scientific reports

Check for updates

OPEN

Uncovering candidate *Nanog*-Helper genes in early mouse embryo differentiation using differential entropy and network inference

Francisco Prista von Bonhorst[1]✉, Olivier Gandrillon[2], Ulysse Herbach[3], Corentin Robert[1,4], Claire Chazaud[5], Yannick De Decker[4], Didier Gonze[1] & Geneviève Dupont[1]

In the preimplantation mammalian embryo, stochastic cell-to-cell expression heterogeneity is followed by signal reinforcement to initiate the specification of Inner Cell Mass (ICM) cells into Epiblast (Epi). The expression of NANOG, the key transcription factor for the Epi fate, is necessary but not sufficient: coincident expression of other factors is required. To identify possible *Nanog*-helper genes, we analyzed gene expression variability in five time-stamped single-cell transcriptomic datasets using differential entropy, a quantitative measure of cell-to-cell heterogeneity. The entropy of *Nanog* displays a peak-shaped temporal pattern from the 16-cell to the 64-cell stage, consistent with its key role in Epi specification. By estimating the entropy profiles of the 21 genes common to all five datasets, we identified three genes - *Pecam1*, *Sox2*, and *Hnf4a* - whose variability in expression patterns mirrors that of *Nanog*. We further performed gene regulatory network inference using CARDAMOM, an algorithm that exploits temporal dynamics and transcriptional bursting. The results revealed that these three genes exhibit reciprocal activation with *Nanog* at the 32-cell stage. This regulatory motif reinforces fate-switching decisions and co-expression states. Our innovative analysis of single-cell transcriptomic data thus uncovers a likely role for *Pecam1*, *Sox2*, and *Hnf4a* as key genes that, when coincidentally expressed with *Nanog*, initiate ICM differentiation.

Cellular differentiation, a key process in the development of multicellular organisms, is controlled by signaling-modulated gene regulatory networks. Variability in gene expression due to biological noise can act as an additional driving force of cellular differentiation[1–4]. Stochastic activation and inactivation of promoters, known as transcriptional bursting[5], can induce significant cell-to-cell heterogeneity in gene expression that initiates lineage segregation[6,7]. For example, an analysis based on single-cell transcriptomic data revealed that retinoic acid-driven differentiation of mouse embryonic stem cells is largely impacted by noise, with a significant increase in gene expression heterogeneity occurring at the exit of pluripotency[8]. The highly noisy expression of the *Dlk1* gene was shown to facilitate the discrimination of two sub-populations of hematopoietic stem cells[9]. The role of noise in driving cell-to-cell heterogeneity has also been highlighted in the differentiation of neuromesodermal progenitors in zebrafish[10], and in the *Yan/Pnt* network driving *Drosophila* eye development[11]. However, the impact of stochasticity has been much less investigated in in vivo developing embryos than in in vitro culture conditions.

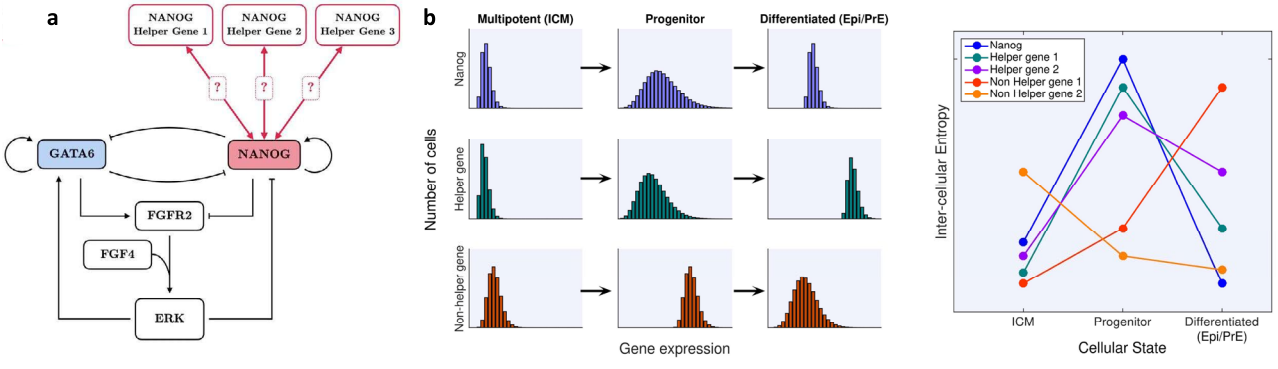
During early mouse embryogenesis, the first cell fate decision gives rise to two distinct populations: the Inner Cell Mass (ICM) and the Trophectoderm (TE). The second cell fate decision starts at the 16-cell stage (E3.0), when ICM cells give rise to Epiblast (Epi) and Primitive Endoderm (PrE) cells. While PrE cells will produce extra-embryonic tissues, Epi cells are at the origin of all the cell types that will constitute the future organism. At the protein level, ICM cells are characterized by the co-expression of NANOG and GATA6[12]. This corresponds to the beginning of the blastocyst formation and takes place between the 16C to 32C stage (E3.0-E3.25). ICM

[1]Unit of Theoretical Chronobiology, Université libre de Bruxelles (ULB), 1050 Brussels, Belgium. [2]Univ Lyon, CNRS, ENS de Lyon, INSERM, UMR5239, LBMC, U1210, 69364 Lyon, France. [3]CNRS, Inria, IECL, Université de Lorraine, F-54000 Nancy, France. [4]Non-linear Physical Chemistry Unit, Université libre de Bruxelles (ULB), 1050 Brussels, Belgium. [5]CNRS, INSERM, GReD Institute, Université Clermont Auvergne, Clermont-Ferrand, France. ✉email: fpristas@ulb.be

cells then asynchronously differentiate between the 32C to 90C stage (E3.25-E3.75) into a random "salt-and-pepper" motif of Epi and PrE cells. Epi cells, some of which already arise between the 32–64C stage (E3.25-E3.5), are characterized by a high level of NANOG and a low level of GATA6, while PrE cells, which arise later at the 64–90C stage (E3.5-E3.75) are characterized by a low level of NANOG and a high level of GATA6[13–15]. Other components also play a crucial role in this decision, such as the FGF/ERK signaling pathway activated by FGF4 secreted by Epi cells. Activation of the FGF/ERK signaling pathway stimulates GATA6 expression and passage to the PrE fate[13,16–20]. Because ICM cells co-express *Nanog* and *Gata6* while Epi/PrE almost exclusively express one of these genes, a key question relates to the source of initial heterogeneity that drives the onset of specification of ICM cells into Epi or PrE cells. This question is particularly relevant since Epi cells are precursors of embryonic stem cells (ES), that are increasingly used in regenerative medicine[21].

The gene regulatory network (GRN) that underlies the ICM to Epi or PrE differentiation is based on a toggle switch between the NANOG and GATA6 transcription factors that also auto-activate (Fig. 1a). Regulation of this core GRN by the ERK pathway, via extra-cellular FGF4 whose production is stimulated by NANOG, plays a key role in the establishment of the salt-and-pepper pattern. Theoretical studies based on dynamical simulations of this GRN modulated by signaling[22–25], or related versions of it[20,26], concluded that heterogeneities are required to trigger differentiation. Slight heterogeneities in extracellular FGF4 at the early stage of the differentiation were shown to be necessary and sufficient to reproduce observations performed on wild-type (WT) and mutant embryos[23,24]. A study analyzing which source of heterogeneity would ensure the most robust differentiation event, in terms of population proportions, led to the prediction that noisy transcription of *Nanog* was the most suitable candidate to initiate these *Fgf4* heterogeneities[25].

Single-cell transcriptomic data indicate that stochastic cell-to-cell expression heterogeneity, followed by signal reinforcement, drives the segregation of ICM cells into Epi and PrE lineages[7,27]. In agreement with model predictions, NANOG plays a key role in the initiation and coordination of other pluripotency factors to initiate Epi specification. By analyzing single-cell transcriptomic data of WT and *Nanog*$^{-/-}$-*Gata6*$^{-/-}$ double knock-out (DKO) embryos, Allègre et al. (2022) indeed found that ICM cells of DKO embryos exhibit an uncoordinated variability in their transcriptome comparable to that of the undifferentiated ICM cells in WT embryos[27]. This highlights the essential role played by NANOG in the initiation of Epi specification. Although NANOG is necessary, it is not sufficient to induce differentiation. Indeed, a subset of ICM cells in wild-type (WT) embryos have high levels of *Nanog* expression at the 16-cell (16 C) and 32-cell (32 C) stages, even though the expression of *Fgf4* - the first marker of binary differentiation - is barely detectable. This indicates that other factors are necessary to induce Epi specification. To verify this hypothesis, the authors analyzed the expression of other pluripotency factors (such as *Sox2*, *Klf4* or *Oct4*), and showed that the interaction levels between these transcription factors were significantly reduced in DKO embryos, supporting an essential role of NANOG in the coordination of these pluripotency factors in Epi specification. Thus, the Epi state seems to be defined by the coordinated expression of Epi/pluripotency markers and NANOG is required to initiate Epi differentiation by



c

	8C	16C	32C	64C	90C
Allègre	/	**18**	**29**	**33**	**24**
Guo	**34**	**39**	**45**	**55**	/
Goolam	**32**	**6**	**6**	/	/
Posfai	/	**33**	**40**	**33**	/
Wang	**64**	**48**	**73**	/	/

Fig. 1. (**a**) Simplified Gene Regulatory Network (GRN) driving the differentiation of ICM cells to Epi/PrE cells as proposed in previous modelling studies[21,22]. (**b**) Leftmost panel: Schematic view of the evolution of the distribution of gene expression of prototype genes. Rightmost panel: Schematic profiles of the evolution of inter-cellular entropy during the differentiation process. We hypothesize that a gene that initiates the coordination of pluripotency markers alongside *Nanog* (so-called "Helper genes"), displays a similar inter-cellular entropy profile. (**c**) Number of cells per cellular stage available for each single cell expression data used in this study.

enabling this coordinated expression. Allègre et al. (2022) suggested that several genes could act as NANOG-helping factors and that the coincident expression of any of these with NANOG would initiate Epi specification, with none of them being required per se[27]. The identity of these factors remains to be determined, and it is not known whether several factors are required simultaneously.

In the present study, we used computational approaches to identify a set of genes that may participate in Epi specification when coincidentally expressed with *Nanog*. Because of the primary role played by cell-to-cell heterogeneity in ICM differentiation, we first estimated the temporal evolution of the inter-cellular entropy of expression of key genes involved in this process[28], using several single-cell datasets. The inter-cellular entropy of a gene driving the differentiation process is expected to first increase and then decrease, reflecting the progressive passage through a pluripotent, progenitor and finally differentiated state. Computation of inter-cellular entropy is generally carried out using the Shannon entropy. This requires a discretization of data, for which several methods have been suggested, including those based on the number of cells in the experiments[3], or the Bayes-Block discretization procedure[29]. These studies revealed an initial increase followed by a decrease in the mean inter-cellular entropy across all genes. Similar conclusions have been reached upon estimation of the changes in digital entropy[30]. Thus, we first checked whether the temporal profile of the entropy of expression of the driver gene *Nanog*, exhibits the same evolution. Then, using the inter-cellular entropy evolution to identify genes whose variability in expression plays a key role in the differentiation of ICM cells, we searched for other genes displaying a similar profile (see rightmost panel of Fig. 1b for a schematic representation).

Although the Shannon entropy is simple to interpret – large values of this entropy reflect a high degree of randomness in the underlying data–, discretization and binning procedures can bias the results when data are limited. Because gene expression datasets for early mammalian development typically involve a small number of cells, we used inter-cellular *differential entropy*, which is a continuous analog to the Shannon entropy. Our computation of differential entropy relies on the fitting of the empirical mRNA distributions to gamma distributions[28,31], which are directly related to a minimal model of transcriptional bursting and involve only two parameters[32,33]. This circumvents the challenging issues of binning required for the Shannon entropy. As illustrated in the leftmost panel of Fig. 1b, the changes in inter-cellular entropy associated with differentiation are expected to be associated but not limited to a transient spreading of the gamma distribution. To strengthen the computations based on relatively low numbers of cells available in each gene expression dataset for the ICM to Epi/PrE differentiation process in mice (Fig. 1c), we computed and analyzed the inter-cellular differential entropy using five available sets of expression data, consisting of single-cell RT-qPCR and single-cell RNA-seq datasets[14,27,34–36]. This allowed us to identify several genes, which we call candidate *Nanog*-helper genes, displaying peak-shaped temporal profiles of inter-cellular differential entropy (henceforth referred to as inter-cellular entropy) similar to that of *Nanog*.

In the second part of this study, we performed network inference to unravel the relationships between *Nanog* and the selected genes using CARDAMOM[37,38]. We found that the candidate *Nanog*-helper genes identified through the examination of the temporal profiles of differential entropy all have a mutually positive influence with *Nanog*. This network motif is known to enforce the decision of fate switching[39] and to promote the occurrence of transient coordinated states of high expression[40]. As these two properties characterize the passage towards the Epi state, we concluded that the genes identified, *Pecam1*, *Sox2* and *Hnf4a*, could act as *Nanog*-helper genes in the specification of ICM cells.

Results

Preliminary analyses

To capture cell-to-cell variability, we rely on single-cell observations[8,41,42], i.e. scRT-qPCR and scRNA-seq datasets[14,27,34–36] for which the number of cells at each cellular stage can be found in Fig. 1c. We focused on the 21 genes that are common to all datasets, listed in Supp. Table 1. This small set of genes includes the main known components of the GRN governing the ICM to Epi/PrE differentiation. The datasets have been pre-processed to allow computation of differential entropy, as described in the *Methods* section. As detailed in the *Transformation to mRNA counts* sub-section, we did not normalize the threshold cycle values (C_t) of the scRT-qPCR data using reference genes. Indeed, as shown in Fig. 2a, the variability on C_t values is similar for reference genes and for genes of interest. Moreover, the level of expression of reference genes vary with the cellular stage. Thus, normalizing the scRT-qPCR data would bias the computation of entropy of the genes of interest by including a component imputable to the variability of the reference genes. For scRNA-seq datasets, we used the raw counts. Because we do not analyze gene expression profiles individually but rather focus on the time evolution of inter-cellular entropy, the absence of normalization does not qualitatively affect the results shown below.

Given the non-standard way of treating the expression data and the reduced number of genes considered, we first checked the ability of the datasets after transformation to capture specification of ICM cells into two distinct populations of Epi and PrE cells using Principal Component Analysis (PCA). A bifurcation event where an initially homogeneous population divides into two sub-populations as the developmental stage increases is visible, except for Goolam and Posfai's dataset, which includes cells from the trophectoderm. For the others, the emergence of two populations is more visible as the number of cells increases. As shown in Supp. Figure 1 and Supp. Figure 2, there are hints of these two populations arising in Goolam and Posfai's dataset as well, although not as clearly. However, in all cases, the two emerging populations correspond to cells expressing *Fgf4* (supposedly Epi cells) and to cells that do not express this gene (supposedly PrE cells) as shown in Supp. Figure 1. Analysis of the two sub-populations in Allègre's dataset reveals that they are composed of two almost orthogonal sets of vectors, one defined by Epi lineage markers (notably *Nanog*, *Fgf4*, *Pecam1*, *Sox2* and *Bmp4*), and the other by PrE lineage markers (notably *Fgfr2*, *Sox17*, *Pdgfra* and *Gata4*) as shown in the rightmost panel of Fig. 2b. These vectors remain similar for Guo and Wang's dataset (Supp. Figure 2). As expected, in Goolam's and Posfai's datasets, the vectors are less clustered, but we note that *Nanog* and *Gata6* point towards opposite

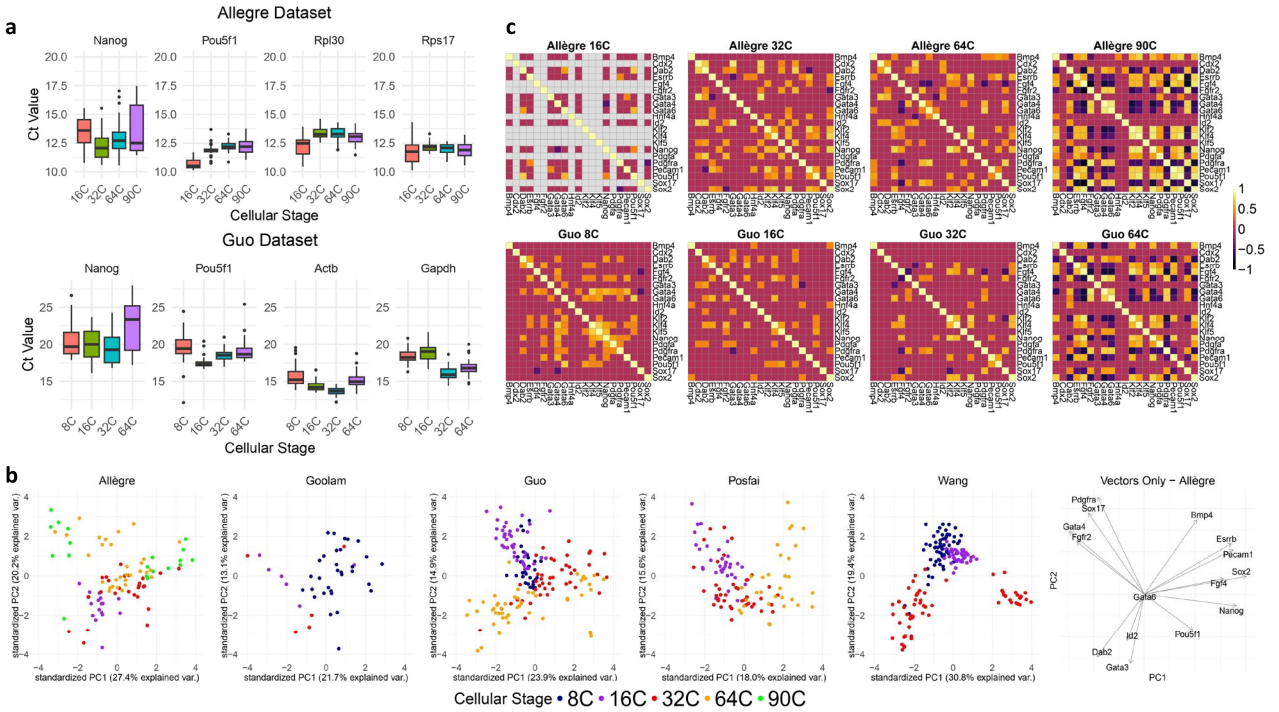


Fig. 2. (**a**) C_t values from the Allègre and Guo scRT-qPCR analyses for *Nanog*, *Pou5f1* and two reference genes used in each study (*Rpl30* and *Rps17*). The variations in the level of expression of the reference genes are of the same order as those of the genes of interest in the two datasets. (**b**) PCA graphs of the two main components for all datasets, considering the 8C, 16C, 32C, 64C and 90C stages for the 21 genes common to all datasets. Except for Goolam and Posfai's dataset, a bifurcation event where an initially homogeneous population divides into two sub-populations as the developmental stage increases is visible, despite the reduced number of data and the minimal pre-processing procedure (see *Methods*). The rightmost panel shows, for Allègre's dataset, that the two sub-populations are composed of two almost orthogonal sets of vectors, one defined by Epi lineage markers (*Nanog*, *Fgf4*, *Pecam1*, *Sox2* and *Bmp4*), the other by PrE lineage markers (*Fgfr2*, *Sox17*, *Pdgfra* and *Gata4*). (**c**) Spearman correlation coefficients for Allègre's and Guo's datasets, using the pre-processing procedure described in the *Methods* section. The increase in the numbers of correlations and anticorrelations aligns with observations reported in the original studies (see text).

directions, supporting the appearance of two sub-populations. We thus conclude that the selection of genes, together with the proposed data transformation, allows capturing cell specification although with a varying degree of robustness.

To further confirm that the transformation of data performed here does not bias the intrinsic interactions between genes, we computed Spearman correlation coefficients as was done in Allègre's and Guo's study (Fig. 2c). As expected, we observed an increase in the number of correlations and anticorrelations as differentiation progresses. Similarly to the observations summarized in Supp. Tables 1&2 in Allègre et al. (2022)[27], we found a positive correlation between *Nanog* and *Pou5f1* at the 16C stage, a positive correlation between *Nanog* and *Fgf4*, *Klf2*, *Klf4*, *Sox2*, *Pou5f1* and *Pecam1* at the 32C stage, and a positive correlation between *Sox2* and *Klf2*, *Klf4* and *Pecam1* at the 32C stage. We also recovered the observed anti-correlation between *Sox2* and *Fgfr2* at the 32C stage. The analysis of Guo's dataset revealed anti-correlations between *Fgf4* and *Fgfr2*, and between *Fgf4* and *Sox17* at the 32C stage. At the 64C stage, computation of the Spearman correlations indicate an anti-correlation between *Fgf4* and *Pdgfra*, *Gata4* and *Gata6*, and between Nanog and *Gata4* and *Gata6*. These observations all align with the analysis performed by Guo et al. (2010) (See Fig. 5a therein)[14]. We thus conclude that our data pre-processing for both the scRT-qPCR and scRNA-seq datasets and the genes used in this study captures the differentiation process of interest. However, PCA and correlation analyses do not allow a clear characterization of possible *Nanog*-helper genes.

Nanog shows a robust temporal profile of inter-cellular entropy

For hematopoietic differentiation, it has been shown that the genes playing a key role in the associated GRN display an increase in Shannon entropy at the exit of the pluripotency stage, followed by a decrease when cells acquire a differentiated state[30,43]. To build on these results while circumventing the complications inherent to the computation of Shannon entropy when data are limited (see *Introduction*), we here use differential entropy that relies on fitting the mRNA distributions to gamma distributions[28,31–33] as detailed in the *Methods* section. Both scRT-qPCR and scRNA-seq data for Epi/PrE specification were well fitted by gamma distributions, as illustrated by four typical examples of distributions and associated cumulative distribution functions (CDF) in

Fig. 3a. Kolmogorov-Smirnov (KS) statistics and p-values for all datasets at each available cell stage are shown in Fig. 3b. In most cases, KS statistic values are relatively small, reflecting a good adequacy between empirical and theoretical cumulative distribution functions. These small KS values are associated with large p-values (> 0.05), which correspond to a small 'non-significant' deviation between the two cumulative distribution functions as can be seen in Fig. 3a. Despite the high accuracy of the fits compared to the empirical distributions, some fits require further inspection. We show examples of such fits, with larger KS statistics and smaller p-values, in Supp. Fig. 3. As can be seen for *Sox2* at the 16C stage in Wang's dataset, or *Sox2* at the 64C stage in Posfai's dataset, the fits are still adequate and capture the behavior of the empirical mRNA distribution. The CDF is, however, more sensitive to outliers because of the presence of many zero expression values. Note that the sparse number of cells implies that a few outliers largely decrease the p-value of an otherwise accurate fit. Nevertheless, comparison between data-based and fitted gamma distribution functions (Fig. 3a and Supp. Fig. 3) shows that fitting gamma distributions to both single-cell RT-qPCR datasets after our pre-processing steps is accurate. The good quality of the fits confirms that this transformation gives rise to a quantity proportional to mRNA counts (see *Methods*).

The inter-cellular differential entropies computed with the gamma distributions of the main components of the GRN driving ICM differentiation (Fig. 1a) are shown for each dataset in Fig. 4a. Accuracies of the predictions are quantified using bootstrapping (see *Methods*). The inter-cellular entropy of *Nanog* shows an increase with a maximum reached at the 32C stage, followed by a decrease onwards, for all datasets involved. This robustness in the temporal profile of *Nanog* cell-to-cell heterogeneity is in line with the main role played by this transcription factor in the emergence of the population of Epi cells. It also confirms that driver genes display a surge in inter-cellular entropy also during in vivo development in mice, as was shown for in vitro erythrocytic differentiation[3]. For *Gata6*, inter-cellular entropy decreases from the 16C stage onwards, in agreement with the proposed mechanism of PrE specification that is triggered by the appearance of Epi cells[22]. As for *Fgfr2* and *Fgf4*, the analysis does not reveal a consistent pattern of entropy evolution.

Identification of candidate *Nanog*-helper genes based on the entropy profiles

We then computed the inter-cellular entropies of the genes common to all datasets at each cellular stage. Results are shown in the form of heatmaps for every dataset in Supp. Fig. 3. By nature, trends in differential entropy can only be compared qualitatively[44]. To identify genes assisting *Nanog* in Epi specification, we focused on the 16C, 32C and 64C stages, since ICM to Epi specification occurs during these stages and then qualitatively compared the entropy profiles of the 20 genes at these stages to that of *Nanog*[27]. The genes with the same qualitative evolution – that is, those whose entropy increases from the 16C to 32C and then decreases from the 32C to 64C, just as *Nanog*'s entropy (as shown in the rightmost panel of Fig. 4a) – were selected as possible candidate genes helping *Nanog* during Epi specification. Due to the limited number of cells in each dataset, we also imposed as

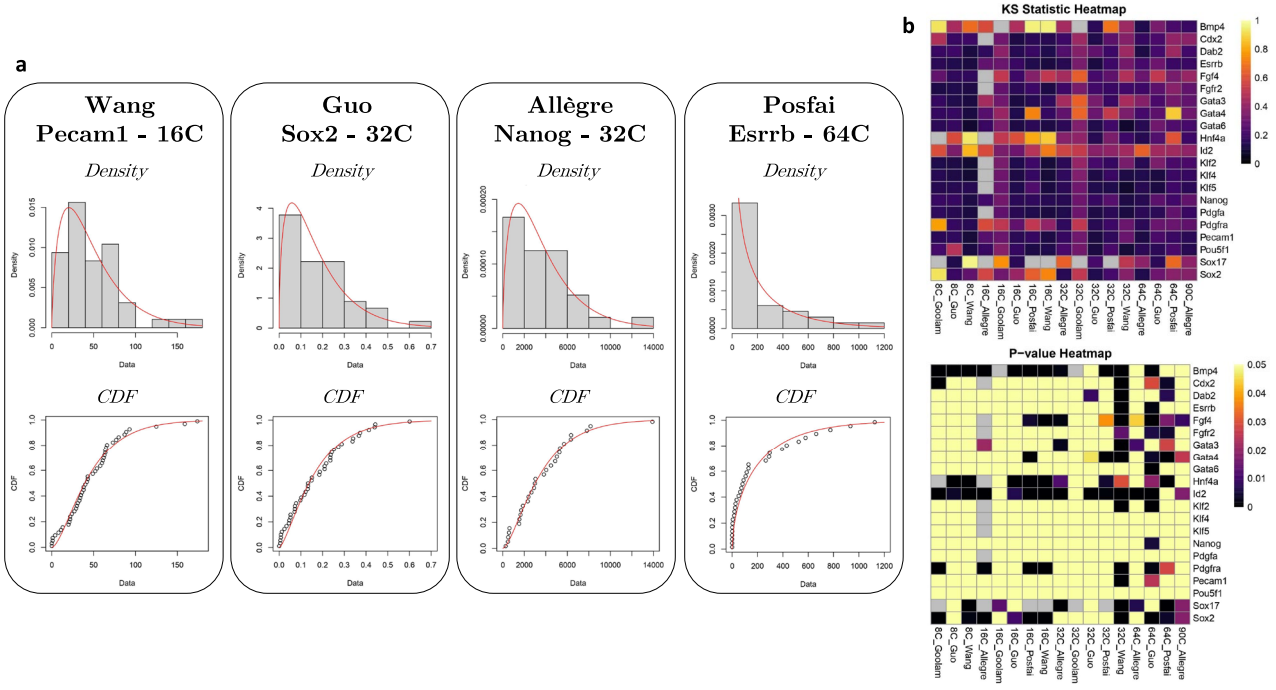


Fig. 3. (**a**) Representative examples of histograms and fits to gamma distributions of the mRNA counts. As shown in the lower panels, the theoretical cumulative distribution functions are close to the empirical ones, indicating the good quality of the fit. (**b**) Kolmogorov-Smirnov (KS) coefficients and p-values of the fits to gamma distributions for the 21 genes considered, for all datasets at each available cell stage. All p-values above 0.05 are shown in the same yellow.

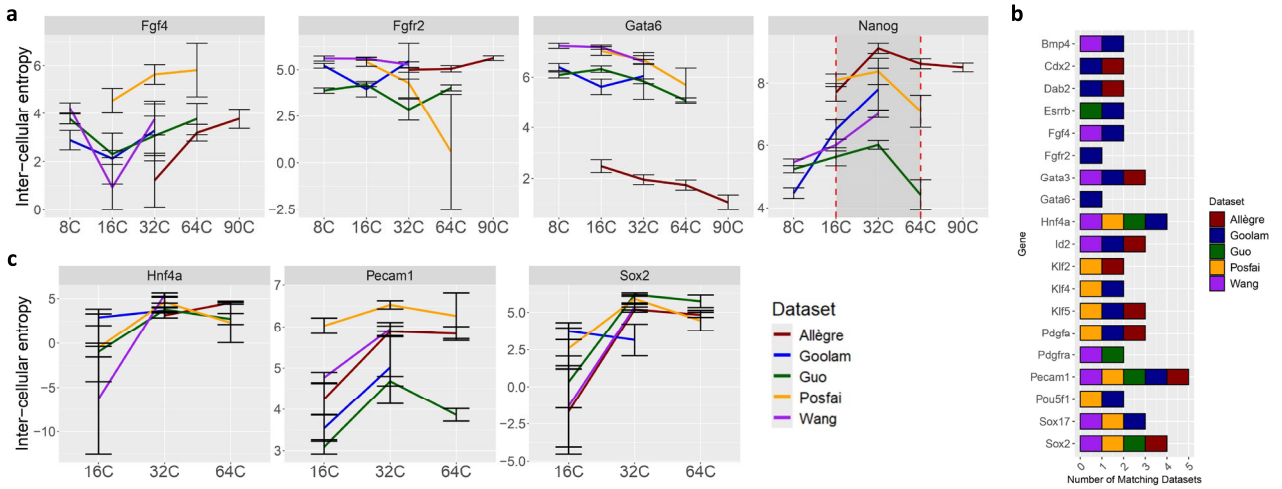


Fig. 4. (**a**) Evolution of the inter-cellular differential entropy of expression of the main genes driving the specification of ICM cells into the Epi or PrE fate. Only *Nanog* displays a peaked shape profile in entropy, in agreement with its driving role in this process. Each dataset is represented by a different color, and error bars correspond to confidence intervals obtained by bootstrapping. The shaded region between the 16C and 64C stage in *Nanog's* entropy profile represents the region of interest where the entropy profiles are compared. (**b**) Comparison of the evolution of inter-cellular differential entropy between *Nanog* and each of the 20 common genes considered, from the 16C to the 64C stage. A square indicates that the entropy profile in a given dataset is similar for the two genes. (**c**) Evolution of the inter-cellular differential entropy of *Hnf4a*, *Pecam1* and *Sox2* between the 16C to 64C stage. Only these three genes show the peak shaped entropy in at least 4 datasets, when considering the mean inter-cellular entropy computed over 1000 bootstraps.

a selection criterion for a candidate helper gene that the evolution of its inter-cellular entropy is consistent in at least 4 datasets, although this probably excludes valid candidates.

Surprisingly, very few genes display the same inter-cellular entropy profile as *Nanog* in this timeframe, and results were dataset dependent, as shown in Fig. 4b. Three genes emerged from this analysis: *Hnf4a*, *Pecam1 and Sox2*. Their entropy profiles between the 16C to 64C stage are shown in Fig. 4c. *Pecam1* and *Sox2* are known to be related to the Epi fate, the former as a marker and the latter as an active player. Indeed, *Sox2* codes for the transcription factor SOX2, which is important for the emergence of the ICM lineage[45]. More generally, it mediates FGF4 signaling and FGFR2 expression[27] while regulating NANOG[46]. It is also a central player in the formation and maintenance of pluripotency in the ICM[47] and in stem cell populations[48]. At the 32C stage, the levels of expression of *Pecam1* and *Sox2* are correlated with those of *Nanog*.

By contrast, *Hnf4a* is not known as an Epi factor. It codes for the zinc finger transcription factor HNF4A and has never been related directly to *Nanog*. It is a marker of the endoderm layer in the implanting blastocyst[49] known to be necessary for the survival of the embryonic ectoderm[50] and the formation of the visceral yolk sac[51]. Furthermore, HNF4A is regulated by GATA6[52]. Interestingly, *Hnf4a* was up regulated in an experiment involving the presence of a FGFR inhibitor, alongside most markers of the Epi cell fate, although it is not known as a Epi marker[14].

In conclusion, based on the analysis of inter-cellular entropy of expression during Epi specification, we found out that three of the common genes robustly display the same temporal profile of variability as *Nanog*. Based on previous observations highlighting the requirement for coincident expression between *Nanog* and other unidentified helper genes for Epi specification[27], we concluded that *Pecam1*, *Sox2* and *Hfn4a* are suitable candidates that may act as *Nanog*-helper genes.

Mutually positive interactions between *Nanog* and candidate *Nanog*-helper genes

To confirm this finding and shed light on the possible mechanism underlying this influence on the specification process, we investigated their impact in the underlying GRN using a network inference method applied to the same datasets. Previous analyses have shown that genes playing a key role in differentiation often display a surge in variability during specification (see *Introduction*). However, it has not yet been tested whether, conversely, key genes can be identified based on their peak-shaped inter-cellular entropy profile. To investigate this question and find out how the genes identified by entropy estimations interact within the minimal GRN shown in Fig. 1a, we used CARDAMOM, a network inference algorithm able to infer GRNs from time-stamped single-cell expression data[37,38]. The method crucially exploits a particular mechanistic model of transcriptional bursting and the resulting gamma distributions. When selecting the main genes of the minimal model shown in Fig. 1a to build the network, CARDAMOM recovered all the interactions proposed in our previous studies (Fig. 5a). It should be noted that the inferred network predicts interactions from the levels of expression of the genes and does not include ERK signaling. As such, inferred interactions on *Fgf4* correspond to influences on *Fgf4* expression. The activation of *Nanog* by *Fgf4* could be seen as a shortcut for other underlying activations that are

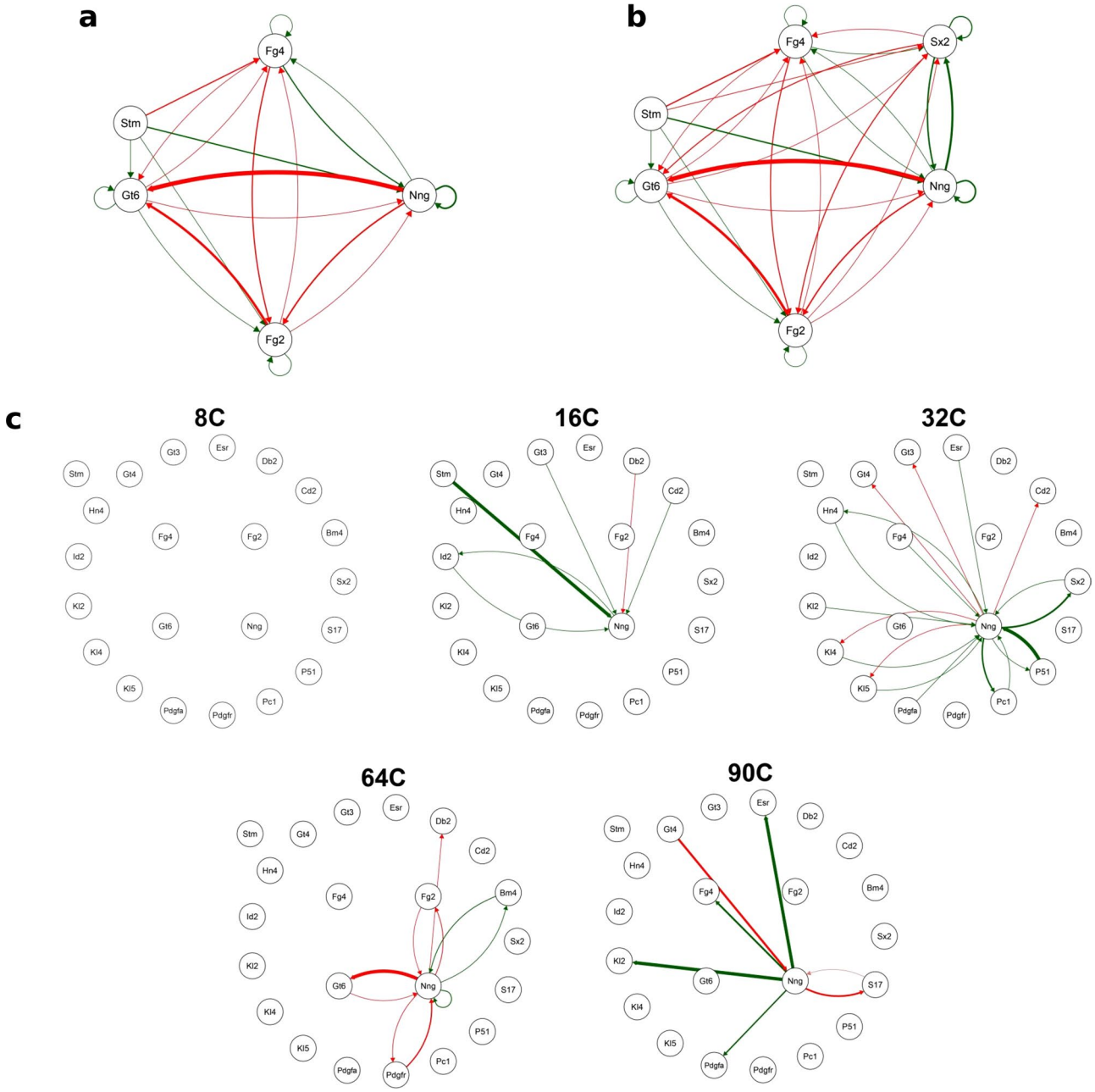


Fig. 5. Inferred Gene Regulatory Networks using CARDAMOM from a combination of the 5 single-cell time-stamped datasets. Green arrows represent activation and red arrows represent inhibitions. The width of the arrows represents interaction strength, with a larger width being associated to a larger interaction strength. Gene abbreviations are described in Supp. Table 2. The 'Stm' node stands for stimulus node and represents a perturbation in the environment of the cells, inducing them to evolve towards a new steady state[37,38]. (**a**) Inferred GRN when using only the principal genes proposed in previous modelling studies *Nanog* - *Gata6* - *Fgf4* and *Fgfr2*[22,23]. Interactions inferred by CARDAMOM are in agreement with original assumptions of the model. (**b**) Inferred GRN when using the principal genes and the candidate *Nanog*-helper gene *Sox2*. *Sox2* appears in a reciprocal activation loop with *Nanog*. The same behavior is found for *Bmp4*, *Fgf4*, *Hnf4a*, *Id2*, *Klf2*, *Pdgfa*, *Pdgfra* and *Pecam1*, as discussed in the main text. (**c**) Time decomposition of the networks inferred by CARDAMOM when considering all the genes common to all data sets and for which the entropy profiles were estimated. For clarity, only interactions with *Nanog* are shown.

not considered in the subset of genes considered to infer the GRN (i.e. *Fgf4* → *Nanog* is inferred instead of *Fgf4* → *other genes* → *Nanog*). In the same way, cross-inhibition between *Gata6* and *Fgf4*, activation of *Nanog* by *Fgf4* and inhibition of *Gata6* by *Fgfr2* cannot be compared with the GRN that considers signaling, and most probably correspond to indirect interactions mediated by other genes not considered in the inference process.

When the proposed helper gene *Sox2* was added to the main genes of the minimal model mentioned above, a reciprocal activation with *Nanog* was revealed (Fig. 5b). Because this small motif enforces the decision of fate switching[39] and promotes the occurrence of transient coordinated states of high expression[40], it confirms that *Sox2* coincident expression with *Nanog* favors the passage to the Epi state. Closer inspection reveals the two activations are simultaneous, i.e. they are detected at the same cell stage, notably the 32C stage, which corresponds to the time of appearance of Epi cells[27]. The same results were obtained when *Bmp4*, *Hnf4a*, *Klf2* or *Pecam1* were individually added to the main GRN, as shown in Supp. Table 2. Other genes, such as *Fgf4*, *Id2*, *Pdgfa and Pdgfra* also show these double activations, although they do not necessarily appear simultaneously at the same stage, nor at the 32C stage specifically as shown in Supp. Table 2.

CARDAMOM was then run considering the whole set of common genes whose entropy profile was determined in the previous section. The interactions inferred between *Nanog* and all the other genes considered are shown separately for the different cell stages in Fig. 5c. *Sox2*, *Pecam1* and *Hnf4a* robustly display a reciprocal activation with Nanog at the 32C stage. Importantly, as shown in Fig. 5c, the double activation patterns between *Nanog* and the helper-genes precede the appearance of the toggle switch between *Nanog* and *Gata6*, which agrees with their roles in driving Epi specification. Another Epi factor, *Pou5f1* was found to activate and be activated by *Nanog* at the 32C stage. This gene codes for the OCT4 transcription factor, which together with SOX2, cooperates with NANOG on enhancers to maintain cell pluripotency in mouse embryonic stem cells[53]. Its role as a *Nanog*-helper gene is thus highly plausible, suggesting that it may not have been detected properly by the method based on the entropy profile.

Other genes identified when considering the smaller network, namely *Bmp4*, *Fgf4*, *Klf2*, *Id2*, *Pdgfa*, and *Pdgfra*, which displayed double activation profiles, either do not exhibit these double activation profiles, or these are not simultaneous, or these are not observed at the 32C stage, when considering the whole set of common genes. The robustness of their identification is thus dependent on the number of genes involved for the inference. Only *Hnf4a*, *Pecam1* and *Sox2* display simultaneous double activation profiles at the 32C stage regardless of how many genes are used for the inference, revealing the robust behavior of the genes identified via the estimation of the inter-cellular entropy.

Discussion

In the present study, we have used single-cell transcriptomic data from scRT-qPCR and scRNA-seq experiments, to analyze the ICM to Epi/PrE differentiation in preimplantation mouse embryos. The aim was to identify genes favoring the transition to the Epi state when coincidentally expressed alongside *Nanog*. Focusing on the genes common to the five datasets considered, PCA on all datasets captured the bifurcation event corresponding to the differentiation process under investigation. Spearman correlations on the transformed scRT-qPCR datasets agreed with the previously reported correlations and anticorrelations between gene expression levels. These preliminary analyses have thus confirmed that the data considered are adequate, despite the limited number of genes and cells involved. Moreover, they have established that the pre-processing methods used, which, for scRT-qPCR datasets do not include normalization of C_t values with respect to reference genes, capture the essential relations between genes during ICM to Epi/PrE differentiation. However, these analyses also revealed significant differences between datasets and did not allow the identification of candidate *Nanog*-helper genes via standard analysis only. To this end, we resorted to the computation of inter-cellular differential entropy. Estimation of the inter-cellular differential entropy revealed a robust peak-shaped temporal profile for *Nanog*, in agreement with the known role of this gene as a driver of Epi specification. Indeed, the precocious Epi cells then induce the specification of neighboring cells in PrE through *Fgf4* signaling. This mechanism results in the salt-and-pepper arrangement of Epi and PrE cells in the blastocyst[22]. In agreement with this scenario, *Gata6*, the transcription factor associated with the PrE fate, does not display a peak-shaped profile in inter-cellular entropy. Three genes among those analyzed displayed an inter-cellular entropy profile similar to that of *Nanog*: *Pecam1*, *Sox2* and *Hfn4a*. Based on the findings of Allègre et al. (2022)[27] suggesting that the coincident expression of *Nanog* and still to be identified factors would initiate Epi specification, we identified *Pecam1*, *Sox2* and *Hfn4a* as candidate *Nanog*-helper genes. Network inference using CARDAMOM has shown that all three are connected to *Nanog* by a reciprocal activation loop, with the two branches being simultaneously active at the 32C stage, preceding the toggle switch between *Nanog* and *Gata6*. Because such mutual reinforcement in expression is known to induce coordinated expression and to support fate switching decisions[39,40], *Pecam1*, *Sox2* and *Hfn4a* most probably correspond to factors triggering Epi specification when coincidently expressed with *Nanog*.

Indeed, Sriram et al.[39] have shown that a reciprocal activation loop – also called mutual activation – permits the enforcement of cell fate choice in *Candida albicans*. Bifurcation analysis of the underlying GRN shows that bistability in the level of the master regulator *Wor1* occurs on a wide range of kinetic parameters. Changing the strength of the mutual regulations directly influences the lower threshold values at which the concentration of the master regulator *Wor1* starts exhibiting bistability. While mutual inhibition increases this threshold, mutual activation decreases it. Mutual activation also extends the domain of bistability by increasing the value of the upper threshold at which bistability disappears. The interplay between the two feedback loops gives plasticity to this system and makes it robustly reproducible. It is straightforward to apply this to the ICM to Epi/PrE differentiation; *Nanog* being the master regulator in this case[27], which forms the negative feedback loop with *Gata6* and the mutual activation loops with the candidate helper-genes *Hnf4a*, *Pecam1* and *Sox2*. Furthermore, we have identified that the reciprocal activation loops with the candidate *Nanog*-helper genes precede the toggle switch with *Gata6*. According to Schuh et al.[40] who performed extensive stochastic simulations of abstract GRNs characterized by different interactions, reciprocal activation loops favor co-bursting and the appearance of transient states of coordinated high expression of the genes involved in the reciprocal activation loops. Thus, the reciprocal activation loops play a double role in 1) the initiation of the differentiation, by creating transient

coordinated states of high expression surpassing the threshold necessary for the initiation of the differentiation process, while 2) enforcing and locking the cellular decision once the threshold has been passed.

To quantify cell-to-cell heterogeneity, we have estimated differential entropy that can be directly computed from the parameters of the gamma distributions fitted to mRNA expression data. Entropy is more commonly computed using the Shannon entropy. Two types of entropy can be computed: intra-cellular or inter-cellular entropy. Intra-cellular entropy relates to the variability of the transcriptome of a single-cell and is connected to the stemness and pluripotency level of a cell[14,28,42,54]. Intra-cellular entropy shows mixed behaviors, with some studies reporting an increase[54,55] and others a decrease[56] during early development. To explain these apparently contradictory results, it has been argued that a decrease in intra-cellular entropy is observed during late stages of differentiation when cells are subject to an increase in regulatory constraints and thus become more specialized. In the earlier stages however, an increase in cell fate pluripotency, associated with an increase in intra-cellular entropy, could be necessary to ensure that all cell fates are attainable[28]. In contrast, inter-cellular entropy, which captures the cell-to-cell variability of expression of a given gene in a cell population, has always been reported to exhibit a peak-shaped behavior for genes driving cell specification. As found here for *Nanog*, it increases from the pluripotent to the progenitor stage and then decreases from the progenitor to the differentiated stage.

Entropy sorting, another mathematical framework based on Shannon entropy, has been used to distinguish genes indicative of cell identity[57]. Like for digital entropy[30], scRNA-seq data are discretized into two groups for each gene: a group representing the active state, where the gene is expressed, and a group representing the inactive state, where the gene is not expressed. The algorithm is particularly successful in clustering cells based on their identity, allowing it to define cell types with great precision at a given time point. However, this method does not consider dynamics.

Analysis of the most delta-entropic genes, i.e. genes for which the entropy difference between two consecutive stages is the largest, could complement the present approach. Indeed, for hematopoietic differentiation, Dussiau et al. (2022) have shown that the most delta-entropic genes are those involved in lineage specification, whereas the genes with the highest variation in mean expression are related to mechanisms of cell survival[43]. Such analysis needs entropy values to be comparable in absolute values, which is not possible for differential entropy, and thus could not be done in this study[44]. Another interesting point would be to compare the temporal profiles of differential entropy of *Nanog* and candidate *Nanog*-helper genes between WT embryos and embryos knocked-out for specific genes.

It has been suggested that Epi factors exhibit redundancy[27,58]. In the same line, we hypothesize that the number of *Nanog*-helper genes is most probably larger than three. As identified by CARDAMOM from the reciprocal activation loop, *Pou5f1* could be one of them although it was not firmly detected on the basis of its entropy profile. Moreover, as our innovative analysis was carried out with the 21 genes common to the five datasets considered, some significant genes are probably missing. From a biological point of view, gene transcription is known to be inactive most of the time[59,60]. The probability of having simultaneous bursts of transcription of two genes is thus expected to be low, which would call for the existence of several *Nanog*-helper genes. In the same line, experimental validation of the candidate *Nanog*-helper genes identified here would be delicate. When knocked-down individually, the absence of a *Nanog*-helper gene is only expected to provoke some delay in the average specification time. The validation of this effect would thus require a fine tuning of the observation window around the 32C stage, when Epi cells start to appear, performed on many embryos.

It would be instructive to include candidate *Nanog*-helper genes in existing models of Epi and PrE specification[20,23,25,61]. Because all models require some source of heterogeneity to initiate specification, simulations of these extended GRN would allow the identification of the conditions under which a coordinated noise pattern in the expression of *Nanog* and another gene would be sufficient to trigger Epi specification. This scenario is also expected to affect the robustness of differentiation in terms of populations and differentiation timings. Altogether, combination of experimental approaches, data analyses and modeling are required to address the critical role of gene expression heterogeneity in driving cell differentiation in early development.

Methods

Transformation to mRNA counts

For scRT-qPCR datasets, we first transform the C_t data into quantities that are proportional to mRNA counts. Based on Richard et al. 2016[3], we define the number $m_{k,l}$ of gene l mRNA molecules in cell k by:

$$m_{k,l} = n_{wells} \times D \times 2^{C_{t,threshold} - A - C_{t_{k,l}}} \tag{1}$$

where n_{wells} is the number of wells in the experiment, D is the sampling coefficient, $C_{t,threshold}$ is the detection threshold, A is a constant number of pre-amplifications, and $C_{t_{k,l}}$ is the normalized C_t value using spikes and references genes:

$$C_{t_{k,l}} = \widehat{C_{t_{k,l}}} - \left(\overline{C_{t_{k,l}}} - \overline{C_{t_0}}\right) \tag{2}$$

In the above equation $\widehat{C_{t_{k,l}}}$ is the raw C_t value for gene j, cell i, $\overline{C_{t_{k,l}}}$ is the i-th cell mean spike value and $\overline{C_{t_0}}$ is the global mean spike value. As shown in Fig. 2a, using reference genes would bias the computation of entropy. Because we do not have access to the spike values, we use the raw values $\widehat{C_{t_{k,l}}}$ to compute the "pseudo-mRNA counts". Furthermore, we do not have access to the sampling coefficient, therefore the final transformation for scRT-qPCR datasets is:

$$m_{k,l} = n_{wells} \times 2^{C_{t,threshold} - A - \widehat{C_{t_{k,l}}}} \quad (3)$$

The values used for each scRT-qPCR dataset are described below in the Datasets and Pre-processing subsection.

For scRNA-seq datasets, we simply use the raw counts.

Datasets and pre-processing

Five datasets of single cell gene expression levels in developing embryos before implantation were used. A brief description of these datasets and the pre-processing procedures are given below. The number of cells available in the analysis is shown in Fig. 1c.

- Allègre et al. (2022) scRT-qPCR dataset[27]. The dataset combines single-cell expression levels of 48 genes in 18 individual cells isolated at the 16-cell (16C) stage and the single-cell expression levels of 98 genes from 98 individual cells at the 32-, 64-, and 90- (32C, 64C, 90C) cell stage. Some expression values are thus missing at the 16C stage. The expression cut-off $C_{t,threshold}$ is 35, n_{wells} is either 48 or 96 depending on the cell stage considered, and $A_{pre-amplification}$ is 18. The pseudo-mRNA counts were computed using Eq. (3) as described above. This transformed dataset was used for PCA, correlation, entropy estimation and CARDAMOM analysis.
- Guo et al. (2010) scRT-qPCR dataset[14]. The dataset contains the single-cell expression values of 48 genes from 387 individual cells isolated at four developmental stages (8C, 16C, 32C and 64C), from mouse embryos encompassing two differentiation events (formation of the TE and ICM, and formation of the Epi and PrE). The expression cut-off $C_{t,threshold}$ is 28, n_{wells} is 48, and $A_{pre-amplification}$ is 18. The pseudo-mRNA counts were computed using Eq. (3) as described above. Since our purpose is to study the ICM to Epi/PrE differentiation event, we further used the supplementary file S5 in Guo et al. 2010 to remove presumed outside cells (and therefore TE progenitors) based on the $C_t$value of *Id2* since *Id2* expression is a marker of TE cells[14]. Most outer cells have $C_{t,Id2}$ levels above 24 and were thus removed. This leaves us with 195 individual cells. This transformed dataset was used for PCA, correlation and entropy estimation. The dataset was slightly transformed to be used in CARDAMOM in order to be *quantitatively* comparable to the other datasets. The *quantitative* differences are probably due to the fact that spikes and sampling coefficients used in Guo's study are not available and would be necessary for the datasets to be *quantitatively* comparable. Thus, for CARDAMOM, n_{wells} is 1, and $A_{pre-amplification}$ is 0.
- Goolam et al. (2016) scRNA-seq dataset[34]. The dataset contains the single-cell expression values of 41 480 genes from 124 individual cells isolated at 5 developmental stages (2C, 4C, 8C, 16C and 32C) encompassing the two first differentiation events (formation of the TE and ICM, and formation of the Epi and PrE). Data were obtained using Smart-seq2. We used TPMs for PCA and correlation analysis, while raw counts were used for entropy estimation and CARDAMOM.
- Posfai et al. (2017) scRNA-seq dataset[35]. The dataset contains the single-cell expression values of 16 379 genes from 106 individual cells isolated at 3 developmental stages (16C, 32C and 64C). The study focuses on the formation of the TE and ICM, but data also includes the formation of Epi and PrE. Data were obtained using Smart-seq2. We used TPMs for PCA and correlation analysis, and raw mRNA counts for entropy estimation and CARDAMOM.
- Wang et al. (2021) scRNA-seq dataset[36]. The dataset contains the single-cell expression values of 37 405 genes from 124 individual cells isolated at 7 developmental stages (1C, 2C, 4C, 4LC, 8C, 16C, 32C and 64C) encompassing two differentiation events (formation of the TE and ICM, and formation of the Epi and PrE). We used TPMs for PCA and correlation analysis, and raw mRNA counts for entropy estimation and CARDAMOM.

We restricted the analysis to genes that are common to all datasets. The full list of these 21 genes is given in Supp. Table 1. We also restricted ourselves to the 8C to the 90C stage in all datasets. For the scRNA-seq datasets, presumed outside cells were not removed from the analysis due to a lack of quantitative criterion to do so in the original studies.

Correlations - PCAs - fitting of mRNA counts to gamma distribution

Spearman correlations were computed using the cor() function of the dplyr package on R version 4.3.0[62]. To generate the PCA plots we used the prcomp() function of the stats package on R. We fitted the mRNA count distribution to gamma distributions using the fitdist() function of the fitdistr package, developed by Delignette-Muller et al. 2015, for R version 4.3.0, using the method of moments estimation. The KS-statistics and their p-values were computed using the ks.test() function available on R.

Entropy Estimation

To estimate the differential entropy of various single-cell RT-qPCR and scRNA-seq gene expression datasets we used the property that gene expression is typically subject to transcriptional bursting[5], leading to mRNA molecules of a gene following a gamma distribution[28,31–33]. The number of mRNA molecules have hence been fitted to a gamma distribution. The entropy of the gamma distribution, determined by its parameters α and β, is given by

$$H_{diff,\gamma} = \alpha - ln(\beta) + ln(\Gamma(\alpha)) + (1-\alpha)\Psi(\alpha) \quad (4)$$

where α is the shape parameter, β the rate parameter of the gamma distribution, Γ(x) is the gamma function, and Ψ(x) is the digamma function. We use the method of moments to obtain the parameters of the distribution. We

computed the mean μ and standard deviation σ of the empirical data. These two moments are related to the parameters of the gamma distribution by

$$\alpha = \frac{\mu^2}{\sigma^2} \text{ and } \beta = \frac{\mu}{\sigma^2} \tag{5}$$

To obtain a measure of uncertainty on the estimation of entropy, we performed bootstrapping. For each dataset, at every cellular stage, and for every gene, we resampled the same number of data (corresponding to the initial number of cells at that stage, in that dataset), computing their mean and standard deviation, relating them to the parameters of the gamma distribution, then computing the inter-cellular entropy. This was repeated 1000 times for every gene, at every cellular stage, in each dataset, to obtain the distributions of inter-cellular entropy. The resulting mean entropy was plotted for these 1000 re-samples, and the error was taken to be the standard variation of the 1000 entropy estimates.

CARDAMOM

The CARDAMOM software was cloned from https://github.com/eliasventre/cardamom as of 10/10/2024. It was run using python 3.9.13.

CARDAMOM inference was performed by merging the 5 single-cell datasets used in the present study. To this end, we transformed Guo et al.'s (2010) dataset[4], as explained in the *Datasets and Pre-processing* subsection, to make it quantitatively similar to other datasets when compared on PCAs and UMAPs.

Data availability

All datasets analyzed are available from their respective initial publications, as described in the Methods Section. The codes for this analysis can be found on github: https://github.com/LeviCarpet/Differential-Entropy-and-Network-Inference/.

Received: 25 March 2025; Accepted: 23 May 2025
Published online: 06 June 2025

References

1. Papili Gao, N., Gandrillon, O., Paldi, A., Herbach, U. & Gunawan, R. Single cell transcriptional uncertainty landscape of cell differentiation. *F1000* **12**, 426 (2023).
2. MacArthur, B. D. & Lemischka, I. R. Statistical mechanics of pluripotency. *Cell* **154**, 484–490 (2013).
3. Richard, A. et al. Single-cell-based analysis highlights a surge in cell-to-cell molecular variability preceding irreversible commitment in a differentiation process. *PLOS Biol.* **14**, e1002585 (2016).
4. Teschendorff, A. E. & Enver, T. Single-cell entropy for accurate Estimation of differentiation potency from a cell's transcriptome. *Nat. Comm.* **8**, 15599 (2017).
5. Tunnacliffe, E. & Chubb, J. R. What is a transcriptional burst? *Trends Genet.* **36**, 288–297 (2020).
6. Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E. & Huang, S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature* **453**, 544–547 (2008).
7. Ohnishi, Y. et al. Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. *Nat. Cell. Biol.* **16**, 27–37 (2014).
8. Semrau, S. et al. Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nat. Comm.* **8**, 1096 (2017).
9. Rosales-Alvarez, R. E. et al. VarID2 quantifies gene expression noise dynamics and unveils functional heterogeneity of ageing hematopoietic stem cells. *Genome Biol.* **24**, 148 (2023).
10. Toh, K., Saunders, D., Verd, B. & Steventon, B. Zebrafish neuromesodermal progenitors undergo a critical state transition *in vivo*. *iScience* **25**, 1005216 (2022).
11. Pelaez, N. et al. Dynamics and heterogeneity of a fate determinant during transition towards cell differentiation. *eLife* **4**, e08924 (2015).
12. Plusa, B., Piliszek, A., Frankenberg, S., Artus, J. & Hadjantonakis, A. K. Distinct sequential cell behaviours direct primitive endoderm formation in the mouse blastocyst. *Development* **135**, 3081–3091 (2008).
13. Chazaud, C., Yamanaka, Y., Pawson, T. & Rossant, J. Early lineage segregation between epiblast and primitive endoderm in mouse blastocysts through the Grb2-MAPK pathway. *Dev. Cell.* **10**, 615–624 (2006).
14. Guo, G. et al. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell.* **18**, 675–685 (2010).
15. Saiz, N., Williams, K. M., Seshan, V. E. & Hadjantonakis, A. K. Asynchronous fate decisions by single cells collectively ensure consistent lineage composition in the mouse blastocyst. *Nat. Comm.* **7**, 13463 (2016).
16. Yamanaka, Y., Lanner, F. & Rossant, J. FGF signal-dependent segregation of primitive endoderm and epiblast in the mouse blastocyst. *Development* **137**, 715–724 (2010).
17. Kang, M., Piliszek, A., Artus, J. & Hadjantonakis, A. K. FGF4 is required for lineage restriction and salt-and-pepper distribution of primitive endoderm factors but not their initial expression in the mouse. *Development* **140**, 267–279 (2013).
18. Azami, T. et al. Regulation of the ERK signalling pathway in the developing mouse blastocyst. *Development* **146**, dev1771339 (2019).
19. Fischer, S. C., Corujo-Simon, E., Lilao-Garson, J., Stelzer, E. H. K. & Munoz-Descalzo The transition from local to global patterns governs the differentiation of mouse blastocysts. *PLoS One*. **15**, e0233030 (2020).
20. Saiz, N. et al. Growth-factor-mediated coupling between lineage size and cell fate choice underlies robustness of mammalian development. *eLife* **9**, e56079 (2020).
21. Cherry, A. B. & Daley, G. Q. Reprogramming cellular identity for regenerative medicine. *Cell* **148**, 1110–1122 (2012).
22. Bessonnard, S. et al. Gata6, Nanog and Erk signaling control cell fate in the inner cell mass through a tristable regulatory network. *Development* **141**, 3637–3648 (2014).
23. De Mot, L. et al. Cell fate specification based on tristability in the inner cell mass of mouse blastocysts. *Biophys. J.* **110**, 710–722 (2016).
24. Tosenberger, A. et al. A multiscale model of early cell lineage specification including cell division. *NPJ Syst. Biol. Appl.* **3**, 1–11 (2017).

25. Robert, C., von Bonhorst, P., De Decker, F., Dupont, Y., Gonze, D. & G., and Initial source of heterogeneity in a model for cell fate decision in the early mammalian embryo. *Interface Focus*. **12**, 20220010 (2022).
26. Huang, S., Guo, Y. P., May, G. & Enver, T. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Dev. Biol.* **305**, 695–713 (2007).
27. Allègre, N. et al. NANOG initiates epiblast fate through the coordination of pluripotency genes expression. *Nat. Commun.* **13**, 3550 (2022).
28. Gandrillon, O. et al. Entropy as a measure of variability and stemness in single-cell transcriptomics. *Curr. Op Syst. Biology*. **27**, 100348 (2021).
29. Stumpf, S. Stem cell differentiation as non-Markov stochastic process. *Cell. Syst.* **5**, 268–282 (2017).
30. Wiesner, K., Teles, J., Hartnor, M. & Peterson, C. Hematopoietic stem cells: entropic landscapes of differentiation. *Interface Focus*. **8**, 20190040 (2018).
31. Albayrak, C. et al. Digital quantification of proteins and mRNA in single mammalian cells. *Mol. Cell*. **61**, 914–924 (2016).
32. Herbach, U., Bonnaffoux, A., Espinasse, T. & Gandrillon, O. Inferring gene regulatory networks from single-cell data: a mechanistic approach. *BMC Syst. Biol.* **11**, 105 (2017).
33. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4**, e309 (2006).
34. Goolam, M. et al. Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. *Cell* **165**, 61–74 (2016).
35. Posfai, E. et al. Position- and Hippo signaling-dependent plasticity during lineage segregation in the early mouse embryo. *eLife* **6**, e22906 (2017).
36. Wang, Y. et al. Single-cell multiomics sequencing reveals the functional regulatory landscape of early embryos. *Nat. Comm.* **12**, 1247 (2021).
37. Ventre, E. Reverse engineering of a mechanistic model of gene expression using metastability and Temporal dynamics. *Silico Biol.* **14**, 89–113 (2021).
38. Ventre, E., Herbach, U., Espinasse, T., Benoit, G. & Gandrillon, O. One model fits all: combining inference and simulation of gene regulatory networks. *PLoS Comput. Biol.* **19**, e1010962 (2023).
39. Sriram, K., Soliman, S. & Fages, F. Dynamics of the interlocked positive feedback loops explaining the robust epigenetic switching in Candida albicans. *J. Theor. Biol.* **258**, 71–88 (2009).
40. Schuh, L. et al. Gene networks with transcriptional bursting recapitulate rare transient coordinated high expression States in cancer. *Cell. Syst.* **10**, 363–378 (2020).
41. Trapnell, C. Defining cell types and States with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
42. Teschendorff, A. E. & Feinberg, A. P. Statistical mechanics Meets single-cell biology. *Nat. Rev. Genet.* **22**, 459–476 (2021).
43. Dussiau, C. et al. Hematopoeitic differentiation is characterized by a transient peak of entropy at a single-cell level. *BMC Biol.* **20**, 60 (2022).
44. Michalowicz, J. V., Nichols, J. M., Bucholtz, F. & & in *Handbook of Differential Entropy*. 28–29 (eds Press, C. R. C.) (Taylor & Francis Group, 2014).
45. Wicklow, E. HIPPO pathway members restrict SOX2 to the inner cell mass where it promotes ICM fates in the mouse blastocyst. *PLoS Genet.* **10**, e1004618 (2014).
46. Kuroda, Y. et al. Octamer and Sox elements are required for transcriptional cis regulation of Nanog gene expression. *Mol. Cell. Biol.* **25**, 2475–2485 (2005).
47. Avilion, A. A. et al. Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev.* **17**, 126–140 (2003).
48. Campolo, F. et al. Essential role of Sox2 for the establishment and maintenance of the germ cell line. *Stem Cells*. **31**, 1408–1421 (2013).
49. Duncan, S. A. et al. Expression of transcription factor HNF-4 in the extraembryonic endoderm, gut, and nephrogenic tissue of the developing mouse embryo: HNF-4 is a marker for primary endoderm in the implanting blastocyst. *Proc. Natl. Acad. Sci. USA* 91,7598–7602 (1994).
50. Chen, W. S. et al. Disruption of the HNF-4 gene, expressed in visceral endoderm, leads to cell death in embryonic ectoderm and impaired gastrulation of mouse embryos. *Genes Dev.* **8**, 2466–2477 (1994).
51. Taraviras, S., Monaghan, A. P., Schütz, G. & Kelsey, G. Characterization of the mouse HNF-4 gene and its expression during mouse embryogenesis. *Mech. Development*. **48**, 67–79 (1994).
52. Morrisey, E. E. et al. GATA6 regulates HNF4 and is required for differentiation of visceral endoderm in the mouse embryo. *Genes Dev.* **12**, 3579–3590 (1998).
53. Gagliardi, A. et al. A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. *EMBO J.* **32**, 2231–2247 (2013).
54. Liu, J., Song, Y. & Lei, J. Single-cell entropy to quantify the cellular order parameter from single-cell RNA-seq data. *Biophys. Rev. Lett.* **15**, 35–49 (2020).
55. Piras, V., Tomita, M. & Selvarajoo, K. Transcriptome-wide variability in single embryonic development cells. *Sci. Rep.* **4**, 7137 (2014).
56. Grün, D. et al. De Novo prediction of stem cell identity using single-cell transcriptome data. *Cell. Stem Cell.* **19**, 266–277 (2016).
57. Radley, A., Corujo-Simon, E., Nichols, J., Smith, A. & Dunn, S. J. Entropy sorting of single-cell RNA sequencing data reveals the inner cell mass in the human pre-implantation embryo. *Stem Cell. Rep.* **18**, 47–63 (2023).
58. Le Bin, G. C. et al. Oct4 is required for lineage priming in the developing inner cell mass of the mouse blastocyst. *Development* **141**, 10011010 (2014).
59. Larsson, A. J. et al. Genomic encoding of transcriptional burst kinetics. *Nature* **565**, 251–254 (2019).
60. Ramsköld, D. et al. Single-cell new RNA sequencing reveals principles of transcription at the resolution of individual bursts. *Nat. Cell. Biol.* **26**, 1725–1733 (2024).
61. Simon, C. S., Hadjantonakis, A. K. & Schröter, C. Making lineage decisions with biological noise: lessons from the early mouse embryo. *WIREs Dev. Biol.* **7**, e319 (2018).
62. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. dplyr: A grammar of data manipulation. (2023). https://www.tidyverse.org

Author contributions

Conceptualization & Investigation: FPvB, OG, UH, CC, CR, YDD, DG, GD. Softwares: FPvB. Manuscript writing: FPvB and GD. Manuscript review: All authors. All authors have read and approved the final version of the manuscript.

Funding

This work was supported by the ARC project "Noise sensitivity of GRN underlying cell fate specification" financed by the Université Libre de Bruxelles (ULB). CC and GD acknowledge support from the INSERM-IRP DiffEpi. GD is Research Director at the Belgian "Fonds de la Recherche Scientifique (FRS-FNRS)".

Declarations

Competing interests
The authors declare no competing interests.

Ethical approval
This article does not contain any studies with human participants or animals performed by any of the authors.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-03956-y.

Correspondence and requests for materials should be addressed to F.P.v.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.