Topography of epithelial-mesenchymal plasticity

Francesc Font-Clos[a], **Stefano Zapperi**[a], **and Caterina A. M. La Porta**[b]

[a]Center for Complexity and Biosystems, Department of Physics, University of Milan, via Celoria 16, 20133 Milano, Italy; [b]Center for Complexity and Biosystems, Department of Environmental Science and Policy, University of Milan, via Celoria 26, 20133 Milano, Italy

This manuscript was compiled on April 4, 2018

The transition between epithelial and mesenchymal states has fundamental importance for embryonic development, stem cell reprogramming and cancer progression. Here, we construct a topographic map underlying epithelial-mesenchymal transitions using a combination of numerical simulations of a Boolean network model and the analysis of bulk and single cell gene expression data. The map reveals a multitude of meta-stable hybrid phenotypic states, separating stable epithelial and mesenchymal states, and is reminiscent of the free energy measured in glassy materials and disordered solids. Our work elucidates not only the nature of hybrid mesenchymal/epithelial states but provides a general strategy to construct a topographic representation of phenotypic plasticity from gene expression data using statistical physics methods.

Epithelial (E) cells can transdifferentiate into mesenchymal (M) cells and vice-versa under a cohort of transcription factors, including the Snail and Zeb families (1). The epithelial-to mesenchymal transition (EMT), associated with the loss of cell-cell adhesion and the gain of invasive traits, is considered to be an hallmark of plasticity within a stem cell population and is particularly relevant for tumors. For this reason, a great effort has been devoted in the past to identify the critical biological functions regulating the EMT and its reverse, the mesenchymal to epithelial transition (MET). Almost 80% of human malignancies origin from epithelial tissues and a transition towards a mesenchymal phenotype is usually associated with a more aggressive potential (2–5). Emerging evidence shows that the EMT is a multiple process where cells express a mix of markers, both characteristic of E and M cells (6–8). These recent results are blurring the rigid distinction between epithelial and mesenchymal phenotypes, indicating that cancer cells can acquire hybrid E/M phenotypes, combining invasive capabilities with intracellular adhesion (9, 10), becoming extremely aggressive and associated to a poor patient outcome (11, 12).

According to an old and influential metaphor due to Waddington (13), the cell phenotype is analogous to a marble rolling over an *epigenetic landscape* and phenotypic plasticity corresponds to the marble crossing a hill separating different valleys. This landscape should correspond to the attractors of the kinetics of gene regulatory networks (14–20) and be encoded in gene expression data (21, 22). Here, we combine numerical simulations of a large Boolean model for the EMT-MET network with the analysis of a wide set of bulk and single cell gene expression data to reconstruct the topography underlying E/M plasticity. Genetic circuits regulating the EMT have been widely investigated theoretically with models ranging from simple switches composed by few genes (23) to large complex networks requiring extensive numerical simulations, both in discrete (24–26) and continuous time (27). Some of these models have provided insights in particular EM transitions, generating hypothesis that have later been experimentally tested (26). We show how these models can be used to rationalize and classify genetic drivers of the EMT and clarify the nature of hybrid E/M states guided by the Waddington picture (13).

Our results reveal that EMT/MET occurs across an extremely complex landscape characterized by a startling number of valleys and mountains organized according to a scale-free hierarchical statistical pattern. We observe a multitude of stable E/M states separated by a series of progressively less stable and more hybrid states that are increasingly prone to phenotypic changes in response to external perturbations. Hence, EMT and MET can take place in widely different locations and across multiple paths, in close analogy with non-equilibrium phase transitions in disordered solids (28, 29).

Model

To reconstruct the topographic landscape of E/M plasticity, we build on the large Boolean network model previously used to investigate EMT in hepatocellular carcinoma (25, 26). Since the model as it stands is hardwired towards EMT and MET is completely suppressed, we add to the model a missing contribution from the LIF/KLF4 pathway whose role for MET has been widely reported (30, 31) (see Fig. S1, Dataset S1 and SI for details). In this way we obtain a network of $N = 72$ nodes , whose state is defined by a string of binary variables $\{s_i\}$, determining if each gene/factor i is expressed/present ($s_i = 1$) or not ($s_i = -1$). Regulatory relations between two nodes i and j are encoded into a (non-symmetric) matrix J_{ij} taking the value $J_{ij} = 1$ if j promotes i and $J_{ij} = -1$ when j inhibits i (see Dataset S2). The network nodes evolves asynchronously according to a simple majority rule, so that the

Significance Statement

Cells can change their phenotype from epithelial to mesenchymal during development and in cancer progression, where this transition is often associated with metastasis and poor disease prognosis. Here we show this process involves the transit through a multitude of meta-stable hybrid phenotypes in a way that is similar to the driven dynamics of disordered materials. Our map, shows that highly aggressive hybrid epithelial/mesenchymal cell phenotypes are located in metastable regions that can easily switch under external and internal perturbations. Our general mapping strategy can be used for other pathways, providing a useful tool to visualize the ever increasing number of gene expression data obtained from single cells and tissues.

FFC analyzed data and performed numerical simulations. SZ and CAMLP designed and coordinated the project. FCC,SZ and CAMLP wrote the paper.

The authors declare no conflicts of interest.

[2]To whom correspondence should be addressed. E-mail: caterina.laportaunimi.it

node is set to $s_i = 1$ if the sum of its promoting interactions is larger than the sum of inhibitory ones (see Fig. 1a) (32). In case of ties, the node is not updated, keeping its present state. This evolution rule is the binary version of the *half-functional rule*[*] recently proposed in (27) to derive continuum kinetic reaction models and can be formally expressed as

$$s_i(t+1) = \text{sign}\left(\sum_j J_{ij} s_j(t)\right), \qquad [1]$$

which is the same equation used to simulate the zero-temperature dynamics in random ferromagnets (28) and spin glasses (29). Guided by this analogy, but keeping in mind that we are dealing with non-symmetric interactions, we show that the pseudo-Hamiltonian $H = -\sum_{i,j} J_{ij}\, s_i s_j$ is lowered under repeated application of the evolution rule Eq. (1) (see SI for full derivations and Figure S9), so that H provides a measure of the stability of a network state, with low-H states being more stable than high-H states.

Results

Simulated E/M topography displays fractal features. A phenotypic landscape associated to our EMT/MET network can be reconstructed by performing a large number ($M_0 = 10^7$) of simulations starting from random initial conditions until the network reaches a steady-state where s_i does not change[†]. In this way, we find a large number of distinct steady-states that can be projected into a two dimensional map using the principal component analysis (PCA). We classify these steady-states according to the expression of E-cadherin (CDH1) which we use as a reporter of the E/M phenotype (see Fig. 1b). The E/M map reconstructed from model shows a clear separation between E and M states with a boundary layer where E and M states coexist in very close proximity. A topographic representation of the stability of the states can be obtained by projecting H on the same two dimensional map (Fig. 1c) showing that the boundary layer is more elevated with respect to pure E/M states, suggesting that those states are less stable. Furthermore, the map displays a very rough topography, with two main valleys separated by a large barrier populated by smaller and smaller valleys.

Given the sheer amount of distinct steady states (see the inset of Fig. 1d), we resort to a statistical analysis and compute the probability distribution $P(a)$ of the relative abundances of the states, where a is the fraction of times we find a given state. Fig. 1d shows that $P(a)$ is a power law distribution indicating that most of the states are very rarely found (when a is small $P(a)$ is large) but few states are found multiple times (when a is large $P(a)$ is small). Alternative functional forms for $P(a)$ are discussed in SI and shown in Figure S10. The presence of a power law is a signature of a scale-free fractal organization of the map, as is also apparent by the correlation matrix of the states. Fig 1e shows the presence of large correlated clusters subdivided into smaller and smaller clusters. In the physics of disordered systems, a hierarchical organization of the states is traditionally revealed by a broad distribution $P(q_{\alpha\beta})$ of states overlap $q_{\alpha\beta} = \sum_i (s_i^\alpha s_i^\beta)/N$, measuring the similarity between two states $\{s_i^\alpha\}$ and $\{s_i^\beta\}$ (33). Hierarchical ground state structures have been observed in short-range Ising spin glasses, see (34, 35). When we restrict the sampling to low H states, $P(q_{\alpha\beta})$ displays a two peak structure indicating the presence of two classes of distinct and separate states (Fig 1f), but when we consider all steady-states the overlap distribution becomes very broad, resembling the one observed in spin glasses, as noticed long time ago for random Boolean networks(36–38).

Simulated phenotypic transitions reveal scale-free stochastic fluctuations. Once the topography associated with the E/M landscape has been established, we investigate how the landscape changes when each one of nodes is held fixed to $s_i = \pm 1$, which simulates overexpression (OE) or knock-down (KD) of the corresponding gene (see SI for details). As an example, Fig 2a and 2b report the one-dimensional projection of the topography under OE or KD of the SNAIL1 gene, a well known inducer of the EMT. SNAIL1 OE leads to a rightward tilt of the landscape, favoring the M phenotype, while under SNAIL1 KD the landscape tilts to the left, favoring the E state. This behavior is reminiscent of the effect of a magnetic field in a disordered magnet, where the free-energy landscape tilts in the direction of the field. If the network is initially in a E state, SNAIL1 OE can induce EMT but the success rate and the trajectory crucially depends on the initial state (see Fig. 2c), with high-H states much more likely to undergo EMT than low-H states (see Fig. S2). The variability in the outcome resulting from the OE/KD of a single gene can also be quantified by measuring the distribution of the number of nodes z affected by the process (see Fig. 2d). The distribution decays as power law $P(z) \sim z^{-\tau}$ up to a cutoff value that increases with the H-value of the initial state (see Fig. 2e), a further indication that high-H states are more susceptible to fluctuations (see also Fig. S2). The avalanche exponent of the power law distribution is $\tau \simeq 3/2$, a value expected for mean-field avalanches in driven disordered systems (28).

Using the model it is possible to perform OE/KD on all the nodes and estimate the probability of each node to induce EMT or MET (see Fig. 2f). Ranking the nodes as a function of their relevance for EMT we recover well known EMT inducers such as SNAIL1, ZEB1 or TGFβ, and MET suppressors such as KLF4 and mir-200. The general pattern is that an inducer of EMT by OE also induces MET by KD, and similarly for MET. We also simulate a transient version of OE/KD where a node is switched ($s_i \to -s_i$) but it is then allowed to eventually relax back to its previous state. The results summarized in Fig. S2 are similar to those obtained under stable OE/KD, for which the node variable is held fixed throughout the simulation, but the probability of EMT/MET is always smaller.

E/M topography inferred from gene expression data agrees with simulations. To confirm that the topographic representation of the E/M landscape obtained through the model provides an accurate representation of cellular phenotype, we examine the large cohort of gene expression data from human tissues provided by the GTEx project(39). In order to directly compare experimental data to the model, we design a simple binarization strategy to decide weather a gene is expressed or not in a particular sample or cell. To calibrate the binarization scale, we use skin cells and fibroblasts as reference E and M states, respectively, and set a threshold based on the expression distribution of each gene in these two data sets (See Fig. 3a and SI). Genes whose expression is above the

[*] Modifications of the model that include random local fields and their relation to network reconstruction errors are discussed in SI. See also Figure S8

[†] No limit cycles are found, see SI for details.

threshold are assigned to $s_i = 1$ and otherwise to $s_i = -1$. The same threshold can then be used to binarize all the 11688 transcriptomes from different tissues present in the GTEx database.

Using the topographic map of the E/M landscape constructed from simulations, we can now localize individual samples projecting their gene expression data on the map as shown in Fig. 3b. We then use the model to infer the stability of each phenotype by computing H associated to each state (Fig. 3c). When we plot skin cells and fibroblasts on a two dimensional map, we see that they correctly fall into E or M regions, respectively (see Fig. 3b), but not all samples have the same value of H (see Fig. 3c). We use the same strategy to localize on the same topographic map the entire set of tissues present in the GTEx database (see Fig. S4 and S5) and show that they cover all the available phase space. Assuming that the GTEx database contains an unbiased random sampling of all the available states –which is a reasonable assumption given that the GTEx project provides multi-tissue gene-expression data from healthy individuals only (39)– we analyze the statistical properties of these states. As shown in Fig. 3d, the abundance distribution derived from GTEx data decays as a power law with an exponent that is very close to the one found numerically (compare with Fig. 1d and see SI and Fig. S4 for technical details). Furthermore, clustering of the states reports a correlation matrix with hierarchical features that are in reasonable agreement with the prediction of the model (compare Fig. 3e with Fig. 1e). Finally, the overlap distribution displays a two peak structure when the statistics is restricted to fibroblasts and skin cells (Fig. 3f), while a single peaked distribution is found when using all the GTEx samples (Fig. 3g). This is in close agreement with the simulations results reported in Fig. 1 and confirms that experimental gene expression data give rise to a topographic landscape quantitatively similar to the one predicted by the model.

Tracing bulk and single cell RNAseq trajectories reveal the nature of hybrid E/M states. The topographic representation of E/M states derived above can be used to visualize and interpret RNAseq data obtained while the cells are undergoing phenotypic transformations. We first consider the classical example of TGF-β induced EMT in a human lung adenocarcinoma cell line (40). Fig. 4a reports the trajectory of the states obtained from the bulk RNAseq data recorded at different time points after TGF-β induction. As expected, the trajectory starts from the E region and crosses over to the M region of the map, as revealed by coloring the map according to the predicted expression of CDH1. Conversely, the trajectory obtained from RNAseq data for DOX induced MET during somatic cell reprogramming starts from the M valley and moves into the E valley of the landscape (30).

Our methodology is even more revealing when applied to single cell RNAseq (scRNAseq) data as shown in Fig 4c reporting the time course of the states obtained from scRNAseq data undergoing EMT during embryonic to endoderm differentiation (41) (see also Fig. S6 illustrating MET during fibroblast to cardiomyocyte reprogramming in single cell and bulk samples (42)). As time goes on, cells originally in the the E region transition to the M region between 24 and 36 hours. After this time even though EMT is apparently completed, the kinetic evolution of the cell population does not stop and the region occupied by single cell states shrinks. If we color the map by the predicted expression of other markers, we observe that the evolution moves cells in a low-KLF4 region (Fig. 4c see also Fig. S7 for similar maps for other markers). Hence, when applied to scRNAseq data our method can reveal subtle features associated with phenotypic transformations.

This last point is best illustrated by an analyzing recent data (6000 single cells) obtained from 18 head and neck squamous cell carcinoma patients (43). The original analysis revealed the presence of an aggressive cancer cell population, associated with metastasis and poor prognosis, described as partial-EMT (pEMT)(43). Classification of cells as pEMT was based on a pEMT score computed from the expression values of a set of 100 genes(43), none of which directly maps into nodes of our model. It is thus particularly remarkable to see that the projection of the scRNAseq data on our map reveals that tumor cells are correctly located into the E region of the map and cells with high pEMT score are typically located on higher ground with respect to low pEMT cells (see Fig. 4d). This is corroborated by the strong correlation between H and the the pEMT score, as reported in Fig. 4e.

Discussion

Our work builds on the premise that cell phenotypic plasticity should emerge from the activity of a complex gene regulatory network. The general assumption is that network activity and the ensuing phenotypes are primarily determined by the topology on the network, rather than the specific values of the rate constants of individual reactions (27). This allows us to rely on relatively simple Boolean networks, where individual nodes are only characterized by the presence or absence of activity (14). Application of this program to the EMT/MET networks unveils the topography of the epigenetic landscape (13) associated with this kind of phenotypic plasticity. The map reconstructed from the model and confirmed analyzing RNAseq data shows a rugged landscape with scale-free fractal-like features that are reminiscent of disordered solids and glassy materials (33).

A direct consequence of the landscape we uncover is that individual cells can be found in an extremely large variety of E or M states with intermediate or mixed states hierarchically organized between two sets of more stable and phenotypically well defined states. Intermediate E/M states are particularly prone to external perturbations which can lead to scale-free distributed avalanches with the potential to trigger extensive phenotypic changes. This extreme phenotypic plasticity is associated with highly aggressive behavior of tumor cells, as we show by analyzing recent scRNAseq data from head and neck carcinoma patients. Our topographic representation provides a quantitative representation of the cell phenotypic plastic potential, encoded here in the value of the pseudo-Hamiltonian H, that correlates extremely well with other independent measures of partial EMT. Furthermore, a topographic representation of the E/M phenotypes allows for a graphical representation of EMT and MET transitions in a variety of different contexts, from cancer to development and stem cell differentiation. Our general methodological strategy is not restricted to EMT but could be readily applied to other gene regulatory networks relevant to understand a variety of physiological functions and pathological conditions. The method appears to be a promising tool to build convenient

and accessible maps to orient ourselves among the exploding amount of single cell sequencing data.

Materials and Methods

Conversion of gene-level expression values to node-level binary states. We compute node-level expression values follows: All nodes except Hypoxia and miR200 are mapped to one or more genes, see Dataset S1. If expression data for more than one gene of a given node is available, we take the average of these for non-complexes and the minimum for complexes. We then binarize the node-level expression data using thresholds computed via a weighted average of the log2 expression of two reference samples (see Datasets section for details). We use a weighted average to avoid subsampling when the reference samples are of unequal size. The statistical significance of the binarization procedure is assessed with the Fisher's exact test. The EMT-MET model takes into account the localization of β-catenin by considering two separate nodes: one for β-catenin located in the nucleus, and one for β-catenin in the membrane. In gene-expression datasets, it is not possible to infer the localization of β-catenin looking only at the expression level of CTNNB1. To circumvent this issue, we consider β-catenin to be in the nucleus if its targets TCF/LEF are expressed, and in the membrane otherwise. If CTNNB1 is not expressed, the state of both nodes is set to -1 independently on the value of TCF/LEF.

Datasets. Data in Figure 3 comes from the GTEx project (39) and was downloaded from the GTEx portal (https://gtexportal.org/home/datasets) on 12/10/2017. We use samples labeled as "Cells - Transformed fibroblasts" and "Skin - Not Sun Exposed (Suprapubic)" as reference samples for binarization. The PCA basis in Figure 3(b,c) was computed using all GTEx samples. All nodes were included in this analysis. TGB-β-induced EMT data in Figure 4(a) was downloaded from the Gene Expression Omnibus, accession number GSE17708 (40), on 25/09/2017. We used $T = 0.5, 1$h and $T = 24, 72$h as reference samples for binarization. A total of 29 nodes with binarization p-value below 0.05 are included in the analysis. We use 10^7 steady states from the model, restricted to such nodes, to compute the PCA basis in Figure 4(a). Dox-induced MET data in Figure 4(b) was downloaded from the Gene Expression Omnibus, accession number GSE21757 (30), on 02/10/2017. We use $T = 0$d and $T = 21$d as reference samples for binarization. With one single sample per time-point, binarization p-values cannot be computed as explained above. As an alternative, we restrict the analysis to 47 nodes with fold-change greater than or equal to 0.5. We use 10^7 steady states from the model, restricted to such nodes, to compute the PCA basis in Figure 4(b). Single-cell data of embryonic-to-endoderm differentiation in Figure 4(c) was downloaded from the Gene Expression Omnibus, accession number GSE75748 (41), on 25/09/2017. We use $T = 0$h and $T = 96$h as reference samples. Given the large number of samples, the PCA basis in Figure 4(c) was computed using the experimental data. All nodes were included in the analysis. Head and neck cancers single-cell data in Figure 4(d,e) was obtained from the Gene Expression Omnibus, accession number GSE103322. We used epithelial and fibroblast samples as reference samples for binarization. The PCA basis was fitted to the single-cell data using all nodes. The pEMT score is computed as the average expression of the 100 genes that constitute the pEMT program in (43). Fibroblast-to-cardiomyocyte differentiation data in Figure S6 was downloaded from the Gene Expression Omnibus, accession numbers GSE98570 (bulk data) and GSE98567 (single-cell data) (42), on 22/11/2017. We used samples labeled as "control" and "reprogramming cells" as reference samples for single-cell data binarization, and samples labeled as "D0" and "D14" for bulk data binarization, and Single-cell data was used to fit the PCA basis in Figure S7.

ACKNOWLEDGMENTS. FFC and SZ are supported by ERC Advanced Grant n. 291002 SIZEFFECTS.

1. Ye X, Weinberg RA (2015) Epithelial-mesenchymal plasticity: A central regulator of cancer progression. *Trends Cell Biol* 25(11):675–86.
2. Huber MA, Kraut N, Beug H (2005) Molecular requirements for epithelial-mesenchymal transition during tumor progression. *Curr Opin Cell Biol* 17(5):548–58.
3. Rhim AD, et al. (2012) Emt and dissemination precede pancreatic tumor formation. *Cell* 148(1-2):349–61.
4. Sarrió D, et al. (2008) Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype. *Cancer Res* 68(4):989–97.
5. Aleskandarany MA, et al. (2014) Epithelial mesenchymal transition in early invasive breast cancer: an immunohistochemical and reverse phase protein array study. *Breast Cancer Res Treat* 145(2):339–48.
6. Bitterman P, Chun B, Kurman RJ (1990) The significance of epithelial differentiation in mixed mesodermal tumors of the uterus. a clinicopathologic and immunohistochemical study. *Am J Surg Pathol* 14(4):317–28.
7. Haraguchi S, Fukuda Y, Sugisaki Y, Yamanaka N (1999) Pulmonary carcinosarcoma: immunohistochemical and ultrastructural studies. *Pathol Int* 49(10):903–8.
8. Paniz Mondolfi AE, et al. (2013) Primary cutaneous carcinosarcoma: insights into its clonal origin and mutational pattern expression analysis through next-generation sequencing. *Hum Pathol* 44(12):2853–60.
9. Revenu C, Gilmour D (2009) Emt 2.0: shaping epithelia through collective migration. *Curr Opin Genet Dev* 19(4):338–42.
10. Yu M, et al. (2013) Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science* 339(6119):580–4.
11. Jolly MK, et al. (2016) Stability of the hybrid epithelial/mesenchymal phenotype. *Oncotarget* 7(19):27067–84.
12. George JT, Jolly MK, Xu S, Somarelli JA, Levine H (2017) Survival outcomes in cancer patients predicted by a partial emt gene expression scoring metric. *Cancer Res* 77(22):6415–6428.
13. Waddington C (1957) *The Strategy of the Genes.* (Allen & Unwin, London).
14. Kauffman SA (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theor Biol* 22(3):437–67.
15. Huang S, Eichler G, Bar-Yam Y, Ingber DE (2005) Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys. Rev. Lett.* 94(12):128701.
16. Huang S, Guo YP, May G, Enver T (2007) Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Dev Biol* 305(2):695–713.
17. Wang J, Xu L, Wang E, Huang S (2010) The potential landscape of genetic circuits imposes the arrow of time in stem cell differentiation. *Biophys J* 99(1):29–39.
18. Wang J, Zhang K, Xu L, Wang E (2011) Quantifying the waddington landscape and biological paths for development and differentiation. *Proc Natl Acad Sci U S A* 108(20):8257–62.
19. Huang S (2012) The molecular and mathematical basis of waddington's epigenetic landscape: a framework for post-darwinian biology? *Bioessays* 34(2):149–57.
20. Li C, Wang J (2013) Quantifying waddington landscapes and paths of non-adiabatic cell fate decisions for differentiation, reprogramming and transdifferentiation. *J R Soc Interface* 10(89):20130787.
21. Scialdone A, et al. (2016) Resolving early mesoderm diversification through single-cell expression profiling. *Nature* 535(7611):289–293.
22. Bargaje R, et al. (2017) Cell population structure prior to bifurcation predicts efficiency of directed differentiation in human induced pluripotent cells. *Proc Natl Acad Sci U S A* 114(9):2271–2276.
23. Jolly MK, et al. (2014) Towards elucidating the connection between epithelial-mesenchymal transitions and stemness. *J R Soc Interface* 11(101):20140962.
24. Li F, Long T, Lu Y, Ouyang Q, Tang C (2004) The yeast cell-cycle network is robustly designed. *Proc. Natl. Acad. Sci. U. S. A.* 101(14):4781–4786.
25. Steinway SN, et al. (2014) Network modeling of tgf-beta signaling in hepatocellular carcinoma epithelial-to-mesenchymal transition reveals joint sonic hedgehog and wnt pathway activation. *Cancer Research* 74(21):5963–5977.
26. Steinway SN, et al. (2015) Combinatorial interventions inhibit tgfβ-driven epithelial-to-mesenchymal transition and support hybrid cellular phenotypes. *NPJ Syst Biol Appl* 1:15014.
27. Huang B, et al. (2017) Interrogating the topological robustness of gene regulatory circuits by randomization. *PLOS Computational Biology* 13(3):1–21.
28. Sethna JP, et al. (1993) Hysteresis and hierarchies: Dynamics of disorder-driven first-order phase transformations. *Phys. Rev. Lett.* 70(21):3347–3350.
29. Pázmándi F, Zaránd G, Zimányi GT (1999) Self-organized criticality in the hysteresis of the sherrington-kirkpatrick model. *Phys. Rev. Lett.* 83(5):1034–1037.
30. Samavarchi-Tehrani P, et al. (2010) Functional genomics reveals a BMP-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming. *Cell Stem Cell* 7(1):64–77.
31. Li R, et al. (2010) A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts. *Cell Stem Cell* 7(1):51–63.
32. Bornholdt S (2008) Boolean network models of cellular regulation: prospects and limitations. *J R Soc Interface* 5 Suppl 1:S85–94.
33. Mezard M, Parisi G, Virasoro MA (1987) *Spin Glass Theory and Beyond.* (World Scientific).
34. Hed G, Hartmann AK, Stauffer D, Domany E (2001) Spin domains generate hierarchical ground state structure in J = +/-1 spin glasses. *Phys. Rev. Lett.* 86(14):3148–3151.
35. Hed G, Young AP, Domany E (2004) Lack of ultrametricity in the low-temperature phase of three-dimensional ising spin glasses. *Phys. Rev. Lett.* 92(15):157201.

36. Derrida B, Flyvbjerg H (1986) Multivalley structure in kauffman's model: analogy with spin glasses. *Journal of Physics A: Mathematical and General* 19(16):L1003.
37. Miranda EN, Parga N (1988) Ultrametricity in the kauffman model: a numerical test. *Journal of Physics A: Mathematical and General* 21(6):L357.
38. Bastolla U, Parisi G (1996) Closing probabilities in the kauffman model: An annealed computation. *Physica D: Nonlinear Phenomena* 98(1):1 – 25.
39. Carithers LJ, et al. (2015) A novel approach to high-quality postmortem tissue procurement: The GTEx project. *Biopreservation and Biobanking* 13(5):311–319.
40. Abnaof K, et al. (2014) Tgf-beta stimulation in human and murine cells reveals commonly affected biological processes and pathways at transcription level. *BMC Systems Biology* 8(1):55.
41. Chu LF, et al. (2016) Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biology* 17(1).
42. Liu Z, et al. (2017) Single-cell transcriptomics reconstructs fate conversion from fibroblast to cardiomyocyte. *Nature* 551(7678):100–104.
43. Puram SV, et al. (2017) Single-Cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* 0(0).

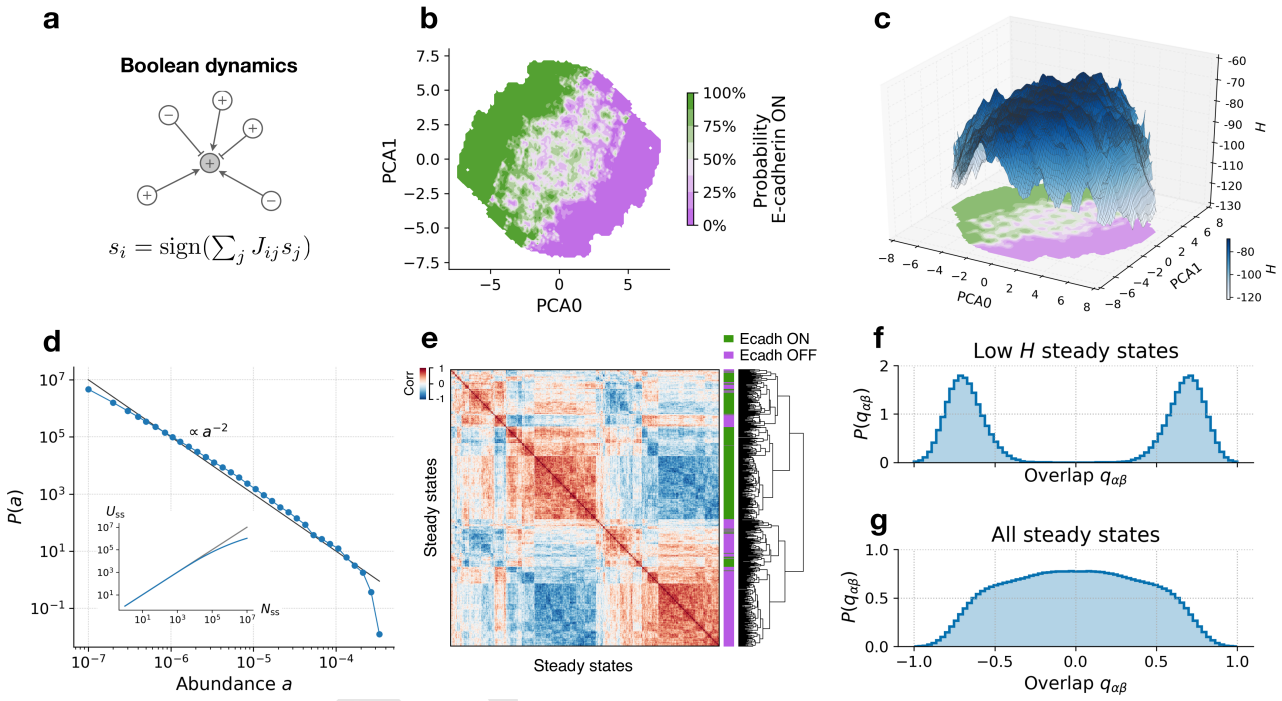

Fig. 1. The topography of E/M states displays a hierarchical complex structure. (a) Illustration of the Boolean update rule. The state of a node s_i depends on the state of its promoters ($J_{ij} = +1, \rightarrow$) and inhibitors ($J_{ij} = -1, \dashv$). (b) PCA projection of 10^6 steady states. Color corresponds to the ratio of steady states that express E-cadherin. The panel shows intricate patterns of transition between areas of high/low Ecadherin expression probability, colored in green/violet shades. (c) 3D reconstruction of topography of EMT. The xy-projection reproduces the data in (b). The z-axis corresponds to the value of H, showing that high-H states (colored in darker blue shades) coincide with the central transition area in (b). (d) Distribution of steady-state abundances, computed from 10^7 steady states of the EMT model (blue symbols). The relative abundance a of a steady state is the fraction of times it is found, starting from random initial conditions. The black line of slope -2 is shown only as a guide to the eye. The inset shows the number of distinct steady states U_{ss} as a function of the total number of steady-states N_{ss} found in simulations. (e) Clustering of steady states, computed using 500 steady states of the model. The heatmap shows correlation between steady states. Colors adjacent to the dendogram mark the expression of E-cadherin (green) or lack of expression (violet). States expressing E-cadherin cluster together but display additional hierarchical organization. (f) Overlap distribution over the 20% of steady states with lowest H. A two-peak distribution marks the presence of two symmetric sets as in disordered magnets. (g) The broad overlap distribution over all steady states resembles the one observed in spin glasses.

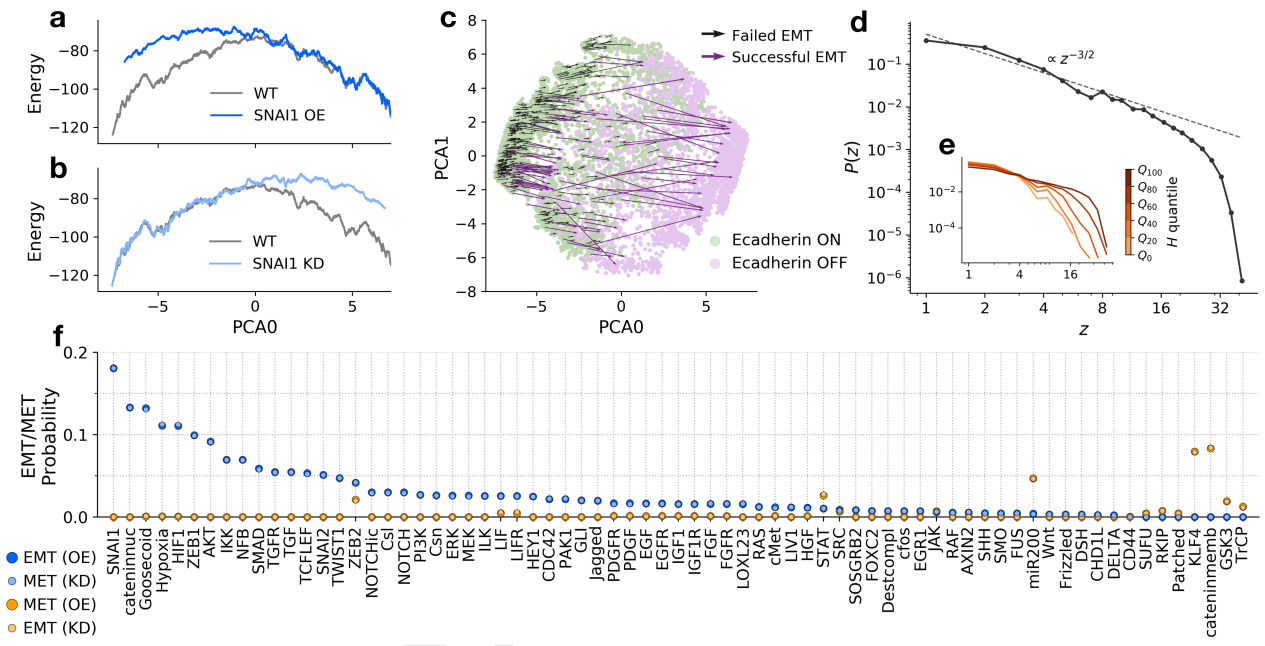

Fig. 2. EMT/MET occurs with different probabilities through multiple paths. The model shows many forms of EMT/MET, and these occur with different probabilities. (a,b) One-dimensional PCA projection of the H landscape where (a) Over-expression (OE) or (b) knock-out (KD) of SNAI1 tilts the landscape towards the M or E regions, respectively. (c) Transition map under SNAI1 OE. The model displays different forms of SNAI1-induced EMT. d) The distribution of gene expression avalanches after individual KD/OE is a power law with exponent $\tau \simeq 1.5$. e) The cutoff of the distribution depends on H, quantified here by quartiles, with high H states producing larger avalanches. (f) EMT/MET probabilities under KD/OE conditions. The model lays out a non-deterministic picture of EMT/MET where well-known factors such as SNAI1 (EMT) or KLF4 (MET) induce phenotypic transitions with higher probability (see Materials and methods and Fig. S2 for further details).

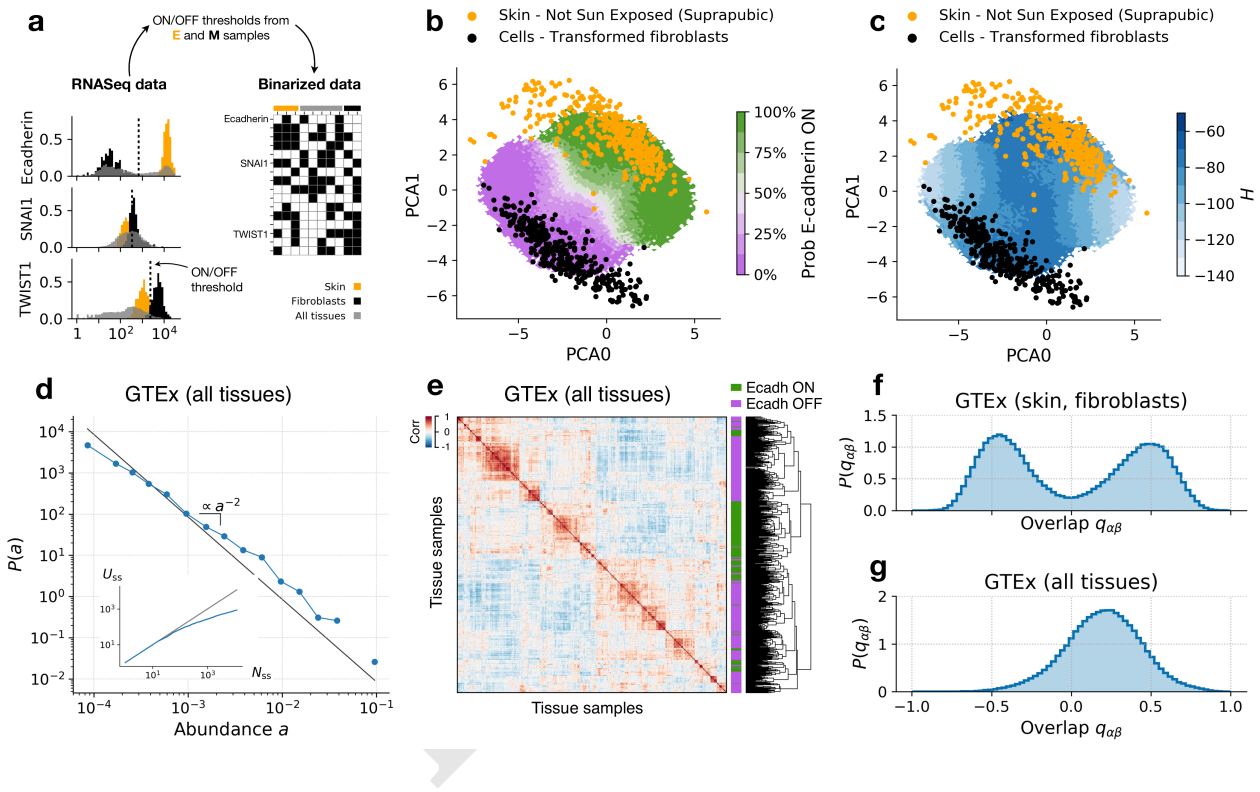

Fig. 3. Multi-tissue gene expression data display statistical features in agreement with simulations. (a) Illustration of the binarization process (see Materials and methods for details). Gene-level expression data is casted into node-level binary data using binarization thresholds, computed using two reference samples (orange and black coloring). (b) Skin (orange) and fibroblasts (black) samples from the GTEx project projected in PCA space. The E-cadherin expression probability in the model is shown with green (100%) to violet (0%) shades. Fibroblast samples tend to be in areas of very low E-cadherin expression probability. (c) Same as (b) but coloring the model steady states by average H. (d) Distribution of abundances, computed using all GTEx binarized samples and the 14 most relevant nodes, see SI and Figure S3 for details. (e) Clustering of 500 GTEx samples (all tissues), displaying a hierarchical structure qualitatively similar to that of the model (compare with Figure 1e). (f) Overlap distribution over skin and fibroblast samples from the GTEx project (compare with Figure 1f). (g) Overlap distribution over all GTEx samples. (compare with Figure 1g).

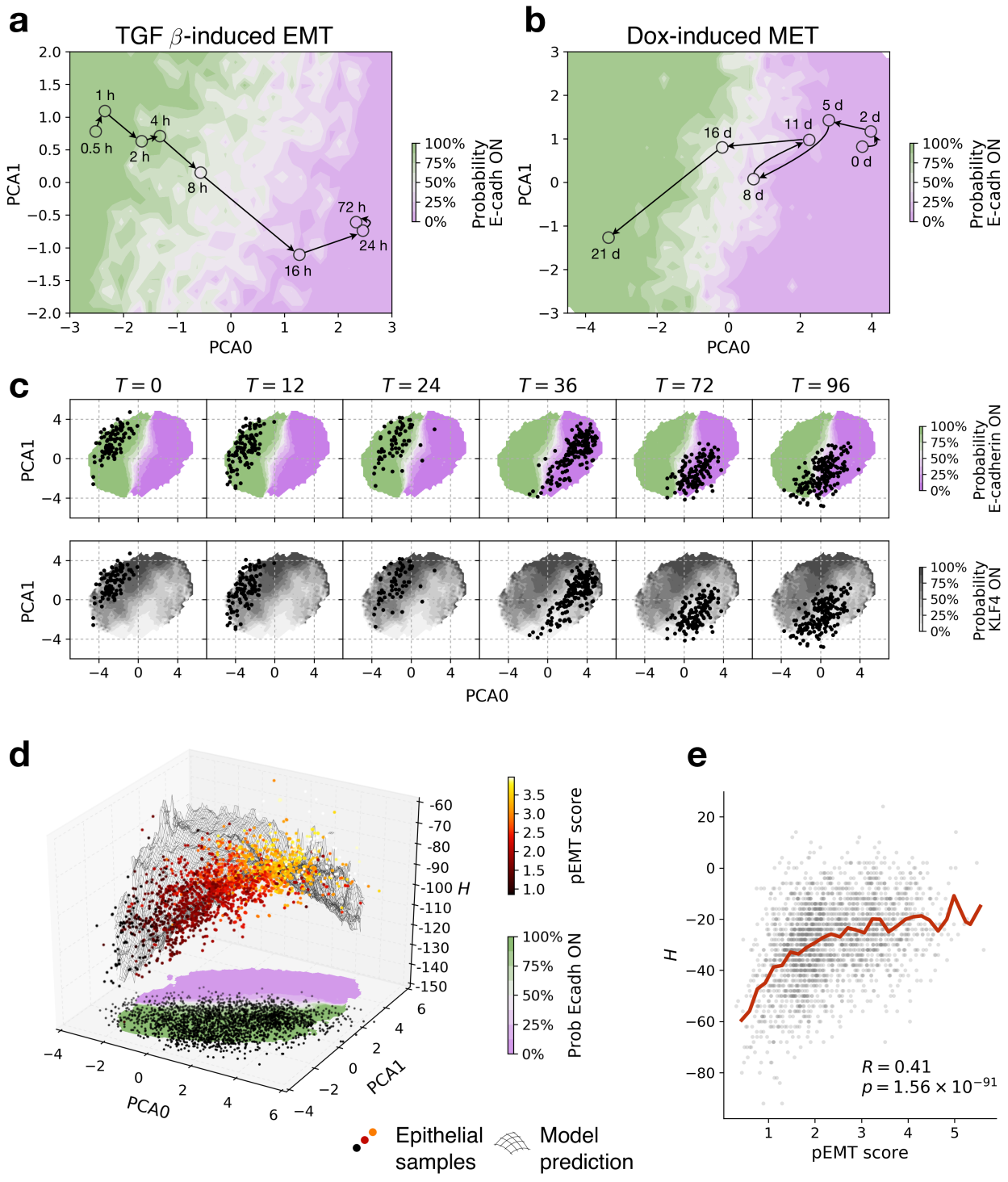

Fig. 4. Single cells and bulk transcriptomic data yield trajectories through the E/M map with putative hybrid states lying on high H regions. (a) Data from TGF-β-treated lung adenocarcinoma cell lines (GSE17708 (40)) yield a trajectory moving from the E to the M region. (b) Data from Dox-induced somatic cell reprogramming (GSE21757 (30)) display a reverse trajectory from M to E. Experimental data are shown as colored symbols with time course marked with arrows. The colored background depends on the ratio of steady states of the model that express E-cadherin at a given location in PCA coordinates, ranging from 100% (green) to 0% (violet). (c). Experimental data from single-cell embryonic-to-endoderm differentiation (GSE75748 (41)) move across the map as cells undergo EMT. The background color indicates the ratio of steady states that express E-cadherin or KLF4 (see Fig. S7 for more markers). (d) Localization of single cell gene expression data from tumor cells obtained from head and neck squamous cell carcinoma patients (43). All tumor cells correctly lie in the E region of the map, with high pEMT-scored cells located towards high-H areas. (e) The pEMT score correlates with H. Each gray dot represents a single cell. R and p denote the Pearson correlation coefficient and its associated p-value (Student's t test, two-tailed). The red line shows the average H