 2006; Wang et al., 2018) but they can be computationally demanding and sensitive to the choice of hyper-parameters (Feizi et al., 2013).

Given the stage of development of human preimplantation embryos and their precious nature, together with the restrictions on such research in some countries, the omics datasets from human blastocysts are very small compared to those from other biological contexts. This makes it challenging to mine these datasets using GRN inference methods, which require a sufficiently large number of cells to produce reproducible results (Kang et al., 2021; Pratapa et al., 2020). Here, we assessed whether the integration of other types of omics datasets with transcriptomic-based predictions could help to reduce indirect TF-gene relationships and thereby produce more reliable gene regulatory networks. We utilised a low number of samples in high quality omics data from early human embryos at the blastocyst stage (6-7 days post-fertilisation) (Blakeley et al., 2015; Petropoulos et al., 2016; Yan et al., 2013) to evaluate the plausibility of our predictions: this has the advantage of being a biological context with only three well-defined cell types (Blakeley et al., 2015) to evaluate the plausibility of our predictions.

The aims of this work were (i) to evaluate whether it is possible to infer reliable cell type-specific regulatory networks for each one of the cell types in the early human embryo in spite of the size of the available omics data, (ii) to determine if the integration of chromatin accessibility data with transcriptome analysis would better inform predictions of GRNs, and (iii) to predict and validate previously unidentified gene regulation in the human blastocyst. Overall, we demonstrate that the available single cell transcriptomic data was most robustly analysed by a mutual information-based inference method which had been refined with chromatin accessibility data (MICA). The resulting Analysis predicted GRNs that were consistent with the transcripts and proteins that are known to be enriched in specific lineages. This method allowed us to discover a novel putative regulatory interaction between the TFAP2C transcription factor AP-2γ and the AP-1 transcription factor subunit proto-oncogene JUND in human preimplantation embryos. Mouse embryos analysed at equivalent stages of development did not have correlated expression of AP-2γ and JUND, suggesting a network interaction that may be distinct between these species. Overall, we propose that MICA will be a useful method to make gene regulatory network predictions in other challenging-to-study biological contexts.

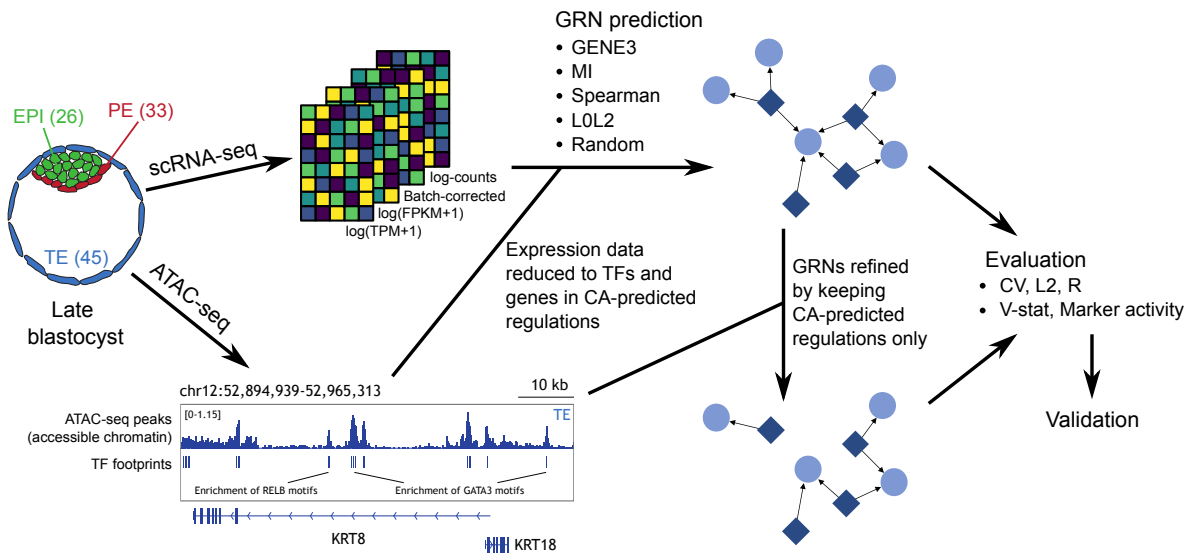


Figure 1. An overview of our regulatory network inference, evaluation, and validation framework. Footprinting analysis was applied to ATAC-seq data from human embryos at the late blastocyst stage to identify potential transcription factor (TF) binding sites and TF-gene regulations. The TFs and genes in this list of chromatin accessibility (CA)-predicted regulations were used to reduce the size of scRNA-seq data from each of the cell types in embryos at the same developmental stage (EPI: Epiblast, 26 samples; PE: Primitive Endoderm, 33 samples; TE: Trophectoderm, 45 samples). Five gene regulatory network (GRN) inference approaches were applied to the transcriptomic data normalised using four different methods for comparison (log(TPM+1), log(FPKM+1), batch-corrected and log-counts). The reproducibility and biological content of the predicted GRNs were evaluated with several metrics. GRNs refined by keeping CA-predicted regulations only were also evaluated.

Results

Assessing the conditions for optimal GRN inference on synthetic data

We initially generated synthetic gene-expression data with known ground-truth GRN structure for a variety of sample sizes (see Methods) to determine the characteristics of the transcriptomic data on which GRN inference methods perform best and, importantly, to compare these conditions with our human blastocyst scRNA-seq datasets (see Fig. 1). We then applied four different GRN prediction strategies to the simulated transcriptomes (see Supplementary Information (SI)-Section 1 and Figs. S1-S2 for more details). 1) We applied GENIE3 (Aibar et al., 2017; Huynh-Thu et al., 2010), a random forests regression approach which takes into account expression levels of the regulated or target gene and is used for regulatory network inference in the SCENIC GRN prediction pipeline (Aibar et al., 2017; Kang et al., 2021; Marbach et al., 2012; Nguyen et al., 2021; Pratapa et al., 2020; Stone et al., 2021). 2) We compared this to the L0L2 sparse penalised regression model (Hazimeh and Mazumder, 2018; Tibshirani, 1996) which minimizes the total size of the linear model coefficients so that a minimal set of regulatory TFs receive the largest linear model coefficients. 3) We also compared with the correlation coefficient as a reference: we used Spearman's rank correlation coefficient (Fieller et al., 1957) as we expect non-linear associations in the data. 4) Lastly, we applied a non-linear alternative method based on mutual information (MI) that we speculated may be more appropriate for non-linear response functions in single-cell data (Faith et al., 2007; Krishnaswamy et al., 2014) (Fig. S2). For the MI method, we used the empirical distribution of the MI values for each gene (Faith et al., 2007) (see Methods for more detail). We also varied the number of transcriptional regulators of each target gene in the synthetic data, to study the impact of this parameter on their performance. Further details of the analysis methods and step-by-step codes are also available at: https://github.com/galanisl/early_hs_embryo_GRNs

Two main conclusions emerged from these simulations. First, using synthetic datasets, the performance of all methods increases from the sample size of n = 10 to n = 1000, as assessed statistically by the area under the receiver operating characteristic curve (AUROC) (Fig. 2). However, we noted that the prediction accuracy of the GRNs only marginally improves as the number of samples surpasses n = 100 (Fig. 2). Therefore, while sample size is important, increasing this beyond a threshold sample size does not further improve the predicted transcriptional regulations. We further note that low sequencing depth (such as in many single-cell studies) may increase this optimal value of n. Second, limiting the number of potential transcriptional regulators of a target gene, positively impacts the ability of a chosen inference method to recover the ground-truth GRN (Fig. 2). From our simulations, 50 or fewer TFs yield the highest AUROC values for n < 100 (Fig. 2). This would be a reasonable prediction for a given gene in a specific cell type based on our analysis of low input chromatin accessibility (CA) data from the human blastocyst (Liu et al., 2019), where the median number of TF motifs per gene is 35 for the inner cell mass (ICM) and 40 for the TE (Fig. S3).

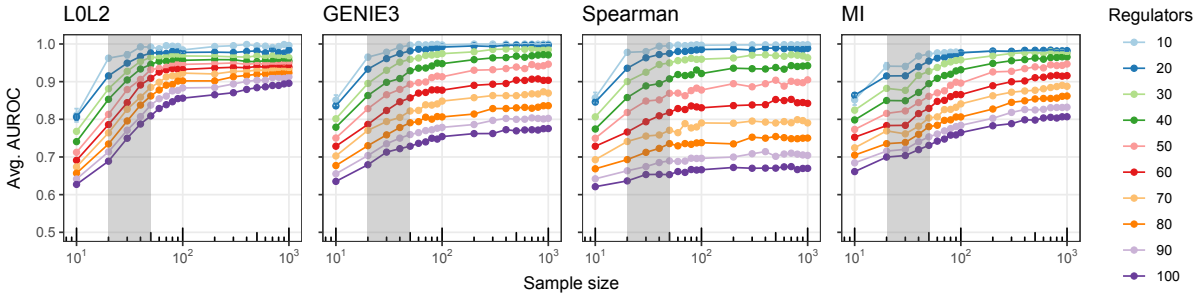


Figure 2. GRN inference on simulated scRNA-seq data. Ability of different GRN inference methods (L0L2, GENIE3, Spearman correlation and MI) to recover the ground-truth GRN structure from simulated gene expression data as measured with the average area under the receiver operating characteristic curve (AUROC) from 100 simulations. The AUROC is shown as a function of different sample sizes (n=10 to 1000) and the number of potential regulators of each gene in the simulated datasets is also varied from 10 to 100. The range of sizes of the three human blastocyst datasets that we analysed is highlighted in grey as a reference. Error bars correspond to standard errors of the mean.

Based on these results and given the size of our human blastocyst datasets (EPI: 26 samples, PE: 33 samples, TE: 45 samples with 1,366 TFs among 25,098 genes (Blakeley et al., 2015; Petropoulos et al., 2016; Yan et al., 2013)), we reasoned that TF-gene interactions predicted from experimentally collected gene expression data could be refined with complementary, context-specific epigenomic datasets. Specifically, we used low-input CA analysis of human blastocyst TE cells and the inner cell mass, comprised of EPI and PE cells, to refine the GRNs of the respective cell types (Liu et al., 2019). Peak calling and annotation were performed with nf-core/atacseq to identify regions of open chromatin (Ewels et al., 2020). TF motif enrichment analysis in these open regions of open chromatin were identified using rgt-hint with TF binding models from HOCOMOCO and JASPAR (Li et al., 2019). Finally, the open chromatin regions enriched for TF motifs along with the predicted downstream target genes were determined based on the nearest transcriptional start sites (most distances ranging from -5kbp to 10kbp). This narrowed down the number of genes (to 12,780 for EPI and PE, 12,981 for TE) and TFs (514) in the datasets, thus bringing our sample sizes and potential TFs per gene to the ranges suggested by our simulations.

Statistical evaluation of inferred human blastocyst GRNs

We next integrated the single-cell transcriptome data from human blastocysts generated using the SMART-seq2 library preparation protocol (Blakeley et al., 2015; Petropoulos et al., 2016; Picelli et al., 2014; Yan et al., 2013) where fewer cells are sequenced but better transcript coverage and sequencing depth is obtained (~15,000 genes and ~7 million reads per cell) compared to more conventional scRNA-seq methods, such as 10X (~3,000 genes and

~50,000 reads per cell) (Wang et al., 2021). Due to the lack of genome-scale experimentally derived GRNs in the human embryo context, we evaluated the robustness of the predictions made by four inference approaches (L0L2 regression, GENIE3, Spearman correlation and MI) using three different strategies (see Methods and SI-Section 1). Since we did not find reports about the impact of gene expression normalisation choice on GRN predictions, we also assessed this parameter and considered the application of GRN predictors to log(TPM+1), log(FPKM+1), log-count and batch-corrected expression data.

For GENIE3, transcription factors are ranked according to the degree of variability in their expression level and how the expression of the putative regulator correlates with a target gene. This ranking is then used to construct the co-expression network for all genes and transcription factors, by thresholding the algorithm's variance reduction score at the tenth percentile of its empirical distribution. The GRN is then inferred as the intersection of the edges in this co-expression network with the edges in the network of all possible gene regulations derived from the chromatin accessibility / DNA binding motif data. This network of all possible gene regulations is defined as all network edges from regulating TF to regulated gene, where an edge represents a DNA binding motif for the TF in regulatory DNA within open chromatin in the regulated gene.

For L0L2 regression, the model automatically chooses a ranked subset of TFs from those predicted as regulators of the target gene in the chromatin accessibility / DNA binding motif data inferences. In this way, each target gene takes its turn for a model to be fitted around that target gene. After the model has been fitted to every target gene, the global gene regulatory network can be constructed by combining the local networks fitted around each target gene.

With the Spearman correlation coefficient, a weighted co-expression network is inferred as the absolute value of the correlation coefficient. The gene regulatory network is then inferred as the intersection of the edges in this co-expression network with the edges in the network of all potential gene regulations derived from the chromatin accessibility / DNA binding motif data.

For MI-based inference, the co-expression network is estimated first, by estimating p-values from a model likelihood based on the empirical distributions of the calculated MI values (Daub et al., 2004; Faith et al., 2007). Again, the gene regulatory network is inferred as the intersection of the edges in this co-expression network with the edges in the network of all possible gene regulations derived from the chromatin accessibility / DNA binding motif data.

We first calculated a reproducibility score R for each putative regulatory interaction following the application of these network inference methods to human blastocyst data (see Methods) and investigated the distribution of the reproducibility values for the top-100,000 predicted edges (Fig. S4). Fig. 3A reports the difference between the median of this distribution and the median of the distribution produced by a predictor that generates a random ranking of all possible TF-gene interactions (ΔR). We found that the most reproducible regulatory interactions were inferred by MI followed by a filtering process in which only TF-gene associations supported by CA data were considered in the final network (MICA) (Fig. 3A). As predicted by our simulations, the size of each cell type-specific dataset had a clear impact on GRN inference with methods such as GENIE3 producing more reproducible interactions in the TE. Interestingly, most inference methods produced better results with gene expression values following log(TPM+1) or log(FPKM+1) normalisation (Fig. 3A).

Noting that the reproducibility score R quantifies robustness at the edge level, we also assessed robustness at the level of network structures, features, or subnetworks. To do so, we randomly split each dataset (EPI, PE and TE) into two groups with the same number of samples, then applied the inference methods to each one (with or without CA refinement), and finally benchmarked them either focussing on the top regulatory interactions (early

recognition) or on all regulatory interactions (see Methods). We repeated this 10 times for 10 random splits of the data. For the first comparison, we identified the top 1% (by network score) of all possible regulatory interactions from each group, and then quantified the overlap of these top interactions between the groups using the area under the precision-recall curve (AUPRC) (Fig. S5). By comparing the differences between the median of ten resulting AUPRCs for each method to the random predictions described above, we found that MICA outperformed the other approaches and gave the highest difference in AUPRC when applied to log(TPM+1)- and log(FPKM+1)-normalised expression values (Fig. 3B). For the second comparison, we divided the data into two portions and then calculated a normalised L2 loss between the network scores over the whole network, rather than just the top 1% of interactions (Fig. S6). We found that the inverse value of this metric also confirms the robustness of MICA (Fig. 3C).

Overall, our statistical evaluation strategies showed that the most robust GRN inference method to analyse the limited number of human blastocyst cells was MICA. Importantly, if transcriptome-based predictions are not refined with CA data, most methods perform just slightly better than random (Figs. S4 and S5), underscoring the importance of integrating multi-omics analysis in inference models. It was also important to assess the impact of gene expression normalisation on GRN prediction because we observed apparent effects in our benchmarks depending on the normalisation method used, with log(TPM+1) or log(FPKM+1) being the most suitable gene expression units in this context.

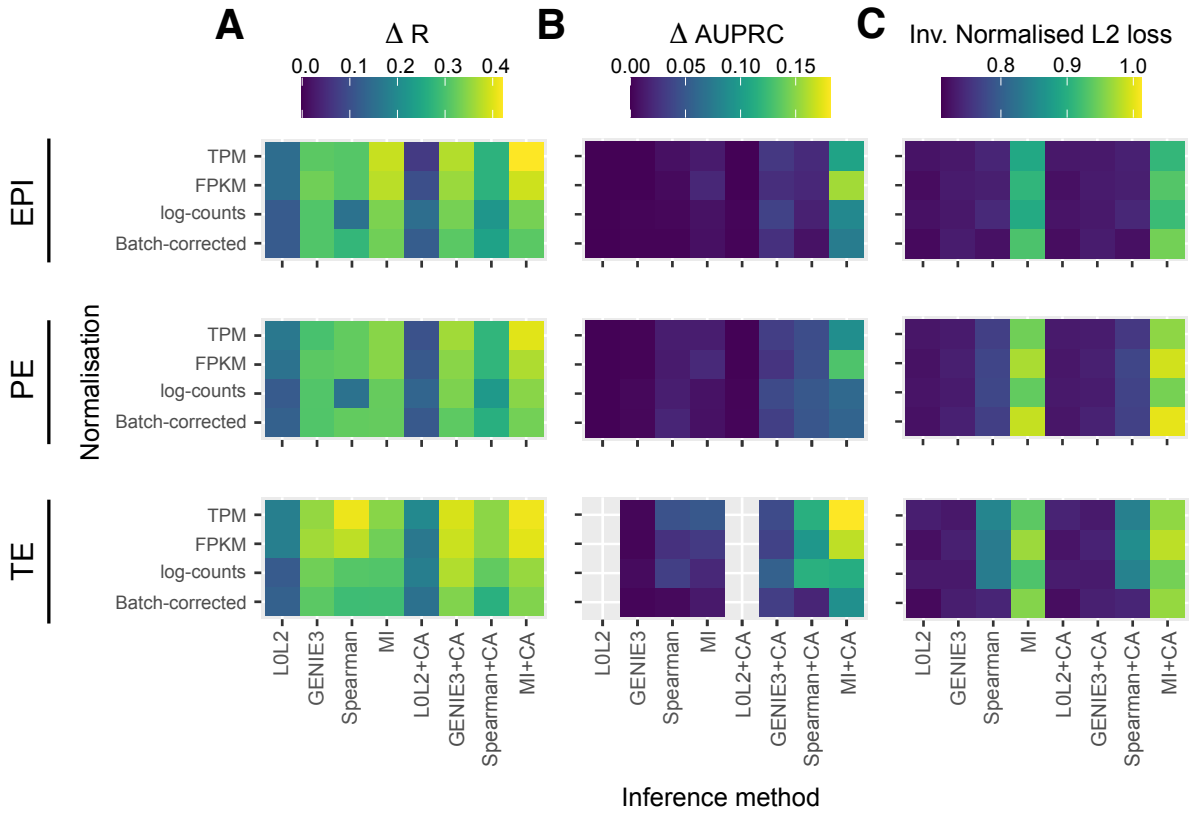


Figure 3. Statistical evaluation of inferred human blastocyst GRNs. Robustness of the GRNs predicted by four inference methods with or without chromatin accessibility (CA) refinement was evaluated with three different metrics. **(A)** ΔR measures the difference between the median edge-level reproducibility for a GRN inference method and a random predictor. **(B)** ΔAUPRC quantifies the extent to which the interactions inferred from half of a dataset coincide with the top-1% interactions inferred from the second half in a 2-fold cross-validation experiment repeated ten times, also with respect to a random predictor. **(C)** Finally, the inverse of the normalised L2 loss between the network scores from the two folds was also computed. L0L2 did not converge when applied to the TE dataset.

Association of inferred GRNs to human blastocyst cell lineages

We next evaluated the inferred GRNs to determine if they could recapitulate interactions of molecular markers of the three cell types that in the human blastocyst (Fig. S7). First, we computed the overlap between the GRN edges predicted by each inference method for each blastocyst cell type to identify interactions associated with the EPI, PE and TE, as well as the interactions common to the three cell types (see Methods). We then used the out-degree of NANOG, GATA4 and CDX2, TFs which respectively mark the EPI, PE and TE (Fig. 4A), as a proxy for their activity in each one of the four networks. Our prediction is that these TFs should be actively regulating genes in the cell type-specific networks and participate in only a few interactions in the common GRN, based on their known expression pattern in the blastocyst and function in other mammalian contexts such as the mouse (Arceci et al., 1993; Chambers et al., 2003; Dietrich and Hiiragi, 2007; Mitsui et al., 2003; Strumpf et al., 2005). The common network did not show marker activity in the predicted GRNs (Fig. 4B and Figs. S8-S11), which is expected given that the expression of these markers is known to be mutually exclusive at this stage. The expected pattern for NANOG (active in the EPI and inactive in the PE and TE, Fig. 4A), was observed in the GRNs with TF-gene interactions supported by CA data and inferred with GENIE3, Spearman correlation and MI when applied to batch-corrected, FPKM and TPM data with some instances of the non-refined L0L2, GENIE3 and Spearman GRNs also matching the expected pattern (Fig. 4B). In the PE, the expected activity pattern for GATA4 (active in the PE and inactive in EPI and TE, Fig. 4A) and lack of detectable networks for NANOG and CDX2, were predicted by applying GENIE3+CA, Spearman correlation+CA and MICA to the FPKM and TPM datasets (Fig. 4B). Finally, the TE-expected pattern for CDX2 (active in the TE and inactive in the EPI and PE) and lack of detectable network for NANOG and GATA4, Fig. 4A) was only observed when using Spearman correlation or MICA on log-counts and batch-corrected data, MICA or MICA+CA on FPKM data and MICA+CA on TPM data (Fig. 4B). L0L2, with or without CA refinement, consistently performed poorly at predicting GRNs compared to the other methods used. Log count normalisation failed to recover GRNs for these TFs that were associated with the specific cell type in which they are known to be exclusively expressed.

We next manually curated a list of gene sets representing the most relevant pathways and biological processes in the EPI (e.g., regulation of pluripotency) and the TE (e.g., embryonic placenta development) to perform a gene set enrichment analysis with the genes involved in the 500 regulatory interactions with the best prediction scores in the EPI and TE GRNs of the 25 most active TFs (Tables S2, S3, Methods). We restricted this analysis to the EPI and TE to compare between these distinct cell types and because the CA data we used were generated for the TE and ICM only. Next, we computed a validation score *V* as the number of relevant gene sets that were enriched at a significance level of 10% ($p<0.1$) in the EPI and TE GRNs. We note that a low significance level was set because the resulting inferences are aggregated, mitigating the effect of false positive hits for individual gene groups. To facilitate the comparison between *V* scores across GRNs from different inference and normalisation methods, we normalised this score to the maximum score attained at the cell type level (Fig. 4C). We found that the CA-refined data agree better with the cell type-specific gene sets for both the top-predicted EPI and TE interactions, i.e. the interactions with the highest prediction scores (Fig. 4B). The GRNs inferred by Spearman+CA produced the highest *V* scores in the EPI and TE, and these were not impacted by the normalisation method employed. The second-best inference method was MICA, with consistent *V* scores across normalisation methods. We note that both MICA and Spearman+CA are non-linear methods which do not involve regression, indicating that advanced regression methods may not be the most effective choice for biological discovery with these restricted sample-sizes.

Taken together, both biological evaluation metrics that we considered confirmed the importance of refining single-cell transcriptome-based GRN inferences with CA data and underscored the robustness of MICA in predicting GRNs from the analysed human blastocyst

datasets. Based on these results, we decided to focus on the GRNs predicted by this method for the analyses presented in the following sections.

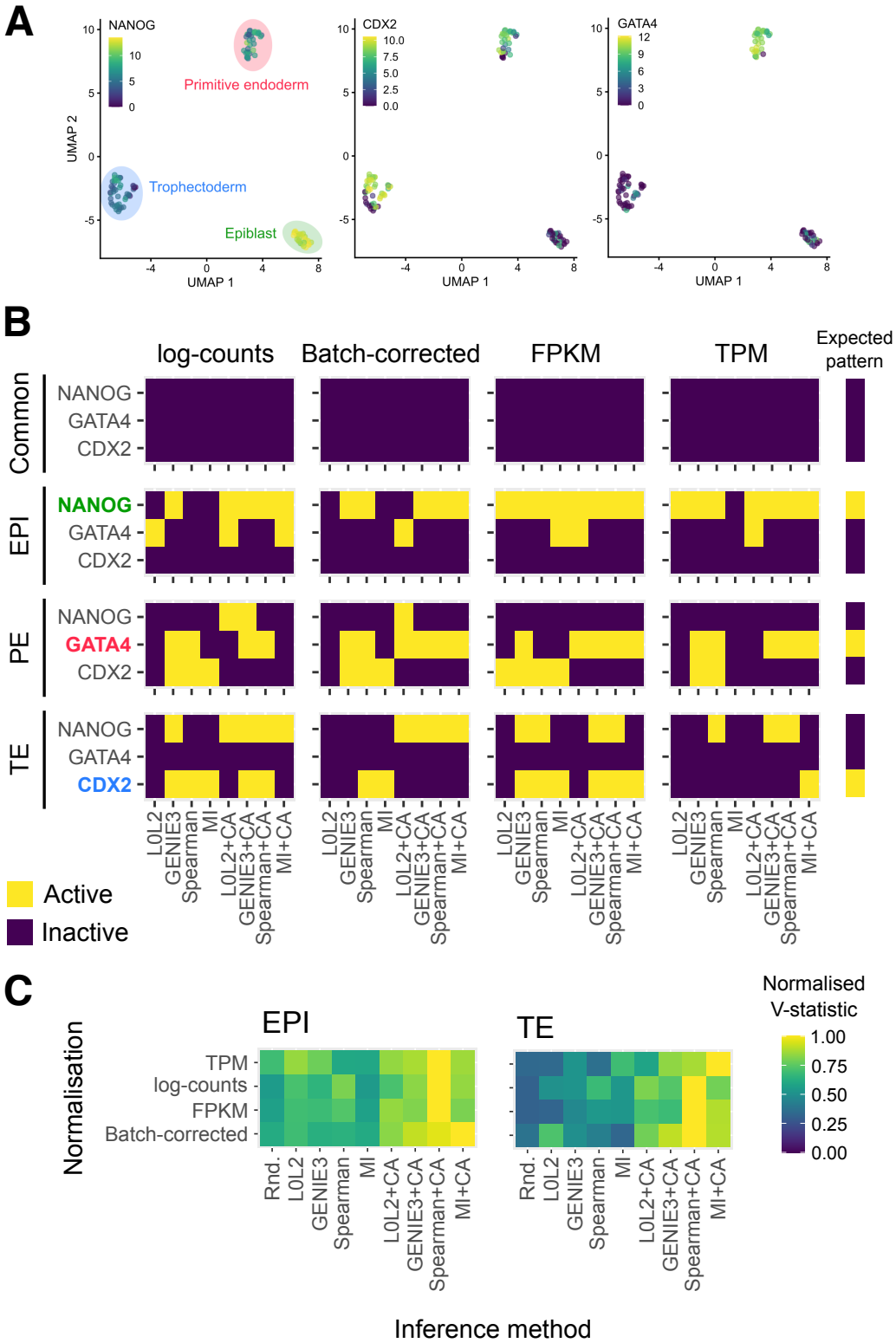


Figure 4. Biological evaluation of inferred human blastocyst GRNs. (A) Two-dimensional UMAP representation of the scRNA-seq data from each of the three cell types in the early human at the blastocyst stage. The expression of marker genes for each cell type is also represented in UMAP space: NANOG is used as an Epiblast marker, CDX2 as a Trophectoderm marker and GATA4 as a Primitive endoderm marker. **(B)** Marker activity in the cell type-specific GRNs (EPI: Epiblast, PE: Primitive endoderm; TE:

Trophectoderm) and the network common to the three cell types in the human blastocyst. The expected pattern, i.e. the marker that was expected to be active in each GRN, is shown for reference. **(C)** Normalised V-statistic across datasets and inference methods. V corresponds to the number of gene sets that were enriched at a significance level of 10% in a gene set enrichment analysis performed with the genes involved in the top-500 regulations by the 25 most-connected TFs in the EPI or TE GRNs.

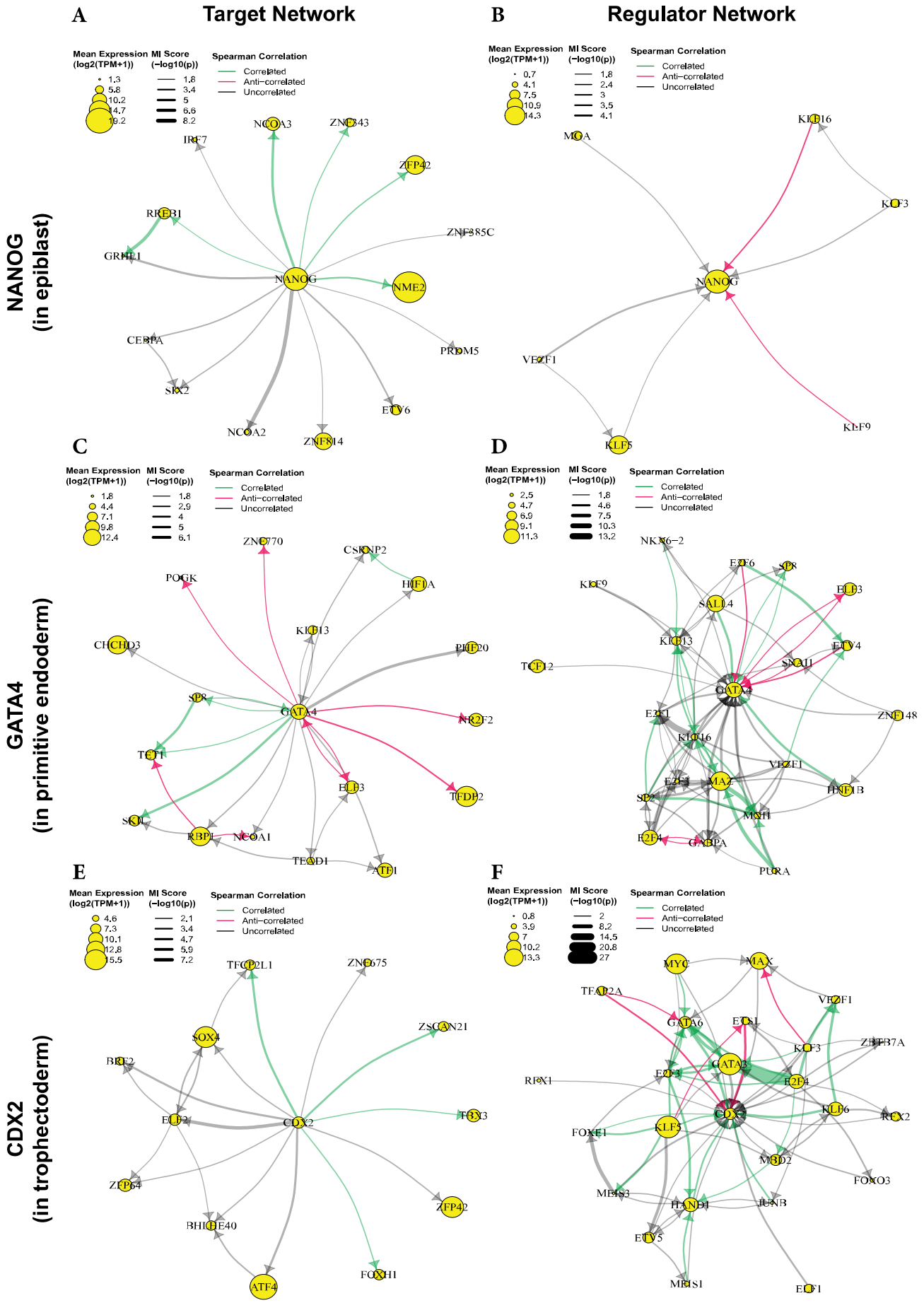
Predicted TF networks for NANOG, GATA4 and CDX2

Using MICA, we constructed GRNs for TFs expressed in the EPI, PE and TE (FigShare: doi.org/10.6084/m9.figshare.21968813). For visualization, the predicted GRN for each of the TFs is separated into target and regulator TF networks. Target networks contain a maximum of 25 top potential target TFs of the hub (or central) TF, whereas regulator networks include a maximum of 25 top TFs that potentially regulate the hub. The average expression of the network members across samples of the cell type of interest is represented by the size of the node. MI scores are represented by the thickness of the edges in the network and edge colour highlights whether the expression of the source and target node is correlated, anti-correlated or uncorrelated across samples. To define these three categories, we used the Spearman's rank correlation coefficient between the expression levels of the source and target nodes across samples of the cell type of interest. Correlated or anti-correlated node pairs correspond to positive or negative Spearman's rank coefficient with p-value smaller than 0.1, whereas node pairs having p-value equals or larger than 0.1 were defined as uncorrelated.

Among the top NANOG targets, TFs RREB1, NCAO3, ZNF343, ZFP42, and NME2 are predicted to be positively regulated by NANOG (Fig. 5A). ZFP42 is a pluripotency marker encoding the REX1 protein and has been shown to be a direct target of NANOG in mouse pluripotent stem cells (Shi et al., 2006). ZNF343 is a less well-characterized TF, but multiple NANOG ChIP-seq datasets in both naïve and primed human embryonic stem cells (ESCs) showed high binding score (MACS2 score > 1000) in the proximal region of the transcription start site (TSS) of ZNF343 (Fig. S13; Barakat et al., 2018; Chovanec et al., 2021; Lyu et al., 2018), which suggests direct regulation of ZNF343 by NANOG. Interestingly, NME2 was previously predicted to be a regulator instead of a target of NANOG in mouse pluripotent stem cells using the TENET GRN inference method (Kim et al., 2021). This inconsistency may be due to the lack of chromatin accessibility data for robust directionality inference in the TENET method, or a species difference. Indeed, some mouse naïve pluripotency regulators, such as ESRRB, which is also a direct target of NANOG in mouse pluripotent stem cells (Festuccia et al., 2012), was not expressed in the human EPI (Fig. S14). Putative regulators of NANOG predicted from our MICA network analysis contain multiple KLF factors, including KLF3, KLF5, KLF9 and KLF16 (Fig. 5B). Spearman correlation analysis suggest that KLF9 and KLF16 potentially down-regulate NANOG expression, based on their anti-correlated expression, whereas KLF3 and KLF5 are not significantly correlated with the NANOG expression profile.

In the PE, GATA4 was predicted to enhance the expression of SP8, TET1, and SKIL while repressing the expression of ELF3, TFDP2 POGK, ZNF770, and NR2F2 (Fig. 5C). It is also predicted to be a target of HNF1B and SALL4 and repressed by ETV4 and E2F6 (Fig. 5D). Most of these interactions have not been experimentally validated or inferred. However, NR2F2 was previously identified as a maturation marker of polar TE (Meistermann et al., 2021). The repression of NR2F2 by GATA4 predicted in the PE suggests the role of GATA4 in maintaining the PE cell identity by inhibiting the polar TE program. Furthermore, it has been shown that SALL4 is required for mouse PE-derived extra-embryonic endoderm cell derivation and knockout of SALL4 in these cells cause down-regulation of GATA4 (Lim et al., 2008). In addition, multiple endoderm genes such as GATA4, GATA6, and SOX17 were shown to be SALL4-bound genes by ChIP-seq (Lim et al., 2008). These findings are consistent with the inferred GATA4 regulatory network.

Interestingly, TBX3 is a target of CDX2, and its role has been implicated in trophoblast cell differentiation (Lv et al., 2019). CDX2 is predicted to be regulated by both KLF5 and KLF6, which are also markers of the trophectoderm lineage (Fig. 5F). KLF5 is necessary for the trophectoderm formation in the mouse pre-implantation embryo and is required for CDX2 expression in mouse ESCs (Ema et al., 2008; Lin et al., 2010). Interestingly, further analysis of the CDX2 target-to-regulator interactions showed that TBX3 and FOXH1 are also potential positive regulators of KLF5 and KLF6. Positive feedback loops are formed between CDX2, TBX3 and KLF5, as well as CDX2, FOXH1 and KLF6 (Fig. S15). Other key developmental regulators including HAND1, GATA3 and GATA6, are also predicted to regulate CDX2 expression. This is consistent with the timing of GATA3 protein expression preceding that of CDX2 (Gerri et al., 2020). Further investigation will be needed to understand the differences between the CDX2-high and CDX2-low TE cells and how the positive feedback loops formed within the CDX2 network enhance and stabilize CDX2 expression in the CDX2-high population.



465 See figure legend on the next page.

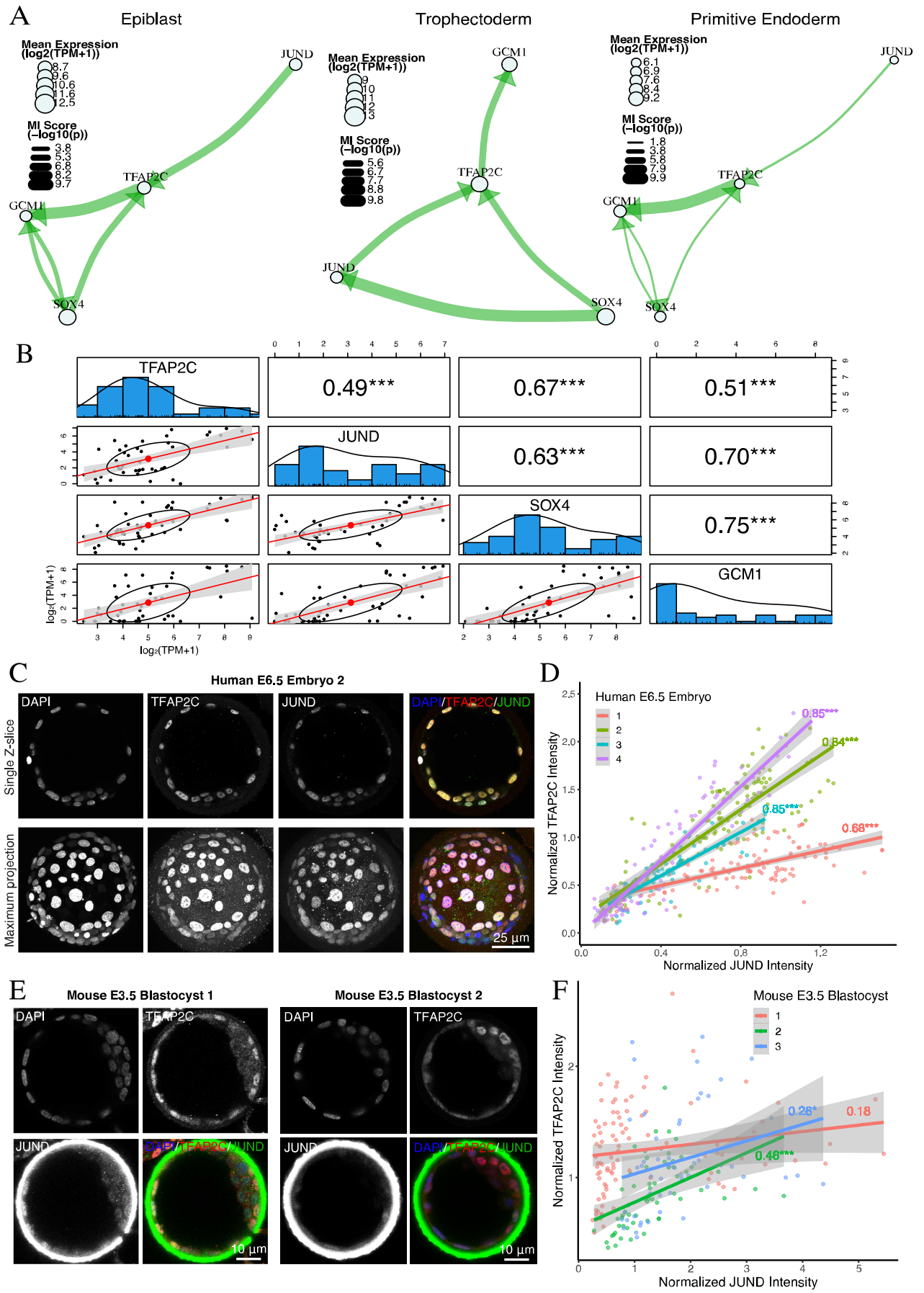
Figure 5. TF-to-TF networks of NANOG, GATA4 and CDX2 inferred by MICA. (A) Network of TFs targeted by NANOG. (B) Network of TFs that regulate NANOG. (C) Network of TFs targeted by GATA4. (D) Network of TFs that regulate GATA4. (E) Network TFs targeted by CDX2. (F) Network of TFs that regulate CDX2.

Maintenance of TFAP2C expression by JUND in all lineages of human blastocyst

We next sought to determine if the MICA network modelling could predict novel interactions or associations between TFs in early human embryo development. We focused on TFAP2C as an example for the network comparison. TFAP2C is a molecular marker that is initially expressed in all the cells at the morula stage and later specifically restricted to the TE at the blastocyst stage in the mouse embryo and not expressed in other lineages (Cao et al., 2015; Gerri et al., 2020). However, in human embryos, TFAP2C is expressed in the TE, EPI and PE at the blastocyst stage (Blakeley et al., 2015) and has been shown to maintain pluripotency in naïve human embryonic stem cells by regulating OCT4 expression (Chen et al., 2018; Pastor et al., 2018). By performing comparisons of TFAP2C networks in human EPI, TE, and PE cells, we found conserved interactions consisting of TFAP2C, JUND, SOX4, and GCM1 (Fig. 6A). In addition, all four factors showed significant positive correlations in their expression (Figs. 6B, S16, S17). The MICA network predicts that JUND and SOX4 regulate TFAP2C, while TFAP2C targets GCM1 in all three lineages. In the EPI and PE, SOX4 and GCM1 formed a feedback loop, resulting in a larger positive feedback loop comprised of TFAP2C, GCM1 and SOX4. Interestingly, in the TE, the correlation between GCM1 and SOX4 is absent, and SOX4 is predicted to target JUND, which in turn may regulate TFAP2C. Our network predicts that interactions between these four TFs maintains TFAP2C expression in all three lineages.

We subsequently performed immunostaining of TFAP2C and JUND initially on naïve human embryonic stem cells, thought to represent the blastocyst EPI (naïve hESCs) and compared the expression pattern to primed hESCs, which are thought to reflect later stages of EPI development (Bredenkamp et al., 2019; Vallier et al., 2009). TFAP2C and JUND showed a correlated expression pattern in the naïve hESCs (Fig. S18A). Although TFAP2C is expressed at a much lower level in the primed H9 cells, nuclei staining could be observed in the cells with JUND expression (Fig. S18A). To determine if TFAP2C and JUND protein expression is positively correlated, we performed immunofluorescence analysis of human preimplantation blastocysts approximately 6.5 days after fertilisation. This showed that TFAP2C and JUND were co-expressed in human blastocysts (Fig. 6C, S18B; n=4). To further characterize the protein expression correlation, we performed 3D nuclear segmentation and calculated the Pearson correlation between TFAP2C and JUND based on the DAPI-normalized protein intensity (Fig. 6D). TFAP2C intensity showed a linear increase with the increase of JUND intensity in all analysed human embryos, and their intensities are highly correlated (0.68-0.85 Pearson correlation; $p < 0.001$; Fig. 6D).

To compare TFAP2C and JUND correlation between species, we performed immunofluorescence analysis and 3D nuclei segmentation on the mouse blastocysts approximately 3.5 days after fertilisation, which is a similar stage to the human embryos analysed. As expected from previous studies, TFAP2C showed specific localisation to the TE, whereas JUND was expressed in both TE and EPI cells (Fig. 6E). Moreover, Pearson correlation analysis showed insignificant or weaker correlations between TFAP2C and JUND intensity in mouse blastocysts (Fig. 6F). Overall, these results indicate a positive correlation between TFAP2C and JUND in human embryos that warrants further investigation. Moreover, the difference between the correlation of these factors in mouse embryos versus human embryos is intriguing. Overall, this demonstrates the informativeness of the MICA network analysis to predict correlations and possible regulatory relationships between transcription factors that can be experimentally tested.



See figure legend on the next page.

Figure 6. Conserved TFAP2C interactions across EPI, TE, and PE in the human blastocyst. (A) Network composed of TFAP2C, JUND, SOX4, and GCM1 in EPI (left), TE (middle), and PE (right). (B) Correlation plots between TFAP2C, JUND, SOX4, and GCM1 RNA expression in trophectoderm. Bottom left: log transformed RNA expression of genes in single cells. Diagonal: Distribution of log transformed RNA expression for TFAP2C, JUND, SOX4 and GCM1. Top right: Spearman correlation between TF pairs. (***: $p < 0.001$). (C) Immunofluorescence staining of TFAP2C and JUND in E6.5 human blastocysts. (D) Correlation plot of TFAP2C and JUND protein intensity in nuclei of human E6.5 blastocysts. Numbers on the plot are Pearson correlation value. (***: $p < 0.001$). (E) Immunofluorescence staining of TFAP2C and JUND in E3.5 mouse blastocysts. (F) Correlation plot of TFAP2C and JUND protein intensity in nuclei of E3.5 mouse blastocysts. Numbers on the plot are Pearson correlation values. (*: $p < 0.05$; ***: $p < 0.001$).

Discussion

The relationship between genome-wide transcriptomic and epigenomic changes and cell fate specification in human embryogenesis is unclear. Studies of human pre-implantation development rely on the donation of surplus embryos derived from assisted reproduction technologies, and the use of such embryos for research is tightly regulated and subject to significant limitations, such as a lack of ability to conduct such research in some jurisdictions (Niakan et al., 2012). In addition, the collection of single cells from such precious embryos is technically challenging and requires specialist expertise and micropipettes to disaggregate microscopic embryos. Therefore, currently available -omics datasets from human blastocysts comprise only a few tens of samples per cell type and are therefore very limited. This contrasts with single-cell analyses in other cellular and developmental contexts that are based on tens of thousands of samples (or more) per cell type (Zheng et al., 2017).

Sample size is one of the most important considerations when selecting or designing statistical methodology, for example to infer networks of regulations of TFs and their target genes. Hence, GRN inference in -omics data from human pre-implantation embryos presents unique statistical challenges. In particular, methodology that can leverage information about gene regulations from small sample-sizes is required for this context. On the other hand, the lack of heterogeneity in early human embryos compared to adult tissue makes this a good context in which to assess GRN inference methodologies, because there is less unmeasured variability arising from environmental factors. To assess which GRN inference method performs the best in this challenging-to-study context, we have systematically compared several popular methodologies. Furthermore, we have tested how incorporation of complementary epigenomic data from ATAC-seq improves GRN inferences. Consistent with other contexts (Argelaguet et al., 2022; Kamimoto et al., 2020), we found that incorporating chromatin accessibility/TF motif analysis together with transcriptional inferences improves the accuracy of GRN inference, by first narrowing down the choice of TF targets from which to infer the mRNA transcript co-expression network, a principle that is likely to be especially important with small sample sizes.

Here we have shown that incorporating complementary epigenomic data with transcriptomic data improves the reproducibility of inferred GRNs. Furthermore, it has enabled us to make predictions about GRNs operational in early human embryos that are consistent with an understanding of the function and association of the regulators in other developmental and stem cell contexts. We suggest that for network inference using advanced regression methods, it may be preferable to pre-filter using epigenomic data to first narrow down the TF targets to choose from in scRNA-seq datasets, because this gives the network inference algorithm an easier time of selecting the regulators from transcriptome data, which is consistent with the experience of others using alternative GRN prediction methods (Fleck et al., 2022; Kamimoto et al., 2020; Kartha et al., 2022) . Alternative methods incorporate epigenomic data after inference of the mRNA transcript co-expression network (González-Blas et al., 2022) and this has been applied successfully to infer GRNs in the fly brain

(Janssens et al., 2022). It will be interesting to determine how the type and sequential order of incorporating multi-omics datasets impacts on GRN predictions.

We also note some limitations on the interpretation of the GRNs predicted. For example, one ATAC-seq peak could cover multiple transcription factor binding sites in a region. In such cases, we include all TFs with motifs mapping to this region as potential regulators of the gene with the closest TSS (transcriptional start site) to the region, for subsequent transcript co-expression network modelling. However, it will be interesting to determine if this can be refined, by applying the recently developed chromVAR-Multiome method to human blastocysts to generate an *in silico* ChIP-seq library for this context (Argelaguet et al., 2022). Moreover, the ATAC-seq data applied in this study analysed the inner cell mass data in bulk, without distinguishing between EPI and PE cells. Therefore, cell-type specific chromatin accessibility could not be considered, and specific interactions may have been missed due to the heterogeneity of the data. In addition, predictions based on the nearest TSS will miss long-range enhancers that are known to be important for gene regulation (Schoenfelder and Fraser, 2019). In the future, integration of single-cell or nuclei matched transcriptome and ATAC-seq chromatin accessibility data, like recent studies in the mouse (Argelaguet et al., 2022), would be preferable to apply in the human blastocyst context.

We also note that so far, we have separately modelled GRN structure to specific pre-implantation embryonic cell types constrained to a single developmental time-point. In the future, we seek to model dynamic GRN structure in transcriptionally distinct human blastocyst lineages. Moreover, we seek to integrate CUT&RUN or CUT&TAG TF-DNA binding analysis (Kaya-Okur et al., 2019; Meers et al., 2019; Skene and Henikoff, 2017) for key putative developmental regulators, such as GATA3, to further narrow down experimentally validated occupancy from the ATAC-seq predictions we used in this study, similarly to a recent application in mouse blastocysts (Hainer et al., 2019; Hayashi and Inoue, 2023), though we note that in this context TF occupancy studies will be restricted to a few TFs with good quality antibodies.

In summary, the MICA network analysis pipeline we developed is a tool that can be applied to challenging-to-study developmental contexts with limited sample size, such as the human blastocyst, to make predictions about TF interactions that can be experimentally tested in the future. As more datasets become available, we anticipate that the networks predicted will be further refined.

Methods

Ethics statement

All human embryo experiments followed all relevant institutional and national guidelines and regulations.

This study was approved by the UK Human Fertilisation and Embryology Authority (HFEA): research licence number R0162, and the Health Research Authority's Research Ethics Committee (Cambridge Central reference number 19/EE/0297).

The process of licence approval entailed independent peer review along with consideration by the HFEA Licence and Executive Committees. Our research is compliant with the HFEA code of practice and has undergone inspections by the HFEA since the licence was granted. Research donors were recruited from patients at Bourn Hall Clinic, Homerton University Hospital, The Bridge Centre and IVF Hammersmith.

Informed consent was obtained from all couples that donated surplus embryos following IVF treatment. Before giving consent, people donating embryos were provided with all of the necessary information about the research project, an opportunity to receive counselling and the conditions that apply within the licence and the HFEA code of practice. Donors were informed that, in the experiments, embryo development would be stopped before 14 days post-fertilization, and that subsequent biochemical and genetic studies would be performed. Informed consent was also obtained from donors for all the results of these studies to be published in scientific journals. No financial inducements were offered for donation. Consent was not obtained to perform genetic tests on patients and no such tests were performed. The patient information sheets and consent document provided to patients are publicly available (https://www.crick.ac.uk/research/a-z-researchers/researchers-k-o/kathy-niakan/hfea-licence/). Donated embryos surplus to the IVF treatment of the patient were cryopreserved and were transferred to the Francis Crick Institute where they were thawed and used in the research project.

ATAC-seq data processing and analysis

Chromatin accessibility profiles from the inner cell mass (ICM) and TE were obtained from the data produced by Liu *et al.* with the LiCAT-seq protocol (Liu et al., 2019). Alignment to the reference genome (GRCh38), peak calling and annotation were performed with nf-core/atacseq v1.1.0 (Ewels et al., 2020). Then, we carried out footprinting followed by TF motif enrichment analysis in the regions of open chromatin using rgt-hint v0.13.0 with TF binding models from HOCOMOCO and JASPAR (Li et al., 2019). Finally, we associated the TFs that exhibited over-represented motifs in these regions with their closest transcription starting sites (most distances ranging from -5kbp to 10kbp).

scRNA-seq data processing and analysis

We integrated scRNA-seq data from three different studies (Blakeley et al., 2015; Petropoulos et al., 2016; Yan et al., 2013) focusing on the three cell types present at the late blastocyst stage (Fig. 1, Fig. S7A). Cell type annotations were taken from the work by Stirparo *et al*. Alignment to the reference genome (GRCh38) and calculation of gene counts and TPM-normalised counts were performed on each dataset separately with nf-core/rnaseq v1.4.2 (Ewels et al., 2020). The resulting gene expression matrices were integrated and normalised (log-counts and batch-corrected counts) using Bioconductor tools (Amezquita et al., 2019). The final set of samples was manually curated based on the UMAP representation of the batch-corrected data (see Fig. S7 and Table S1). The list of 25,098 genes in the expression matrices was reduced to the unique set of TFs and transcription starting sites (TSS) derived from the motif enrichment analysis applied to the ATAC-seq data. We used the ICM TFs and TSS for the EPI and PE matrices (12,780 genes) and the TE TFs and TSS for the TE matrices (12,981 genes).

Network inference methods

For network inference, we compared the best performing strategy in the DREAM5 challenge, GENIE3 (Huynh-Thu et al., 2010), with a non-linear alternative based on MI (Faith et al., 2007), the Spearman's rank correlation coefficient and the L0L2 sparse regression method that we applied using recent advances in sparse multivariate statistical modelling (Bartlett et al., 2019; Hazimeh and Mazumder, 2018). We also considered a “random predictor” that outputs a random ordering of all the possible TF-gene interactions. See SI-Section 1 and Figs. S1-S3 for more details.

The putative regulatory links predicted by each of these methods using the scRNA-seq data were evaluated *as is* but also subjected to a filtering process in which only TF-gene

associations supported by the chromatin accessibility data were considered in the final network. To identify these refined predictions, we added the *+CA* suffix to the name of the GRN inference methods. In both cases, to generate the final network, we selected the top 100,000 edges by ranking edges according to their network score. We define network score as the absolute correlation coefficient or linear regression coefficient (in the case of Spearman's correlation and sparse regression respectively), or as -log10*p* for mutual information, or the GENIE3 score.

Simulation study

For our simulation study, we used the `daggitty` package in `R` to generate synthetic gene expression data with a known ground-truth GRN structure, as follows. These synthetic datasets were generated with a range of sample sizes n and number of potential regulators of each node (gene) p using linear structural equation models (SEMs) (Eq. 1 in SI-Section 1) based on pre-defined GRNs (with known ground-truth structure) generated with network edge density ρ=0.07. We note that this value of ρ was based on an empirical estimate from the available ATAC-seq+TF motif data. For each combination of n and p, we generated 100 network instances. Inferences from the different network methods we considered (GENIE3, L0L2 sparse regression, MI, Spearman correlation) were then compared with the known ground-truth. This comparison was assessed by the AUC statistic, the 'area under the ROC (receiver-operator characteristic) curve'. This statistic was averaged across the 100 network instances we generated.

Statistical evaluation metrics

We calculated a reproducibility score R as a bootstrap estimate of the posterior probability of observing an edge E given the dataset D, $R=P(E|D)$ (Pe'er, 2005), for the top 100,000 predicted edges. We also carried out 2-fold cross-validation in two ways as follows. We randomly split each dataset (EPI, PE and TE) in two groups with the same number of samples ten times (repeated 2-fold cross-validation), applied the inference methods (with or without chromatin accessibility refinement), took the top-1% of all possible regulatory interactions from one group and quantified the extent to which the top interactions inferred from the other group coincide with that reference using the area under the precision-recall curve (AUPRC). In addition, we calculated a normalised L2 loss comparing all the network inferences from the two network fits on the 2 folds of the data (at the level of the network scores). The normalised L2 loss is then defined as the L2 loss comparing the network scores from each of the 2 data folds, divided by the product of the square roots of the L2 norms of the network scores of each of the 2 data folds.

Biological evaluation metrics

For each combination of inference method and normalisation approach, we computed the intersection between the GRN edges predicted to identify interactions that are specific to the EPI, PE and TE, as well as the interactions common to the three cell types. In set notation this corresponds to EPI-specific = (EPI \ PE) \ TE, PE-specific = (PE \ EPI) \ TE, TE-specific = (TE \ EPI) \ PE and Common = EPI $\cap$ PE $\cap$ TE. Then, we used the out-degree of NANOG, GATA4 and CDX2 (marker TFs of the EPI, PE and TE, respectively) as a proxy for their activity in each one of the four networks. We considered that a marker was active if its normalised out-degree (i.e., the proportion of all genes it regulates in that GRN) was at least the median normalised out-degree across all networks.

We calculated a validation score V as the number of relevant gene sets that were enriched at a significance level of 10% (p<0.1) in a gene-set enrichment analysis performed with the genes regulated by the 25 most connected TFs and that were part of the 500 interactions with the

739 highest prediction scores produced by the GRN inference methods for the EPI and TE. The
740 list of gene sets for the EPI and TE were manually curated and represent the most relevant
741 pathways and biological processes in each one of these cell types (see Tables S2 and S3).
742 The starting list of gene sets comprised all the pathways and Gene Ontology Biological
743 Process (GO BP) terms collated by the Bader Lab at the University of Toronto on 01 October
744 2020
745 (http://download.baderlab.org/EM_Genesets/October_01_2020/Human/symbol/Human_GO
746 BP_AllPathways_no_GO_iea_October_01_2020_symbol.gmt). To facilitate the comparison
747 between *V* scores from different GRN inference methods within a cell type, we normalised *V*
748 to the maximum score attained by a network predictor for that cell type (i.e., EPI or TE).
749

750 **Human and Mouse Embryo Culture**

751
752 Human embryos were cultured as previously described (Gerri et al., 2020). Vitrified human
753 blastocyst stage embryos were thawed using Kitazato Thawing Media (VT602, order number
754 91121) following the manufacturer's instructions.
755
756 All animal research was performed under the project licence PP8826065 according to UK
757 Home Office regulations. The project passed ethical review by the Francis Crick Institute
758 Animal Welfare Review Board in 2019. Six-week-old CBA female mice were super-ovulated
759 using injection of 5 IU of pregnant mare serum gonadotrophin (PMSG; Sigma-Aldrich). Forty-
760 eight hours after PMSG injection, 5 IU of human chorionic gonadotrophin (HCG; Sigma-
761 Aldrich) was administered. Superovulated females were set up for mating with eight-week-old
762 or older (C57BL6) males. F1 hybrid mouse zygotes were isolated in FHM and cumulus cells
763 were removed with hyaluronidase (Sigma-Aldrich; H4272).
764
765 Mouse embryos and human embryos were cultured in drops of pre-equilibrated Global
766 medium (LifeGlobal; LGGG-20) supplemented with 10% human serum albumin (LifeGlobal;
767 GHSA-125) and overlaid with mineral oil (Origio; ART-4008-5P). Preimplantation embryos
768 were incubated at 37 °C and 5% CO_2 in an EmbryoScope+ time-lapse incubator (Vitrolife) and
769 cultured for up to 24hrs for human blastocyst and 4 days for mouse blastocysts.
770

771 **Cell Culture**

772
773 Primed hESC culture
774 Primed hESCs were cultured on growth factor-reduced Matrigel-coated dishes (BD
775 Biosciences; 356231) in humidified incubators at 37°C, 5% CO_2. Cells were fed with fresh
776 mTeSR1 medium (Stem Cell Technologies; 85850) every day and passaged as clumps at a
777 ~1:10 ratio using ReLeSR (Stem Cell Technologies; 05872) every 5-6 days.
778
779 Naïve hESC culture
780 Naïve hESCs were cultured on mitotically inactivated DR4 mouse embryonic fibroblasts
781 (MEFs; density = 1×106 per well of a six-well plate) in humidified incubators at 37°C, 5% O_2
782 and 5% CO_2. Cells were fed with fresh PXGL medium [N2B27 basal medium (Takara; Y40002)
783 + 1 uM PD03 (Cambridge Bioscience; 13034) + 2 uM XAV939 (Sigma; X3004) + 2 uM Go6983
784 (Bio-Techne; 2285) + 10 ng/mL hLIF (PepROTECH; 300-05-25)] every 2 days and passaged
785 as single cells at ~1:5 ratio using Accutase (ThermoFisher; A11105-01) every 4-5 days. 10
786 μM ROCK inhibitor (Tocris Bioscience; Y-27632) were added into the medium for 24 hours
787 after passaging to increase the cell survivability.
788

789 **Immunofluorescence staining**

790
791 Cells were cultured on μ-Slide 8-well dishes (Ibidi; 80826) before immunofluorescence
792 staining. Embryos and cell cultures were fixed with freshly prepared 4% paraformaldehyde in

PBS. Fixation was performed for 20 min at room temperature for embryos and overnight at 4 °C for cell cultures. Embryos and cell cultures were then washed 3 times in 1× PBS with 0.1% Tween-20 to remove residual paraformaldehyde. Permeabilization was performed with 1× PBS with 0.5% Tween-20 and followed by blocking in blocking solution (10% donkey serum in 1× PBS with 0.1% Tween-20) for 1 h at room temperature on a rotating shaker. Then, antibody incubation was performed with primary antibodies diluted in blocking solution overnight at 4 °C on rotating shaker. The following day, embryos and cell cultures were washed in 1× PBS with 0.1% Tween-20 for 3 times, and then incubated with secondary antibodies diluted in blocking solution for 1 h at room temperature on a rotating shaker in the dark. Next, embryos and cell cultures were washed in 1× PBS with 0.1% Tween-20 for 3 times. Finally, embryos were placed in 1× PBS with 0.1% Tween-20 with Vectashield and DAPI mounting medium (Vector Lab; H-1200) (1:30 dilution). Embryos were placed on μ-Slide 8-well dishes (Ibidi; 80826) for confocal imaging. The antibodies and concentrations used are reported in Table S4.

Confocal imaging

Confocal immunofluorescence images were taken with Leica SP8 confocal microscopes for cell cultures and embryos. 2-μm-thick optical sections were collected for embryo Z-stack imaging.

Nuclei segmentation and quantification

Nuclei segmentation and quantification were performed on Z-stack confocal images taken at 2-μm-thick optical sections. Stardist (Weigert et al., 2020) was used for nuclei segmentation followed by CellProfiler (Stirling et al., 2021) for nuclei tracking and fluorescence intensity quantification. Detailed pipeline and scripts can be found here: https://github.com/galanisl/early_hs_embryo_GRNs

Data and code availability

scRNA-seq and ATAC-seq data were obtained from the European Nucleotide Archive [accession numbers: PRJNA153427, PRJNA277181, PRJEB11202 and PRJNA494280]. Data pre-processing and analysis scripts are available at https://github.com/galanisl/early_hs_embryo_GRNs.

Acknowledgements

We thank Zimeng Chen for contributing to the initial analysis of the synthetic datasets. We thank members of the Niakan lab for helpful discussions and feedback on the manuscript. Work in the laboratory of KKN was supported by the Wellcome Trust (221856/Z/20/Z). Work in the laboratory of KKN was also supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (CC2074), the UK Medical Research Council (CC2074) and the Wellcome Trust (CC2074). The work carried out by TEB was partly supported by MRC grant MR/P014070/1. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Competing Interests

No competing interests declared

Author contributions

GAL, TEB and KKN conceived the project. GAL and TEB developed the analytical network interference and statistical methods and analysed data. QH developed the network model representation pipeline, analysed data and performed immunostaining analysis and quantification. AM performed human and mouse embryo thaw and culture. CS provided advice on image quantification and analysed data. KKN analysed data and funded the project. KE, PS and LC coordinate the donation of embryos to the research project. GAL, TEB, QH, CS and KKN wrote the manuscript.

References

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nature Methods* 14, 1083–1086.

Amezquita, R. A., Lun, A. T. L., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., et al. (2019). Orchestrating single-cell analysis with Bioconductor. *Nature Methods* 17, 137–145.

Arceci, R. J., King, A. A., Simon, M. C., Orkin, S. H. and Wilson, D. B. (1993). Mouse GATA-4: a retinoic acid-inducible GATA-binding transcription factor expressed in endodermally derived tissues and heart. *Mol Cell Biol* 13, 2235–2246.

Argelaguet, R., Lohoff, T., Li, J. G., Nakhuda, A., Drage, D., Krueger, F., Velten, L., Clark, S. J. and Reik, W. (2022). Decoding gene regulation in the mouse embryo using single-cell multi-omics. *bioRxiv* 2022.06.15.496239,.

Barakat, T. S., Halbritter, F., Zhang, M., Rendeiro, A. F., Perenthaler, E., Bock, C. and Chambers, I. (2018). Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell Stem Cell* 23, 276-288.e8.

Bartlett, T. E., Kosmidis, I. and Silva, R. (2019). Two-way sparsity for time-varying networks, with applications in genomics. *arXiv* arXiv:1802.08114,.

Barzel, B. and Barabási, A.-L. (2013). Network link prediction by global silencing of indirect correlations. *Nature Biotechnology* 31, 720–725.

Blakeley, P., Fogarty, N. M. E., del Valle, I., Wamaitha, S. E., Hu, T. X., Elder, K., Snell, P., Christie, L., Robson, P. and Niakan, K. K. (2015). Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development* 142, 3151–3164.

Brandman, O. and Meyer, T. (2008). Feedback Loops Shape Cellular Signals in Space and Time. *Science* 322, 390–395.

Bredenkamp, N., Yang, J., Clarke, J., Stirparo, G. G., von Meyenn, F., Dietmann, S., Baker, D., Drummond, R., Ren, Y., Li, D., et al. (2019). Wnt Inhibition Facilitates RNA-Mediated Reprogramming of Human Somatic Cells to Naive Pluripotency. *Stem Cell Reports* 13, 1083–1098.

885 Cahan, P., Li, H., Morris, S. A., Lummertz da Rocha, E., Daley, G. Q. and Collins, J. J. (2014).
886 CellNet: Network biology applied to stem cell engineering. *Cell* 158, 903–915.

887 Cao, Z., Carey, T. S., Ganguly, A., Wilson, C. A., Paul, S. and Knott, J. G. (2015). Transcription
888 factor AP-2γ induces early Cdx2 expression and represses HIPPO signaling to specify
889 the trophectoderm lineage. *Development* 142, 1606–1615.

890 Chai, L. E., Loh, S. K., Low, S. T., Mohamad, M. S., Deris, S. and Zakaria, Z. (2014). A review on
891 the computational approaches for gene regulatory network construction. *Comput*
892 *Biol Med* 48, 55–65.

893 Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S. and Smith, A. (2003).
894 Functional expression cloning of Nanog, a pluripotency sustaining factor in
895 embryonic stem cells. *Cell* 113, 643–655.

896 Chan, T. E., Stumpf, M. P. H. and Babtie, A. C. (2017). Gene Regulatory Network Inference
897 from Single-Cell Data Using Multivariate Information Measures. *Cell Systems* 5, 251-
898 267.e3.

899 Chen, S. and Mar, J. C. (2018). Evaluating methods of inferring gene regulatory networks
900 highlights their lack of performance for single cell gene expression data. *BMC*
901 *Bioinformatics* 19, 232.

902 Chen, D., Liu, W., Zimmerman, J., Pastor, W. A., Kim, R., Hosohama, L., Ho, J., Aslanyan, M.,
903 Gell, J. J., Jacobsen, S. E., et al. (2018). The TFAP2C-Regulated OCT4 Naive Enhancer
904 Is Involved in Human Germline Formation. *Cell Rep* 25, 3591-3602.e5.

905 Chovanec, P., Collier, A. J., Krueger, C., Várnai, C., Semprich, C. I., Schoenfelder, S., Corcoran,
906 A. E. and Rugg-Gunn, P. J. (2021). Widespread reorganisation of pluripotent factor
907 binding and gene regulatory interactions between human pluripotent states. *Nat*
908 *Commun* 12, 2098.

909 Daub, C. O., Steuer, R., Selbig, J. and Kloska, S. (2004). Estimating mutual information using
910 B-spline functions - an improved similarity measure for analysing gene expression
911 data. *BMC Bioinformatics* 5, 118.

912 Davidson, E. H. and Erwin, D. H. (2006). Gene regulatory networks and the evolution of
913 animal body plans. *Science* 311, 796–800.

914 Dietrich, J.-E. and Hiiragi, T. (2007). Stochastic patterning in the mouse pre-implantation
915 embryo. *Development* 134, 4219–4231.

916 Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G. and West, M. (2004). Sparse graphical
917 models for exploring gene expression data. *Journal of Multivariate Analysis* 90, 196–
918 212.

919 Ema, M., Mori, D., Niwa, H., Hasegawa, Y., Yamanaka, Y., Hitoshi, S., Mimura, J., Kawabe, Y.,
920 Hosoya, T., Morita, M., et al. (2008). Krüppel-like factor 5 Is Essential for Blastocyst

921 Development and the Normal Self-Renewal of Mouse ESCs. *Cell Stem Cell* 3, 555–
922 567.

923 Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di
924 Tommaso, P. and Nahnsen, S. (2020). The nf-core framework for community-curated
925 bioinformatics pipelines. *Nature Biotechnology* 38, 276–278.

926 Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins,
927 J. J. and Gardner, T. S. (2007). Large-Scale Mapping and Validation of Escherichia coli
928 Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biology*
929 5, e8.

930 Feizi, S., Marbach, D., Médard, M. and Kellis, M. (2013). Network deconvolution as a general
931 method to distinguish direct dependencies in networks. *Nature Biotechnology* 31,
932 726–733.

933 Festuccia, N., Osorno, R., Halbritter, F., Karwacki-Neisius, V., Navarro, P., Colby, D., Wong, F.,
934 Yates, A., Tomlinson, S. R. and Chambers, I. (2012). Esrrb Is a Direct Nanog Target
935 Gene that Can Substitute for Nanog Function in Pluripotent Cells. *Cell Stem Cell* 11,
936 477–490.

937 Fieller, E. C., Hartley, H. O. and Pearson, E. S. (1957). Tests for rank correlation coefficients. I.
938 *Biometrika* 44, 470–481.

939 Fiers, M. W. E. J., Minnoye, L., Aibar, S., Bravo González-Blas, C., Kalender Atak, Z. and Aerts,
940 S. (2018). Mapping gene regulatory networks from single-cell omics data. *Brief Funct*
941 *Genomics* 17, 246–254.

942 Fleck, J. S., Jansen, S. M. J., Wollny, D., Zenk, F., Seimiya, M., Jain, A., Okamoto, R., Santel,
943 M., He, Z., Camp, J. G., et al. (2022). Inferring and perturbing cell fate regulomes in
944 human brain organoids. *Nature*.

945 Gerri, C., McCarthy, A., Alanis-Lobato, G., Demtschenko, A., Bruneau, A., Loubersac, S.,
946 Fogarty, N. M. E., Hampshire, D., Elder, K., Snell, P., et al. (2020). Initiation of a
947 conserved trophectoderm program in human, cow and mouse embryos. *Nature* 587,
948 443–447.

949 González-Blas, C. B., De Winter, S., Hulselmans, G., Hecker, N., Matetovici, I., Christiaens, V.,
950 Poovathingal, S., Wouters, J., Aibar, S. and Aerts, S. (2022). SCENIC+: single-cell
951 multiomic inference of enhancers and gene regulatory networks. *bioRxiv*
952 2022.08.19.504505.

953 Hainer, S. J., Bošković, A., McCannell, K. N., Rando, O. J. and Fazzio, T. G. (2019). Profiling of
954 Pluripotency Factors in Single Cells and Early Embryos. *Cell* 177, 1319-1329.e11.

955 Haury, A.-C., Mordelet, F., Vera-Licona, P. and Vert, J.-P. (2012). TIGRESS: Trustful Inference
956 of Gene REgulation using Stability Selection. *BMC Systems Biology* 6, 145.

957 Hayashi, R. and Inoue, A. (2023). Low-Input CUT&RUN for Mouse Oocytes and
958 Preimplantation Embryos. *Methods Mol Biol* 2577, 83–92.

959 Hazimeh, H. and Mazumder, R. (2018). Fast Best Subset Selection: Coordinate Descent and
960 Local Combinatorial Optimization Algorithms. *arXiv* arXiv:1803.01454,.

961 Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. and Geurts, P. (2010). Inferring Regulatory
962 Networks from Expression Data Using Tree-Based Methods. *PLoS ONE* 5, e12776.

963 Iacono, G., Massoni-Badosa, R. and Heyn, H. (2019). Single-cell transcriptomics unveils gene
964 regulatory network plasticity. *Genome Biology* 20, 110.

965 Janssens, J., Aibar, S., Taskiran, I. I., Ismail, J. N., Gomez, A. E., Aughey, G., Spanier, K. I., De
966 Rop, F. V., González-Blas, C. B., Dionne, M., et al. (2022). Decoding gene regulation in
967 the fly brain. *Nature* 601, 630–636.

968 Kamimoto, K., Hoffmann, C. M. and Morris, S. A. (2020). CellOracle: Dissecting cell identity
969 via network inference and in silico gene perturbation. *bioRxiv* 2020.02.17.947416,.

970 Kang, Y., Thieffry, D. and Cantini, L. (2021). Evaluating the Reproducibility of Single-Cell Gene
971 Regulatory Network Inference Algorithms. *Front. Genet.* 12, 617282.

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

975 Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., Pledger, E. S., Bryson, T. D., Henikoff, J. G.,
976 Ahmad, K. and Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of
977 small samples and single cells. *Nat Commun* 10, 1930.

978 Kim, J., T. Jakobsen, S., Natarajan, K. N. and Won, K.-J. (2021). TENET: gene network
979 reconstruction using transfer entropy reveals key regulatory factors from single cell
980 transcriptomic data. *Nucleic Acids Research* 49, e1–e1.

981 Krishnaswamy, S., Spitzer, M. H., Mingueneau, M., Bendall, S. C., Litvin, O., Stone, E., Pe'er,
982 D. and Nolan, G. P. (2014). Conditional density-based analysis of T cell signaling in
983 single-cell data. *Science* 346, 1250689.

984 Li, L., Guo, F., Gao, Y., Ren, Y., Yuan, P., Yan, L., Li, R., Lian, Y., Li, J., Hu, B., et al. (2018).
985 Single-cell multi-omics sequencing of human early embryos. *Nature Cell Biology* 20,.

986 Li, Z., Schulz, M. H., Look, T., Begemann, M., Zenke, M. and Costa, I. G. (2019). Identification
987 of transcription factor binding sites using ATAC-seq. *Genome Biology* 20,.

988 Lim, C. Y., Tam, W.-L., Zhang, J., Ang, H. S., Jia, H., Lipovich, L., Ng, H.-H., Wei, C.-L., Sung, W.
989 K., Robson, P., et al. (2008). Sall4 Regulates Distinct Transcription Circuitries in
990 Different Blastocyst-Derived Stem Cell Lineages. *Cell Stem Cell* 3, 543–554.

991 Lin, S.-C. J., Wani, M. A., Whitsett, J. A. and Wells, J. M. (2010). Klf5 regulates lineage
992 formation in the pre-implantation mouse embryo. *Development* 137, 3953–3963.

993 Liu, L., Leng, L., Liu, C., Lu, C., Yuan, Y., Wu, L., Gong, F., Zhang, S., Wei, X., Wang, M., et al.
994 (2019). An integrated chromatin accessibility and transcriptome landscape of human
995 pre-implantation embryos. *Nature Communications* 10, 364.

996 Lv, B., An, Q., Zeng, Q., Zhang, X., Lu, P., Wang, Y., Zhu, X., Ji, Y., Fan, G. and Xue, Z. (2019).
997 Single-cell RNA sequencing reveals regulatory mechanism for trophoblast cell-fate
998 divergence in human peri-implantation conceptuses. *PLoS Biol* 17, e3000187.

999 Lyu, X., Rowley, M. J. and Corces, V. G. (2018). Architectural Proteins and Pluripotency
1000 Factors Cooperate to Orchestrate the Transcriptional Response of hESCs to
1001 Temperature Stress. *Molecular Cell* 71, 940-955.e7.

1002 Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R.,
1003 The DREAM5 Consortium, Kellis, M., Collins, J. J., et al. (2012). Wisdom of crowds for
1004 robust gene network inference. *Nature Methods* 9, 796–804.

1005 Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D. and
1006 Califano, A. (2006). ARACNE: An Algorithm for the Reconstruction of Gene
1007 Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics* 7, S7.

1008 Materna, S. C. and Davidson, E. H. (2007). Logic of gene regulatory networks. *Curr Opin*
1009 *Biotechnol.* 4, 351–354.

1010 Meers, M. P., Bryson, T. D., Henikoff, J. G. and Henikoff, S. (2019). Improved CUT&RUN
1011 chromatin profiling tools. *eLife* 8,.

1012 Meistermann, D., Bruneau, A., Loubersac, S., Reignier, A., Firmin, J., François-Campion, V.,
1013 Kilens, S., Lelièvre, Y., Lammers, J., Feyeux, M., et al. (2021). Integrated pseudotime
1014 analysis of human pre-implantation embryo single-cell transcriptomes reveals the
1015 dynamics of lineage specification. *Cell Stem Cell* 28, 1625-1640.e6.

1016 Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M.,
1017 Maeda, M. and Yamanaka, S. (2003). The homeoprotein Nanog is required for
1018 maintenance of pluripotency in mouse epiblast and ES cells. *Cell* 113, 631–642.

1019 Mochida, K., Koda, S., Inoue, K. and Nishii, R. (2018). Statistical and Machine Learning
1020 Approaches to Predict Gene Regulatory Networks From Transcriptome Datasets.
1021 *Front Plant Sci* 9, 1770.

1022 Morris, S. A., Cahan, P., Li, H., Zhao, A. M., San Roman, A. K., Shivdasani, R. A., Collins, J. J.
1023 and Daley, G. Q. (2014). Dissecting engineered cell types and enhancing cell fate
1024 conversion via CellNet. *Cell* 158, 889–902.

1025 Nguyen, H., Tran, D., Tran, B., Pehlivan, B. and Nguyen, T. (2021). A comprehensive survey of
1026 regulatory network inference methods using single cell RNA sequencing data.
1027 *Briefings in Bioinformatics* 22, bbaa190.

1028 Niakan, K. K., Han, J., Pedersen, R. A., Simon, C. and Pera, R. A. R. (2012). Human pre-
1029 implantation embryo development. *Development* 139, 829–841.

1030 Pastor, W. A., Liu, W., Chen, D., Ho, J., Kim, R., Hunt, T. J., Lukianchikov, A., Liu, X., Polo, J.
1031 M., Jacobsen, S. E., et al. (2018). TFAP2C regulates transcription in human naive
1032 pluripotency by opening enhancers. *Nat Cell Biol* 20, 553–564.

1033 Pe'er, D. (2005). Bayesian Network Analysis of Signaling Networks: A Primer. *Science*
1034 *Signaling* 2005, pl4–pl4.

1035 Peter, I. S. and Davidson, E. H. (2011). Evolution of Gene Regulatory Networks Controlling
1036 Body Plan Development. *Cell* 144, 970–985.

1037 Petropoulos, S., Edsgärd, D., Reinius, B., Deng, Q., Panula, S. P., Codeluppi, S., Plaza Reyes,
1038 A., Linnarsson, S., Sandberg, R. and Lanner, F. (2016). Single-Cell RNA-Seq Reveals
1039 Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* 165,
1040 1012–1026.

1041 Picelli, S., Faridani, O. R., Björklund, A. K., Winberg, G., Sagasser, S. and Sandberg, R. (2014).
1042 Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* 9, 171–181.

1043 Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. and Murali, T. M. (2020). Benchmarking
1044 algorithms for gene regulatory network inference from single-cell transcriptomic
1045 data. *Nature Methods* 17, 147–154.

1046 Schoenfelder, S. and Fraser, P. (2019). Long-range enhancer-promoter contacts in gene
1047 expression control. *Nat Rev Genet* 20, 437–455.

1048 Shi, W., Wang, H., Pan, G., Geng, Y., Guo, Y. and Pei, D. (2006). Regulation of the
1049 Pluripotency Marker Rex-1 by Nanog and Sox2. *Journal of Biological Chemistry* 281,
1050 23319–23325.

1051 Shu, H., Zhou, J., Lian, Q., Li, H., Zhao, D., Zeng, J. and Ma, J. (2021). Modeling gene
1052 regulatory networks using neural network architectures. *Nature Computational*
1053 *Science* 1, 491–501.

1054 Skene, P. J. and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-
1055 resolution mapping of DNA binding sites. *eLife* 6, e21856.

1056 Stirling, D. R., Swain-Bowden, M. J., Lucas, A. M., Carpenter, A. E., Cimini, B. A. and
1057 Goodman, A. (2021). CellProfiler 4: improvements in speed, utility and usability. *BMC*
1058 *Bioinformatics* 22, 433.

1059 Stone, M., Li, J., McCalla, S. G., Siahpirani, A. F., Periyasamy, V., Shin, J. and Roy, S. (2021).
1060 Identifying strengths and weaknesses of methods for computational network
1061 inference from single cell RNA-seq data. *bioRxiv* 2021.06.01.446671,.

1062 Strumpf, D., Mao, C.-A., Yamanaka, Y., Ralston, A., Chawengsaksophak, K., Beck, F. and
1063 Rossant, J. (2005). Cdx2 is required for correct cell fate specification and
1064 differentiation of trophectoderm in the mouse blastocyst. *Development* 132, 2093–
1065 2102.

1066 Thompson, D., Regev, A. and Roy, S. (2015). Comparative analysis of gene regulatory
1067 networks: from network reconstruction to evolution. *Annu Rev Cell Dev Biol* 31, 399–
1068 428.

1069 Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal*
1070 *Statistical Society: Series B (Methodological)* 58, 267–268.

1071 Tipping, M. E. (2001). Sparse Bayesian Learning and the Relevance Vector Machine. *Journal*
1072 *of Machine Learning Research* 1, 211–244.

1073 Vallier, L., Touboul, T., Chng, Z., Brimpari, M., Hannan, N., Millan, E., Smithers, L. E., Trotter,
1074 M., Rugg-Gunn, P., Weber, A., et al. (2009). Early cell fate decisions of human
1075 embryonic stem cells and mouse epiblast stem cells are controlled by the same
1076 signalling pathways. *PLoS One* 4, e6082.

1077 Wang, B., Pourshafeie, A., Zitnik, M., Zhu, J., Bustamante, C. D., Batzoglou, S. and Leskovec,
1078 J. (2018). Network enhancement as a general method to denoise weighted biological
1079 networks. *Nature Communications* 9,.

1080 Wang, J., Ma, A., Ma, Q., Xu, D. and Joshi, T. (2020). Inductive inference of gene regulatory
1081 network using supervised and semi-supervised graph neural networks.
1082 *Computational and Structural Biotechnology Journal* 18, 3335–3343.

1083 Wang, X., He, Y., Zhang, Q., Ren, X. and Zhang, Z. (2021). Direct Comparative Analyses of 10X
1084 Genomics Chromium and Smart-seq2. *Genomics Proteomics Bioinformatics* 19, 253–
1085 266.

1086 Weigert, M., Schmidt, U., Haase, R., Sugawara, K. and Myers, G. (2020). Star-convex
1087 Polyhedra for 3D Object Detection and Segmentation in Microscopy. In *2020 IEEE*
1088 *Winter Conference on Applications of Computer Vision (WACV)*, pp. 3655–3662.
1089 Snowmass Village, CO, USA: IEEE.

1090 Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., et al.
1091 (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and
1092 embryonic stem cells. *Nature Structural & Molecular Biology* 20, 1131–1139.

1093 Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B.,
1094 Wheeler, T. D., McDermott, G. P., Zhu, J., et al. (2017). Massively parallel digital
1095 transcriptional profiling of single cells. *Nat Commun* 8, 14049.

1096
1097
1098
1099
1100
1101
1102
1103