bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pattern analysis of pluripotency reveals the molecular basis of naïve, primed and formative states

Patrick S Stumpf[1,2,*] and Ben D MacArthur[1,2,3]

[1] *Centre for Human Development, Stem Cells and Regeneration, Faculty of Medicine, University of Southampton, SO17 1BJ, United Kingdom*

[2] *Institute for Life Sciences, University of Southampton, SO17 1BJ, United Kingdom*

[3] *Mathematical Sciences, University of Southampton, SO17 1BJ, United Kingdom*

[*] *Correspondence to ps.stumpf@soton.ac.uk*

Abstract

The molecular regulatory network underlying stem cell pluripotency has been intensively studied, and we now have a reliable ensemble model for the 'average' pluripotent cell. However, evidence of significant cell-to-cell variability suggests that the activity of this network varies within individual stem cells, leading to differential processing of environmental signals and variability in cell fates. Here, we adapt a method originally designed for face recognition to infer regulatory network patterns within individual cells from single-cell expression data. Using this method we identify three distinct pluripotency configurations – corresponding to naïve, primed and 'formative' states – and associate these configurations with particular combinations of regulatory network activity archetypes that govern different aspects of the cell's response to environmental stimuli, cell cycle status and core information processing circuitry. These results show how variability in cell identities arise naturally from alterations in underlying regulatory network dynamics and demonstrate how methods from machine learning may be used to better understand single cell biology, and the collective dynamics of cell communities.

Introduction

The pluripotent epiblast exists transiently in the developing embryo and is the founding tissue for all somatic and germ cells in the adult mammalian organism [1;2]. Because of this remarkable ability there has been sustained interest in deciphering the molecular regulatory mechanisms that underpin pluripotency [3]. From these studies, it has become increasingly clear that the functional state of pluripotency emerges in a complex, and as yet incompletely understood, way from the collective dynamics of underpinning molecular regulatory networks, which involve numerous protein-protein, protein-DNA, epigenetic and signaling interactions [4–10].

The nature of the regulatory relationships in these underlying networks have accordingly become a focus of increasing research attention [11;12]. Typically, regulatory interactions are inferred from measurements taken

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from cellular aggregates, usually containing many thousands of cells, and therefore provide an ensemble view that characterizes those interactions that are typical for the 'average' pluripotent cell [13]. These ensemble models have been tremendously useful in dissecting the molecular basis of pluripotency and have become successively refined in recent years [4;5] to include, for example, the processing logic of combinatorial interactions [11;12].

However, although undoubtedly powerful tools to understand pluripotency, these networks are fundamentally derived from bulk cell measurements and there is now a need to better understand how these ensemble models relate to regulatory processes within individual pluripotent cells [14–17].

The relationship between ensemble and individual cell regulatory networks are particularly relevant to the study of pluripotency for two reasons.

Firstly, it is now well observed that apparently functionally homogeneous pluripotent cells exhibit substantial cell-to-cell variability in gene/protein expression patterns, suggesting that pluripotency as a function is compatible with a variety of different molecular configurations [18–20]. This has led to acceptance that there are numerous alternate states of pluripotency – most notably naïve and primed states corresponding to the epiblast of the blastocyst, and the epiblast in the egg cylinder of the mouse post-implantation embryo respectively – each with subtly different developmental potential. Our understanding is such that propagation of these alternate pluripotent states *in vitro* is now routine, using different cocktails of growth factor supplementation [21–26]. Importantly, these distinct populations can each contribute to all principal embryonic lineages *in vitro* and are apparently inter-convertible [25;27;28], suggesting a remarkable plasticity in the dynamics of the underlying regulatory networks. It seems likely that as our understanding of pluripotency develops, other varieties of pluripotency will be discovered and sustained *in vitro*. Indeed, it has recently been proposed that pluripotent cells also progress through an important *formative* state, in which the naïve regulatory network is partially dissolved and cells become competent for lineage allocation [29;30].

Secondly, the epiblast appears insensitive to the removal or addition of cells [1], suggesting a level of functional redundancy between individual cells that is supportive of the notion that pluripotent cell populations *in vivo* behave more like a 'collection of transition cells' [1], than a defined developmental state *per se*. This collective behavior presumably also emerges from the dynamics of the underlying regulatory networks, although the mechanisms by which such collective dynamics are regulated by intracellular regulatory networks is still largely mysterious [31]. Taken together, these findings suggest that the regulatory network underlying pluripotency exists in a number of interchangeable configurations, although the nature of these different configurations, and their relationships to one another, are not yet fully understood [14;16].

Here, we sought to develop a method to interpret single cell data to better understand how alterations in regulatory network activity within individual cells gives rise to variability within pluripotent cell populations.

To approach this problem, we were inspired by a method from the early days of face recognition, which de-constructs facial images into facial archetypes, known as *eigenfaces*, that are learned from a training set of portraits, and reconstructs unseen faces as weighted sums of these learned eigenfaces [32;33] (see Fig. 1). Although face recognition methods are now highly sophisticated, the original implementation of the eigenface routine is essentially an ingenious, although mathematically straightforward, implementation of principal component analysis (PCA) that relies on the fact that each facial image may be considered as a matrix of numbers, and therefore reshaped to a vector and associated with a point in a high-dimensional space. Thus, given a set

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

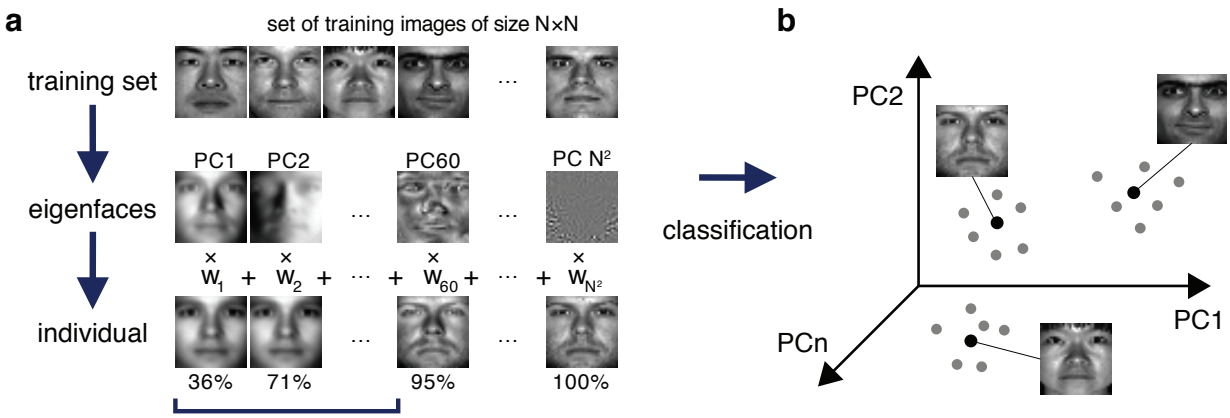


Figure 1: **Eigenfaces for face recognition.** (**a**) A training set of portrait images of size $N \times N$ is used to extract the facial archetypes (eigenfaces) encoded by the N^2 principal components of the training set. A small subset of eigenfaces explains most of the variability in facial features between individuals. In this specific example from The Extended Yale Face Database B[34;35], a recognizable version of an original test image can typically be reconstructed from a weighted sum of the first 5.9% (60 out of 1024) eigenfaces, which explain 95% of the variance in the data. (**b**) Each test face may be reconstructed as a weighted sum of eigenfaces, and thereby efficiently encoded by a weight vector, which may be thought of as a point in a much lower dimensional space than the original feature space. In this case although each face is initially associated with a point in a 1024 dimensional space (corresponding to the 1024 pixels in the original image), a recognizable version may be reconstructed in just 60 dimensions (the corresponding weightings). Different images of the same person typically occupy a region in the principal component space around a central characteristic image.

of training portrait images, PCA may be used to extract the characteristic features – the eigenvectors of the training covariance matrix, also known as principal components – that capture significant variation within the training set (Fig. 1a). By transforming these eigenvectors back into matrices of the same dimension as the images in the training set they can be visualized as facial archetypes (or 'eigenfaces') of the training set (Fig. 1a). Remarkably, it was observed that only a small number of eigenfaces (typically $\sim 5\%$) is sufficient to explain 95% of facial details, and therefore unseen portrait images can be reliably reconstructed as a weighted sum of a very small number of eigenfaces (Fig. 1a). Importantly, this means that the vector of weights alone (i.e. $\sim$5 numbers) is typically sufficient to recognize an individual from their portrait, thus significantly reducing the dimension of the recognition problem (Fig. 1b).

While this does not immediately appear to relate to the study of pluripotency we surmised that a similar approach could be used to reconstruct pluripotent cell identities from single cell data, as a weighted sum of regulatory network archetypes. Furthermore, just as portrait images can be efficiently encoded using the eigenface weight-vector, we wanted to determine if complex patterns of gene/protein expression within individual cells could similarly be encoded by a low-dimensional representation in terms of the activity of these network archetypes, thereby facilitating more accurate classification of cell identities from noisy expression data.

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Integrating regulatory interactions with single cell data

We first sought to obtain a reliable training dataset of protein expression patterns in pluripotent cells across multiple intracellular information levels, including the protein abundance of core transcription factors[4;5], the phosphorylation status of signaling pathways[6–8] and global transcriptional activity based on histone acetylation[9;10]. Such systems-level proteomic information at single-cell resolution is currently only available through immunolabeling followed by mass-cytometry, a highly specialized technique that is available to only a small number of groups[36]. Thus, we sourced a relevant training dataset from the literature[37]. In total this training data consists of expression patterns of 34 proteins and protein modifications in 31,876 pluripotent cells from two mouse embryonic stem cell (mESC) lines (Nanog-GFP [NG] mESCs and Nanog-Neo [NN] mESCs that express green fluorescent protein [GFP] or a Neomycin resistance gene respectively from the endogenous Nanog locus[38]), grown in low-serum medium supplemented with Leukemia Inhibitory Factor (LIF; 0i conditions). In addition, this dataset also contains expression levels of the same features in 15,540 NG mESCs and 15,752 NN mESCs grown in medium supplemented further with a GSK3β inhibitor and a MEK inhibitor (known as 2i conditions, which support the pluripotent 'ground' state[39]), as well as expression time-course data containing 834,548 secondary mouse embryonic fibroblasts (MEFs) generated from both cell lines that express Yamanaka reprogramming factors[40] under the control of a doxycycline (dox) inducible promoter[38].

To interrogate this data, we sought to supplement it by constructing a directed regulatory network specific to the features (transcription factors, surface epitopes, phosphorylation, etc.) that had been quantified (Fig. 2). Features (that is, proteins profiled) in this signed, directed regulatory network are represented as nodes and regulatory interactions between features are represented as edges between pairs of nodes (an edge is positive if it is activating, and negative if the edge is inhibiting). Evidence for node interactions was extracted from transcription factor binding data from ChIPBase 2.0[41], and information on other known interactions were sourced from the Kyoto Encyclopedia of Genes and Genomes (KEGG)[42] and Reactome[43] (see Table S1 for details). Unconnected nodes, such as the inert GFP reporter, and cell cycle markers pH3 and IdU were removed from the analysis. The resulting network G contains 27 nodes, connected by 124 edges (Fig. 2a).

The overall structure of G is conveniently encoded in the network adjacency matrix,

$$A_{ij} = \begin{cases} s, & \text{if nodes } i \text{ and } j \text{ are connected} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $s = +1$ for activating interactions, and $s = -1$ for inhibitory interactions.

The first step in our process consists of combining this regulatory network with the single cell expression training set. Trivially, the expression data represents the activity of the nodes in the network within each cell, but does not take into account regulatory interactions between nodes. To incorporate this information, we assumed that the activity of each edge within the network is determined by the signal intensities of both interaction partners within the individual cell. Accordingly, denoting the vector of expression values in a given

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

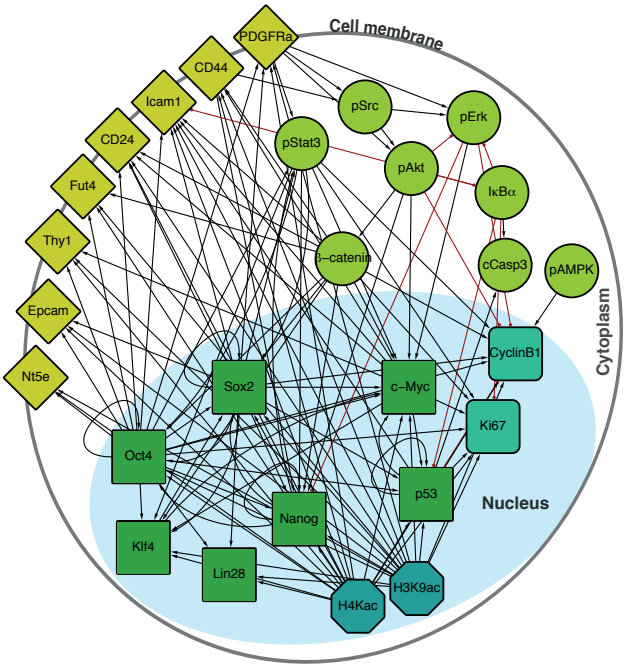


Figure 2: **Integrated regulatory network derived from the literature.** Schematic shows the structure of the inferred regulatory network between the factors profiled, derived from the literature (see Table S1). The network accounts for multiple molecular information processing mechanisms, at multiple different spatial locations in the cell, including interactions between: transcriptional regulators (green squares), chromatin modifiers (petrol octagons), cell cycle factors (sea green rounded squares), signalling cascades (light green circles), and surface molecules (yellow diamonds).

cell by $\boldsymbol{v}$, we created a weighted adjacency matrix $\boldsymbol{W}$

$$W_{ij} = \begin{cases} v_i \times v_j^s & \text{if nodes } i \text{ and } j \text{ are connected} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Thus, we associated a high weight to a positive edge if both the source and the target were highly expressed, and a high weight to a negative edge if the source was highly expressed and the target was expressed at a low level. Informally, this representation may be thought of as assigning high confidence that a given edge is expressed within an individual cell if its source and target nodes are expressed consistently with the sign of the edge relating them. The resulting weighted adjacency matrix $\boldsymbol{W}$ is a simple measure of the extent to which the network G is expressed in the cell given the expression patterns observed in that cell. By analogy with the face recognition problem, $\boldsymbol{W}$ may be considered as the 'image' of the cell.

As with the eigenface routine, this matrix may be easily restructured as a vector. In this case, $\boldsymbol{W}$ may be coerced into a vector of length m (where m is the number of edges in the network, here 124), by first reshaping it to a vector of length n^2 (where n is the number of nodes in the network, here 27), and then squeezing out all entries for which $A_{ij} = 0$. This procedure effectively injects the expression data with prior knowledge of the network structure, leading to an expansion of the original feature space from $\mathbb{R}^n$ to $\mathbb{R}^m$ (generically a connected network will have more edges than nodes, unless it is a tree). Using this method, we inferred the activity of the regulatory network G within each of the $\sim 9 \times 10^5$ individual cells, and treated NG mESCs cultured in 0i

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

conditions as a training dataset and held back the remaining data to test the model.

Regulatory networks characterize alternate states of pluripotency

Once the training data had been produced, we conducted principal component analysis. In the same way that the principal components (PCs) in the eigenface routine may be reshaped and interpreted as facial archetypes from which individual portraits may be reconstructed, the principal components here may be reshaped and interpreted as network archetypes from which pluripotent cell identities may be reconstructed. However, while only $\sim 5\%$ of the PCs are required for accurate face recognition, we found that (for both NG and NN mESCs) $\sim 23\%$ of the PCs were required to explain 95% of the variance in our training data (Fig. 3a). The larger number of PCs required is not unexpected, and is reflective of the high levels of noise that are characteristic of high-throughput single cell data[44]. Therefore, rather than using the proportion of variance explained to determine the appropriate number of PCs to retain for subsequent analysis, we sought to identify the minimal number needed to preserve the natural clustering structure in the data.

We found that four distinct clusters of cells were readily identifiable in the full dataset (natural clustering structure was obtained by fitting to a Gaussian mixture model to the data and selecting the most appropriate model, here with four components, by minimizing the Bayesian information criterion [BIC], see Fig. 3d and Fig. S1d). This natural clustering was robustly retained when projecting the data onto the first three PCs (Fig. 3b); higher components only added noise to this basic clustering structure. This analysis suggests that PCs 1-3 account for the biological variability present in the data, while higher components primarily correspond to technical variability.

Since the PCs are linear combinations of the underlying features (here, network edges) each one may be thought of as regulatory network archetype, and the expression pattern of each cell in the training data may therefore be reconstructed as a weighted sum of these archetypes. By analogy with eigenface routine, we will call these network archetypes *eigen-networks*. Since PCs 1-3 account for the biological variability in the data, the structure of the eigen-networks associated with these components are of particular interest. The first eigen-network (PC1 in Fig. 4a) naturally separated cells into two subsets (compare Fig. 3b), based upon overall activity of regulatory interactions (Fig. 3c). A subset of cells with low overall edge expression (cluster 1 in Fig. 3b) primarily contained apoptotic cleaved Casp3-positive cells (Fig. 3c) and cell cycle arrested cells (Fig. S1g) that lacked expression of the hallmarks of pluripotency, such as Oct4, Nanog and Klf4 (Fig. 3c). In contrast to this small subset, the majority of cells displayed high overall expression of pluripotency related-factors, including Oct4 (Fig. 3c).

The majority pluripotent population identified by the first eigen-network naturally separated into 2 distinct further sub-populations (clusters 2, 3 in Fig. 3b) by expression of the second eigen-network (PC2 in Fig. 4a), which broadly captures the strength of connection between the cell's signaling pathway activity and its core transcriptional regulatory circuitry, including activity of β-catenin (Wnt-signalling), Stat3-phosphorylation (LIF-signalling) and Erk-phosphorylation (FGF/MEK-signalling). This component therefore captures integration of the primary axes of extrinsic control of the pluripotent ground state[39], and distinguishes cells in the pluripotent ground state (cluster 3), which are characterized by high Nanog, Oct4 and Klf4 expression and strong integration of signaling and core transcriptional regulatory programs, from those in the primed pluripo-

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

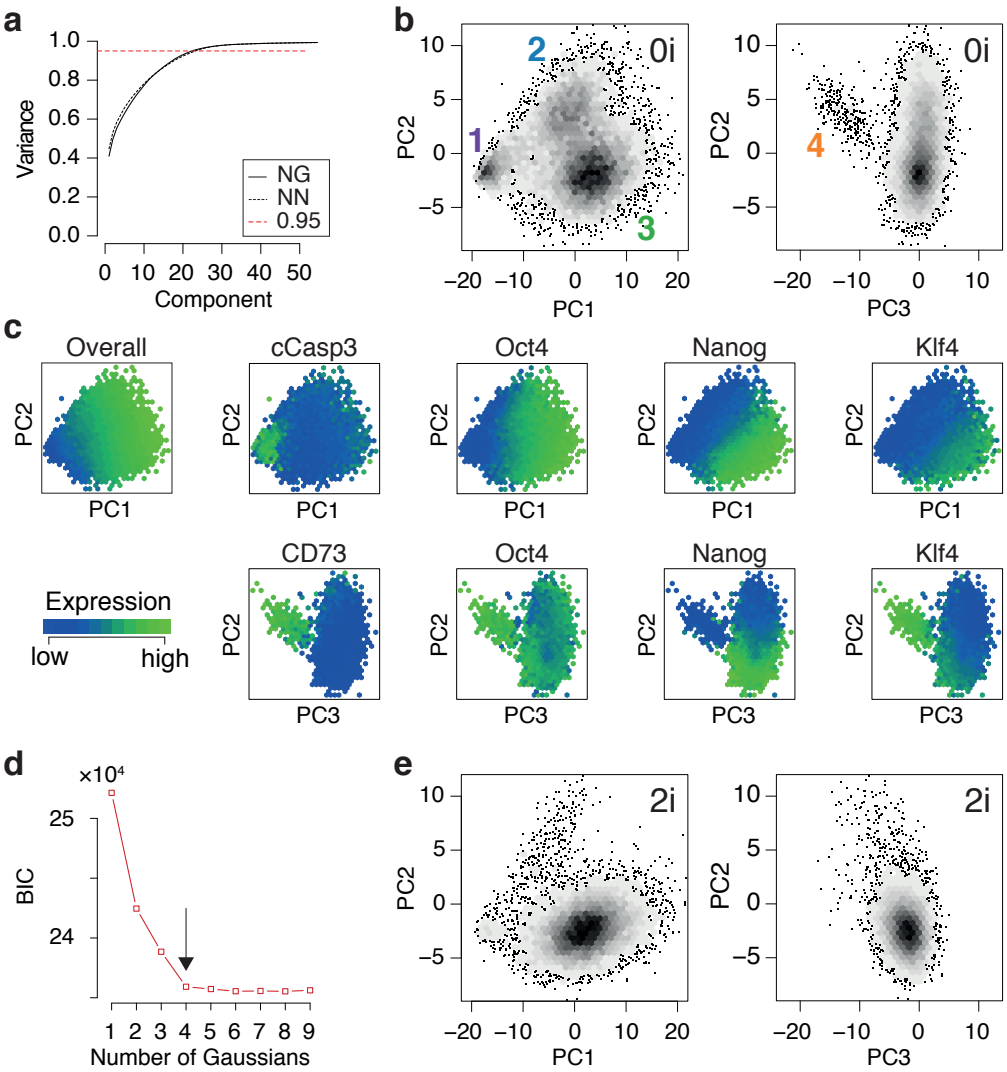


Figure 3: **PCA identifies three distinct pluripotent states within ES cells cultured in 0i conditions.** (**a**) Cumulative proportion of variance explained by principal components (PCs) of the training data for NG mESCs and NN mESCs respectively. The dotted red line marks the commonly used threshold value of 0.95. (**b**) Density plot of training data from NG mESCs projected onto the first three components. Four clear clusters are apparent, labeled 1-4, corresponding to distinct states of network activity. Each hexagonal bin contains at least 5 cells. (**c**) Heat map of expression of important nodes in NG mESCs projected onto PCs 1-3. Mean expression values are displayed for each hexagonal bin. Distinct alternate states of pluripotency are apparent, based upon edge co-expression patterns. (**d**) Bayes information criterion (BIC) as a function of the number of Gaussian mixture components fitted to the first three principal components. The arrow marks the elbow in the plot, indicating the optimal number of components (here 4). (**e**) Projection of a test dataset of expression patterns from NG mESCs cultured in 2i conditions onto the training PCs (panel **b**). Panels b-e show data from NG mESCs, corresponding data for NN mESCs is shown in Fig. S1.

tent state (cluster 2), which are characterized by low Nanog and Klf4 expression, and more sporadic connectivity between signaling and transcriptional controls (Fig. 3c).

In addition to these primary populations we also observed small subset of cells ($\sim$ 2%) that could be

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

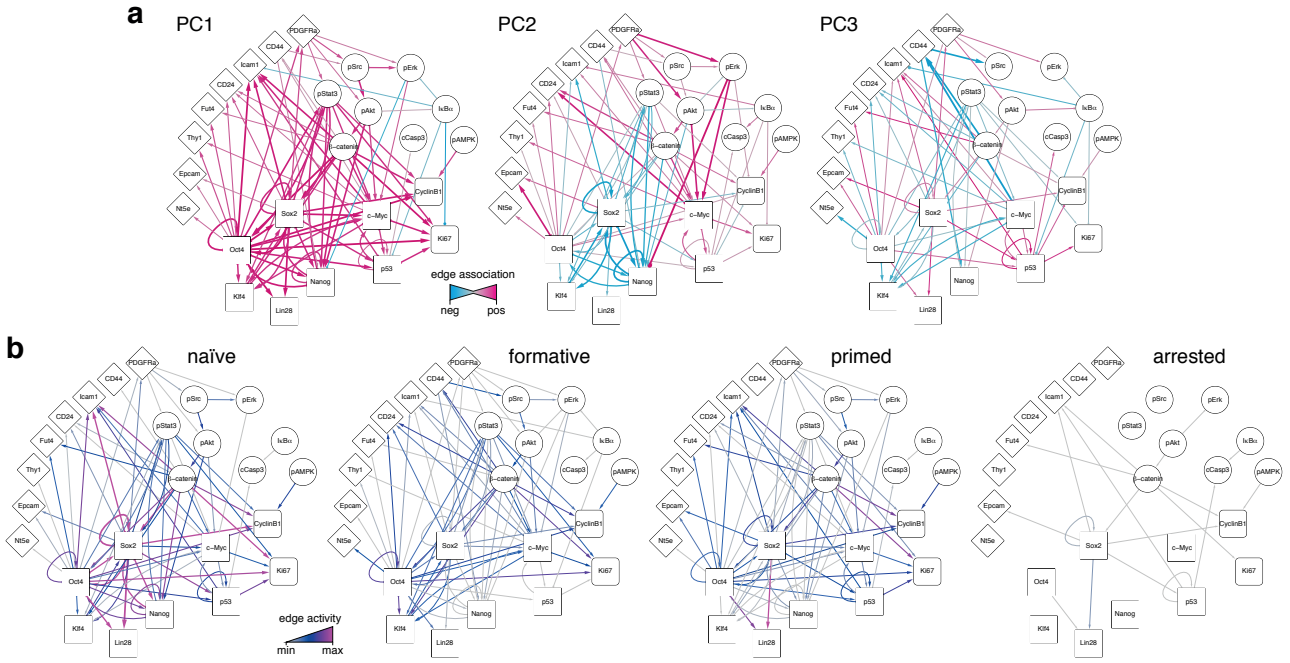


Figure 4: **Regulatory network activity archetypes define alternate pluripotency states.** (**a**) Graphical representation of the first three PCs, interpreted as regulatory network archetypes. Color and edge width indicate signed deviation from the mean. (**b**) Representative regulatory network states for naïve, formative, and primed pluripotency states. The network corresponding to arrested/apoptotic cells (cluster 1) in Fig. 3b is also shown for reference.

distinguished from the primed and naïve pluripotent states based on expression of the third eigen-network (see population 4 in Fig. 3b). This fourth population is similar to the primed state (population 2) with respect to expression of Nanog (both low; see Fig. 3c) and similar to the naïve state (population 3) with respect to expression of Klf4 (both high; see Fig. 3c). However, it is quite distinct with respect to a number of surface markers. Notably cells in cluster 4 are CD73^{high} (Nt5e; Fig. 3c), and CD44^{high} and CD54^{low} (not shown), suggesting an increased interaction with the extracellular matrix. These differences are not simply a manifestation of mitosis or cell cycle arrest, since the proportion of M-phase cells in this population is comparable to both the naïve and primed states and the proportion of G0-phase cells is comparable to the primed state (Fig. S1f-g). Rather, since Klf4 has been implicated in driving the transition from an epiblast stem cell phenotype to an embryonic stem cell phenotype[27], we conjecture that this population is an intermediate between the naïve and primed pluripotent states. In accordance with this notion, we observe that it has the highest total within cluster variance, indicating the presence of substantial cell-cell variation (see Fig. S1e), which is typically found in cells transitioning from one state to another[45]. While the full nature of this state has yet to be determined, it is consistent with the recently proposed 'formative' phase of pluripotency, characterized by dissolution of core pluripotency sustaining mechanisms and increased interaction with the extracellular matrix[30].

To investigate this possibility further we constructed representative networks for each of the four identified states using the first three eigen-networks and the weight vector corresponding to the centroid for each cluster (see Fig. 4b). The resulting networks may be thought of as representations of the characteristic patterns of

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

network activity within each of the four states we identified. These networks show that: (1) the naïve state is characterized by strong co-regulatory activity between members of the core transcriptional circuit and strong integration of signalling pathways with this core sub-network (Fig. 4b). (2) By contrast, the formative state is characterized by partial dissolution of the core transcriptional circuit (in particular a loss of Nanog, Sox2 and p53 activity), which is accompanied by changes in cell-cell (CD54) and cell-matrix (CD73, CD44) mediated signaling. However, cells in this state continue to perceive environmental signals via the LIF/Stat3 signaling pathway (Fig. 4b), indicating continued receptivity to environmental cues. (3) The subsequent primed state is marked by a further dissolution of the core transcriptional circuit, including the loss of Klf4 regulatory activity (Fig. 4b) and a decrease in LIF/Stat3 signaling (Fig. 4b), suggesting that these cells are transitioning out of pluripotency. Accordingly, the primed state is also marked by the positive regulation of EpCAM (Fig. 4b), suggesting the onset of cell polarization, as is observed in the epiblast of the egg cylinder *in vivo*[46].

In summary, this analysis revealed the presence of four distinct cellular communities, each characterized by different levels of activity of regulatory network archetypes, within mouse ES cell populations cultured in 0i conditions. To determine how general these results were we also examined network expression patterns mESCs cultured in 2i conditions, which stimulate Wnt signaling activity and reduce Erk-phosphorylation using small molecule inhibitors of MEK, and thereby shield the core transcriptional circuitry from extrinsic differentiation cues[39]. In accordance with the nature of these conditions we found that populations 1, 2 and 4 (corresponding to differentiated, primed and formative pluripotency states) were comprehensively depleted in mESCs cultured in 2i conditions, while cluster 3 (corresponding to the naïve state) was robustly maintained (Fig. 3e). These results re-affirm the potency of these conditions to purify the ground state of pluripotency, and provide mechanistic insight into the molecular mode of action of these conditions.

Taken together these results indicate the presence of three distinct states of pluripotency each distinguished by characteristic patterns of regulatory network activity within individual cells.

Individual cells transition through distinct network activity states during reprogramming

To further investigate the biological importance of the regulatory network archetypes we had identified we then sought to determine their temporal expression during cellular reprogramming of somatic cells to pluripotency.

During cellular reprogramming, pluripotency regulatory network activity is typically initially established through the ectopic expression of four trans-genes, Oct4, Sox2, Klf4 and c-Myc (OSKM)[40]. Subsequently, the concerted action of these core reprogramming factors leads to profound changes to the cellular phenotype, ultimately re-instating a self-sustaining pluripotent identity in a small proportion of cells. The dynamics of this process are thought to be initially driven by low frequency stochastic events followed by the deterministic progression through a series of characteristic intermediate, partially reprogrammed, expression states[47]. It is presumed that these intermediate partially reprogrammed states correspond to partial re-configurations of the pluripotency regulatory network[48]. However, the relationships between regulatory network reconfigurations and the dynamics of reprogramming are not well understood.

To address this issue, we considered data from a reprogramming time-course in which the expression of ectopic OSKM transgenes were induced in secondary MEFs by doxycycline (dox) supplementation of the MEF

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

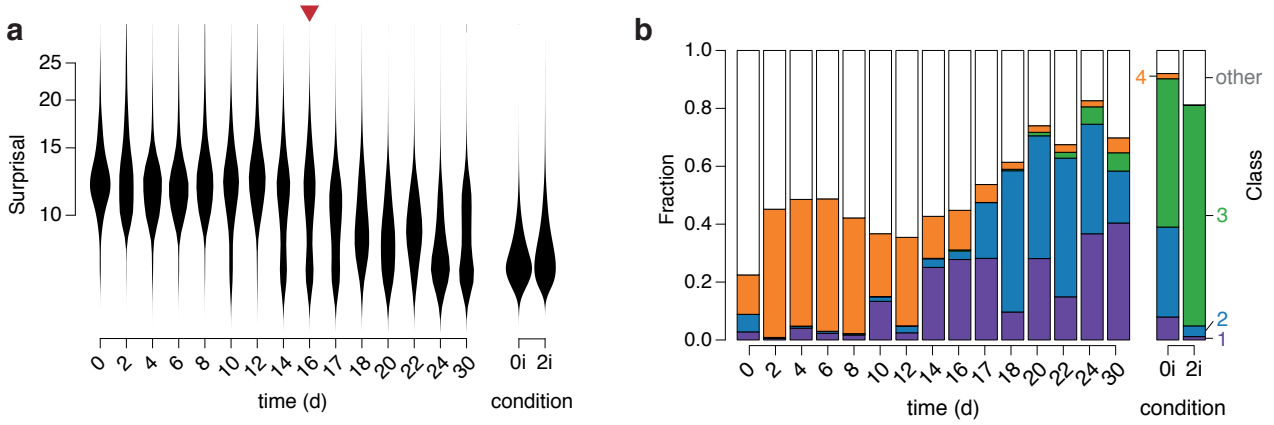


Figure 5: **Dynamics of regulatory network activity during cellular reprogramming.** (**a**) Violin plots of changes in surprisal (Eq. (3)) over time. A gradual decrease in surprisal in the population accompanies cellular reprogramming. The red arrow marks the end of doxycycline treatment. (**b**) The fraction of cells classified into each of the four clusters identified in the training data. Class labels are as in Fig. 3b.

culture medium for 16 days, followed by a further 14 days in 0i conditions without dox[37].

To analyze this data we first fit our training data (expression patterns of NG mESCs cultured in 0i conditions) projected onto the first three eigen-networks (as described above) with a Gaussian mixture model (GMM) with four components. This GMM may be thought of as an estimate of the joint probability density function $\mathcal{P}(\boldsymbol{x})$ for the training data, projected onto the first three PCs (where $\boldsymbol{x} \in \mathbb{R}^3$ identifies points in PC space). We then projected the reprogramming time-course data onto the first three PCs from the training data and used the fitted GMM to estimate the likelihood of observing the expression patterns seen in the reprogramming time-course within the pluripotent cell population. That is, if $\boldsymbol{v}$ is the expression pattern of a given cell in the reprogramming time-course projected onto PCs 1-3 from the training data, we calculated $\mathcal{P}(\boldsymbol{v})$ as a measure of the likelihood of observing $\boldsymbol{v}$ in the training population. The negative logarithm of this probability

$$S(\boldsymbol{v}) = -\log_2 \mathcal{P}(\boldsymbol{v}) \tag{3}$$

is the amount of information imparted by observation $\boldsymbol{v}$ with respect to the probability measure $\mathcal{P}$[49]. Informally, $S(\boldsymbol{v})$ is a measure of the 'surprisal' of observing the expression pattern $\boldsymbol{v}$ in a pluripotent population: cells that express proteins in a pattern similar to that often seen in pluripotent cells have a low surprisal; while cells that express proteins in a pattern that is unusual for pluripotent cells have a high surprisal. To obtain assessment of the dynamics of reprogramming, we calculated the surprisal for each of the 263,692 NG cells in the reprogramming time-course, and monitored how the distribution of surprisal in the population changed over time during reprogramming.

We first observed that the surprisal remained high, and approximately constant, for the first 10-12 days of reprogramming (Fig. 5a), indicating that cells in the starting population (in this case NG MEFs) consistently exhibited expression patterns that are unusual for pluripotent cells, as expected. However, around days 10-12 the population split into two distinct sub-populations: a majority sub-population in which the surprisal remained high, and a minority sub-population in which the surprisal was substantially reduced, suggesting the emergence

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of population of pioneer partially reprogrammed cells (Fig. 5a). Over the next $\sim$ 20 days the proportion of cells in the low surprisal sub-population gradually increased, indicating the consolidation and proliferation of a robustly pluripotent population of cells (Fig. 5a).

To better understand the identity of this emerging pluripotent sub-population we sought to relate it to the three alternate pluripotency states we had identified (see Fig. 5). To do so we used our fitted GMM to classify each cell in the time-course into one of the four populations identified in the training data (Fig. 5b). Since numerous cells, particularly at the beginning of the time-course, did not resolve well onto any of the clusters in the training data (which is to be expected, since they are not pluripotent) we also incorporated a fifth class to capture those cells with network activity states that were distinct from those found in the training data (for details see Methods).

This analysis revealed that specific instances of regulatory network activity define distinct phases of the reprogramming process (Fig. 5b).

Initially, while the majority of cells were unclassified, indicating lack of similarity to all of the pluripotent training populations, a small proportion of cells were associated with the fourth cluster, corresponding to the formative state of pluripotency. This observation is not unexpected as cells in the formative state express endogenous OSKM in addition surface markers associated with mesenchymal states (see Fig. 3c). Similarly, in the presence of dox, MEFs express the OSKM transgenes and begin undergoing the mesenchymal-to-epithelial transition[50], at which point they start to adopt a regulatory configuration similar to the formative state with respect to the limited number of features quantified in this data.

This initial phase is followed by the emergence of a population of cells in cluster 1 (corresponding to arrested or apoptotic cells that are frequently observed in reprogramming[51]) from day 10-14, followed closely by the emergence of a population of cells in cluster 2 (corresponding to the primed pluripotent state) from day 17 and lastly, the emergence of a small population of fully reprogrammed cells in cluster 3 (corresponding to the naïve pluripotent state) after 22 days.

These data suggest that reprogrammed cells do not emerge in significant numbers until after after dox is withdrawn, at which point the pluripotency regulatory network begins to assume a more natural configuration. These observations are in accordance with the notion that activation of the OSKM transgenes prevent cells from entering a stabilization phase of reprogramming in which the pluripotent state becomes fully established[48]. Notably, at around the same time there is an apparent reduction in the frequency of cluster 4 cells, which are marked by low Sox2 and p53 activity, indicating that these cells only exist transiently during reprogramming. Since this population is more variable than the naïve and primed pluripotent populations, it may also mark the handover from the early stochastic phase of reprogramming, in which the activation of OSKM transgenes initiate transformation, to the late deterministic phase, in which the pluripotent cell identities are consolidated by endogenous regulatory mechanisms[47].

Taken together these results indicate that reprogrammed MEF cells enter pluripotency via the formative state. It remains to be seen if this is a general characteristic of reprogramming, or if this particular route is due to the fact that the MEF starting population has a mesenchymal origin that happens to be more similar to the formative state than it is to the other pluripotent identities.

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

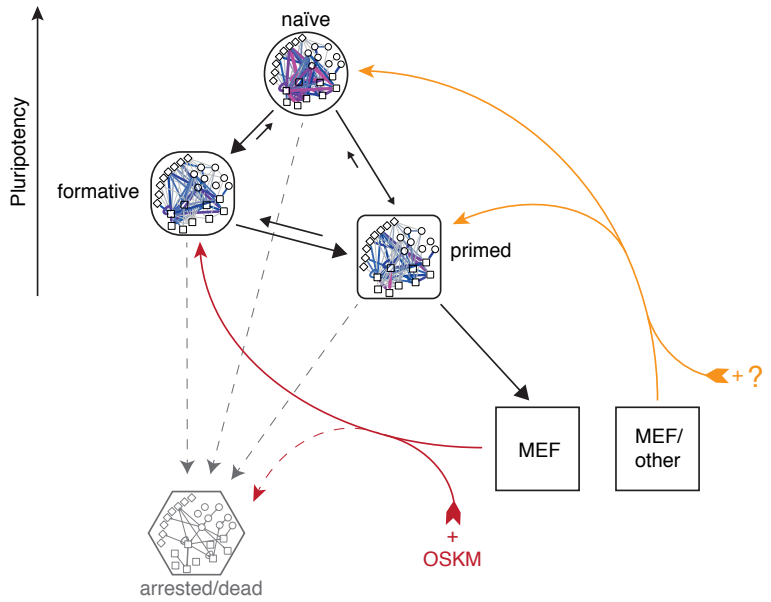


Figure 6: **Proposed topography of pluripotency states.** Pluripotent stem cells descend a natural hierarchy of distinct regulatory network activity states (black arrows). Despite this natural hierarchy, 0i culture conditions permit the inter-conversion of these different network configurations *in vitro*. Somatic cell reprogramming re-establishes a network configuration similar to that of the formative state, from which cells replenish the remaining pluripotency states (red arrows). A different cocktail of reprogramming factors from different starting populations may enable direct reprogramming to the primed or naïve states (orange arrows). Subset of cells undergo cell cycle arrest or apoptosis (dashed arrows).

Discussion

The notion that there is a single well-defined pluripotent stem cell identity has been rapidly eroded by advances in single cell analysis methods, which are now revealing ever greater varieties of pluripotency [18–20;39]. Collectively, these results suggest that pluripotency is not a single phenotype but instead is a property that spans a continuum of observable cell states [1;30;39;52–54]. This is in part because the densely connected pluripotency regulatory network is rich in feedback loops which both stabilize pluripotency, and endow pluripotent cells with a remarkable phenotypic plasticity [5;55]. Hence, to fully understand pluripotency, strategies to decipher regulatory networks at single cell resolution are needed.

There have been a number of notable advances to this end, particularly with regard to methods for inferring and analyzing regulatory networks directly from single cell data, which can reveal aspects of regulatory control that are inaccessible to study with ensemble techniques such as ChIP-Seq [14;47;54;56]. For example, Trott and co-workers have inferred regulatory network activity from correlation patterns in single cell data in different stem cell sub-populations, and related these different activity patterns to different aspects of the stem cell identity [14]. Similarly, Stumpf (not the current author) and colleagues have used powerful notions from information theory to more precisely identify regulatory interactions from single cell time-course data [56]. However, single cell data is inherently noisy, and consequently large numbers of cells are needed to gain the statistical power to accurately distinguish functional from spurious interactions [56].

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To circumvent this problem here we have presented a method that incorporates prior knowledge of regulatory interactions directly into single cell expression patterns, rather than inferring regulatory interactions from the data itself, and uses this prior knowledge to dissect the regulatory processes that give rise to different states of pluripotency. This approach is similar to that taken by Teschendorff and colleagues, who, by projecting single cell data onto a known regulatory network, find that pluripotency can be remarkably well related to systems-level emergent network properties[17]. We anticipate that as single cell profiling methods develop we will see concurrent advances in the statistical methods needed to investigate and interrogate the resulting data: indeed, new statistical advances will be *essential* to fully realize the power of these new and emerging technologies. We expect that Bayesian methods, which use known regulatory interactions as a prior to guide learning of functional interactions directly from single cell data, will combine the benefits of the two approaches to this problem and may therefore be particularly powerful.

In summary, we have adapted a simple image analysis method to infer the presence of four distinct patterns of pluripotency, based on the activity patterns of three regulatory network archetypes within individual cells. The power of our method is not due to its mathematical or computational sophistication – indeed, it is mathematically and computationally straightforward – but rather in the biological interpretation it allows. As such it provides a simple example of how methods from machine learning may be easily adapted to address biological questions in an intuitive way. In particular, using this method we have identified a novel pluripotent state, which appears to be an intermediate between the well-known naïve and primed states (see Fig. 6) and shares many of the putative properties of a recently proposed 'formative' state[30]. Cells in this state are still dependent on LIF/Stat3 signaling yet are characterized by partial dissolution of the core transcriptional regulatory circuit and distinct changes in cell-cell and cell-matrix interactions. Furthermore, these cells only appear at low frequency in 0i culture conditions and transiently during the early stages of cellular reprogramming of MEFs to pluripotency. Taken together these results suggest that this 'formative' state is a temporary intermediate in which the feedback mechanisms that stabilize the core pluripotency circuit become weakened and cells begin to become competent for lineage allocation. It remains to be seen how the population we have identified relates to recent observations of formative pluripotency characterized by loss of Rex1 expression and genome wide reorganization[57]. We anticipate that the coming years will see greater advances in single cell profiling and analysis methods that will enable us to address this question, and identify with greater precision the regulatory networks that control the maintenance and exit from pluripotency.

Materials and methods

Single-cell expression data

Expression data from Zunder *et al.* (2015)[37] was retrieved from the Cytobank repository (accession no. 43324). In summary, these data contain measurements of 46 features taken at the single-cell level by mass cytometry, from two separate engineered mouse embryonic stem cell (mESC) lines NG (Nanog-GFP) and NN (Nanog-Neomycin). Each mESC line contains doxycycline (dox) inducible gene cassettes for *Oct4*, *Sox2*, *Klf4* and *c-Myc* used for secondary reprogramming to pluripotency from somatic mouse embryonic fibroblasts (MEFs). Data includes the expression profiles of mESCs in steady state pluripotent stem cell culture conditions containing

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

either Serum/LIF (denoted 0i) or Serum/LIF supplemented with 3μM GSK3 inhibitor CHIR-99021 and 1μM MEK inhibitor PD-0325901 (denoted 2i). Furthermore, time-course data comprised of snapshots of MEFs undergoing 16 days of dox treatment in MEF medium (DMEM, 10% serum) followed by 14 days without dox (123 medium + LIF)[37]. De-barcoded raw data was processed in R[58] version 3.3.2 using the flowCore[59] package version 1.40.4. Relevant features were logicle-transformed with parameters $w = 0.6$, $t = 10,000$ and $m = 4.5$.

Cell cycle analysis

Classification of cell cylce status was performed based on the expression levels of Ki67 (absence indicates G0), phosphorylation of Histone H3 (presence indicates M) as described in Figure 4c of Zunder *et al.* (2015)[37]. Classification of G1-, G2- and S-phase was not possible due to a lack of discernible modes for marker IdU.

Ensemble regulatory network

An ensemble model of binary node interactions (valid for an abstract average cell) was derived from publicly available data. Transcription factor binding data was derived from ChIPBase 2.0[41], and information on other known interactions were sourced from KEGG[42] and Reactome.org (see **Table S1**).

Statistical analysis

Principal components analysis

Principal components analysis of scaled and centered training data (expression from mouse ES cells cultured in 0i conditions, see above) was conducted in R[58] using the *prcomp* function.

Gaussian mixture model

Gaussian mixture models were constructed in R[58] using the *Mclust* package version 5.2.2[60]. Fit quality was assessed using the Bayesian information criterion (BIC). Minimum BIC indicates the best model fit, however, models with a higher number of parameters often only provide marginally better fits and the overall quality approaches a natural limit. Optimal trade off between increased parameters and quality of fit was obtained by selecting the model corresponding to the 'elbow' in the plot of fit quality against number of components.

Density estimation

Estimate of the probability density function corresponding to the GMM identified above was obtained using the *densityMclust* function in R[58]. Probability density estimates were calculated using the *predict* method in R[58].

Classification

The GMM identified above was used for classification of data into either of four categories based on the highest posterior probability in combination with a reject option to avoid misclassification of vastly dissimilar phenotypes. Thus, points outside the 90th percentile for all individual multivariate Gaussian distributions were rejected as outliers.

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Software and computer code

Analyses were performed in R[58] version 3.3.2. Computer code used in this study is available as a R-markdown file from https://github.com/passt/Eigen-Networks.

Data availability

Data used in this study is available from Cytobank (accession 43324).

References

[1] Gardner, R. L. & Beddington, R. S. Multi-lineage 'stem' cells in the mammalian embryo. *J Cell Sci* **10**, 11–27 (1988).

[2] Boroviak, T., Loos, R., Bertone, P., Smith, A. & Nichols, J. The ability of inner-cell-mass cells to self-renew as embryonic stem cells is acquired following epiblast specification. *Nat Cell Biol* **16**, 516–528 (2014).

[3] Li, M. & Belmonte, J. C. I. Ground rules of the pluripotency gene regulatory network. *Nat. Rev. Genet.* **18**, 180–191 (2017).

[4] Loh, Y.-H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* **38**, 431–440 (2006).

[5] Kim, J., Chu, J., Shen, X., Wang, J. & Orkin, S. H. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132**, 1049–1061 (2008).

[6] Niwa, H., Burdon, T., Chambers, I. & Smith, A. Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. *Genes Dev* **12**, 2048–2060 (1998).

[7] Sato, N., Meijer, L., Skaltsounis, L., Greengard, P. & Brivanlou, A. H. Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor. *Nat Med* **10**, 55–63 (2004).

[8] Kunath, T. *et al.* FGF stimulation of the Erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment. *Development* **134**, 2895–2902 (2007).

[9] Azuara, V. *et al.* Chromatin signatures of pluripotent cell lines. *Nat Cell Biol* **8**, 532–538 (2006).

[10] Meshorer, Eran *et al.* Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. *Dev Cell* **10**, 105–116 (2006).

[11] Dunn, S-J, Martello, G, Yordanov, B, Emmott, S & Smith, A G. Defining an essential transcription factor program for naïve pluripotency. *Science* **344**, 1156–1160 (2014).

[12] Xu, H., Ang, Y.-S., Sevilla, A., Lemischka, I. R. & Ma'ayan, A. Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells. *PLoS Comput Biol* **10**, e1003777 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[13] Gerstein, Mark B *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **488**, 91–100 (2012).

[14] Trott, Jamie, Hayashi, Katsuhiko, Surani, Azim, Babu, M Madan & Martinez Arias, Alfonso. Dissecting ensemble networks in ES cell populations reveals micro-heterogeneity underlying pluripotency. *Mol Biosyst* **8**, 744–752 (2012).

[15] Filipczyk, Adam *et al.* Network plasticity of pluripotency transcription factors in embryonic stem cells. *Nat Cell Biol* **17**, 1235–1246 (2015).

[16] Stumpf, P. S., Ewing, R. & MacArthur, B. D. Single-cell pluripotency regulatory networks. *Proteomics* **16**, 2303–2312 (2016).

[17] Teschendorff, A. E. & Enver, T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nat Commun* **8**, 15599 (2017).

[18] Singer, Z. S. *et al.* Dynamic heterogeneity and DNA methylation in embryonic stem cells. *Mol. Cell* **55**, 319–331 (2014).

[19] Kumar, Roshan M *et al.* Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* **516**, 56–61 (2014).

[20] Guo, G. *et al.* Serum-Based Culture Conditions Provoke Gene Expression Variability in Mouse Embryonic Stem Cells as Revealed by Single-Cell Analysis. *CellReports* **14**, 956–965 (2016).

[21] Martin, G. R. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *PNAS* **78**, 7634–7638 (1981).

[22] Evans, M. J. & Kaufman, M. H. Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292**, 154–156 (1981).

[23] Tesar, P. J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448**, 196–199 (2007).

[24] Brons, I. G. M. *et al.* Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature* **448**, 191–195 (2007).

[25] Chou, Y.-F. *et al.* The growth factor environment defines distinct pluripotent ground states in novel blastocyst-derived stem cells. *Cell* **135**, 449–461 (2008).

[26] Weinberger, L., Ayyash, M., Novershtern, N. & Hanna, J. H. Dynamic stem cell states: naive to primed pluripotency in rodents and humans. *Nat Rev Mol Cell Biol* **17**, 155–169 (2016).

[27] Guo, Ge *et al.* Klf4 reverts developmentally programmed restriction of ground state pluripotency. *Development* **136**, 1063–1069 (2009).

[28] Greber, B. *et al.* Conserved and divergent roles of FGF signaling in mouse epiblast stem cells and human embryonic stem cells. *Cell Stem Cell* **6**, 215–226 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[29] Kalkan, T. & Smith, A. Mapping the route from naive pluripotency to lineage specification. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **369**, 20130540–20130540 (2014).

[30] Smith, A. Formative pluripotency: the executive phase in a developmental continuum. *Development* **144**, 365–373 (2017).

[31] MacArthur, B. D. & Lemischka, I. R. Statistical mechanics of pluripotency. *Cell* **154**, 484–489 (2013).

[32] Sirovich, L. & Kirby, M. Low-dimensional procedure for the characterization of human faces. *J Opt Soc Am A* **4**, 519–524 (1987).

[33] Turk, M. & Pentland, A. Eigenfaces for recognition. *J Cogn Neurosci* **3**, 71–86 (1991).

[34] Georghiades, A. S., Belhumeur, P. N. & Kriegman, D. J. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Trans Pattern Anal Mach Intell* **23**, 643–660 (2001).

[35] Lee, K.-C., Ho, J. & Kriegman, D. J. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans Pattern Anal Mach Intell* **27**, 684–698 (2005).

[36] Spitzer, M. H. & Nolan, G. P. Mass Cytometry: Single Cells, Many Features. *Cell* **165**, 780–791 (2016).

[37] Zunder, E. R., Lujan, E., Goltsev, Y., Wernig, M. & Nolan, G. P. A continuous molecular roadmap to iPSC reprogramming through progression analysis of single-cell mass cytometry. *Cell Stem Cell* **16**, 323–337 (2015).

[38] Wernig, M. *et al.* A drug-inducible transgenic system for direct reprogramming of multiple somatic cell types. *Nat Biotechnol* **26**, 916–924 (2008).

[39] Ying, Qi-Long *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519–523 (2008).

[40] Takahashi, Kazutoshi & Yamanaka, Shinya. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).

[41] Zhou, K.-R. *et al.* ChIPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data. *Nucleic Acids Res* gkw965–8 (2016).

[42] Ogata, H. *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **27**, 29–34 (1999).

[43] Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res* **44**, D481–7 (2016).

[44] Graf, T. & Stadtfeld, M. Heterogeneity of embryonic and adult stem cells. *Cell Stem Cell* **3**, 480–483 (2008).

[45] Bargaje, R. *et al.* Cell population structure prior to bifurcation predicts efficiency of directed differentiation in human induced pluripotent cells. *PNAS* **114**, 2271–2276 (2017).

[46] Bedzhov, I. & Zernicka-Goetz, M. Self-organizing properties of mouse pluripotent cells initiate morphogenesis upon implantation. *Cell* **156**, 1032–1044 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[47] Buganim, Yosef *et al.* Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150**, 1209–1222 (2012).

[48] Golipour, A. *et al.* A late transition in somatic cell reprogramming requires regulators distinct from the pluripotency network. *Cell Stem Cell* **11**, 769–782 (2012).

[49] Cover, T. M. & Thomas, J. A. *Elements of Information Theory.* Wiley Series in Telecommunications (John Wiley & Sons, Inc., New York, USA, 1991).

[50] Li, Ronghui *et al.* A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts. *Cell Stem Cell* **7**, 51–63 (2010).

[51] Smith, Z. D., Nachman, I., Regev, A. & Meissner, A. Dynamic single-cell imaging of direct reprogramming reveals an early specifying event. *Nat Biotechnol* **28**, 521–526 (2010).

[52] Silva, J. & Smith, A. Capturing pluripotency. *Cell* **132**, 532–536 (2008).

[53] Morgani, S., Nichols, J. & Hadjantonakis, A.-K. The many faces of Pluripotency: in vitro adaptations of a continuum of in vivo states. *BMC Dev. Biol.* **17**, 1–20 (2017).

[54] Stumpf, P. S. *et al.* Stem Cell Differentiation as a Non-Markov Stochastic Process. *Cell Syst* **5**, 268–282.e7 (2017).

[55] MacArthur, B. D. *et al.* Nanog-dependent feedback loops regulate murine embryonic stem cell heterogeneity. *Nat Cell Biol* **14**, 1139–1147 (2012).

[56] Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Syst* **5**, 251–267.e3 (2017).

[57] Kalkan, T. *et al.* Tracking the embryonic stem cell transition from ground state pluripotency. *Development* **144**, 1221–1234 (2017).

[58] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria (2016). URL `https://www.R-project.org/`. R version 3.3.2 (2016-10-31).

[59] Ellis, B. *et al.* *flowCore: flowCore: Basic structures for flow cytometry data* (2017). URL `http://bioconductor.org/packages/flowCore/`. R package version 1.40.4.

[60] Fraley, C. & Raftery, A. E. Model-Based Clustering, Discriminant Anaxlysis, and Density Estimation. *J Am Stat Assoc* **97**, 611–631 (2002).

Author Information

Acknowledgments

This research was funded by the Biotechnology and Biological Sciences Research Council, United Kingdom, grant number BB/L000512/1 and by the Medical Research Council, United Kingdom, grant number MC_PC_15078.

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Author contributions

Conceptualization, P.S.S. and B.D.M; Methodology and Investigation, P.S.S. and B.D.M.; Writing – Original Draft, P.S.S, Writing – Review & Editing, P.S.S., B.D.M.; Supervision, B.D.M.

Competing interests

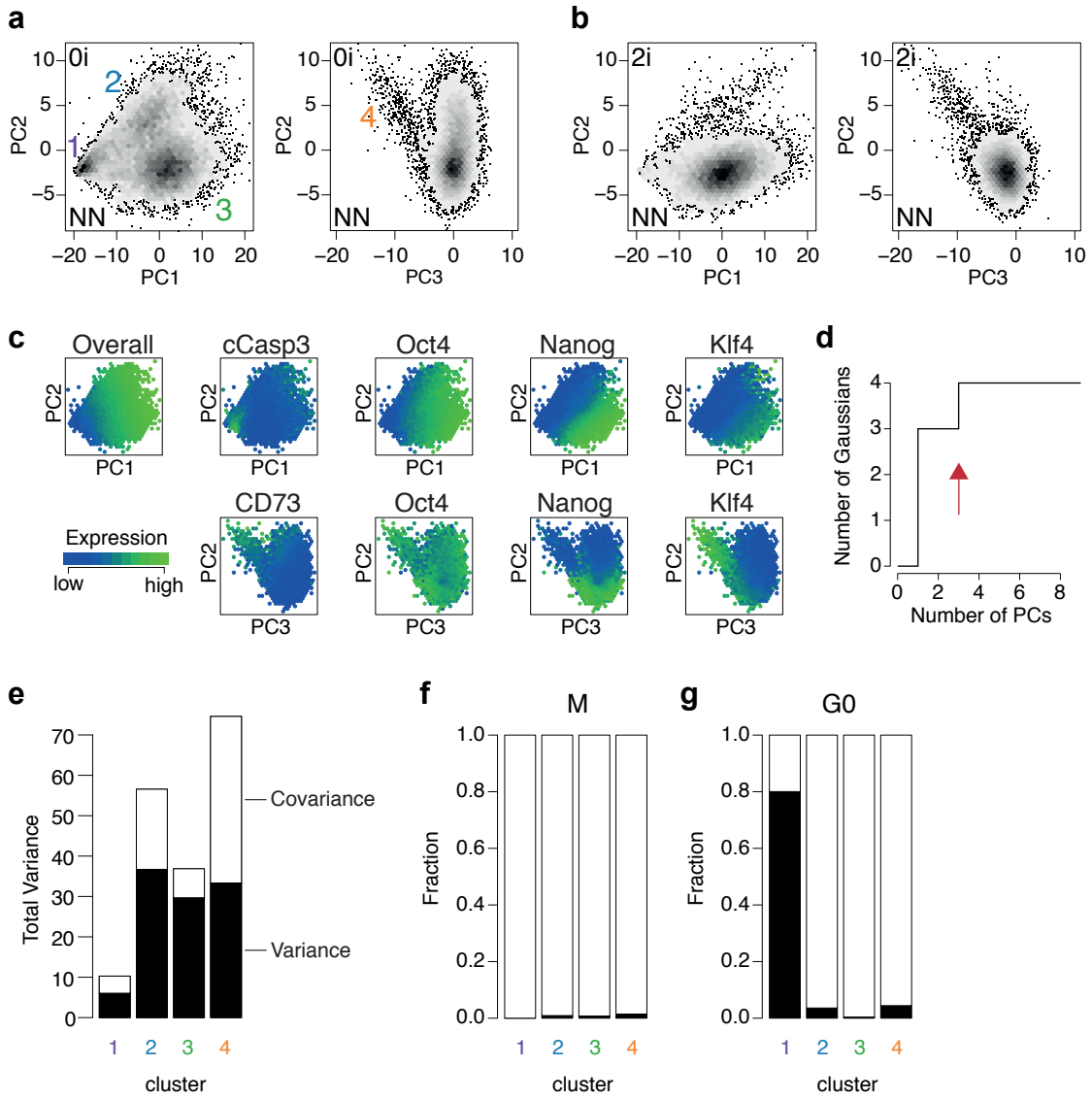
The authors declare no competing interests.

Corresponding Author

Correspondence to P.S. Stumpf.

bioRxiv preprint doi: https://doi.org/10.1101/208470; this version posted October 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supporting information



Supplementary Figure 1: **Sub-states of regulatory network activity.** (**a-c**) Projection of NN mESC onto the same principal component space derived from NG mESCs (shown in Fig. 3). NN mESC display qualitatively the same population structure and corresponding node expression levels as NG mESCs. (**d**) Relationship between number of multivariate Gaussian distributions required to fully represent population structure, given the number of Principal Components used to represent network activity state. (**e**) Total variance/covariance within each sub-population (estimated from trace of the covariance matrix and the sum of the off diagonal elements of the covariance matrix for the respective fitted multivariate Gaussian models). (**f**) Fraction of cells of each cluster in M-phase of the cell cycle. (**g**) Fraction of cells of each cluster in G0-phase of the cell cycle.