

Universality of cell differentiation trajectories revealed by a reconstruction of transcriptional uncertainty landscapes from single-cell transcriptomic data

Nan Papili Gao¹, Olivier Gadrillon^{2,3}, András Páldi⁴, Ulysse Herbach⁵, Rudiyanto Gunawan^{6,*}

¹ Institute for Chemical and Bioengineering, ETH Zurich, 8093 Zurich, Switzerland

² Univ Lyon, ENS de Lyon, Univ Claude Bernard, CNRS UMR 5239, INSERM U1210, Laboratory of Biology and Modelling of the Cell, F69007 Lyon, France

³ Inria Team Dracula, Inria Center Grenoble Rhône-Alpes, France

⁴ Ecole Pratique des Hautes Etudes, PSL Research University, St-Antoine Research Center, Inserm U938, 34 rue Crozatier, 75012, Paris, France

⁵ Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France

⁶ Department of Chemical and Biological Engineering, University at Buffalo, Buffalo, NY 14260, USA

*Corresponding author (email: rgunawan@buffalo.edu)

ABSTRACT

We employed our previously-described single-cell gene expression analysis CALISTA (Clustering And Lineage Inference in Single-Cell Transcriptional Analysis) to evaluate transcriptional uncertainty at the single-cell level using a stochastic mechanistic model of gene expression. We reconstructed a transcriptional uncertainty landscape during cell differentiation by visualizing single-cell transcriptional uncertainty surface over a two dimensional representation of the single-cell gene expression data. The reconstruction of transcriptional uncertainty landscapes for ten publicly available single-cell gene expression datasets from cell differentiation processes with linear, single or multi-branching cell lineage, reveals universal features in the cell differentiation trajectory that include: (i) a peak in single-cell uncertainty during transition states, and in systems with bifurcating differentiation trajectories, each branching point represents a state of high transcriptional uncertainty; (ii) a positive correlation of transcriptional uncertainty with transcriptional burst size and frequency; (iii) an increase in RNA velocity preceding the increase in the cell transcriptional uncertainty. Finally, we provided biological interpretations of the universal rise-then-fall profile of the transcriptional uncertainty landscape, including a link with the Waddington's epigenetic landscape, that is generalizable to every cell differentiation system.

Keywords: single cell, cell differentiation, epigenetic landscape, gene expression, heterogeneity

MAIN

Cell differentiation is the process through which unspecialized stem cells become more specialized. Because of its importance in development, cellular repair, and organismal homeostasis, the molecular mechanisms of cell differentiation has been the subject of intense scrutiny. Since roughly 50 years ago – along with the promulgation of the central dogma of molecular biology by Francis Crick and the characterization of the lactose operon by François Jacob and Jacques Monod – the existence of a genetic program has become a prevailing explanation for the cell differentiation process. Although the details were originally not defined, at least not formally, such a genetic program purports a constellation of master genes (i.e., transcription factors) that orchestrate the transcription of downstream target genes in a precise spatiotemporal fashion, resulting in long-lasting alterations in the gene expression patterns¹⁻³. A notable experimental evidence substantiating this view is the overexpression of myoD inducing a myogenic phenotype in seemingly naive cells⁴. Over the past few decades, the repertoire of such master genes across numerous stem cell systems, such as Nanog, Oct4, Sox2, BATF and MyoD, begin to coalesce⁵⁻⁷.

Recent advances in single-cell technologies has revealed new aspects of the cell differentiation that are incompatible with the idea of ordered and programmed (i.e., deterministic) gene expression. More specifically, single-cell data paint a stochastic differentiation process that increases cell-to-cell variability of gene expression. Such an observation has been made for a wide variety of cell differentiation systems, including chicken erythroid progenitors⁸, erythroid myeloid lymphoid (EML) cells⁹, mouse embryonic stem cells (mESCs)^{10,11}, and human CD34+ cells¹². Interestingly, a similar increase of gene expression variation was also observed during the de-differentiation of somatic cells into iPSCs¹³. Stochastic gene expression also appears to have functional role beyond cell differentiation systems. For example, an increase in cell-to-cell variability of gene expression has been reported during a forced adaptation of budding yeast cells to unforeseen challenges¹⁴.

Based on the observations of single-cell data above, a different view of cell differentiation begins to materialize. Instead of stem cells following an identical genetic program, the cell differentiation is akin to a dynamic exploratory process. More specifically, in this view, the cell differentiation is thought to proceed as follows¹⁴⁻¹⁷:

I) extrinsic and/intrinsic internal stimuli, such as a medium change or the addition of new molecules in the external medium, trigger a cellular response that destabilizes the initial high potency cell state;

II) each cell alters its internal cell state and engages an exploratory dynamics through a combination of the inherent stochastic dynamics of gene transcription and the emergence of new stable cell state(s). At the cell population level, we observe a rise in the cell-to-cell variability of gene expression;

III) a physiological selection / commitment to one stable lineage among possibly multiple lineages that arise from the degeneracy of the system;

IV) finally, a reduction in the exploratory dynamics commences along with the establishment of stable cell state(s) corresponding to differentiated cell type(s).

The above view is compatible with the idea that cell phenotype transition results from the dynamics of an underlying stochastic molecular network^{18,19}. In 1957, Conrad Waddington proposed the presently well-known epigenetic landscape that likens the cell differentiation process to a ball rolling on a downward sloping surface, starting from a state of high cell potency and ending at one of possibly several states of low cell potency. The landscape itself is shaped by the action of the genes and gene network – depicted in the less-frequently-shown part B of Waddington's original figure as a network of ropes that are tied to the surface, creating valleys and hills. Although the epigenetic landscape was originally proposed only as a metaphor of how gene regulation governs the cell differentiation process, this landscape has been formalized within the framework of dynamical systems theory²⁰. The valleys in the Waddington's epigenetic landscape are equated to stable states of a dynamical system, called attractors, while the hills are often interpreted as energetic barriers.

The Waddington's metaphor has further been re-examined in a series of works, where the behaviour of the underlying dynamical gene regulatory network (GRN) drives cell fate determination. Some of these studies employed stochastic simulations of simple gene networks with bi-/tri-stable states to describe the role of non-genetic cell heterogeneity in cell differentiation processes^{21–23}. Here, to quantitatively reconstruct the epigenetic landscape, an underlying GRN structure must be assumed *a priori*. In reality, the GRN driving cell fate determination is complex and proved to be challenging to infer from data^{6,24–28}. Even when the complete regulatory system is known, the high-dimensionality of the parameter space makes the landscape generation computationally prohibitive²¹, especially in the absence of an analytical solution, which are slowly emerging²⁹. Thus, the above approaches provide more of a conceptual understanding than a mechanistic or molecular explanation.

A number of recent studies provided a graphical representation of the differentiation process based on single-cell transcriptomic data that conforms with the Waddington's epigenetic landscape^{30–34}. More specifically, these studies reconstructed the epigenetic landscape from single-cell gene expression data using probabilistic and quasi-

potential methods, for example by applying Hopfield neural networks^{30,31}, a cell-density based strategy³³, network entropy measurements³⁴ or more recently Large Deviation Theory²⁹. However, with the exception of Fard et al.³¹ and Lv et al., 2014²⁹, the aforementioned studies produced monotonic descent passages during cell differentiation, mimicking closely the Waddington's epigenetic landscape metaphor (see for example^{22,34}). Also, none of the above studies consider directly the cellular mechanism that generates stochastic gene transcriptional bursts.

In the present work, we aimed to shed light on the gene transcriptional mechanism behind the rise-then-fall trajectory of cell-to-cell variability in gene expression observed during the cellular differentiation process. To this end, we analyzed a collection (8) of published single-cell transcriptomic datasets from various cell differentiation systems, comprising both single-cell RT-qPCR (scRT-qPCR)^{8,10,12,35-37} and single-cell RNA-sequencing (scRNA-seq)^{38,39}. We employed a likelihood-based analysis using a recent method CALISTA (Clustering And Lineage Inference in Single-cell Transcriptomics Analysis)⁴⁰. The analysis relied on a mechanistic model of the stochastic gene transcriptional bursts to characterize single-cell gene expression distribution. We defined a new concept, called transcriptional uncertainty landscape, based on the cell likelihood value from CALISTA analysis to characterize the stochastic dynamics of the gene transcription process during cell differentiation. The stochastic gene transcriptional model enabled identifying the specific parameters or mechanisms that explain the observed changes in the gene transcriptional uncertainty at the single-cell level⁴¹. For two additional single-cell datasets, we also evaluated the single-cell RNA-velocity using the recently published Velocyto method⁴². The two-state model parameter analysis, combined with RNA-velocities, provided insights into the mechanism regulating cell fate decisions, specifically on the role of stochastic gene transcriptions in the differentiation processes and on the possible mechanism generating this stochasticity.

RESULTS

Single-cell Transcriptional Uncertainty Landscape using CALISTA

In this work, we used CALISTA, a likelihood-based bioinformatics toolbox designed for an end-to-end analysis of single-cell gene expression data, to evaluate the transcriptional uncertainty of each individual cell based on its gene expression data⁴⁰. CALISTA uses the two-state model of stochastic gene transcription bursts to characterize the distribution of mRNA counts in individual cells⁴³. In the model, a gene promoter stochastically switches between ON and OFF state, and only in the ON state can gene transcription occur. The distribution of

mRNA depends on four model parameters: θ_{on} (the rate of promoter activation), θ_{off} (the rate of promoter inactivation), θ_t (the rate of mRNA production when the promoter is on the ON state), and θ_d (the rate constant of mRNA degradation)^{27,44} (see Fig. 1). For example, when $\theta_{off} \gg \theta_{on}$ and $\theta_{off} \gg \theta_d$, keeping θ_t / θ_{off} fixed, mRNA are produced through bursts of short but intense transcription, which is a typical case observed for gene transcriptions in single cells^{45,46}.

Our analysis procedure using CALISTA involves four main steps: cell clustering, lineage inference, and calculation of single-cell transcriptional uncertainty. In the analysis, each cell is assigned a likelihood value, which is computed from the probability of its gene expression (mRNA counts) based on the mRNA distribution given by the two-state model above. In the single-cell clustering, we employed a greedy algorithm to find single-cell clusters that maximizes the total likelihood value for the cells. In the lineage progression inference, we connected cell clusters sharing similar gene expression distributions – by computing cluster distances based on cell likelihood – to form a hierarchical network. Here, we assigned each cell to an edge in the lineage progression network that is pointing to or emanating from its respective cluster, such that its single-cell likelihood is maximized. Finally, we computed single-cell transcriptional uncertainty as the negative logarithm of the single-cell likelihood (NLL) value. The full detail of CALISTA can be found in a recent publication⁴⁰.

The single-cell likelihood value reflects the joint probability of its gene expression repertoire. A cell with a low likelihood value may indicate that the gene expression of the cell is different from its neighboring cells, i.e. the cell is an outlier. But, more interestingly, a low likelihood value may also correspond to a cell state of high uncertainty in the gene expression. The group of cells in such high uncertainty state have gene expressions that are dissimilar to each other, and thus, the gene expression distributions will have high entropy. As mentioned above, in this work we used the negative logarithm of the single-cell likelihood value as a metric of single-cell transcriptional uncertainty. By plotting the single-cell transcriptional uncertainty over the two-dimensional projection of the single-cell transcriptomics data – for example, using the first two principal components from Principal Component Analysis (PCA), we constructed a transcriptional uncertainty landscape in the form of a surface plot of the NLL values. By visualizing the single-cell transcriptional uncertainty as a surface plot, we can study the landscape of transcriptional uncertainty during cell differentiation at single-cell resolution. On the single-cell transcriptional uncertainty surface, an aberrant cell can be easily distinguished from a cell of high uncertainty state, since an aberrant cell will appear isolated from its nearby cells and will be located on a region with high NLL.

Transcriptional uncertainty landscape of iPSC cell differentiation to cardiomyocytes

In the following, we demonstrated an application of our procedure described above to a single-cell transcriptional dataset from cardiomyocytes differentiation from human induced pluripotent stem cells (iPSCs)³⁵. The single-cell clustering of CALISTA returned five clusters⁴⁰, in good agreement with the number of cell types reported in the original study. CALISTA identified one bifurcation event in the lineage progression, which led to two cell lineages³⁵. Fig. 2a-b gives the single-cell transcriptional uncertainty landscape showing cells exiting the initial epiblast state that is characterized by a valley in the landscape, passing through a hill of high transcriptional uncertainty corresponding to primitive streak (PS)-like progenitor state, before ending up at one of the low transcriptional uncertainty terminal states corresponding to either mesodermal (desired) or endodermal (undesired) fate. As depicted in Fig. 2c, the intermediate cell cluster (cluster 2) comprising PS-like cells have higher cell uncertainty (lower single-cell likelihood) than the other clusters. Figures 2d and 2e give the moving-averaged uncertainty values for pseudotemporally ordered cells using a moving window of 10% of the total cells for both endodermal and mesodermal paths, respectively. The moving-averaged transcriptional uncertainty for the two differentiation paths follows a rise-then-fall trajectory where the peak of uncertainty coincides with the lineage bifurcation event.

We explored whether the rise-then-fall in uncertainty is an artefact from using the two-state model to evaluate the cell likelihood values. To this end, we implemented a modified version of the algorithm for ordering cells by calculating the cell likelihood values using the empirical (observed) distribution, instead of the analytical distribution from the two-state model. As shown in Supplementary Fig. S1, the transcriptional uncertainty landscape from the modified implementation shows a strong resemblance to the original one. We also investigated whether the number of clusters may affect the landscape, in which using too few of clusters may artificially inflate the uncertainty due to mixing of cells from different states. We reran CALISTA by using a higher number of clusters (set to nine based on the eigengap heuristic⁴⁷). The hill in the uncertainty landscape is again seen around the bifurcation event upon using a higher number of cell clusters (Supplementary Fig. S2). Finally, we used a different algorithm to cluster cells, specifically using a Laplacian-based clustering algorithm SIMLR⁴⁸, to test whether the shape of the transcriptional uncertainty landscape changes with the clustering algorithm. The single-cell clusters can be interpreted as the transitional states that the differentiating cells go through. Starting with the result of SIMLR cell clustering, we then generated the lineage progression and estimated the cell likelihood values using CALISTA. The transcriptional uncertainty landscape from SIMLR cell clustering has the same shape as that

in Fig. 2a-b, demonstrating that the transcriptional uncertainty landscape observed above is not dependent on using CALISTA for cell clustering (Supplementary Fig. S3).

To further elucidate the role of specific genes in shaping the transcriptional uncertainty landscape, we looked at the transcriptional uncertainty associated with individual genes. Fig. 2f depicts the NLL distribution of each gene for the five single-cell clusters. As expected, cell in cluster 2 have generally higher NLL than those in the other clusters. Fig. 2f clearly illustrates that within cluster 2, some genes show higher NLL values than the others (Supplementary Figure S4). To identify the important genes related to transcriptional uncertainty, we identified genes with NLL values exceeding a threshold δ for at least 30% of the cells in each cluster, where δ is set to 3 standard deviation above the overall mean NLL for all cells and genes in the dataset (see Methods Eq. (2)). None of the genes in clusters 1, 4 and 5 have NLL above the threshold. Meanwhile, 16 and 8 genes in clusters 2 and 3, respectively, pass the above criterion for high uncertainty with 4 common genes between the two gene sets (Supplementary Table S1). Genes with high transcriptional uncertainty in cluster 2 may have functional roles in cell fate determination. The gene set of cluster 2 includes known genes upregulated only in the PS-like state (e.g. EOMES, GSC, MESP1 and MIXL1), as well as markers of mesodermal and endodermal cells (e.g. BMP4, HAND1, and SOX17)^{35,40} (Supplementary Figure S5). Meanwhile, the main contributors to cell uncertainty in cluster 3 (e.g. BMP4 and MYL4^{35,40}) are known transition genes between PS-like cells and the final mesoderm fate (Supplementary Figure S6). Fig. 2g depicts the protein-protein interaction (PPI) network related to the gene set of cluster 2 using STRING (minimum required interaction score of 0.4)⁴⁹, indicating that these genes form a strongly interconnected hub of known transcription factors and molecules involved in the signal transduction of embryonic development (Supplementary Table S1).

Transcriptional uncertainty landscapes of cell differentiation

We further applied the procedure above to seven additional single-cell transcriptomic datasets that were generated using scRT-qPCR^{8,10,12,36,37} and scRNA-sequencing^{38,39}, to assess the universality of the rise-then-fall feature of single-cell transcriptional uncertainty landscape during cell differentiation. The first of these datasets came from 405 cells during mouse embryonic fibroblast (MEF) reprogramming into induced neural (iN) and myogenic (M) cells³⁸. Like the iPSC differentiation above, the lineage progression has a single bifurcation point. As depicted in Figure 3a, the single-cell transcriptional uncertainty increases from the initial MEF state and reaches a peak around the bifurcation before decreasing toward two end-point cell fates. The rise-then-fall of transcriptional uncertainty in the MEF reprogramming is in good agreement with what we observed in the iPSCs

differentiation above. Higher entropy of gene expression distribution in a cell population has also been reported in the reprogramming of iPSCs¹³.

Next, we analyzed datasets from cell differentiation processes without a lineage bifurcation and with multiple lineage bifurcations. Three scRT-qPCR datasets came from differentiation systems without bifurcation, including Richard et al. study on chicken erythrocytic differentiation of T2EC cells⁸, Stumpf et al. study on differentiation of mouse embryonic stem cells (ESC) to neural progenitor cells (NPC)¹⁰, and Moussy et al. study during CD34+ cell differentiation¹². The single-cell clustering and lineage progression by CALISTA produced the expected cell differentiation trajectory (see Supplementary Figure S7-9). The single-cell transcriptional uncertainty landscapes of these three differentiation systems, as shown in Figure 3b, exhibit a rise-then-fall profile, creating a hill that the cells traverse through in the differentiation process. A transitory increase in single-cell gene expression uncertainty was reported either directly or indirectly in the original publications. In Richard et al.⁸ and Stumpf et al¹⁰, the authors adopted the Shannon entropy to quantify cell-to-cell variability (uncertainty), while Moussy et al.¹² reported an unstable transition state with ‘hesitant cells’ flipping their morphology between polarized and round shapes before committing to the common myeloid progenitors-like fate. Morphological uncertainty therefore corresponded to a higher transcriptional uncertainty. Note that Moussy et al. study looked at only the initial phase of the (hematopoietic) cell differentiation, and thus, it is likely that the differentiation process had not completed for the cells in the dataset.

The next set of single-cell gene expression data came from differentiation systems with multi-branching lineage, including Guo et al. study during mouse embryo development from zygote to blastocyst³⁶, Nestorowa et al.³⁹ and Moignard et al.³⁷ studies on hematopoietic stem cell differentiation. Figure 3c show the single-cell transcriptional landscape for each of the datasets. For Guo et al. study, we identified 7 cell clusters and identified two bifurcations in the lineage. Here, we observed two hills in the transcriptional uncertainty landscape, each coinciding with a bifurcation event in the lineage progression – one at 32-cell stage (cluster 2 to cluster 3 and 4) and another at 64-cell stage (cluster 4 to cluster 6 and 7) (see Supplementary Figure S10). For Nestorowa et al.³⁹ (Supplementary Figure S11) and Moignard et al.³⁷ (see Methods and⁴⁰) datasets, we again observed peaks in the transcriptional uncertainty landscape that colocalize with the bifurcation points in the lineage progression.

The use of the two-state mechanistic gene transcriptional model within CALISTA enabled us to probe into a mechanistic explanation for the observed shape of the transcriptional uncertainty landscape. Table 1 show the pairwise Pearson correlations between the cell-averaged NLL of each cluster with two biologically interpretable model parameters, namely transcriptional burst size (number of transcripts generated in each burst) and burst

frequency (occurrence of burst per unit time)⁵⁰ (see Methods). The Pearson correlations indicate that the single-cell gene expression uncertainty increases with higher burst size and burst frequency (p -value ≤ 0.01). Higher transcriptional burst size and frequency are associated with a lower θ_{off} – a lower rate of promoter turning off – and a greater θ_{on} – higher rate of promoter turning on. One possible explanation for such change in model parameters is a higher chromatin accessibility during the transition period of cell differentiation. This finding is consistent with the view that stem cells increase its gene expression uncertainty or stochasticity by adopting a more open chromatin state to enable the exploration of the gene expression space^{33,50–52}.

Coupling cell uncertainty with RNA velocity

In a recent paper⁴², La Manno and colleagues introduced the concept of RNA velocity, which involves computing the rate of change of mRNA from the ratio of unspliced to spliced mRNA. A positive RNA velocity indicates an induction of gene expression, while a negative RNA velocity indicates a repression of gene expression. La Manno et al. demonstrated that RNA velocities are able to predict the trajectory of cells undergoing a dynamical transition, such as in circadian rhythms or cell differentiation. In the following, we explored the relationship between RNA velocities and single-cell transcriptional uncertainty.

We evaluated the single-cell transcriptional uncertainty and RNA velocity for two single-cell gene expression datasets that were previously analyzed in La Manno et al.^{42, 42}. The first dataset came from human glutamatergic neurogenesis which has a linear (non-bifurcating) lineage progression. Figure 4 (top row) depicts the cell clustering, single-cell transcriptional uncertainty, and RNA velocities (see also Supplementary Figure S12). The single-cell transcriptional uncertainty landscape again has the rise-then-fall shape, as in the other cell differentiation systems discussed above. Interestingly, the same rise-then-fall profile is also seen in the RNA velocity. As illustrated in Figure 4, the increase and decrease of the RNA velocity precede the transcriptional uncertainty, and the peak of RNA velocity occurs prior to those of the transcriptional uncertainty (see Supplementary File S1 for an animated illustration). Furthermore, a gene-wise cross-correlation analysis confirms a positive correlation between RNA velocity and single-cell transcriptional uncertainty with a delay for individual genes (see Supplementary Figure S13).

We also compared RNA velocity and single-cell transcriptional uncertainty for another dataset from mouse hippocampal neurogenesis with a multi-branching lineage⁵³. Figure 4 (bottom row) shows that like in the neurogenesis dataset earlier, the RNA velocity increases and then decreases during cell differentiation, and the change in the RNA precede that of the transcriptional uncertainty (see Supplementary File S2). Also, the RNA

velocity peaks take place before the transcriptional uncertainty peaks. The rise-then-fall dynamic of the RNA velocity seen in the two datasets above is consistent with the view that cells engage in an exploratory stochastic dynamics as they leave the progenitor state, and disengage this explorative mode as they reach toward the final cell state.

DISCUSSION

Although Waddington's epigenetic landscape was originally proposed only as a metaphor, the landscape has helped stem cell researchers to conceptualize the cell differentiation processes through canalization of cell lineages. As mentioned earlier, much of the existing literature on the analytical reconstruction of the epigenetic landscape relied on either a dynamical system theory applied to a simple gene network, or a thermodynamic interpretation based on the potential energy of a reaction²¹⁻²³. In our study, we did not make any prior assumptions on the gene regulatory network driving the differentiation process nor on the characteristics of the landscape, such as the existence of a stable valley or that of an energetic barrier (hill). Here, we assumed that the gene transcription at the single-cell level occurs as stochastic transcriptional bursts, in which mRNA counts follow the equilibrium distribution that is given by the two-state gene transcription model⁴³. Therefore, the transcriptional uncertainty landscape in our study is a reflection of the dynamic evolution of gene transcriptional stochasticity within individual cells during the cell differentiation process.

The reconstruction of the transcriptional uncertainty landscapes from 10 single-cell transcriptomic datasets of various cell differentiation processes in our study reveals a universal rise-then-fall trajectory in which cells start from a high potency state with a uniform gene expression pattern in the cell population, then progress through transitional cell state(s) marked by increased transcriptional uncertainty (i.e., higher cell-to-cell variability), and eventually reach one of possibly several final cell states with again a uniform gene expression pattern among the cells. Furthermore, the peaks of the transcriptional uncertainty landscape colocalize with forks in the cell lineage. The rise-then-fall in cell uncertainty agrees well with other reports from different cell differentiation systems^{8-12,54,55}, suggesting that stem cells go through a transition state of high gene expression uncertainty before committing to a particular cell fate. The existence of a hill or barrier during the intermediate stage of cell differentiation has also been proposed in previous studies^{14,31,56}. In particular Moris and colleagues compared this transition state to the activation energy barrier in chemical reactions⁵⁶. We noted however, that a hill in our transcriptional uncertainty landscape is a reflection of a peak in the cell-to-cell gene expression variability, and thus does not represent a resistance or barrier that a cell has to overcome.

In the analysis of iPSCs differentiation into cardiomyocytes ³⁵, the key genes regulating cardiomyocyte differentiation are among the largest contributors to the overall transcriptional uncertainty at or around the peak in the landscape, supporting the idea that dynamic cell-to-cell variability has a functional role in cell-fate decision making processes ^{21,57,58}. Such an idea would be in congruence with the recent demonstration that, in a physiologically relevant cellular system, gene expression variability is functionally linked to differentiation ^{57,58}.

The rise-then-fall trajectory in the transcriptional uncertainty landscape are more pronounced in some datasets than in others. For example, in Nestorowa ³⁹ and Moignard ³⁷ datasets (see Figure 3c), peaks in the transcriptional uncertainty landscape are less noticeable than in the other differentiation systems. We noted that cells in the Nestorowa ³⁹ and Moignard ³⁷ studies were pre-sorted by using flow cytometry based on the expression of surface protein markers. We posited that at least some cells in the transition state(s) might have been lost during the cell pre-sorting since such cells might not express the chosen surface markers strongly.

Further, the correlation analysis between the cell transcriptional uncertainty and biologically meaningful rates of the stochastic gene transcription model showed strong positive correlations with transcriptional burst size and frequency. In agreement with the result of the correlation analysis, several studies have reported an increase in gene transcriptional bursts during transition states in cell differentiation and other recent studies have suggested that both burst frequency and burst size regulate gene expression levels ^{33,51,52}. Importantly, our comparison of the single-cell transcriptional uncertainty and the single-cell RNA velocity revealed that an increase (decrease) in RNA velocity predicts an increase (decrease) in transcriptional uncertainty after a short delay, and that a peak of RNA velocity precedes that of the transcriptional uncertainty.

The aforementioned observations, while correlative in nature, points to possible biological mechanisms underlying the universal dynamic feature of single-cell transcriptional uncertainty during cell differentiation. At the start of the differentiation process, cells engage an exploratory search dynamics in the gene expression space by increasing stochastic transcriptional burst size and burst frequency. The putative objective of such a stochastic search is to optimize the cell's gene expression pattern given its new environment. The engagement of this stochastic exploratory mode is supported by the observed increased in the overall RNA velocity and its expected-but-delayed effect in elevating the cell-to-cell gene expression variability (i.e. higher transcriptional uncertainty). Increased transcriptional burst size and frequency suggest a promoter state of stochastic gene transcription that favors the ON state (higher θ_{on} and lower θ_{off}).

A possible mechanism behind this exploratory search dynamics is an increase in chromatin mobility, driven by metabolic alterations in early differentiation ⁵⁹. Multiple studies have demonstrated that a mismatch between the

intracellular state of stem cells and their immediate environment can lead to metabolic reorganization^{60–62}.

More specifically, a change in the balance between glycolysis and OXPHOS metabolism has been associated to numerous differentiation processes (see Richard et al.⁶³ and references therein).

Furthermore, changes in the metabolic flux state in early differentiation can modulate the activity of chromatin modifying enzymes through their metabolic co-factors⁶⁴, or in more direct fashion⁶⁵ and alter the cell differentiation outcome. A more dynamic state of the chromatin is associated with more variable gene expressions due to the changes in the opening-closing dynamics (breathing) of the chromatin⁶⁶. As the cells approach the final state, cells disengage the exploratory search mode, as the cells approach an optimal gene expression and metabolic state associated with a chosen cell type.

The findings of our analysis fit within the paradigm of a stochastic stem cell differentiation process as proposed in the introduction. The disordered gene expression pattern during the transition period can be seen as an exploratory dynamics to find the optimal pattern(s)^{14,17}. The transcriptional uncertainty in our analysis can be interpreted as the width of the valley in Waddington's epigenetic landscape. If one considers the epigenetic landscape as a depiction of the accessible gene expression subspace through stochastic single-cell trajectories pass during differentiation, a wider valley indicates a more variable gene expression pattern. While in the original Waddington's epigenetic landscape the valley naturally widens around the branching point in the cell lineage, our analysis shows that a widening of the valley (an increase in transcriptional uncertainty) also occurs in non-branching lineage. In other words, the increase in transcriptional uncertainty appears to be a universal feature of the cell differentiation process, one that arises from the engagement of exploratory mode through increased stochasticity in transcriptional bursts, as explained above.

METHODS

Main steps of CALISTA workflow

Herein, we briefly describe the main steps involved in the calculation of single-cell transcriptional uncertainty using CALISTA (for further details see⁴⁰).

Pre-processing

Given an $N \times G$ single-cell expression matrix M , where N denotes the number of cells and G the number of genes, the pre-processing in CALISTA involves two steps: a normalization of the expression data $m_{n,g}$ – i.e. the number of transcripts of gene g in the n -th cell, and a selection of the most variable genes⁴⁰.

Cell clustering

CALISTA clustering follows a two-step procedure. The first step involves a greedy optimization strategy to find cell clustering that maximizes the total cell likelihood, i.e. the sum of the likelihood value for all cells. The single-cell likelihood value is computed as the joint probability of the cell's gene expression data, which is set equal to the product of the probabilities of the mRNA counts for the selected genes based on the mRNA distribution from the two-state stochastic gene transcription model. To avoid issues with numerical overflow, we use the logarithm of the cell likelihood. By performing the greedy optimization multiple times, a consensus matrix containing the number of times two cells in the dataset are put in the same cluster, is generated.. In the second and final step, CALISTA generates the cell cluster assignments by using k -medoids clustering based on the consensus matrix. The final outcome of CALISTA's clustering is the assignment of cells into K clusters and the optimal model parameters for the two-state gene transcription model: $\theta^*(g, k) = \{\theta_{on}^*, \theta_{off}^*, \theta_t^*\}_g^k$ for each gene g in cluster k (for simplicity, we set $\theta_d^* = 1$ and scale the other parameters by the degradation rate)⁴⁰.

Lineage progression inference

In CALISTA, cell lineage progression is inferred based on cluster distances – a measure of dissimilarity between two clusters. The cluster distance of two cell clusters is defined as the average decrease in the cell likelihood value if the cells from these two clusters are grouped as one cluster, as opposed to the original clustering. The lineage progression graph is built by adding transition edges between pairs of clusters in increasing magnitude of cluster distance until all clusters are connected to at least one other cluster, or based on user-specified criteria.

Single-cell transcriptional uncertainty

The last step in our analysis is to compute the final single-cell likelihood. Briefly, for each cell, we consider all edges in the lineage progression graph that are adjacent to the cell's respective cluster, i.e. edges that emanate from or pointing to the cluster to which the cell belongs. The likelihood of a cell along an edge is evaluated by interpolating the likelihood values of the cell's gene expression using the mRNA distributions from the two adjacent clusters. Each cell is then assigned to the edge along which its interpolated likelihood value is maximum, and the final cell likelihood is set to this maximum value. As mentioned above, the single-cell transcriptional uncertainty is evaluated as the negative logarithm of the cell likelihood value (NLL).

Pseudotimes calculation

For we can evaluate the pseudotimes for the cells according to the following procedure. First, a pseudotime is given to each cluster with a value between 0 (initial cell state) and 1 (final cell fate). Subsequently, we determine the linear fractional position of each cell along its respective edge at which its interpolated likelihood value is maximum (see *Single-cell transcriptional uncertainty*). The pseudotime of a cell is computed by a linear interpolation of the pseudotimes of the two clusters adjacent to its assigned edge according to the cell's linear fractional position on this edge.

Epigenetic landscape reconstruction

To visualize the 3D transcriptional uncertainty landscape, we apply dimensional reduction techniques such as Principal Component Analysis (PCA) or t-SNE on the z-scored expression data, to project the gene expression of each individual cell on two dimensional axis, which gives the x-y axis of the landscape plot. For the z axis, we plot the NLL values. The transcriptional uncertainty landscape surface is reconstructed by estimating local approximation of individual cell 3D coordinates on a regular 30x30 grid by using a publicly available Matlab surface fitting code called GRIDFIT⁶⁷.

Pre-processing and analysis of single-cell expression datasets

Bargaje et al. scRT-qPCR dataset.

The dataset includes the expression profiles of 96 genes from 1896 single cells at 8 different time points (day 0, 1, 1.5, 2, 2.5, 3, 4, 5) during the differentiation of human pluripotent stem cells (iPSCs) into either mesodermal (M) or endodermal (En) fate³⁵. By employing CALISTA, we obtained five cell clusters and detected a bifurcation event, which gives rise to the two final cell fates. After lineage inference, we pseudotemporally ordered cells along the inferred differentiation paths (for more details, see⁴⁰).

Treutlein et al. scRNA-sequencing dataset.

The dataset includes the gene expression profiles of 405 cells during reprogramming of mouse embryonic fibroblast (MEF) into a desired induced neural (iN) and an alternative myogenic (M) cell fates³⁸. We pre-processed the data using CALISTA to select the 40 most variable genes (10% of the number of cells) for the transcriptional uncertainty analysis. CALISTA identified four different subpopulations and successfully recovered the bifurcation event (for more details, see⁴⁰)

Richard et al. scRT-qPCR dataset.

The dataset contains the expression profile of 91 genes measured from 389 cells at 6 distinct time points (0, 8, 24, 33, 48, 72 h) during the differentiation of primary chicken erythrocytic progenitor cells (T2EC)⁸. Following CALISTA pre-processing step, we removed cells in which less than 75% of the genes are expressed. Then, we selected the subset of genes with at least one non-zero expression values. A total of 354 cells and 88 genes were considered in the transcriptional uncertainty analysis. Based on eigengap heuristics^{40,47}, we grouped cells into 6 optimal clusters and ordered cells along the inferred linear trajectory (see Supplementary Figure S7).

Stumpf et al. scRT-qPCR dataset.

The dataset comprises the single-cell expression of 97 genes at 7 time points (0, 24, 48, 72, 96, 120, 168 h) during neural differentiation of mouse embryonic stem cells (E14 cell line)¹⁰. In the data pre-processing, we excluded cells in which less than 70% of genes are expressed. Then, we selected genes with at least one non-zero expression values. A total of 276 cells and 93 genes were considered for the transcriptional uncertainty analysis. Based on eigengap heuristics⁴⁰, we grouped cells into five optimal clusters and ordered cells along the inferred linear trajectory (Supplementary Figure S8).

Moussy et al. scRT-qPCR dataset.

The single-cell expression dataset includes normalized Ct values for 91 genes in 435 cells captured at 5 distinct time points (0, 24, 48, 72, 96 h) during human cord blood-derived CD34+ differentiation¹². We employed CALISTA to group cells into 7 clusters, reconstruct the developmental trajectory and calculate pseudotimes (Supplementary Figure S9).

Guo et al. scRT-qPCR dataset.

The dataset comprises the single-cell expression values of 48 genes from 387 individual cells isolated at 4 distinct developmental cell stages, from 8-cell stage mouse embryos to 64-blastocyst³⁶. By applying CALISTA, we identified seven different subpopulations along the differentiation process, and the inferred lineage hierarchy pinpointed two bifurcations events at 32- and 64-cell stage (Supplementary Figure S10). The timing of the lineage bifurcations coincides with two well-known branching points: one at 32-cell stage when totipotent cells differentiate into trophectoderm (TE) and inner cell mass (ICM), and another at 64-cell stage when ICM cells differentiate into primitive endoderm (PE) and epiblast (E).

Nestorowa et al. scRNA-sequencing dataset.

The dataset comprises single-cell gene expression of 1656 cells from mouse hematopoietic stem cell differentiation³⁹. We pre-processed the data by removing genes with non-zero values in less than 10% of the cells. Then, we selected 433 most variable genes, which is 10% of the number of genes after the previous pre-processing step, for the transcriptional uncertainty analysis⁴⁰. We set the optimal number of clusters based on the original study³⁹, which reported six different subpopulations and two bifurcation events: the first one producing common myeloid progenitor (CMP) from lymphoid-primed multipotent progenitors (LMPP), and the second one generating granulocyte–monocyte progenitors (GMP) from megakaryocyte–erythroid progenitors (MEP) (Supplementary Figure S11).

Moignard et al. scRT-qPCR dataset.

The dataset contains the single-cell expression level of 18 transcription factors measured in a total of 597 mouse bone marrow cells during hematopoietic differentiation. By applying CALISTA, we successfully identified the five subpopulations and the two branching points detected in the original study³⁷: long-term hematopoietic stem cells (HSC) differentiating into megakaryocyte–erythroid progenitors (PreM) or lymphoid-primed multipotent progenitors (LMPP); LMPP cells differentiating into granulocyte–monocyte progenitors (GMP) and common lymphoid progenitors (CLP) (for details see⁴⁰).

Pairwise correlation analysis of transcriptional uncertainty and transcriptional burst size and frequency.

We defined gene transcriptional burst size and burst frequency using the two-state model parameters, as follows:

$$\text{Burst size} = \frac{\theta_t}{\theta_{off}} \quad (1)$$

$$\text{Burst Frequency} = \theta_{on} \quad (2)$$

The burst sizes and burst frequencies are evaluated using the parameters $\theta^*(g, k) = \{\theta_{on}^*, \theta_{off}^*, \theta_t^*\}_g^k$ obtained from single-cell clustering analysis of CALISTA. Meanwhile, the average gene-wise NLL values for each single-cell cluster was computed as:

$$NLL_g^k = \frac{\sum_{n=1}^{N_k} NLL_g^n}{N_k} \quad (3)$$

where NLL_g^n is the negative log-likelihood of cell n based only on the expression of gene g , and N_k is the total number of cells in cluster k .

RNA VELOCITY ANALYSIS

Cells and genes were first filtered based on the pre-processing strategy in the original publication by La Manno and colleagues⁴², which resulted in a total of 1720 cells and 1448 genes from human glutamatergic neurogenesis, and a total of 18140 cells and 2141 genes from mouse hippocampus dataset. We further reduced the number of genes to only the top 500 highly variable genes for the transcriptional uncertainty analysis. The cell cluster assignments generated by Velocyto – the algorithm for computing RNA velocity from the original publication – were considered, instead of using CALISTA. Based on the clustering, we employed CALISTA to generate the lineage progression and cell pseudotimes (Supplementary Figure S12). The RNA velocity and transcriptional uncertainty values for the top 500 genes were calculated by employing Velocyto and CALISTA, respectively. The cell-wise RNA velocity was set to the Euclidean norm of the vector of RNA velocities for each cell, while the cell-wise NLLs was computed according to:

$$NLL_n^k = \frac{\sum_{g=1}^{500} NLL_g^n}{500} \quad (4)$$

CODE AVAILABILITY

A MATLAB and R version of CALISTA used in this study is freely available from the following website: <https://www.cabselab.com/calista>. All additional data are available from the corresponding author upon request.

DATA AVAILABILITY

All the public single cell data sets analysed in this study are available from the original publications.

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

This work was supported by the Swiss National Science Foundation (grant number 157154 and 176279) and ANR research grant SinCity (grant number ANR-17-CE12-0031-01).

ACKNOWLEDGEMENTS

We would like to thank all members of the SBDM team for lively discussions. We also thank the BioSyL Federation and the LabEx Ecofect (ANR-11-LABX-0048) of the University of Lyon for inspiring scientific events.

REFERENCES

1. Ohno, S., Nagai, Y., Ciccarese, S. & Iwata, H. Testis-Organizing H-Y Antigen and the Primary Sex-Determining Mechanism of Mammals. *Proc. 1978 Laurentian Horm. Conf.* 449–476 (1979) doi:10.1016/B978-0-12-571135-7.50014-X.
2. Herskowitz, I. A regulatory hierarchy for cell specialization in yeast. *Nature* **342**, 749–757 (1989).
3. Lewis, E. B. Clusters of Master Control Genes Regulate the Development of Higher Organisms. *JAMA J. Am. Med. Assoc.* **267**, 1524 (1992).
4. Davis, R. L., Weintraub, H. & Lassar, A. B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* **51**, 987–1000 (1987).
5. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–19 (2013).
6. Papili Gao, N., Ud-Dean, S. M. M., Gandrillon, O. & Gunawan, R. SINCERITIES: Inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* (2017) doi:10.1093/bioinformatics/btx575.
7. Sartorelli, V. & Puri, P. L. Shaping Gene Expression by Landscaping Chromatin Architecture: Lessons from a Master. *Mol. Cell* **71**, 375–388 (2018).
8. 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).
9. Mojtahedi, M. *et al.* Cell Fate Decision as High-Dimensional Critical State Transition. *PLOS Biol.* **14**, e2000640 (2016).
10. Stumpf, P. S. *et al.* Stem Cell Differentiation as a Non-Markov Stochastic Process. *Cell Syst.* **5**, 268–282.e7 (2017).
11. Semrau, S. *et al.* Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nat. Commun.* **8**, 1096 (2017).
12. Moussy, A. *et al.* Integrated time-lapse and single-cell transcription studies highlight the variable and dynamic nature of human hematopoietic cell fate commitment. *PLOS Biol.* **15**, e2001867 (2017).
13. Buganim, Y. *et al.* Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase. *Cell* **150**, 1209–1222 (2012).
14. Braun, E. The unforeseen challenge: from genotype-to-phenotype in cell populations. *Reports Prog. Phys.* **78**, 036602 (2015).

15. Kupiec, J. J. A chance-selection model for cell differentiation. *Cell Death Differ.* **3**, 385–90 (1996).
16. Kupiec, J. J. A Darwinian theory for the origin of cellular differentiation. *Mol. Gen. Genet. MGG* **255**, 201–208 (1997).
17. Paldi, A. Stochastic gene expression during cell differentiation: order from disorder? *Cell. Mol. Life Sci.* **60**, 1775–1778 (2003).
18. Gupta, P. B. *et al.* Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell* **146**, 633–44 (2011).
19. Thomas, P., Popović, N. & Grima, R. Phenotypic switching in gene regulatory networks. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 6994–9 (2014).
20. Huang, S. Non-genetic heterogeneity of cells in development: more than just noise. *Development* **136**, 3853–3862 (2009).
21. Rebhahn, J. A. *et al.* An animated landscape representation of CD4+ T-cell differentiation, variability, and plasticity: Insights into the behavior of populations versus cells. *Eur. J. Immunol.* **44**, 2216–2229 (2014).
22. Bhattacharya, S., Zhang, Q. & Andersen, M. E. A deterministic map of Waddington's epigenetic landscape for cell fate specification. *BMC Syst. Biol.* **5**, 85 (2011).
23. Brackston, R. D., Lakatos, E. & Stumpf, M. P. H. Transition state characteristics during cell differentiation. *PLOS Comput. Biol.* **14**, e1006405 (2018).
24. Ud-Dean, S. M. M. & Gunawan, R. Ensemble inference and inferability of gene regulatory networks. *PLoS One* **9**, e103812 (2014).
25. Bonnaffoux, A. *et al.* WASABI: A dynamic iterative framework for gene regulatory network inference. *BMC Bioinformatics* **20**, 220 (2019).
26. Ud-Dean, S. M. M. & Gunawan, R. Optimal design of gene knockout experiments for gene regulatory network inference. *Bioinformatics* **32**, 875–883 (2016).
27. 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).
28. Qiu, X. *et al.* Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe. *Cell Syst.* **10**, 265–274.e11 (2020).
29. Lv, C., Li, X., Li, F. & Li, T. Constructing the energy landscape for genetic switching system driven by intrinsic noise. *PLoS One* **9**, e88167 (2014).

30. Guo, J. & Zheng, J. HopLand: single-cell pseudotime recovery using continuous Hopfield network-based modeling of Waddington's epigenetic landscape. *Bioinformatics* **33**, i102–i109 (2017).
31. Fard, A. T., Srihari, S., Mar, J. C. & Ragan, M. A. Not just a colourful metaphor: modelling the landscape of cellular development using Hopfield networks. *npj Syst. Biol. Appl.* **2**, 16001 (2016).
32. Zwijsen, M. & Lawrence, N. D. Topslam: Waddington Landscape Recovery for Single Cell Experiments. *bioRxiv* 057778 (2017) doi:10.1101/057778.
33. Zhang, J. & Zhou, T. Topographer Reveals Stochastic Dynamics of Cell Fate Decisions from Single-Cell RNA-Seq Data. *bioRxiv* 251207 (2018) doi:10.1101/251207.
34. Shi, J., Teschendorff, A. E., Chen, W., Chen, L. & Li, T. Quantifying Waddington's epigenetic landscape: a comparison of single-cell potency measures. *Brief. Bioinform.* (2018) doi:10.1093/bib/bby093.
35. Bargaje, R. *et al.* Cell population structure prior to bifurcation predicts efficiency of directed differentiation in human induced pluripotent cells. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 2271–2276 (2017).
36. Guo, G. *et al.* Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell* **18**, 675–85 (2010).
37. Moignard, V. *et al.* Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat. Cell Biol.* **15**, 363–72 (2013).
38. Treutlein, B. *et al.* Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature* **534**, 391–395 (2016).
39. Nestorowa, S. *et al.* A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128**, e20–31 (2016).
40. Papili Gao, N., Hartmann, T., Fang, T. & Gunawan, R. CALISTA: Clustering and LINEAGE Inference in Single-Cell Transcriptional Analysis. *Front. Bioeng. Biotechnol.* **8**, 18 (2020).
41. Coulon, A., Gandrillon, O. & Beslon, G. On the spontaneous stochastic dynamics of a single gene: complexity of the molecular interplay at the promoter. *BMC Syst. Biol.* **4**, 2 (2010).
42. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
43. Peccoud, J. & Ycart, B. Markovian Modeling of Gene-Product Synthesis. *Theor. Popul. Biol.* **48**, 222–234 (1995).
44. Kim, J. & Marioni, J. C. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol.* **14**, R7 (2013).
45. Munsky, B., Neuert, G. & Van Oudenaarden, A. Using gene expression noise to understand gene

regulation. *Science* vol. 336 183–187 (2012).

46. Suter, D. M. *et al.* Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. *Science* (80-.). **332**, 472–474 (2011).

47. von Luxburg, U. A tutorial on spectral clustering. *Stat. Comput.* **17**, 395–416 (2007).

48. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Publ. Gr.* **14**, (2017).

49. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447-52 (2015).

50. Nicolas, D., Phillips, N. E. & Naef, F. What shapes eukaryotic transcriptional bursting? *Mol. Biosyst.* **13**, 1280–1290 (2017).

51. Antolović, V., Miermont, A., Corrigan, A. M. & Chubb, J. R. Generation of Single-Cell Transcript Variability by Repression. *Curr. Biol.* **27**, 1811-1817.e3 (2017).

52. Fritzsch, C. *et al.* Estrogen-dependent control and cell-to-cell variability of transcriptional bursting. *Mol Syst Biol* **14**, 7678 (2018).

53. Kriegstein, A. & Alvarez-Buylla, A. The Glial Nature of Embryonic and Adult Neural Stem Cells. (2009) doi:10.1146/annurev.neuro.051508.135600.

54. Han, X. *et al.* Construction of a human cell landscape at single-cell level. *Nature* (2020) doi:10.1038/s41586-020-2157-4.

55. Ye, Y., Yang, Z. & Lei, J. Using single-cell entropy to describe the dynamics of reprogramming and differentiation of induced pluripotent stem cells. *bioRxiv* 2020.04.13.040311 (2020) doi:10.1101/2020.04.13.040311.

56. Moris, N., Pina, C. & Arias, A. M. Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.* **17**, 693–703 (2016).

57. Moris, N. *et al.* Histone Acetyltransferase KAT2A Stabilizes Pluripotency with Control of Transcriptional Heterogeneity. *Stem Cells* **36**, 1828–1838 (2018).

58. Guillemin, A., Duchesne, R., Crauste, F., Gonin-Giraud, S. & Gandrillon, O. Drugs modulating stochastic gene expression affect the erythroid differentiation process. *PLoS One* **14**, e0225166 (2019).

59. Paldi, A. What makes the cell differentiate? *Prog. Biophys. Mol. Biol.* **110**, 41–43 (2012).

60. Argüello-Miranda, O., Liu, Y., Wood, N. E., Kositangool, P. & Doncic, A. Integration of Multiple Metabolic Signals Determines Cell Fate Prior to Commitment. *Mol. Cell* **71**, 733-744.e11 (2018).

61. Gu, W. *et al.* Glycolytic Metabolism Plays a Functional Role in Regulating Human Pluripotent Stem Cell State. *Cell Stem Cell* **19**, 476–490 (2016).
62. Folmes, C. D. L. *et al.* Somatic oxidative bioenergetics transitions into pluripotency-dependent glycolysis to facilitate nuclear reprogramming. *Cell Metab.* **14**, 264–71 (2011).
63. Richard, A. *et al.* Erythroid differentiation displays a peak of energy consumption concomitant with glycolytic metabolism rearrangements. *PLoS One* **14**, e0221472 (2019).
64. Moussaieff, A. *et al.* Glycolysis-Mediated Changes in Acetyl-CoA and Histone Acetylation Control the Early Differentiation of Embryonic Stem Cells. *Cell Metab.* **21**, 392–402 (2015).
65. Zhang, D. *et al.* Metabolic regulation of gene expression by histone lactylation. *Nature* **574**, 575–580 (2019).
66. Zwaka, T. P. Breathing chromatin in pluripotent stem cells. *Developmental Cell* vol. 10 1–2 (2006).
67. John D'Errico. Surface Fitting using gridfit. *MATLAB Central File Exchange* <https://www.mathworks.com/matlabcentral/fileexchange/8998-surface-fitting-using-gridfit> (2020).

FIGURE AND TABLE LEGENDS

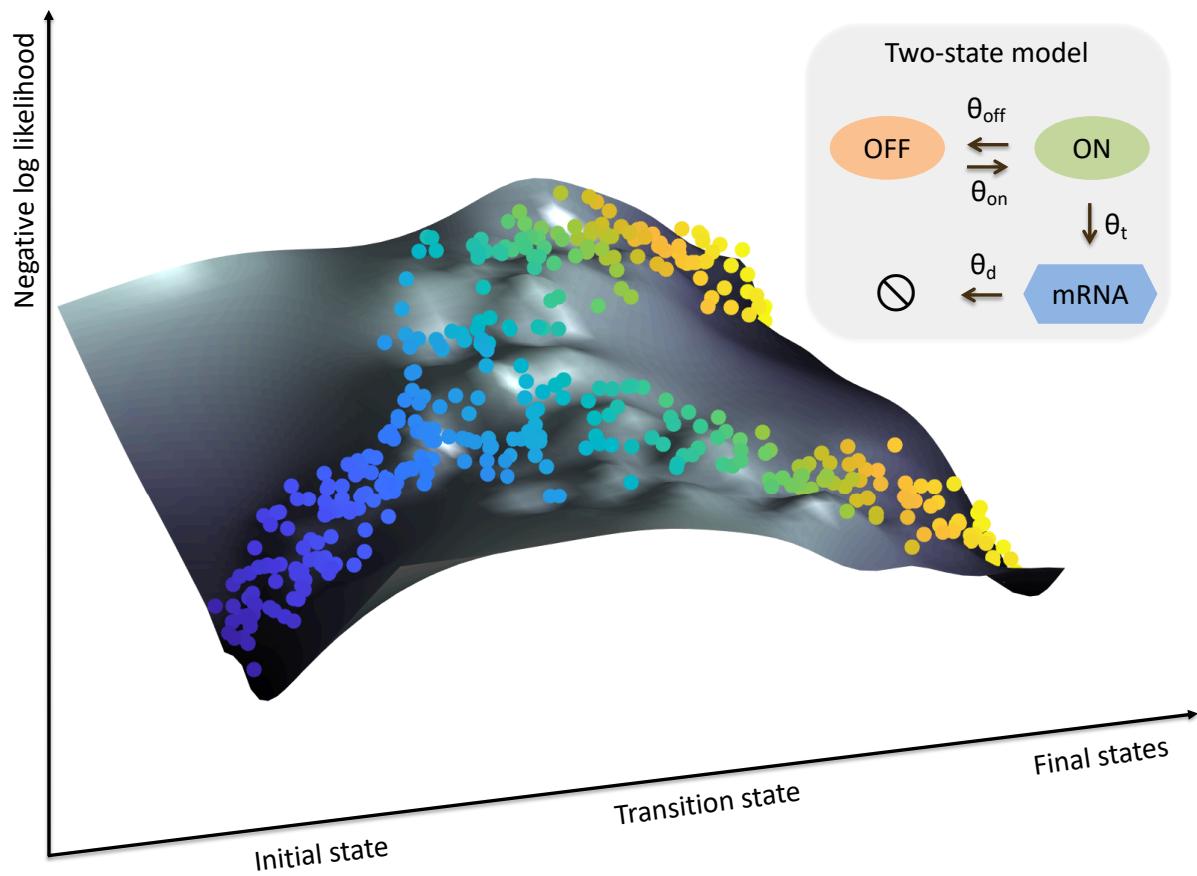


Figure 1. Single-cell transcriptional uncertainty landscape. The illustration depicts the landscape of single-cell transcriptional uncertainty during a differentiation process over the (pseudo) time (from blue to yellow). Each dot corresponds to a cell in the single-cell transcriptomic dataset. Cells start their journey from a valley in the landscape, through a hill, before ending at one of the final valleys / states.

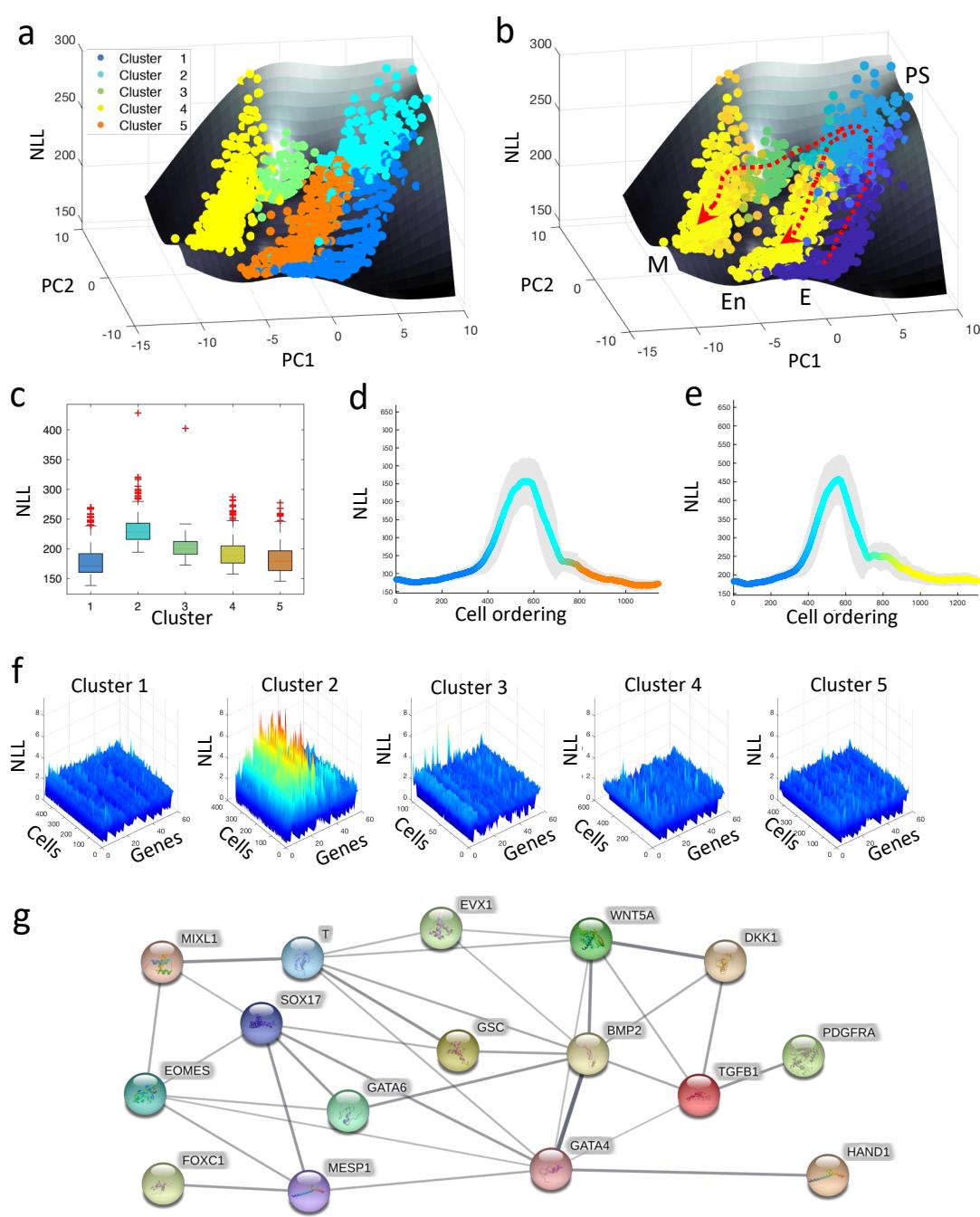


Figure 2. Analysis of single-cell transcriptional profiles during iPSC differentiation into cardiomyocytes.

The single-cell gene expression dataset was taken from the study of Bargajeet al.³⁵ (a-b) Epigenetic landscape plots estimates by CALISTA. Each dot on the landscape represents a cell where the colors indicate (a) the cell cluster, (b) pseudotime (from dark blue to yellow). The x-y axes of the landscape plots correspond to the first and second principal component (PC) coordinates, respectively, of the single-cell transcriptomic data. (c) Boxplots of NLL values for each single-cell cluster. (d-e) Moving-window average NLL along (d) endoderm and (e) mesoderm fate trajectory. (f) NLL of each gene and cell for every single-cell cluster. (g) Protein-protein interaction network of top variable genes inferred by STRING⁴⁹. Blue nodes represent transcription factors, while red nodes denote proteins involved in signal transduction. The width of the edges denotes the confidence for the inferred relationship (thicker edge = higher confidence).

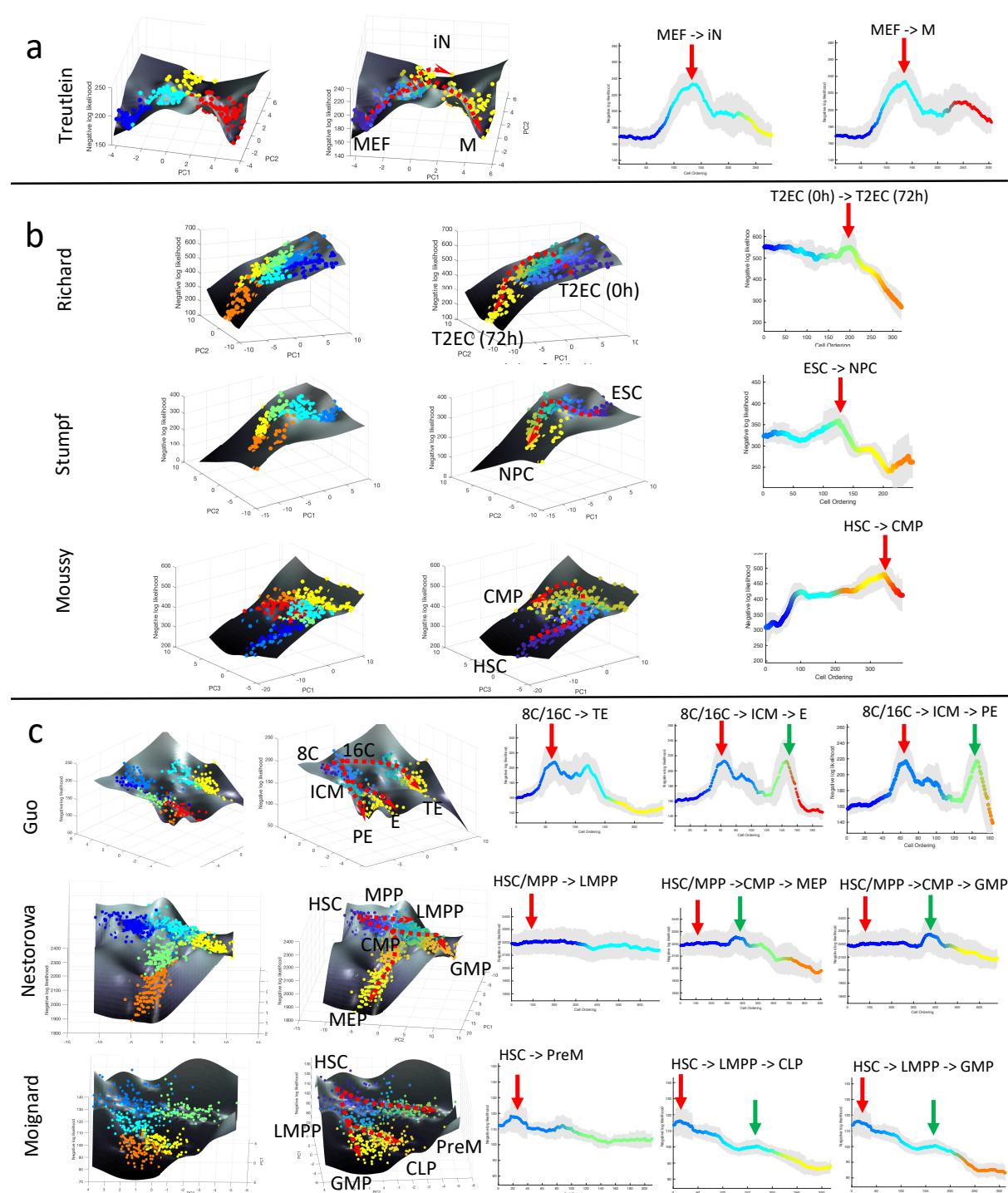


Figure 3. CALISTA analysis of single-cell expression data. (a-c) Landscape plots (based on cell clusters and pseudotime) and moving-averaged NLLs for each differentiation path of (a) single-branching trajectory (Treutlein dataset ³⁸), (b) linear trajectories (Richard ⁸, Stumpf ¹⁰, and Moussey ¹² datasets), (c) multi-braching trajectories (Guo ³⁶, Nestorowa ³⁹, and Moignard ³⁷ datasets). Green and red vertical arrows in moving-averaged NLL plots indicate the first and second peak in cell uncertainty, respectively.

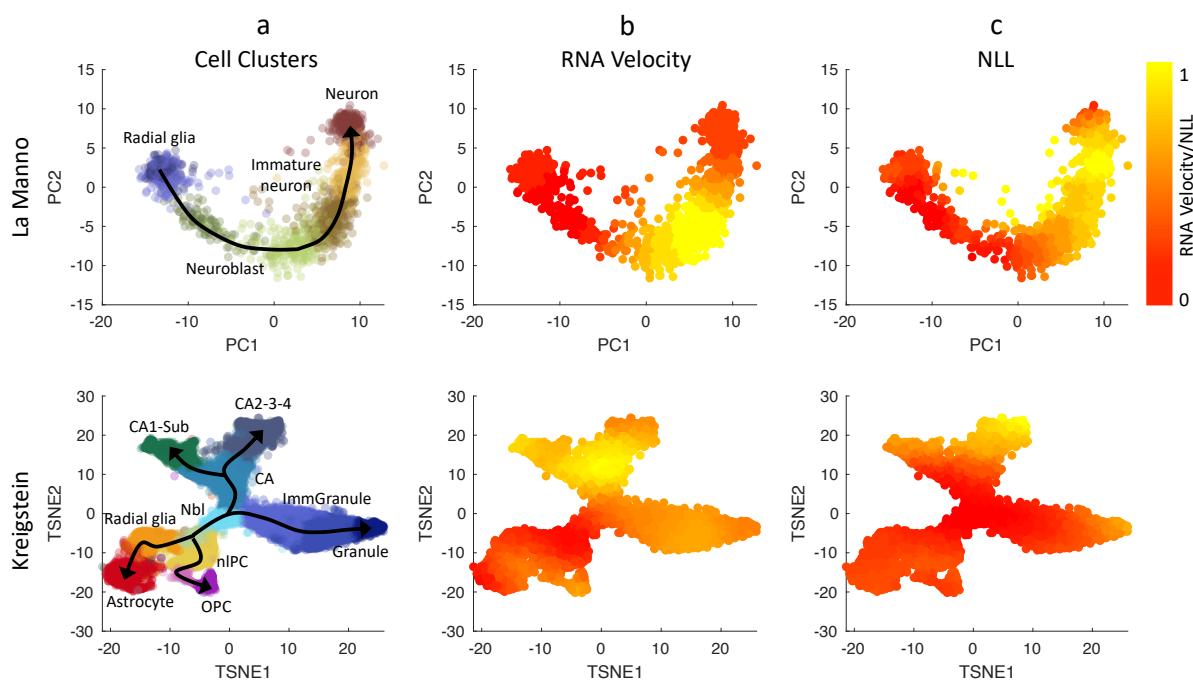


Figure 4. Comparison between RNA Velocities estimated using Velocyto⁴² and CALISTA NLL values.

(Top row) Human glutamatergic neurogenesis in La Manno et al. study⁴². (Bottom row) Mouse hippocampal neurogenesis in Kriegstein et al. study^{53,53}. (First column) Cell clustering assignments evaluated from Velocyto. Normalized values for Euclidean norm of RNA velocities (2nd column), CALISTA single-cell transcriptional uncertainty (NLL; 3rd column). The colors in the first column indicate the cell clusters, and those in the second-third columns indicate the normalized cell-wise RNA velocities and NLL values respectively.

Table 1. Pairwise correlation coefficients between transcriptional uncertainty and transcriptional burst frequency / burst size.

	Correlation with Transcriptional Uncertainty (p-value ≤ 0.01 in red boldface):	
	Burst Frequency	Burst Size
Bargaje et al.	0.74	0.70
Treutlein et al.	0.71	0.64
Richard et al.	0.68	0.55
Stumpf et al.	0.68	0.88
Moussy et al.	0.40	0.81
Guo et al.	0.75	0.82
Nestorowa et al.	0.73	0.71
Moignard et al.	0.04	0.83
La Manno et al.	0.78	0.55
Kriegstein et al.	0.77	0.32