Tipping-point analysis uncovers critical transition signals from gene expression profiles

Running title: **Significant Critical-Transitional Signal**

Xinan H Yang[1,*], Zhezhen Wang[1], Andrew Goldstein[2], Yuxi Sun[1], Dannie Griggs[1], Antonio Feliciano[1], Yanqiu Wang[3], John M Cunningham[1]

[1] Department of Pediatrics, [2] Department of Statistics, The University of Chicago, Chicago, USA, [3] Stanford University, Stanford, USA

*: xyang2@uchicago.edu

Highlights:

- Shrinkage estimation reduces errors in the empirical correlation matrix thus inferring tipping point accurately
- Refined feature selection and random walk isolate the early-warning signal for tipping point from background noise
- Identification of the intermediate-risk stage in neuroblastoma as the most vulnerable and susceptible to treatment
- First tipping-point interpretation of Etv2 oscillation that guides the haemato-endothelial bifurcation

Abstract
Tipping points existing in many biological processes. Identifying gene expression signals in the presence of entropic background is essential to decipher biological tipping points. We introduce a new method BioTIP which accurately, robustly, and reliably identifies tipping points and their significant gene signals. Through application to three benchmark datasets and comparison to previous analyses, we demonstrate that not only can BioTIP be applied to understanding of disease and developmental, but also to data at the cellular or organismal level. We applied BioTIP to two exemplary contexts on which tipping-point analyses had never been done and validated the statistical significance of the identified gene signals using independent datasets. By a literary investigation, we showed how the identified gene signals provide insight into regulatory networks of master transcription factors modulating those transitions. Our open-source software allows for future investigations using tipping point methodology to better characterize the plasticity, heterogeneity, and rapid switches in biological processes.

Key words: cell differentiation / computational biology / correlation estimation / critical transition / disease progression

Introduction

The tipping-point theory has been used to characterize critical transitions in ecosystems, climates, and other complex systems in various disciplines (Clements, McCarthy et al., 2019, Lenton & Livina, 2016). Within this theory, a system with discrete states can exhibit abrupt and often irreversible shifts between two stable states. These shifts, or critical transitions, occur when small perturbations build up to a tipping point, whereby the system moves from one stable state to another. Biological state transitions are often associated with a rapid change in phenotypes (Scheffer, Carpenter et al., 2012). Since these rapid changes are often irreversible in a dynamic process, it is crucial to identify where these tipping points lie and what drives them.

In biological systems, transcription tipping points have implications from the organismal to cellular level. On an organismal level, bulk cells (or tissues) are grouped into different states based on their disease state or more generally, their macroscopic phenotypes (Chen, Liu et al., 2012, Liu, Chang et al., 2017, Oku & Aihara, 2018, Yang, Li et al., 2015). On a cellular level, individual cells are grouped based on their stage of development (Ando, Kato et al., 2019, Mojtahedi, Skupin et al., 2016, Zhao & Choi, 2019). The theoretical underpinning for critical transitions is the same at both levels. Samples of the same state are biological replicates or statistical ensembles, typically presenting similar gene expression patterns and thus low between-sample variance. Another way to conceptualize this is that each phenotype represents a stable state where gene expression regulations are in equilibrium (e.g., maintained by negative feedbacks responding to perturbations). If a perturbation crosses the threshold of maintaining the previous equilibrium, the system will transition into different states. Impending a transitional

threshold, it is common to see a group of genes whose expression profiles start to exhibit larger between-sample variance and correlation (**Fig 1a**, top embedded red panel) (Chen et al., 2012, Clements et al., 2019). The expression patterns of this group of genes thus serve as early-warning signals for an impending critical transition. Detecting early-warning signals in disease systems is of prime interest for forecasting disease onset. In cellular systems, detection of early-warning signals indicates which stages of development are adaptable to multiple lineage fates and therefore have great treatment potential (Pijuan-Sala, Griffiths et al., 2019).

There are two arguments for how transcriptional tipping points should be theorized. The first argument, upon which this investigation is based, is that transcriptomic oscillation during tipping points is significant and should be conceptualized as dynamic expression regulations (Bockamp, Fordham et al., 1998, Hu, Krause et al., 1997, Sheikh & Groom, 2020). The other body of research argues the fluctuation in gene expression as the increased entropy during tipping points thus is completely stochastic (Antolovic, Lenn et al., 2019, Buganim, Faddah et al., 2012). Here, we provide evidence for the former argument -- when there are transcriptional oscillations driving state transitions, they can be extracted from the chaotic expression patterns.

Within this conceptualization of expression regulations, transcription factors (**TFs**) are very important. TFs are sequence-specific DNA-binding factors that regulate gene expression. The activities of certain keys, called master TFs, are known to initiate and maintain stable cell states (Bockamp et al., 1998, Garry, 2016, Sheikh & Groom, 2020). Tipping points occur when one master TF loses activity and another gains activity; during this type of transition, the system is vulnerable to perturbations (**Fig 1a,** bottom).

A practical bottleneck in studying the transcriptional tipping point, however, has been the lack of suitable computational methods. Classical models require a time series in which the same sample (cell or tissue) is collected at various points in time. This type of longitudinal transcriptomic data is rarely available. In the absence of longitudinal data, we have renamed the historical 'early-warning indicator for tipping point' (which was developed for time-series data) as a 'Critical-Transition Signal' (**CTS**). The term CTS helps us consider the application of tipping-point theory for broad biological significance with cross-sectional data analysis.

There exist a few expression analyses that can be used to predict tipping points. Methods like pseudo-ordering analysis of cellular states (Lummertz da Rocha, Rowe et al., 2018, Wang, Karikomi et al., 2019), single-cell clustering using bifurcation analysis (Marco, Karp et al., 2014), unstable-cluster marking using the silhouette-width method (Rousseeuw, 1987), and cellular entropy changes (Mojtahedi et al., 2016, Teschendorff & Enver, 2017) can identify bifurcation, but they are unable to extract significant CTS underpinning state transitions.

The most promising yet flawed method for identifying expressional CTS is a method called Dynamic Network Biomarker (**DNB**) (Chen et al., 2012, Liu et al., 2017, Yang, Li et al., 2018, Yang et al., 2015, Zhang, Liu et al., 2019). However, DNB is designed for statistical assemblies of states with high expressional dispersion. It takes the average across all samples per state and pre-selects genes that are differentially expressed between states. Consequently, DNB may omit

important genes with large expression variance from individual to individual. Another key method has been the critical state-transition index (**Ic**) (Mojtahedi et al., 2016) which was a breakthrough in studying unstable tipping points. However, Ic has a bias towards identifying small-sized states as tipping points (for which we document in this study).

We introduce a comprehensive biological tipping-point (**BioTIP**) scoring scheme to combine tipping point and CTS identifications into one analytic pipeline. Built on significant CTS identifications, we can infer master TFs which play significant roles in regulating phenotypic state transition. BioTIP's performance was evaluated through application to three benchmark datasets (**Table S1**) and comparison to previous analyses, demonstrating that not only can BioTIP be applied to questions of disease and developmental, but also to data at the cellular or organismal level. BioTIP was also applied to two subjects on which tipping point analyses had never been done before, and the statistical significance of the results were validated using independent datasets. A literary investigation into the results of these novel analyses showed how a significant CTS provides insight into regulator networks of master transcription factors. Thus, BioTIP opens a way for characterizing critical transitions from dynamic expression profiles of diseases or developmental biology.

Results

BioTIP overview

We developed BioTIP's pipeline to robustly identify biological tipping points and significant CTS. The strategy is based on the mathematical approaches that quantify three 'early-warning' features of critical state transitions: increased expression variability, increased between-gene correlation, and decreased between-replicate correlation. To this end, we model the probability distribution of discrete states under two common premises: 1) the system of the individual replicate (sample) has a 'dissipative' structure (e.g., having discrete states including the one showing multi-stability); and 2) each state has a characteristic gene expression profile and thus presents a distinct molecular phenotype. BioTIP applies to data meeting these premises, including both single-cell and bulk-cell transcriptomes.

There are five steps to conduct BioTIP analysis, grouped into three main components (**Fig 1b**). The first component (step i, **Fig EV1**) is to predict tipping points from entropic background signals, using an improved Ic score we have coined Ic*. The second component (steps ii-iv) detects significant CTSs, which improves upon existing DNB methods. Finally, the last component (step v) evaluates the significance of each identified CTS (for a tipping point) using the Delta statistics.

Ic* Detects Tipping Points More Accurately Than the Original Ic Scoring System

We first discussed two shortcomings of the existing Ic method for tipping-point identification, using statistical-resampling based approaches. First, Ic is calculated per state and compared

among states. By calculating between-gene correlations over samples in each state, Ic has a bias towards small states, sometimes making wrong predictions. We characterize this bias in two examples. We first applied the Ic method to a benchmark single-cell RNA-seq dataset (**scRNA-seq**), where the largest state size was triple that of the smallest state size. We then repeated the same Ic analysis to the same dataset but after down sampling the data to equal-sized states. Unintendedly, the earlier full-sized analysis and the later equal-sized analysis yielded different results (**Fig 2, b vs a**). This discrepancy is explained by the fact that Ic's algorithm inflates between-gene correlation within a small-sized state (**Fig EV2, b vs a**). Ic's bias was also demonstrated by the fact that applying MCI and Ic to the same dataset can yield disparate results. When both methods were applied to a dataset of disease progression, MCI indicated a significant tipping point at the 'LIR' state (described later) which was rejected by Ic and its Delta scoring (**Fig 2, c and d**). This disagreement can be attributed to the fact that Ic calculations were incorrect by the up to 9-fold difference in population sizes.

The novel **Ic*** score corrects both issues. In the first example described above, Ic* predictions were found to be consistent regardless of the inputted cell numbers per state (**Figs 2, f vs e; EV2 d vs c**). Applying **Ic*** to the second example detected the LIR state identified by MCI and was significant when compared to random gene values (**Fig 2g**) or to random sample assemblies (**Fig 2h**, P<0.001). Thus, we conclude that **Ic***predicts transcriptional tipping points accurately, in tissue and single-cell samples.

Note that although **Ic*** of randomly-selected genes happens to predict tipping points' phenomenon of increasing entropy (**Fig 2, e-f**, boxes, Wilcox P<2e-16), it could be unpractical in heterogeneous expression analysis, e.g., human tumor data (**Fig 2g**, boxes).

Unlike DNB, BioTIP identifies only non-random CTS

We validate BioTIP's CTS identification in two datasets with known tipping points. To identify significant CTS, it is necessary to partition genes into co-expressed modules. Among many developed methods, we choose RW due to its reported great and fast performance (26) and documented the practice of RW in defining co-expressed modules (**Fig 3a**). Comparing the isolated subnetworks with different methods, RW performed equally well as the most popular methods – Partitioning Around Medoids (PAM) and hierarchical clustering (HC) – without requiring the pre-setting of the number of subclusters (**Fig EV3a**).

We next showed that BioTIP is superior to DNB in parsing out which CTSs are significant within noisy background signals, in benchmark datasets about disease progression or response to injury. Each dataset (lymphoma progressions: GSE6136 and toxic lung injury: GSE2565) was previously analyzed using DNB methods. Among the three published DNB signals, two (n=27, 148, respectively) (Chen et al., 2012, Liu, Chen et al., 2015) were insignificant when compared to random gene sets at the reported tipping points (**Fig 3, b-c**, P>0.05); only one (n=259) was significant (**Fig EV3b**, p<0.001) (Chen et al., 2012).

In contrast, BioTIP's CTS identifications exhibited significant MCI scores at the reported tipping-point states, which were further validated by a significant Ic* (**Fig 3, d-e**, P<0.001).

Robustness and specification of the BioTIP identification

To examine the robustness of BioTIP given a variety of feature-preselection choices, we revisited the benchmark bulk-cell dataset (GSE6136). In this dataset, murine lymphomas were originally grouped into five states according to their clinical presentation, pathology, and flow cytometry (Lenburg, Sinha et al., 2007); the normal-tumor-transitional 'activated B cell' (P2) state was a known tipping point (Chen et al., 2012). Conducting analyses on the 1% highest variable genes confirmed this known tipping point, P2; Ic* scores of both the BioTIP-identified and randomly selected 35 genes surged at P2 and were significantly higher than random scores (**Fig 4a**, P<0.001). These results were reproducible with 5% highest variable genes, resulting in 141 CTS genes at P2. Genes in CTSs from the two parameter settings (1% or 5%) overlapped significantly (**Fig 4b**), supporting the reproducibility of the algorithm.

Using the same dataset, we tested whether BioTIP is indeed sensitive to violations of modeling assumptions that assembled samples are repeats of their nominated system states. Although early lymphoma (P3) phenotypically resembles activated B cells (P2) more than aggressive lymphoma (P5), global gene expression patterns of one P5 and one P3 sample were similar to P2 samples (**Fig 4c**). Therefore, we re-assigned this margining P3 sample into P2 to generate an 'almost true' state ensemble; then, we re-assigned the distinct P5 sample into P2 to simulate a misclassified state ensemble. As expected, BioTIP recaptured the P2 state from the almost-true but not from the misclassified ensemble (**Fig 4, d-e**). These CTSs of P2 had common genes but the 'misclassified CTS' was unique (**Fig 4f**). We conclude that BioTIP yields robust CTS identifications when applying to data with an (almost) true ensemble of states.

Single-Cell Data Applications

For single-cell level (scRNA-seq) applications, we first explore how the existing Ic metric works when applied to a dataset with a known cellular state transition (GSE52583) (Treutlein, Brownfield et al., 2014). This dataset profiles individual cells at four lung epithelial developmental points (E14.5, E16.5, E18.5, and mature AT2 lineage cells). At E14.5 the cells are classified as epithelial progenitors; across the next three timepoints, the cells differentiate into either mature alveolar type 1 (**AT1**) and or type 2 (**AT2**) cells with distinct expression patterns (Treutlein et al., 2014). Among these four timepoints, a known lineage bifurcation occurs at E18.5, when cells co-express certain "bipotent progenitor" markers of either AT1 or AT2 (Guo, Du et al., 2019a, Treutlein et al., 2014). Applying the original Ic score to this data failed to detect this critical transition at E18.5, and instead identified E16.5 as the tipping point (Mojtahedi et al., 2016) (**Fig 2b**).

In contrast, BioTIP not only detected E18.5 as the tipping point but also identified two sets of significant CTSs. First, among all 196 cells that passed quality control (**S method, Section 3**),

the E18.5 population yielded the highest Ic* score (P<0.001, **Fig 2f**), defining 205 genes as one CTS. Then, according to marker genes' expression (Treutlein et al., 2014), we split the E18.5 cells into AT1-lineage (n=38, *Ager* and *S100a6* high, and *Lyz2* low) and AT2-lineage (n=15, *Sftpc* high, and *Pdpn* low) (**Fig EV2 g-h**). Subsequent analysis of the cells along the AT2 trajectory (**Fig EV2i**) identified 267 genes at E18.5 (**Fig 4g**). This identified CTS was validated by a significantly higher Ic* score than that of random genes (P<0.001, **Fig 4h**) or that of reshuffled cell-labels (P<0.001, **Fig 4i**). Genes in both CTSs for E18.5 overlap significantly (n=188, P<2e-16, **Fig 4j**).

To characterize the 188 CTS genes that BioTIP constantly and significantly identified at E18.5, we performed an Ingenuity Pathway Analysis (Kramer, Green et al., 2014). We found 'Axonal Guidance Signaling' and 'Ephrin Receptor Signaling' (FDR<0.05, counts>8, **Fig 4k**). Two of these CTS genes are *EphrinB2* and *EphB4*, which regulate lung alveolar development and repair (Vadivel, van Haaften et al., 2012) and are required for proper patterning of the vascular morphogenesis (Adams, Wilkinson et al., 1999). Therefore, our results may reflect a critical transition in lung alveolar development that passes morphological criteria (Schittny, 2017, Treutlein et al., 2014).

However, it is notable that we were unable to validate the identified CTSs using independent data (insignificant Ic* at all 25 cellular subtypes, **Fig 4l**). There are three interpretations. 1) Inherent differences exhibit between the original dataset (assembled timepoints) and the validation dataset (without single-cell lineage-tracking data, we used the snapshotted cellular subtypes (Guo, Du et al., 2019b)). 2) Even at the same 'biological time' and after sorting (by CD45-/EpCAM+ to enrich epitheliums), the cellular heterogeneity of each biological timepoint is visualizable in gene expression: three embryonic states were poorly separated cell clusters with mostly negative silhouette-widths (**Fig 4m**). 3) Other than lineage differentiation, morphological and molecular changes underway at E18.5 could be causing entropic transcriptomic signals (**Fig 4h**, boxes) that lead to an identification of non-lineage-driven CTS from transcriptomics (Guo et al., 2019a, Treutlein et al., 2014).

Overall, BioTIP yields robust and reproducible CTS identifications when applied to single cell datasets and understanding of developmental cell trajectories. To do so, analysis must be done on cells of a discrete states and the results should be validated alongside independent datasets but of concordant cellular states.

Application to Disease Data Reveals Reproducible Transition State

We applied BioTIP analysis to questions of oncogenesis by analyzing gene expression data from children with neuroblastoma (**NB**). NB represents a complex disease system that requires a data-driven model to understand its oncogenesis, given that there is limited knowledge on its heterogeneous mechanisms (e.g., few common variants) (Maris, Hogarty et al., 2007, Ritenour, Randall et al., 2018). Analysis of transcriptional regulatory trajectories among differing NB risk states could provide new guidance on treatment strategies -- Though no

transition between risk states has been reported on a patient level, high Myc oncogene activity and low neuronal differentiation is associated with poor NB outcomes at all risk levels (Fredlund, Ringner et al., 2008). This suggests an underexplored abnormality in transcriptional regulatory networks. To study NB's transcriptional networks while accounting for tumor heterogeneity, we analyzed and compared two populations – the event-free (**EF**) and event-occurring (**E**) populations of NB (n=315 and 182 respectively, GSE49711) (Zhang, Yu et al., 2015).

From both populations, we identified a critical transition appearing at the Low-or-Intermediate-Risk (**LIR**) state (**Figs 5a, EV4, a-d, S Method, section 3**), and LIR is the only state whose CTS overlap significantly (**Fig EV4e**). Evaluation using the Ic* model with gene permutation confirmed the significance of the CTS (**Figs 2g, 5b**, P<0.001), and with patient-label reshuffling confirmed that the 'tipping point' at LIR was significant (**Figs 2h, 5c**, P<0.001, 0.002).

Both CTSs were significantly depleted with differentially expressed transcripts (**Fig 5d**), confirming critical transitions are different from its classical population average view (Chen et al., 2012, Richard, Boullu et al., 2016). Significant reproducibility of the CTS at LIR indicates a common transcriptional configuration associating with phenotypic transition in NB (**Fig 5e**, P<2e-16). Therefore, we expect the CTS identification to reveal previously unknown NB-related gene regulation.

We first confirmed the reproducibility of the LIR CTS. We cross-validated the 119 CTS genes detected from the E patients using the EF data. A significantly surged Ic* was resulted from an increased PCCg* but decreased PCCs* at LIR (**Fig 5f**). Similarly, we cross-validated the 137 CTS genes detected from EF patients using the E data. Both observations suggest a selective and shared CTS expression across all LIR tumors.

We performed two functional analyses to understand the identified CTS. First, a motif-enrichment analysis at promoters of the two CTSs (including coding genes and non-coding transcripts) found two similar motifs: PPARα and HNF4α (**Fig 5g**, adjusted P<0.05); both control lipid metabolism and NB progression (Martinez-Jimenez, Kyrmizi et al., 2010, Song, Li et al., 2020, Valentiner, Carlsson et al., 2005, Vella, Conaldi et al., 2016, Yao, Chen et al., 2017). Second, among 147 coding genes that were identified as CTS of either population, the pathway 'plasma lipoprotein assembly, remodeling, and clearance' (REACT_6823) was enriched (Fisher's exact test, P=3.5e-8, OR=19); eight CTS member genes in this pathway were *ALB*, *APOA1/2/4/5*, *APOC3*, *APOB*, and *CREB3L3*. These results acknowledge further research into the role of lipid metabolism in NB (Zirath, Frenzel et al., 2013).

Furthermore, we validated the identified CTS using independent NB profiles (n=379, GSE120572, **S Method, section 3**) (Ackermann, Cartolano et al., 2018). Note that this available validation profiles are of only coding genes. Nevertheless, we confirmed that the identified CTS genes indicated a 'tipping point' at LIR (**Fig 5h,** red line) and that they were significant (P<0.001). A regulatory role of PPARA and PPARG (but not HNF4A) was suggested by its higher expression in the LIR than other three risk states (Wilcox test P<0.005, **Fig 5i**). These

suggest that the identified CTS is a transcriptional response to phenotypic transition likely related to lipid metabolism instigated by PPAR activity.

Taken together, BioTIP enhances the identification of significant and reproducible CTSs from transcriptomes of complex disease. Regardless of sample heterogeneity, reproducible CTSs (from independent profiles) allow for the characterization of critical transitions in the phenotypic state space.

Application to ScRNA-seq Data Identifies a Haemato-endothelial (HE) Branchpoint and its Regulatory Etv2

To further demonstrate applicability, we shifted our evaluation to embryonic developmental bifurcations. We hypothesize that tipping-point analysis could capture Etv2, a master TF whose transient activation specifies haemato-endothelial (**HE**) lineages but suppresses cardiac and other mesodermal lineages (17). We chose Etv2 because it's regulation of hemangiogenesis has been observed to be threshold-dependent (Koyano-Nakagawa & Garry, 2017, Zhao & Choi, 2017) and its initial activation is autoactivated (Koyano-Nakagawa, Shi et al., 2015). This literature creates a hypothesis that Etv2-modulated HE bifurcation should be recognized in tipping-point analysis, given that thresholds and basins of attraction which is often explained by self-amplifying perturbations are key ingredients of a tipping point (Scheffer et al., 2012).

To test this hypothesis, we studied the mouse gastrulation *in-vivo* dataset (Pijuan-Sala et al., 2019), focusing on early organogenesis embryonic day (E) 8.25 when precursor cells of major organs have been formed (Kaufman & Bard, 1999). Among 19 sub-populations (GSE87038, 7,240 cells, **Fig 6a**, **S Methods**), S13 and S10 states standout, given that they both are the sub-states of HE progenitors with up-regulated *Kdr, Etv2*, and *Lmo2* (**Fig EV5a**), the hematopoietically expressed homeobox gene *Hhex*, and the endothelial marker *Slc25a5*, suggesting the multipotentiality. States S13 and S10 also have notable differences in gene expression: S13 has a higher average expression of *Kdr* than the S10 (**Figs 6b**, **EV5a**); S13 was a closely clustered group but S10 was a poorly separated cluster (**Fig EV5b**). Negative silhouette-width suggested multiple clusters intermingling at S10 in expression space, corresponding to its role of branching between hematopoietic and endothelial lineages in trajectory analysis (**Fig 6a**).

Applying BioTIP to all 19 cell states, we detected four candidate CTSs at S13 and three other states, respectively, where they exhibited significantly higher Ic* scores (**Fig EV5, c-e**). The S13 finding is particular firstly because it sits in the middle of the studied trajectory; this allows us to conduct analysis on pseudotime points before and after to assess biological significance. Secondly, S13 is the only state where Etv2 was not only a CTS member gene but also an upstream TF regulator for the CTS (Homer analysis, adjusted $P < 1e-4$).

At S13, we identified 60 CTS genes (**Figs 6c,** MCI's $p < 0.001$). Since Ic* is designed to evaluate the significance of the MCI-indicated tipping point, Ic* of the 60 CTS genes surged at S13 was expected and confirmed by Delta score ($P=0.044$, **Fig 6d**, right). Unexpectedly, Ic* was

also significantly elevated at adjacent branchpoints S3 and S10 (**Fig 6d**, left). To understand this observation, we examined the two components of Ic* calculation (PCC_s^* and PCC_g^*). At branchpoints, the Ic* was elevated by the appearance of new cellular states that decreased intercellular similarity (i.e., decreased PCC_s^*); at the hematoendothelial progenitors S10 and S13, the Ic* was elevated by the increased PCC_g^* (P<0.001, **Fig EV5f**). The latter suggests the identified 60 genes are selectively incorporated and sustained in hematoendothelial progenitors to guide the HE bifurcation. Therefore, only the higher Ic* at S13 is of interest.

The *Etv2* gene was central to the transcriptional regulatory network of the 60 CTS genes. On one hand, Etv2-binding motif is enriched among the CTS promoters (**Fig 6e**), and 60 CTS genes significantly overlap with Etv2 targets in the hematopoietic and endothelial lineages (P=3.8e-7, **Fig 6f**) (Liu, Li et al., 2015). Both denote Etv2 as an upstream regulator of the CTS. At the same time, 60 CTS genes significantly overlap with reported regulators of Etv2 during hemangiogenic cell lineage development (Zhao & Choi, 2017) (P=3.2e-11, **Fig 6f**), suggesting Etv2 is a downstream target of the 60 CTS genes. Together, we conclude that Etv2 is implicated in the CTS through transcriptional regulatory loops.

To understand these Etv2-involved regulatory networks, we checked the pairwise correlation between *Etv2* and other CTS genes. Over S13 cells, *Tal1* (also named *Scl*) was correlated with *Etv2* positively (R>0.4, P<1e-5), the other genes (eg, *Rasip1*) negatively. The observed expression dynamics (**Fig 6g**) go along with literature that Etv2 auto-transacts during the initial expression of Etv2 (Koyano-Nakagawa et al., 2015) and that the Etv2-Tal1 loop is induced upon the activation by *Etv2* (Zhao & Choi, 2017). Hence, we ration that Etv2 autoregulation drivers the identified CTS thus regulating the coming bifurcation in hematopoietic development (**S Methods**).

To validate independently, we analyzed eight states of *in-vitro* differentiated human embryoid bodies (**EB**s) (**Fig 6h,** GSE130146, 1,531 cells, **S Method**) (Zhao & Choi, 2019). The *in-vitro* S14, impending the HE bifurcation while after the branchpoint between hemangiogenic and mesenchymal (smooth muscle) lineages, analogizes that *in-vivo* S13 with respect to pseudo-orders and mark gene expression (**Fig 6i** vs **Fig 6g**, **S Methods**). As expected, Ic* of the CTS (identified from *in-vivo* S13) peaked at the *in-vitro* S14 thus validated the significance (**Fig 6j**). Additionally, this Ic* increased towards the hemangiogenic lineage selectively and significantly (p<0.001, **Fig 6j**), validating the incorporating effect of CTS in a guided trajectory passing the HE bifuraction (**Fig 6d**).

To further understand the biology underlying these 60 CTS genes, we performed an Ingenuity functional analysis. We found two significant upstream regulators (Vegf and Tnf, >6 overlaps, FDR<0.002), consistent with the consensus that VEGF signaling plays an instructive role in promoting *Etv2* threshold expression (Zhao & Choi, 2017). Literature reported a strong cooperation between the Etv2 targets *Tal1* and *Lyl1* during induced HE transition (Guibentif, Ronn et al., 2017), and both CTS members *Rasip1* and *Rhoj* are endothelium specific and Etv2-

regulated (Palikuqi, Nguyen et al., 2020, Singh, Sierra-Pagan et al., 2020). Collectively, we propose a model in which transient Etv2 autoactivation triggers oscillation in the expression of *Etv2, Tal1, Lyl1, Rasip1, Rhoj*, and other HE-decisive genes, characterizing the multipotential and unstable HE progenitors. Then, negative feedback loops take place to repress *Etv2* and maintain the system at differentiated, stable states (**Fig 6k**). This model interprets the observed expression oscillation of several marks impending the HE bifurcation.

Together, this is the first tipping-point interpretation for the transient Etv2 expression that is required for HE bifurcation. This effort moves one step towards eliminating chaotic signals and understanding the dynamic TF regulation, in which the transient induction of a regulator concurs with critical state changes. The above results provide an impetus to adopt tipping-point analysis to uncover the roles of transcriptional mosaicism with high resolution.

Discussion

In this paper, we detail BioTIP, a new workflow that uses mathematical tipping-point models and expression profiling to identify and characterize biologically critical transitions. We propose that animate gene-regulation processes are subjected to the same quantitative laws as inanimate non-equilibrium systems. According to this paradigm, biological systems exhibit discrete states and phenotypes. When the system approaches a critical state transition, the underlying gene expression signals become more chaotic, before resolving into a new state with a new expression profile. BioTIP improves upon existing methods to parse out noisy transcriptional background signals and identify key gene regulatory networks which are directing the transition.

Previous mathematical studies have contributed importantly to our method development, including the Ic score to predict tipping points (Mojtahedi et al., 2016), the DNB score to detect CTS (Chen et al., 2012), and the accurate estimate of high-scale correlation metrics (Schafer & Strimmer, 2005). In this work, BioTIP achieves accurate and reproducible identifications with three major methodological contributions:

1) BioTIP enables transcriptomic analysis on sample ensembles of macroscopic phenotypes, allowing for both time-series and cross-sectional studies with one tool. The method's focus on critical transitions is novel in current analysis of disease-progression and cell-state dynamics.

2) We bring the most accurate tipping-point anticipation after reducing the sample-size related bias in the existing tipping-point models. By correcting the estimation of large-scale correlation matrices (Schafer & Strimmer, 2005), BioTIP explicitly compares Ic* scores over states despite sample size (often variable and smaller than the gene size), thus more accurately inferring tipping points.

3) The outputs include significant CTSs. These gene features inform key factors of biological progression that were previously unexplored. With the proposed feature selection,

parameter-free module construction (random walk), and a new evaluation score (Delta), BioTIP enables the identification of CTS to be independent with the chaotic background signals.

Existing biological tipping-point studies did not produce software packages. The only available tool, *earlywarnings* (Dakos, Carpenter et al., 2012) (early-warning-signals.org), has been evaluated and accurately predicts the epithelial–hybrid-mesenchymal determination (Sarkar, Sinha et al., 2019); however it only allows for univariate data analysis. Because biological data is usually multivariate, BioTIP represents a new multivariate tipping-point analytic tool for the computational biology community.

In this study, we applied BioTIP to bulk tissue samples and were able to recapture known transition events: the normal-lymphoma transitional state, toxic lung injury, and the intermediate-risk stage in neuroblastoma progression. BioTIP was also applied to a single-cell data where it picked out a known lung epithelial lineage branching point and a hematopoietic-endothelial bifurcation. These findings demonstrate how BioTIP's ability to identify tipping points has applications across types of transcriptome data and fields of biology.

BioTIP correctly identified tipping points and CTSs despite the presence of entropy (background gene signals). After comparing random gene signals across states, Ic* was able to capture significant CTS for the correct tipping points in the one-third of our examples (**Figs 2f, 4a, and 4h**). For the other two-thirds of cases (**Figs 2g, 3e, 5b, 5h, 6d, and 6j**), the results inform that distinct mechanisms (rather than expression entropy) triggered critical transitions and incorporated with expression changes. To unlock those mechanisms besides of or coexisting with TFs, future work will be a comprehensive integration of multi-omics profiles (Stuart, Butler et al., 2019, Su, Ko et al., 2020).

Our work implicates a theoretical debate in the field of computational biology regarding significant CTS identification: whether TFs can play key roles in critical transition (Sarkar et al., 2019, Zhou, Yui et al., 2019), or whether gene signals during transition are entirely stochastic (Mojtahedi et al., 2016, Wu, Chen et al., 2018). The novel conceptualization is that the lineage specifications, at least partly, are not just a simple gradual silencing of the progenitor cells with simple upregulation of lineage marker genes (Zhou et al., 2019). For instance, Etv2's role over the transition had been previously overlooked by static analyses because its expression is induced, fluctuating, and extinguished rapidly throughout the haemato-endothelial trajectory (Koyano-Nakagawa & Garry, 2017, Koyano-Nakagawa et al., 2015). Here, BioTIP analyses allowed for Etv2 to be identified as a CTS driver and for regulatory loops it's involved in the haemato-endothelial specification to be characterized.

Broadly speaking, CTS and involved TFs have the potential for halting disease progressions by applying pharmacological intervention during the vulnerable window of state transition. For example, intermediate-risk neuroblastomas have been shown to have high cure rates and are

susceptible to chemotherapy (Baker, Schmidt et al., 2010). BioTIP identified intermediate-risk group as a critical transition, not with the goal of defining a transition in individual disease progression, but rather to identify which disease state is most vulnerable and susceptible to treatment. Among the predicted regulators of the CTS were peroxisome proliferator-activated receptors (PPARs), the therapeutic targets for high-risk neuroblastoma (Valentiner et al., 2005, Vella et al., 2016, Yao et al., 2017). A pivotal role of PPARs in energy metabolism alterations during neuronal maturation (D'Angelo, Antonosante et al., 2018) may be helpful for the future preventive interventions in NB.

The BioTIP approach is not universally practical. First, this work only detects tipping points based on certain early-warning signs such as increased variance and correlation, but there are many other warning signals associated with other kinds of tipping points. Second, transcriptional CTS cannot be significant when the noise in gene expression is larger than biological CTS (Oku & Aihara, 2018). Additionally, attractors are features of dissipative dynamical systems, but not all biological systems are dissipative.

In conclusion, we have developed the BioTIP algorithm and tool to identify multivariate CTS from gene expression profiles. We illustrate the utility and efficiency of BioTIP across multiple profiling technologies, system levels, and biological contexts. We suggest that transcriptomic tipping-point analysis for a deeper understanding of the plasticity, heterogeneity, and phenotypic changes during disease or developmental systems.

Materials and Methods

Notation

Let X denote the $p \times n$ matrix of the expression levels of p genes ($\{g_1, g_2, \dots g_p\}$) in rows and n samples ($\{s_1, s_2, \dots s_n\}$) in columns. When the $n = n_1 + \cdots + n_R$ samples are divided into R distinct states, we can group the columns of X to have $X = [X^1 | \cdots | X^R]$, where X^r denotes the $p \times n_r$ submatrix for samples in a state $r \in \{1,2, \dots, R\}$ and n_r denotes the number of samples observed in the r-th state.

Given a state r, let g_j^r be the expression vector for gene j among samples in the state (i.e., the j-th row of X^r), and let g_j^{-r} be the expression vector for gene j among samples outside the state (i.e., the j-th row of X where we remove X^r). For a group of q genes (called a '**module**') indexed by $m = \{m_1, m_2, \dots m_q\} \subset \{1,2, \dots, p\}$, let X_m^r denote the $q \times n_r$ submatrix of X^r which selects the rows corresponding to the q genes in the module.

Given a gene expression matrix of interest Z, $PCC_g(Z)$ denotes the Pearson Correlation Coefficient matrix calculated between the genes (rows of Z), and $PCC_s(Z)$ is calculated between the samples (columns of Z) (i.e., $PCC_s(Z) = PCC_g(Z^T)$).

Additionally, let $\langle\cdot\rangle$ take the average of the off-diagonal entries in a square matrix, and let $sd(\cdot)$ be the operator that calculates the standard deviation of a vector.

Five steps in three components of the BioTIP workflow

Component A: Predicting tipping point. **Theory**: The original Ic calculated for each state as the ratio of average PCC_g (correlation coefficient matrix between genes) to average PCC_s (correlation coefficient matrix between samples) (Mojtahedi et al., 2016). Throughout states during a dynamic course, Ic peaks when the system approaches a critical transition because of the following two 'early-warning' features, exampled in the following single-cell scenarios:

a) The average PCC_g increases (**Fig 1a, S method, Section 1**). This is because PCC_g calculates the ratio of the between-gene variability to the average gene variability. Before transition, both of these variabilities are dominated by the noise around the attractor of the stable state. In contrast, when approaching transition, the decaying range-restriction effect limits the average gene variability within the unstable state – a decrease in the denominator of PCC_g. Meanwhile, there is an onset of distinct states among the assembled cells. This new layer of fluctuations contributes to a higher between-gene variability – an increase in the numerator of PCC_g. Together, PCC_g increases impending the transition.

b) The average PCC_s throughout the genes decreases. This is because the cell ensemble becomes destabilized when approaching a tipping point, resulting in greater heterogeneity and thus weaker between-sample co-expression within the transition state.

Problem: The statistic ($Ic(X^r)=\frac{\langle|PCC_g(X^r)|\rangle}{\langle PCC_s(X^r)\rangle}$) estimates the true correlation present in the data X^r with the empirical correlation matrixes. Given a fixed number of genes q, the data varies in sample sizes per state; the numerator of Ic exhibits serious defects with X^r that describes few samples ($n_r \ll q$) because the empirical correlation matrix suffers from high variance, which inflates the numerator of the Ic for states with small sample sizes. Consequently, Ic has an undesirable performance on the comparison between $Ic(X^r)$: inaccuracy in its tipping-point prediction towards states r with a small n_r.

Method: To address this problem, we first introduce a 'shrinkage of correlation' method to PCC_g (see (Schafer & Strimmer, 2005), the model named 'Target D' for details). This method shrinks the most extreme coefficients in the observed PCC_g towards more central value(s), thereby systematically reducing estimation error where it matters most (Ledoit & Wolf, 2004). Given a shrinkage target matrix T_g,this method then weighs both matrices T_g and PCC_g to get an improved approximation (PCC_g^*) of the true (unknown) correlation matrix (**S Method, section 1**):

$$PCC_g^* = \lambda T_g + (1-\lambda_g)PCC_g \qquad \textbf{(Formula 1)}$$

We set T_g to be the identity matrix -- so that we shrink the partial correlation coefficients in PCC_g towards 0 due to the globally low gene covariance in stable states; and the ones on the main diagonal represent the self-correlation of every gene. The parameter $\lambda_g \in [0,1]$ is a value to be estimated from X^r and T_g, controlling how strongly the estimate PCC_g^* is shrunk towards T_g; λ_g also controls the tradeoff between bias and variance when estimating the true correlation matrix (Schafer & Strimmer, 2005). When the sample size n_r is small, the optimal value for λ_g will be larger, leading to a stronger level of shrinkage of the partial correlations towards 0, which in turn shrinks the average of the PCC_g towards 0, limiting the effect of the problem mentioned above.

Similarly, we shrink the PCC_s to get an improved approximation (PCC_s^*) of the true (unknown) between-sample correlation matrix (see (Schafer & Strimmer, 2005), the model named 'Target F' for details). Particularly, we set the target matrix T_s with 1 on the diagonal and $\langle PCC_s(X^r)\rangle$ elsewhere, so that for each state, we shrink the correlations towards the average value. The choice of average not only reflects the steady expression pattern in a stable state but also preserves the pattern difference across states. Note that a different value of λ (λ_g^r or λ_s^r) is estimated when we calculate the PCC_g* or PCC_s* respectively for a state r.

Step i) **Finding tipping point**. We define a refined Index of critical transition (Ic*) scoring system, using this updated PCC_g^* and PCC_s* estimates:

$$Ic^*(X^r) = \frac{\langle|\mathrm{PCC}_g^*(X^r)|\rangle}{\langle\mathrm{PCC}_s^*(X^r)\rangle} = \frac{\langle|\lambda_g^r T_g + (1-\lambda_g^r)\mathrm{PCC_g}(X^r)|\rangle}{\langle\lambda_s^r T_s + (1-\lambda_s^r)\mathrm{PCC_s}(X^r)\rangle} \qquad \textbf{(Formula 2)}$$

An increased $PCC_g^*(X^r)$, contributes to an increased Ic* regardless of sample size, and substantially enhances the reproducibility of the tipping-point identification from restricted sample numbers.

Simple algebra reveals a linear relationship between the original Ic and our improved Ic*: $Ic^*(X^r) = \left(1 - \lambda_g^r\right)Ic(X^r)$ (**S Method**, **section 1**). As a sample size n_r raises, the calculated value for λ_g^r decreases to 0 and therefore we recover the traditional Ic score. Consequently, this correction enhances comparisons among states with variable sample sizes.

Note: Not all tipping points are predictable by Ic (using randomly selected genes) for which we document in this Result section. Therefore, the hybrid of Ic* with the following CTS identification is not trivial.

Component B: Identifying significant CTS. **Theory**: In the tipping-point model, each attractor, e.g., fixed point (phenotype), has a restricted effect (basin of attraction), e.g., expression regulation. The gradient of the basin of the attraction determines the rate at which a system turns back to the attractor from small perturbations. Thus, one indicator of tipping-point is a sudden 'slowing down' in which the system's response to perturbations away from the current attractor decays just prior to the state transition (Scheffer, Bascompte et al., 2009). Not

only is a tipping point characterized by an increase in PCC_g of a group of genes reflecting the onset of a new state, but their cell-to-cell variability also surge corresponding to the system's loss of resilience (Richard et al., 2016). DNB has successfully quantified these characters from bulk-cell (Chen et al., 2012, Liu, Chang et al., 2019) also single-cell (Richard et al., 2016) expression profiles.

Problem: In previous DNB studies, significance, robustness, and reproducibility of the CTS identification are overlooked.

Method: Here, the goal is to identify a subset of genes that "drive" or respond to the state's change at the tipping point. For expression profiles, we introduce new gene-feature selection, gene-module definition, and a module-size-adjusted DNB score (Chen et al., 2012) such that one can infer the underlying TF from the robust and significant CTS identification. With these refinements, BioTIP applies the concept of the DNB approach, but we rename the score as 'module-criticality index' (**MCI**) to highlight its state-specificity.

Step ii) **Feature (gene) preselection**. The purpose of this step is to pre-select robustly informative transcripts and minimize the intrastate dispersions caused by sample outliers. This step is essential in bulk-cell data analysis because we want to capture the interstate dispersions rather than the intrastate dispersions (mostly caused by variable cell aggregations). This step is also non-trivial for single-cell expression analysis as it picks the most informative genes regardless of pseudo-orders. To both ends, we estimated a gene's variation in state *r* relative to other states using a relative transcript fluctuation (**RTF**) score, for each state as follows:

$$RTF_r(\mathrm{g_j}) = \frac{sd(\mathrm{g_j^r})}{\mathrm{sd}(\mathrm{g_j^{-r}})} \qquad \textbf{(Formula 3)}.$$

To ensure robustness to outliers, we used saturation (i.e., statistical-resampling) based approaches to optimize the RTF estimation. This optimization is done by 1) randomly drawing $n_r \times b\%$ samples from a state *r* to estimate the $RTF_r(\mathrm{g_j})$, and 2) repeating this drawing to select the $p \times a\%$ genes that have top RFT scores. Parameter *b* is set to 80% to preserve the data structure, thus being optional and only recommended for large systems (n_r>10 per state). By default, we set *a* up to 10% to focus on 5k or less genes with highest expressional fluctuation in a state (**Fig EV1**, step 2.2). By adjusting *b*, we allow no less than about 200 genes per state for the following analysis. We document in the Result section that BioTIP is robust to this parameter.

Step iii) **Network partition**. Gene expressions are modulated by complex regulatory feedback loops; thus, the gene expression of a system can be conceptualized as an interconnected network. These networks are equilibrium for stable states -- if even one gene's expression is perturbed, the effects could be kept in an attractor's potential well throughout its connectivity in the network. When the perturbation passes a threshold of equilibrium, critical transitions happen (Scheffer et al., 2012). Therefore, it is essential to uncover which genes are

most influential within these co-expressed networks or '**modules**.' Decomposing pre-selected genes per state into several modules will serve as the inputs of the next CTS-searching step.

For each state, we construct co-expressed gene modules based on random walks (**RW**) (Pons & Latapy, 2005) (**S Method**, **Section 2**). The choice of RW is based on a previous evaluation over 42 module-detection algorithms -- RW is not only among the top 20% showing best performance but also the fastest one (Saelens, Cannoodt et al., 2018).

Step iv) **Identifying non-random CTS**. The MCI score adjusts the DNB's bias towards small modules with a scaling factor equal to $\sqrt{module\ size}$ proposed recently (Yang et al., 2018). Comparing expressional deviation, the connectivity and homogeneity in this module relative to its complement set, and the scaling factor gives:

$$MCI(X^r, m) = \mathrm{Avg}_{\mathrm{i}\in\mathrm{m}} sd(g_i) \times \frac{\langle|\mathrm{PCC_g}(X_m^r)|\rangle}{Avg_{i\in m, j\notin m}(|PCC(g_i^r, g_j^r)|)} \times \sqrt{size(m)} \qquad \textbf{(Formula 4)}$$

To identify the candidate CTS, the modules $\widehat{m^r}$ with the higher-than-expected MCI scores in the state r, we estimate random MCI scores from X^r by bootstrapping $\sqrt{size(\widehat{m^r})}$ genes from the background.

Three notes: **1)** An MCI-derived CTS is the genes members of $\widehat{m^r}$ which are exclusively fluctuating and interactive in a tipping point r. **2**) When DNB works on the same networks preselected from the gene background, the state with highest DNB score indicates a tipping point state (Chen et al., 2012, Liu et al., 2017, Liu et al., 2019). This tipping-point prediction should agree with the prediction in Component A (if there is a prediction) because both methods identify increasing co-expression. **3**) The CTS-indicated tipping-point should be evaluable in the same data using different method, introduced in Component C.

Component C: Two-way evaluation. **Theory**: The early-warning features of a true tipping point should be captured by both Ic* and MCI scoring systems. With the same transcriptome data, we can evaluate the significance of the CTS and its indicated tipping point by the Ic* scoring system.

Method: The purpose is to define a tipping point by a significantly increased Ic* throughout discrete states. We design a Delta score to quantify abrupt changes in Ic* between states. Given a candidate CTS, the $\widehat{m^r}$ detected in state $\hat{r}$, calculating its $Ic^*(X_{\widehat{m^r}}^r)$ per state gives a vector $Ic^*|\widehat{m^r} = \left(Ic^*(X_{\widehat{m^r}}^1), \ldots, Ic^*(X_{\widehat{m^r}}^{\hat{r}}), \ldots, Ic^*(X_{\widehat{m^r}}^R)\right)$, which should peak at the state $\hat{r}$ significantly. To quantify the module's between-state changes, we propose a Delta score, the distance between the largest and the second-largest scores in $Ic^*|\widehat{m^r}$:

$$\Delta Ic^*(\widehat{m^r}) = largest(Ic^*|\widehat{m^r}) - second.largest(Ic^*|\widehat{m^r}) \qquad \textbf{(Formula 5)}$$

Step v) **Simulation study using Delta score**. Given a CTS identification, comparing the observed *Delta* score of the observed Ic* to the simulated *Delta* scores (of random Ic*s using randomly selected genes) gives empirical p-value. From the transcriptome X, the

random Delta scores are calculated by fixing the state labels but randomly drawing the expressional values of *p* genes.

For a CTS-indicated tipping point state, comparing the observed *Delta* score to state-irrelevantly simulated *Delta* scores gives another empirical p-value. These random Delta scores are calculated by fixing the CTS genes but randomly assembling all *n* samples.

Note: In complex systems, the identified signal may be a mix of CTS along the trajectory of interest with unknown effects. Therefore, we recommend a validation using independent samples of concordant states on the same trajectory.

Finally, we comprise BioTIP into a comprehensive bioinformatics toolset for identifying transcriptomic tipping points and for describing critical transitions with non-random CTSs.

Software availability

https://github.com/xyang2uchicago/BioTIP

Data and code availability

https://github.com/xyang2uchicago/BioTIP_application

Acknowledgements

This work was supported by NIH grant 5R21LM012619 (XY, ZW), the University of Chicago: Biological Sciences Collegiate Division Quantitative Biology Fellowship (DG), Micro-Metcalf Program Project (YS), and Graham School of Continuing Liberal and Professional Studies, Master of Science in Biomedical Informatics Capstone Project (Biniam Feleke for building the Vignette, Qier An and Ieva Tolkaciovaite for testing functions for the initial package). We thank Susan Cohn, Helen Salwen, and Fangming Tang for discussion on the neuroblastoma data, and Linsin Smith for discussion on figure legends. We thank UChicago Center for Research Informatics for supporting high-performance computing services.

Author contributions

XHY designed the study, managed the project, performed data analysis, contributed to the package, interpreted the results, and wrote the manuscript, with comments from JMC. ZW collected data, analyzed data, created the R package, and contributed to result interpretation. AG contributed to the Ic* algorithm, method writing, and the initial R functions. YS wrote the online method part, updated the vignette, and maintains the package. DA edited the manuscript and contributed to the initial scRNA-seq data analysis. AF tested the package and analyzed data. YW wrote the online method part.

Competing interests

The authors declare no competing interests.

Reference

Ackermann S, Cartolano M, Hero B, Welte A, Kahlert Y, Roderwieser A, Bartenhagen C, Walter E, Gecht J, Kerschke L, Volland R, Menon R, Heuckmann JM, Gartlgruber M, Hartlieb S, Henrich KO, Okonechnikov K, Altmuller J, Nurnberg P, Lefever S et al. (2018) A mechanistic classification of clinical phenotypes in neuroblastoma. Science 362: 1165-1170

Adams RH, Wilkinson GA, Weiss C, Diella F, Gale NW, Deutsch U, Risau W, Klein R (1999) Roles of ephrinB ligands and EphB receptors in cardiovascular development: demarcation of arterial/venous domains, vascular morphogenesis, and sprouting angiogenesis. Genes Dev 13: 295-306

Ando T, Kato R, Honda H (2019) Identification of an early cell fate regulator by detecting dynamics in transcriptional heterogeneity and co-regulation during astrocyte differentiation. Npj Systems Biology and Applications 5

Antolovic V, Lenn T, Miermont A, Chubb JR (2019) Transition state dynamics during a stochastic fate choice. Development 146

Baker DL, Schmidt ML, Cohn SL, Maris JM, London WB, Buxton A, Stram D, Castleberry RP, Shimada H, Sandler A, Shamberger RC, Look AT, Reynolds CP, Seeger RC, Matthay KK, Children's Oncology G (2010) Outcome after reduced chemotherapy for intermediate-risk neuroblastoma. N Engl J Med 363: 1313-23

Bockamp EO, Fordham JL, Gottgens B, Murrell AM, Sanchez MJ, Green AR (1998) Transcriptional regulation of the stem cell leukemia gene by PU.1 and Elf-1. J Biol Chem 273: 29032-42

Buganim Y, Faddah DA, Cheng AW, Itskovich E, Markoulaki S, Ganz K, Klemm SL, van Oudenaarden A, Jaenisch R (2012) Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell 150: 1209-22

Chen L, Liu R, Liu ZP, Li M, Aihara K (2012) Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. Sci Rep 2: 342

Clements CF, McCarthy MA, Blanchard JL (2019) Early warning signals of recovery in complex systems. Nat Commun 10: 1681

D'Angelo M, Antonosante A, Castelli V, Catanesi M, Moorthy N, Iannotta D, Cimini A, Benedetti E (2018) PPARs and Energy Metabolism Adaptation during Neurogenesis and Neuronal Maturation. Int J Mol Sci 19

Dakos V, Carpenter SR, Brock WA, Ellison AM, Guttal V, Ives AR, Kefi S, Livina V, Seekell DA, van Nes EH, Scheffer M (2012) Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. PLoS One 7: e41010

Fredlund E, Ringner M, Maris JM, Pahlman S (2008) High Myc pathway activity and low stage of neuronal differentiation associate with poor outcome in neuroblastoma. Proc Natl Acad Sci U S A 105: 14094-9

Garry DJ (2016) Etv2 IS A MASTER REGULATOR OF HEMATOENDOTHELIAL LINEAGES. Trans Am Clin Climatol Assoc 127: 212-223

Guibentif C, Ronn RE, Boiers C, Lang S, Saxena S, Soneji S, Enver T, Karlsson G, Woods NB (2017) Single-Cell Analysis Identifies Distinct Stages of Human Endothelial-to-Hematopoietic Transition. Cell Rep 19: 10-19

Guo M, Du Y, Gokey JJ, Ray S, Bell SM, Adam M, Sudha P, Perl AK, Deshmukh H, Potter SS, Whitsett JA, Xu Y (2019a) Single cell RNA analysis identifies cellular heterogeneity and adaptive responses of the lung at birth. Nat Commun 10: 37

Guo M, Du Y, Gokey JJ, Ray S, Bell SM, Adam M, Sudha P, Perl AK, Deshmukh H, Potter SS, Whitsett JA, Xu Y (2019b) Single cell RNA analysis identifies cellular heterogeneity and adaptive responses of the lung at birth. Nature Communications 10: 37

Hu M, Krause D, Greaves M, Sharkis S, Dexter M, Heyworth C, Enver T (1997) Multilineage gene expression precedes commitment in the hemopoietic system. Genes Dev 11: 774-85

Kaufman M, Bard J (1999) The Anatomical Basis of Mouse Development. Academic Press, San Diego

Koyano-Nakagawa N, Garry DJ (2017) Etv2 as an essential regulator of mesodermal lineage development. Cardiovasc Res 113: 1294-1306

Koyano-Nakagawa N, Shi X, Rasmussen TL, Das S, Walter CA, Garry DJ (2015) Feedback Mechanisms Regulate Ets Variant 2 (Etv2) Gene Expression and Hematoendothelial Lineages. J Biol Chem 290: 28107-19

Kramer A, Green J, Pollard J, Jr., Tugendreich S (2014) Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics 30: 523-30

Ledoit O, Wolf M (2004) Honey, I shrunk the sample covariance matrix - Problems in mean-variance optimization. J Portfolio Manage 30: 110-+

Lenburg ME, Sinha A, Faller DV, Denis GV (2007) Tumor-specific and proliferation-specific gene expression typifies murine transgenic B cell lymphomagenesis. J Biol Chem 282: 4803-11

Lenton TM, Livina VN (2016) Detecting and Anticipating Climate Tipping Points. Geophys Monogr 214: 51-62

Liu F, Li D, Yu YY, Kang I, Cha MJ, Kim JY, Park C, Watson DK, Wang T, Choi K (2015) Induction of hematopoietic and endothelial cell program orchestrated by ETS transcription factor ER71/ETV2. EMBO Rep 16: 654-69

Liu R, Chen P, Aihara K, Chen L (2015) Identifying early-warning signals of critical transitions with strong noise by dynamical network markers. Sci Rep 5: 17501

Liu X, Chang X, Liu R, Yu X, Chen L, Aihara K (2017) Quantifying critical states of complex diseases using single-sample dynamic network biomarkers. PLoS Comput Biol 13: e1005633

Liu XP, Chang X, Leng SY, Tang H, Aihara K, Chen LN (2019) Detection for disease tipping points by landscape dynamic network biomarkers. Natl Sci Rev 6: 775-785

Lummertz da Rocha E, Rowe RG, Lundin V, Malleshaiah M, Jha DK, Rambo CR, Li H, North TE, Collins JJ, Daley GQ (2018) Reconstruction of complex single-cell trajectories using CellRouter. Nat Commun 9: 892

Marco E, Karp RL, Guo G, Robson P, Hart AH, Trippa L, Yuan GC (2014) Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. Proc Natl Acad Sci U S A 111: E5643-50

Maris JM, Hogarty MD, Bagatell R, Cohn SL (2007) Neuroblastoma. Lancet 369: 2106-20

Martinez-Jimenez CP, Kyrmizi I, Cardot P, Gonzalez FJ, Talianidis I (2010) Hepatocyte nuclear factor 4alpha coordinates a transcription factor network regulating hepatic fatty acid metabolism. Mol Cell Biol 30: 565-77

Mojtahedi M, Skupin A, Zhou J, Castano IG, Leong-Quong RY, Chang H, Trachana K, Giuliani A, Huang S (2016) Cell Fate Decision as High-Dimensional Critical State Transition. PLoS Biol 14: e2000640

Oku M, Aihara K (2018) On the covariance matrix of the stationary distribution of a noisy dynamical system. Ieice Nonlinear Theo 9: 166-184

Palikuqi B, Nguyen DT, Li G, Schreiner R, Pellegata AF, Liu Y, Redmond D, Geng F, Lin Y, Gomez-Salinero JM, Yokoyama M, Zumbo P, Zhang T, Kunar B, Witherspoon M, Han T, Tedeschi AM, Scottoni F, Lipkin SM, Dow L et al. (2020) Adaptable haemodynamic endothelial cells for organogenesis and tumorigenesis. Nature 585: 426-432

Pijuan-Sala B, Griffiths JA, Guibentif C, Hiscock TW, Jawaid W, Calero-Nieto FJ, Mulas C, Ibarra-Soria X, Tyser RCV, Ho DLL, Reik W, Srinivas S, Simons BD, Nichols J, Marioni JC, Gottgens B (2019) A single-cell molecular map of mouse gastrulation and early organogenesis. Nature 566: 490-495

Pons P, Latapy M (2005) Computing communities in large networks using random walks. Computer and Information Sciences - Iscis 2005, Proceedings 3733: 284-293

Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C (2017) Reversed graph embedding resolves complex single-cell trajectories. Nat Methods 14: 979-982

Richard A, Boullu L, Herbach U, Bonnafoux A, Morin V, Vallin E, Guillemin A, Papili Gao N, Gunawan R, Cosette J, Arnaud O, Kupiec JJ, Espinasse T, Gonin-Giraud S, Gandrillon O (2016) Single-Cell-Based Analysis Highlights a Surge in Cell-to-Cell Molecular Variability Preceding Irreversible Commitment in a Differentiation Process. PLoS Biol 14: e1002585

Ritenour LE, Randall MP, Bosse KR, Diskin SJ (2018) Genetic susceptibility to neuroblastoma: current knowledge and future directions. Cell Tissue Res 372: 287-307

Rousseeuw PJ (1987) Silhouettes - a Graphical Aid to the Interpretation and Validation of Cluster-Analysis. J Comput Appl Math 20: 53-65

Saelens W, Cannoodt R, Saeys Y (2018) A comprehensive evaluation of module detection methods for gene expression data. Nat Commun 9: 1090

Sarkar S, Sinha SK, Levine H, Jolly MK, Dutta PS (2019) Anticipating critical transitions in epithelial-hybrid-mesenchymal cell-fate determination. Proc Natl Acad Sci U S A

Schafer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 4: Article32

Scheffer M, Bascompte J, Brock WA, Brovkin V, Carpenter SR, Dakos V, Held H, van Nes EH, Rietkerk M, Sugihara G (2009) Early-warning signals for critical transitions. Nature 461: 53-9

Scheffer M, Carpenter SR, Lenton TM, Bascompte J, Brock W, Dakos V, van de Koppel J, van de Leemput IA, Levin SA, van Nes EH, Pascual M, Vandermeer J (2012) Anticipating critical transitions. Science 338: 344-8

Schittny JC (2017) Development of the lung. Cell Tissue Res 367: 427-444

Sheikh AA, Groom JR (2020) Transcription tipping points for T follicular helper cell and T-helper 1 cell fate commitment. Cell Mol Immunol

Singh BN, Sierra-Pagan JE, Gong W, Das S, Theisen JWM, Skie E, Garry MG, Garry DJ (2020) ETV2 (Ets Variant Transcription Factor 2)-Rhoj Cascade Regulates Endothelial Progenitor Cell Migration During Embryogenesis. Arterioscler Thromb Vasc Biol 40: 2875-2890

Song H, Li D, Wang X, Fang E, Yang F, Hu A, Wang J, Guo Y, Liu Y, Li H, Chen Y, Huang K, Zheng L, Tong Q (2020) HNF4A-AS1/hnRNPU/CTCF axis as a therapeutic target for aerobic glycolysis and neuroblastoma progression. J Hematol Oncol 13: 24

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, Hao Y, Stoeckius M, Smibert P, Satija R (2019) Comprehensive Integration of Single-Cell Data. Cell 177: 1888-1902 e21

Su Y, Ko ME, Cheng H, Zhu R, Xue M, Wang J, Lee JW, Frankiw L, Xu A, Wong S, Robert L, Takata K, Yuan D, Lu Y, Huang S, Ribas A, Levine R, Nolan GP, Wei W, Plevritis SK et al. (2020) Multi-omic single-cell snapshots reveal multiple independent trajectories to drug tolerance in a melanoma cell line. Nat Commun 11: 2345

Teschendorff AE, Enver T (2017) Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. Nat Commun 8: 15599

Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA, Quake SR (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 509: 371-5

Vadivel A, van Haaften T, Alphonse RS, Rey-Parra GJ, Ionescu L, Haromy A, Eaton F, Michelakis E, Thebaud B (2012) Critical role of the axonal guidance cue EphrinB2 in lung growth, angiogenesis, and repair. Am J Respir Crit Care Med 185: 564-74

Valentiner U, Carlsson M, Erttmann R, Hildebrandt H, Schumacher U (2005) Ligands for the peroxisome proliferator-activated receptor-gamma have inhibitory effects on growth of human neuroblastoma cells in vitro. Toxicology 213: 157-68

Vella S, Conaldi PG, Florio T, Pagano A (2016) PPAR Gamma in Neuroblastoma: The Translational Perspectives of Hypoglycemic Drugs. PPAR Res 2016: 3038164

Wang S, Karikomi M, MacLean AL, Nie Q (2019) Cell lineage and communication network inference via optimization for single-cell transcriptomics. Nucleic Acids Res 47: e66

Wu FY, Chen XL, Zheng YY, Duan JQ, Kurths J, Li XF (2018) Levy noise induced transition and enhanced stability in a gene regulatory network. Chaos 28

Yang B, Li M, Tang W, Liu W, Zhang S, Chen L, Xia J (2018) Dynamic network biomarker indicates pulmonary metastasis at the tipping point of hepatocellular carcinoma. Nat Commun 9: 678

Yang XH, Li M, Wang B, Zhu W, Desgardin A, Onel K, de Jong J, Chen J, Chen L, Cunningham JM (2015) Systematic computation with functional gene-sets among leukemic and hematopoietic stem cells reveals a favorable prognostic signature for acute myeloid leukemia. BMC Bioinformatics 16: 97

Yao PL, Chen L, Dobrzanski TP, Zhu B, Kang BH, Muller R, Gonzalez FJ, Peters JM (2017) Peroxisome proliferator-activated receptor-beta/delta inhibits human neuroblastoma cell tumorigenesis by inducing p53- and SOX2-mediated cell differentiation. Mol Carcinog 56: 1472-1483

Zhang F, Liu X, Zhang A, Jiang Z, Chen L, Zhang X (2019) Genome-wide dynamic network analysis reveals a critical transition state of flower development in Arabidopsis. BMC Plant Biol 19: 11

Zhang W, Yu Y, Hertwig F, Thierry-Mieg J, Zhang W, Thierry-Mieg D, Wang J, Furlanello C, Devanarayan V, Cheng J, Deng Y, Hero B, Hong H, Jia M, Li L, Lin SM, Nikolsky Y, Oberthuer A, Qing T, Su Z et al. (2015) Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. Genome Biol 16: 133

Zhao H, Choi K (2017) A CRISPR screen identifies genes controlling Etv2 threshold expression in murine hemangiogenic fate commitment. Nat Commun 8: 541

Zhao H, Choi K (2019) Single cell transcriptome dynamics from pluripotency to FLK1(+) mesoderm. Development 146

Zhou W, Yui MA, Williams BA, Yun J, Wold BJ, Cai L, Rothenberg EV (2019) Single-Cell Analysis Reveals Regulatory Gene Expression Dynamics Leading to Lineage Commitment in Early T Cell Development. Cell Syst 9: 321-337 e9

Zirath H, Frenzel A, Oliynyk G, Segerstrom L, Westermark UK, Larsson K, Munksgaard Persson M, Hultenby K, Lehtio J, Einvik C, Pahlman S, Kogner P, Jakobsson PJ, Henriksson MA (2013) MYC inhibition induces metabolic changes leading to accumulation of lipid droplets in tumor cells. Proc Natl Acad Sci U S A 110: 10258-63

Figure Legends

Figure 1

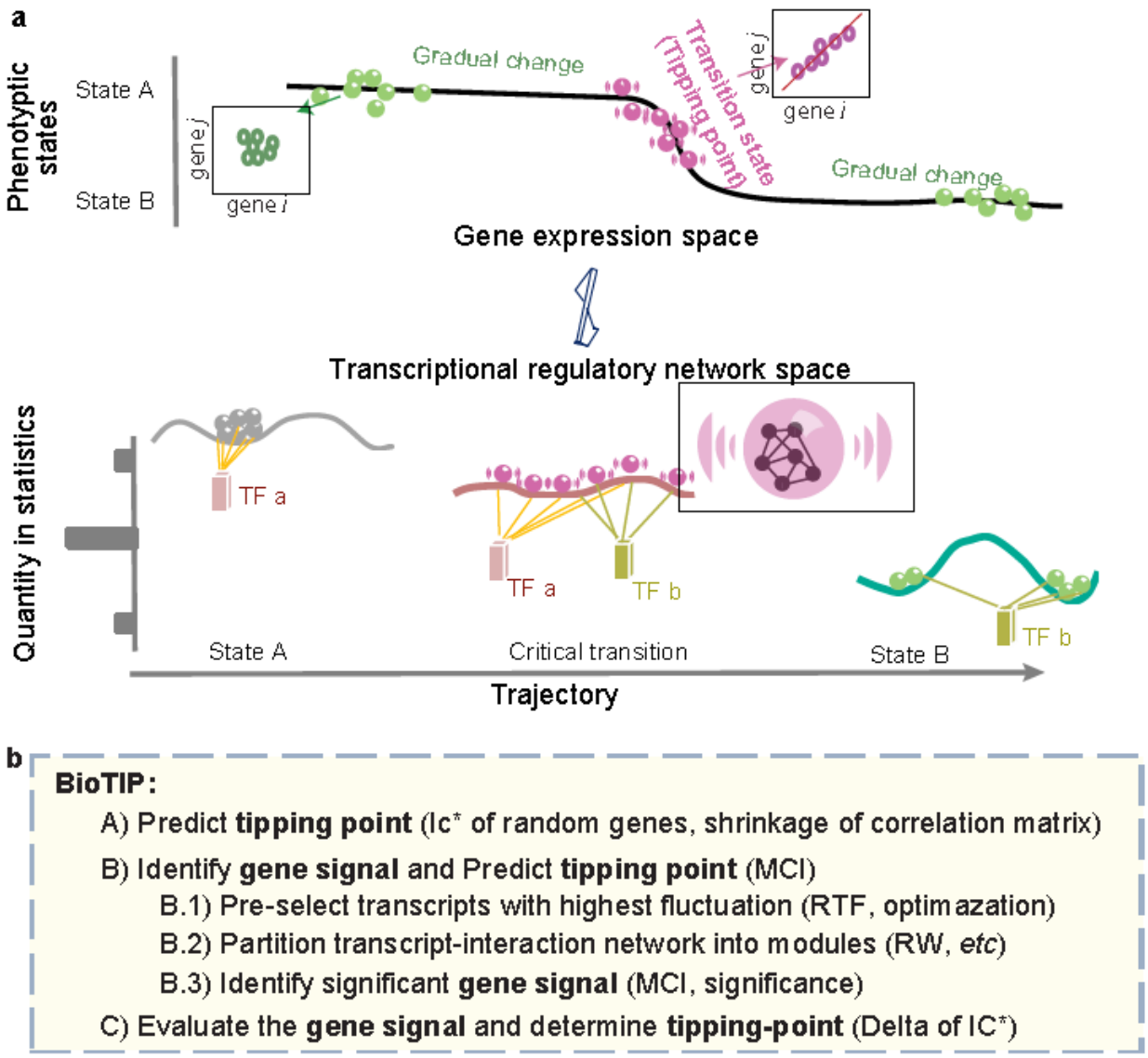

Figure 1. Tipping-point theory and adaption to gene expression analysis.

a, Schematic of the trajectory of transcriptional regulation during phenotypic state changes. Top: A view in the gene expression space: The pink points demonstrate how correlated changes in a small set of genes (embedded box) precipitate the disequilibrium system (red shaking balls), indicating a tipping point in disease states. Bottom: A view in the space of transcriptional-regulatory networks, in which master transcription factors (TFs) maintain a stable cell state A: During the dynamic course, the flat red line shows the effect of multiple decayed attractions at the tipping point, representing a vulnerable regulation. Then, this unstable state is resolved into distinct cellular states and phenotypes, shown by the green line. The y-axis quantifies the tipping point with a scoring system.

b, BioTIP's three-step analytic pipeline for identifying tipping points and corresponding non-random gene signals. RTF: relative transcript fluctuation; RW: random walk; MCI: module-criticality index, adapted from the dynamic network biomarker (DNB) method; Ic: the index of critical state transition. See also **Figure S1**.

Figure 2

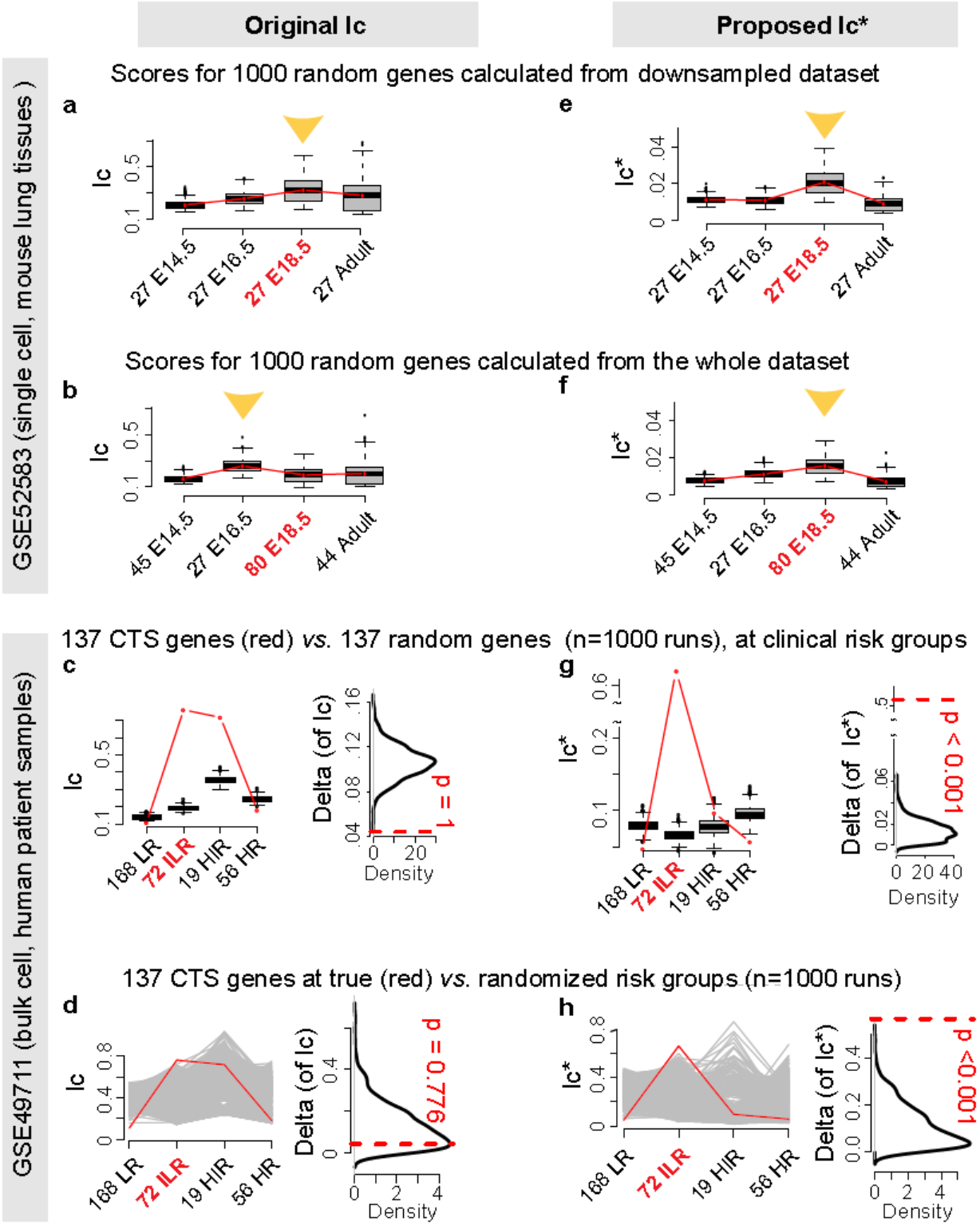
Original Ic
Proposed Ic*
GSE52583 (single cell, mouse lung tissues)
Scores for 1000 random genes calculated from downsampled dataset
a
e
Ic
Ic*
27 E14.5
27 E16.5
27 E18.5
27 Adult
Scores for 1000 random genes calculated from the whole dataset
b
f
45 E14.5
27 E16.5
80 E18.5
44 Adult
GSE49711 (bulk cell, human patient samples)
137 CTS genes (red) vs. 137 random genes (n=1000 runs), at clinical risk groups
c
g
168 LR
72 ILR
19 HIR
56 HR
Delta (of Ic)
Delta (of Ic*)
Density
p = 1
p < 0.001
137 CTS genes at true (red) vs. randomized risk groups (n=1000 runs)
d
h
p = 0.776
p <0.001

Figure 2. Ic* scoring yields more consistent results than the original Ic to compare states with variable sample sizes.

a, The Ic is the ratio between average between-gene correlation (PCCg) and average between-sample correlation (PCCs). The yellow triangle points to the prediction (E18.5). From each time-point, cells were downsampled to 27 and 1000 random genes were sampled to calculate Ic. Red line links the median values.

b, When applied to the complete data that sample sizes vary between states, Ic identified a different point at E16.5.

c, Applying Ic to bulk RNA-seq data (GSE49711, event-free risk stages). Left: Evaluation of the CTS by comparing its Ic (red line) to that of random transcripts (black boxes). Right: The Delta score of the observed Ic (red dashed line), compared to that of random genes (black density) is empirically insignificant.

d, Evaluation of the tipping point. Left: The observed Ic (red line) compared to random Ics generated by shuffling the risk assignments (grey lines). Right: The Delta score of the observed Ic (red dashed line), compared to the Delta scores of random Ics (black density), is insignificant.

e, Similar to panel a, Ic* conducted on equal-size states predicted the tipping-point at E18.5. Note the average value of an estimated PCCs* equals the average value of the observed PCCs when shrinkage PCCs towards to average (**S Methods, Section 1**).

f, In contrast to panel b, Ic* conducted on all cells recaptured E18.5 as the tipping point.

g, Similar to panel c but using Ic*. Here, the random Ic* scores (black boxes) for the smallest population (HIR, n=19) are similar to that for other populations. This stands in contrast to the misleadingly increased random Ic scores of HIR in panel e. Consequently, the identified CTS passes the significance evaluation with Delta scores.

h, Similar to panel d but using Ic*. Consequently, the CTS-indicated tipping point (ILR) passes the significance evaluation.

Figure 3

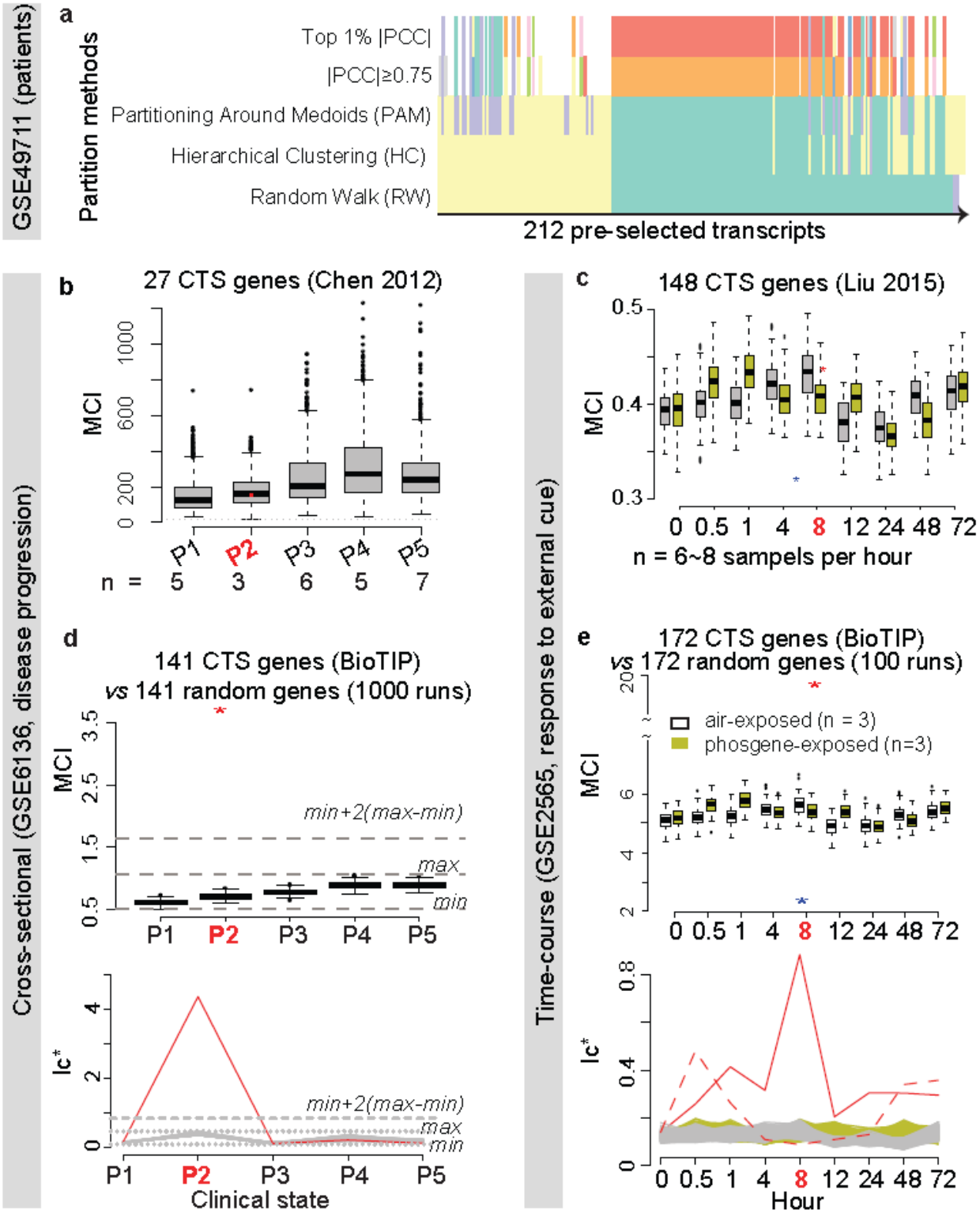
GSE49711 (patients)
a
Partition methods
Top 1% |PCC|
|PCC|≥0.75
Partitioning Around Medoids (PAM)
Hierarchical Clustering (HC)
Random Walk (RW)
212 pre-selected transcripts
Cross-sectional (GSE6136, disease progression)
b
27 CTS genes (Chen 2012)
MCI
P1 P2 P3 P4 P5
n = 5 3 6 5 7
d
141 CTS genes (BioTIP)
vs 141 random genes (1000 runs)
min+2(max-min)
max
min
Ic*
Clinical state
Time-course (GSE2565, response to external cue)
c
148 CTS genes (Liu 2015)
n = 6~8 sampels per hour
e
172 CTS genes (BioTIP)
vs 172 random genes (100 runs)
air-exposed (n = 3)
phosgene-exposed (n=3)
Hour

Figure 3. Gene-module construction and the significance of CTS identification (in contrast to the DNB identifications in Fig EV3)

a, Heatmap comparing five partition methods. Pre-selected transcripts are colored by module identities in NB transcriptomes (from dataset GSE49711, containing data on event-free, low-or-intermediate-risk patients).

b, Comparing the MCI score of 27 DNB identifications (red star) to that of 27 randomly-selected genes (grey boxes) in a mouse lymphoma model (GSE6136). A 'tipping point' at P2 state was reported with the DNB scoring system, but its MCI score falls among that of random genes (1000 runs).

c, MCI score of 148 DNB-identification (red star) in a time-course model (GSE2565). The expression of two groups of CD-1 male mice was profiled after exposure to air (as control) and phosgene, respectively. A tipping-point identification at 8 hours was reported. However, the MCI score of this published signature lacks significance (P>0.05). The observed score (red star) is compared to its size-controlled random transcripts (yellow box) in phosgene-exposed; or in control mice (blue star vs. white box) per state.

d, The BioTIP-identification in contrast to the panel **b** (GSE6136). Top: The observed MCI score (red star) is compared to that of random genes (grey boxes). The top horizontal dashed line indicates a proposed significance threshold. Bottom: Ic* (red line) is compared to that of random genes (grey lines, n=1000). Both scores peak at the tipping point (P2).

e, The BioTIP-identification in contrast to the panel **c** (GSE2565). Top: The MCI score (red star) is significantly higher than that of random genes with phosgene-exposure (yellow boxes, n=1000). Bottom: The Ic* of random genes are flat in either phosgene-exposed (yellow lines) or air-exposed (grey lines, n=1000) mice; the observed Ic* peaked at the tipping point in phosgene-exposed (solid red line) but not air-exposed mice (dashed red line).

Figure 4

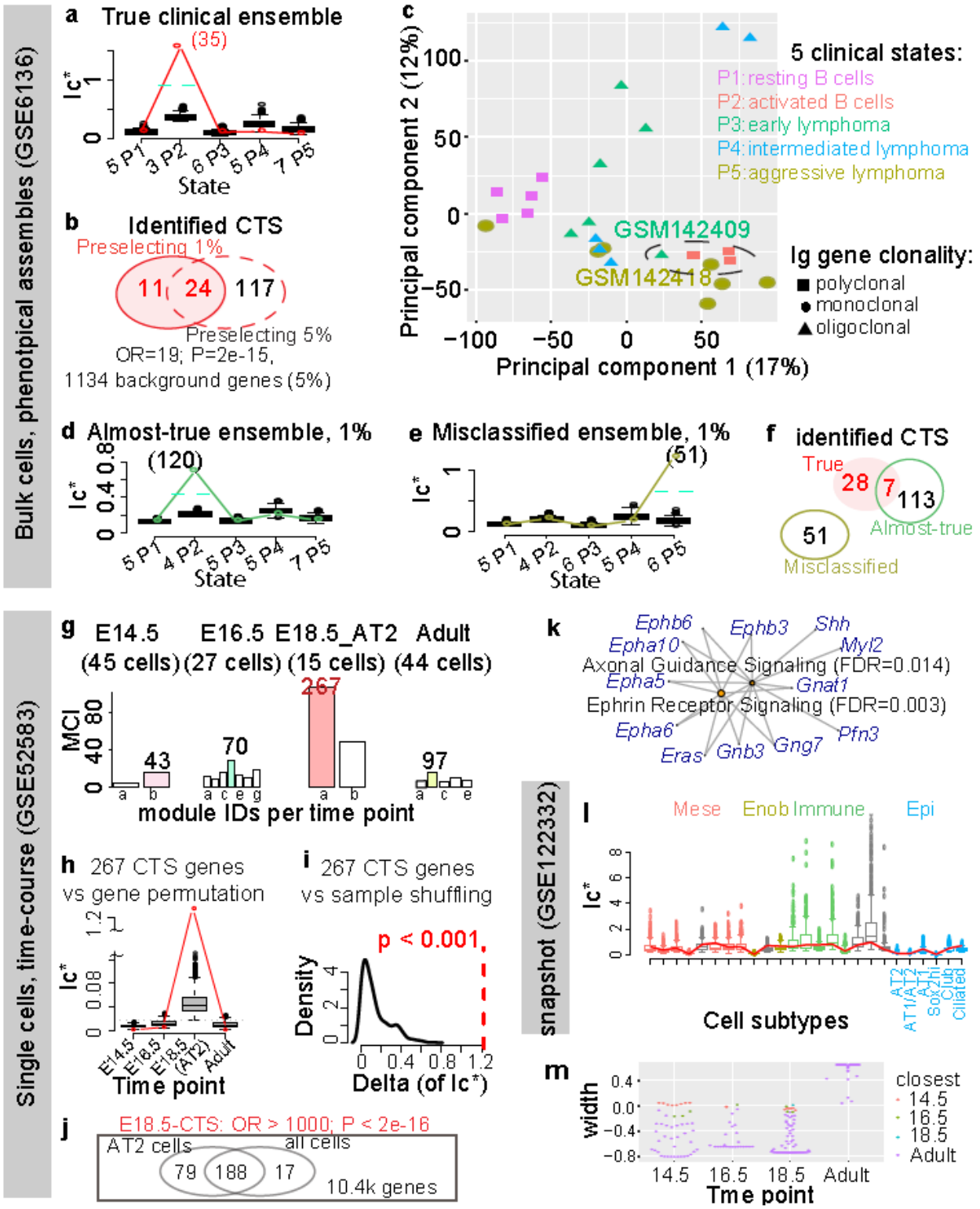
Bulk cells, phenotpical assembles (GSE6136)
a True clinical ensemble
(35)
Ic*
State
5 P1
3 P2
6 P3
5 P4
7 P5
b Identified CTS
Preselecting 1%
11
24
117
Preselecting 5%
OR=19; P=2e-15,
1134 background genes (5%)
c
Principal component 2 (12%)
Principal component 1 (17%)
5 clinical states:
P1:resting B cells
P2:activated B cells
P3:early lymphoma
P4:intermediated lymphoma
P5:aggressive lymphoma
GSM142409
GSM142418
Ig gene clonality:
polyclonal
monoclonal
oligoclonal
d Almost-true ensemble, 1%
(120)
4 P2
5 P3
e Misclassified ensemble, 1%
(51)
6 P3
6 P5
f identified CTS
True
28
7
113
Almost-true
51
Misclassified
Single cells, time-course (GSE52583)
g E14.5 E16.5 E18.5_AT2 Adult
(45 cells) (27 cells) (15 cells) (44 cells)
MCI
43
70
267
97
module IDs per time point
h 267 CTS genes
vs gene permutation
Time point
E14.5
E16.5
E18.5 (AT2)
Adult
i 267 CTS genes
vs sample shuffling
p < 0.001
Density
Delta (of Ic*)
j E18.5-CTS: OR > 1000; P < 2e-16
AT2 cells
all cells
79
188
17
10.4k genes
k
Ephb6
Ephb3
Shh
Epha10
Myl2
Axonal Guidance Signaling (FDR=0.014)
Epha5
Gnat1
Ephrin Receptor Signaling (FDR=0.003)
Epha6
Pfn3
Eras
Gnb3
Gng7
snapshot (GSE122332)
l
Mese
Enob
Immune
Epi
Cell subtypes
AT2
AT1/AT2
AT1
Sox2hi
Club
Ciliated
m
width
closest
14.5
16.5
18.5
Adult
Tme point

Figure 4. Robustness of the CTS identifications.

a, BioTIP identifying P2, with 35 CTS genes, as the tipping point. X-axis labels the states with sample size. The Ic* of observation (red line) is compared to that of 35 random genes (boxes, n=1000). The horizontal dashed line marks the significance cutoff, twice of the maximum from the minimal value of random Ic*s.

b, Venn-diagram comparing CTSs with alternative gene-preselection settings.

c, Principal component visualizing of samples, based on 18k expressed genes. Color denotes five states. A dashed-circle surrounds three P2, GSM142409 of P3, and GSM142418 of P5 that have similar expression patterns.

d-e, Results to be compared with panel a, given 1% gene-preselection but almost-true (d) or misclassified (e) sample ensembles.

f, Venn-diagram comparing three above-identified CTSs.

g, Bar plots illustrating the MCI scores along the AT2-lineage trajectory. Each state has unique gene modules. Dominating modules (the highest MCI per state) are colored with module size on top.

h, Ic* of the identified CTS genes (red line) is significantly higher than that of 267 random genes (boxes, 1000 runs) at E18.5. Random Ic*s are higher at E18.5 than at other states.

i, Evaluation of the tipping-point identification. The Delta score of the observed Ic* (red line) is significantly higher than that of random cell assemblies (density, 1000 runs).

j, Venn-diagram showing the gene overlap between two E18.5-CTS identifications.

k, Network view of the two enriched pathways and their CTS member genes.

l, Red line comparing the Ic* across lung-cell subtypes for the 180 shared E18.5-CTS genes (8 were not expressed the validation dataset). In contrast, boxplots showing the simulated Ic* (1000 runs) for 180 random genes. Box color decodes four major cell types and contaminants in grey.

m, Distribution of the approximate silhouette width (y-axis) across cellular assemblies (x-axis). Each dot represents a cell and colored with the identity of its own cluster if width is positive, or that of the closest other cluster if width is negative.

In b, f, and j, the numbers in the Venn count the CTS genes. OR: Odds ratio. Fisher's Exact test.

Figure 5

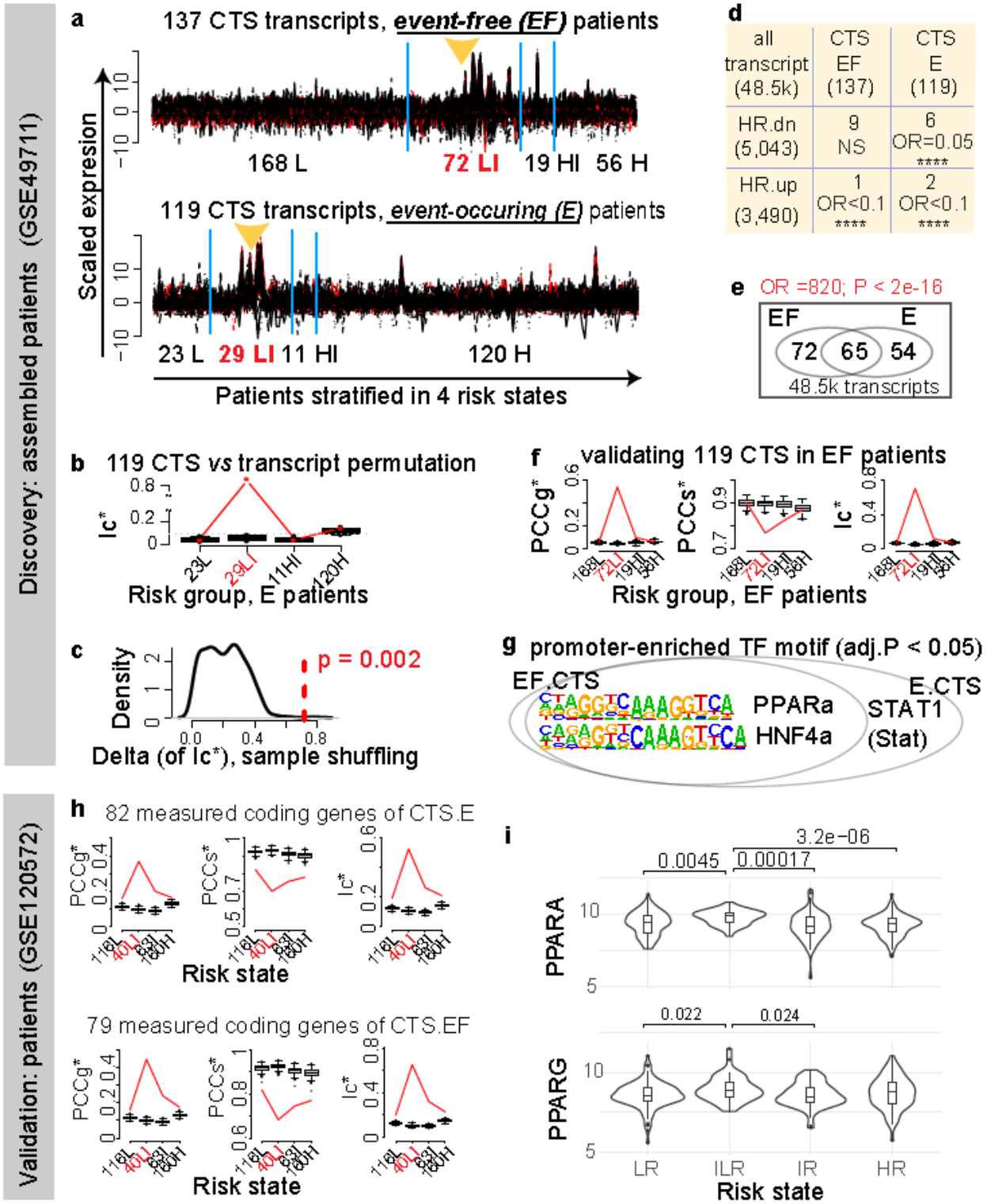
Discovery: assembled patients (GSE49711)
a
137 CTS transcripts, event-free (EF) patients
Scaled expression
168 L
72 LI
19 HI
56 H
119 CTS transcripts, event-occuring (E) patients
23 L
29 LI
11 HI
120 H
Patients stratified in 4 risk states
d
all transcript (48.5k)
CTS EF (137)
CTS E (119)
HR.dn (5,043)
9 NS
6 OR=0.05 ****
HR.up (3,490)
1 OR<0.1 ****
2 OR<0.1 ****
e
OR =820; P < 2e-16
EF
E
72
65
54
48.5k transcripts
b
119 CTS vs transcript permutation
Ic*
Risk group, E patients
f
validating 119 CTS in EF patients
PCCg*
PCCs*
Ic*
Risk group, EF patients
c
Density
p = 0.002
Delta (of Ic*), sample shuffling
g
promoter-enriched TF motif (adj.P < 0.05)
EF.CTS
E.CTS
PPARa
HNF4a
STAT1
(Stat)
Validation: patients (GSE120572)
h
82 measured coding genes of CTS.E
PCCg*
PCCs*
Ic*
Risk state
79 measured coding genes of CTS.EF
Risk state
i
3.2e−06
0.0045
0.00017
PPARA
0.022
0.024
PPARG
LR
ILR
IR
HR
Risk state

Figure 5. Applying BioTIP to bulk-cell expression disclosed a neuroblastoma risk-group transition and a relevant transcription factor PPAR family

a, Scaled CTS expression showing increased fluctuation at the identified tipping-point (yellow triangle). NB patients have been split into event-free (EF) and event-occurring (E) populations, and are separated (blue vertical lines) into four states with sample sizes listed below. L: low-risk; LI: Low-or-Intermediate-risk. HI: High-or-Intermediate-risk; H: High-risk.

b, Comparing the Ic* of the 119 CTS transcripts identified from the E population (red line) to that of 119 random transcripts (boxes).

c, Evaluation of the 119 CTS transcripts in the E patients. The observed Delta score (red dashed line) is compared to the Delta scores after sample-reshuffling (density).

d, Table showing the depletion (OR<0.5) between CTS and differentially expressed transcripts (FC>2, FDR<0.05, comparing high-risk to non-HR patients) in the same dataset. HR.dn (or HR.up): down-regulated (or up-regulated) in high-risk (HR) than non-HR groups, limma test with significance criteria: mean value>1, fold-change >2, and FDR<0.05.

e, Venn-diagram comparing the transcript overlap between the CTSs identified from EF and E patients, counting both genes and non-coding RNAs. OR: Odds Ratio. Fisher's exact test.

f, Cross validation of the 119 CTS transcripts in the independent EF population. PCCg*: the average gene-gene correlation after correlation-matrix shrinkage; PCCs*: the average sample-sample correlation.

g, Venn-diagram showing 2 commonly enriched TF-binding motifs at the two sets of CTS promoters ([-500, 200] around TSS). Significance criteria is Benjamini-adjusted P < 0.05, Homer analysis.

h, Evaluation of the coding component of 119 CTS (top) or the 137 CTS (bottom) in an independent population using a different array platform (Agilent microarray, GSE120572). Showing are scores of the CTS genes whose symbols are matchable in the new dataset, respectively. In both cases, the observed Ic* (red line) peaks at the LI risk stage significantly.

i, Violin-box plot showing the PPARA (top) and PPARG (bottom) expression levels on log2 scale, per state. Sample sizes are the same as panel h. Showing are the significant P-values, using the Wilcoxon rank sum statistic for the comparisons between any two states.

Figure 6

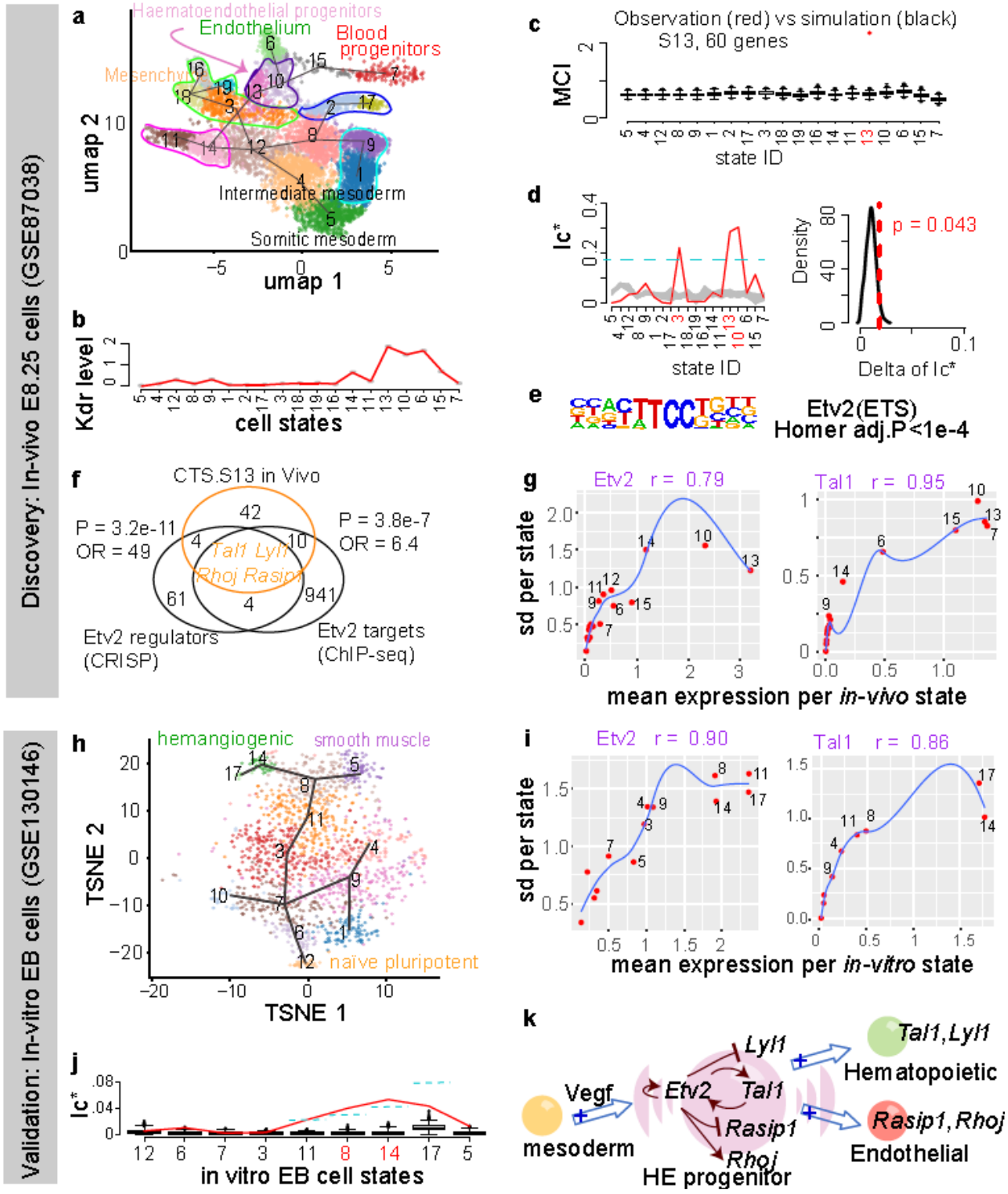
Discovery: In-vivo E8.25 cells (GSE87038)
a
Haematoendothelial progenitors
Endothelium
Blood progenitors
Mesenchyme
Intermediate mesoderm
Somitic mesoderm
umap 2
umap 1
b
Kdr level
cell states
c
Observation (red) vs simulation (black)
S13, 60 genes
MCI
state ID
d
Ic*
state ID
Density
p = 0.043
Delta of Ic*
e
Etv2(ETS)
Homer adj.P<1e-4
f
CTS.S13 in Vivo
P = 3.2e-11
OR = 49
P = 3.8e-7
OR = 6.4
Tal1 Lyl1
Rhoj Rasip1
Etv2 regulators (CRISP)
Etv2 targets (ChIP-seq)
g
Etv2 r = 0.79
Tal1 r = 0.95
sd per state
mean expression per in-vivo state
Validation: In-vitro EB cells (GSE130146)
h
hemangiogenic
smooth muscle
naïve pluripotent
TSNE 2
TSNE 1
i
Etv2 r = 0.90
Tal1 r = 0.86
sd per state
mean expression per in-vitro state
j
Ic*
in vitro EB cell states
k
Vegf
mesoderm
Etv2
Lyl1
Tal1
Rasip1
Rhoj
HE progenitor
Tal1,Lyl1
Hematopoietic
Rasip1,Rhoj
Endothelial

Figure 6. Applying BioTIP to single-cell expression revealed Etv2-regulated haemato-endothelial branchpoint

a, UMAP showing 7,240 E8.25 cells, colored and numbered into 19 cell states. Colored circles identify published cell types. Black lines illustrate the Monocle2-constructed pseudo-orders.

b, Lines showing the average expression levels of *Kdr* in each of the 19 cell states.

c, Bar plots illustrate the MCI scores in 19 states (S). The observed score for a CTS of 60 genes identified from S13 (red dot) is compared to that of 60 random genes (grey boxes, 1000 runs).

d, Evaluation of the identified CTS. Left: Comparing the Ic* of the 60 CTS genes (red line) to that of 1000 random scores (light grey, of 60 random genes and random cell labels). Blue dashed line indicates significance threshold. Right: The Delta of the observed Ic* (red line) is empirically significant.

e, Enriched TF-binding motifs at the identified CTS promoters ([-500, 200]nt around TSS). Significance criteria is Benjamini-adjusted $P < 0.05$, Homer test.

f, Venn-diagram showing the significant overlap between the 60 CTS genes and the published Etv2 upstream regulators or Etv2 downstream target genes.

g, Scale-dot plot comparing the gene expression in average (x-axis) vs standard deviation (sd, y-axis) per state. Each state is a red point, and blue line is the loess-regression fit. Pearson correlation coefficient is shown. Results of Etv2 (left) and Tal1 (right) are shown, respectively.

h, TSNE showing 1,531 in-vitro cells, colored into 17 states. Black lines illustrate the constructed pseudo-orders.

i, Same as panel g, but with the *in-vitro* dataset.

j, Ic* calculated for the 60 CTS genes (red line) in the in-vitro dataset, compared to the Ic* of 60 random genes (black boxes, 1000 runs). Blue dashed lines at four states (with relatively high Ic*) indicate state-specific significance thresholds.

k, A model of Etv2 autoregulation in the dynamic system. The yellow ball indicates the initial state. Increasing fluctuation in CTS-gene expressions (pink) indicate a developmental tipping point. After entering later stable states, some oscillated CTS genes gain enriched expression.

In panels c, d, and j, the red font in the x-label indicates significant states.

Expanded View Figure legends

Figure EV1

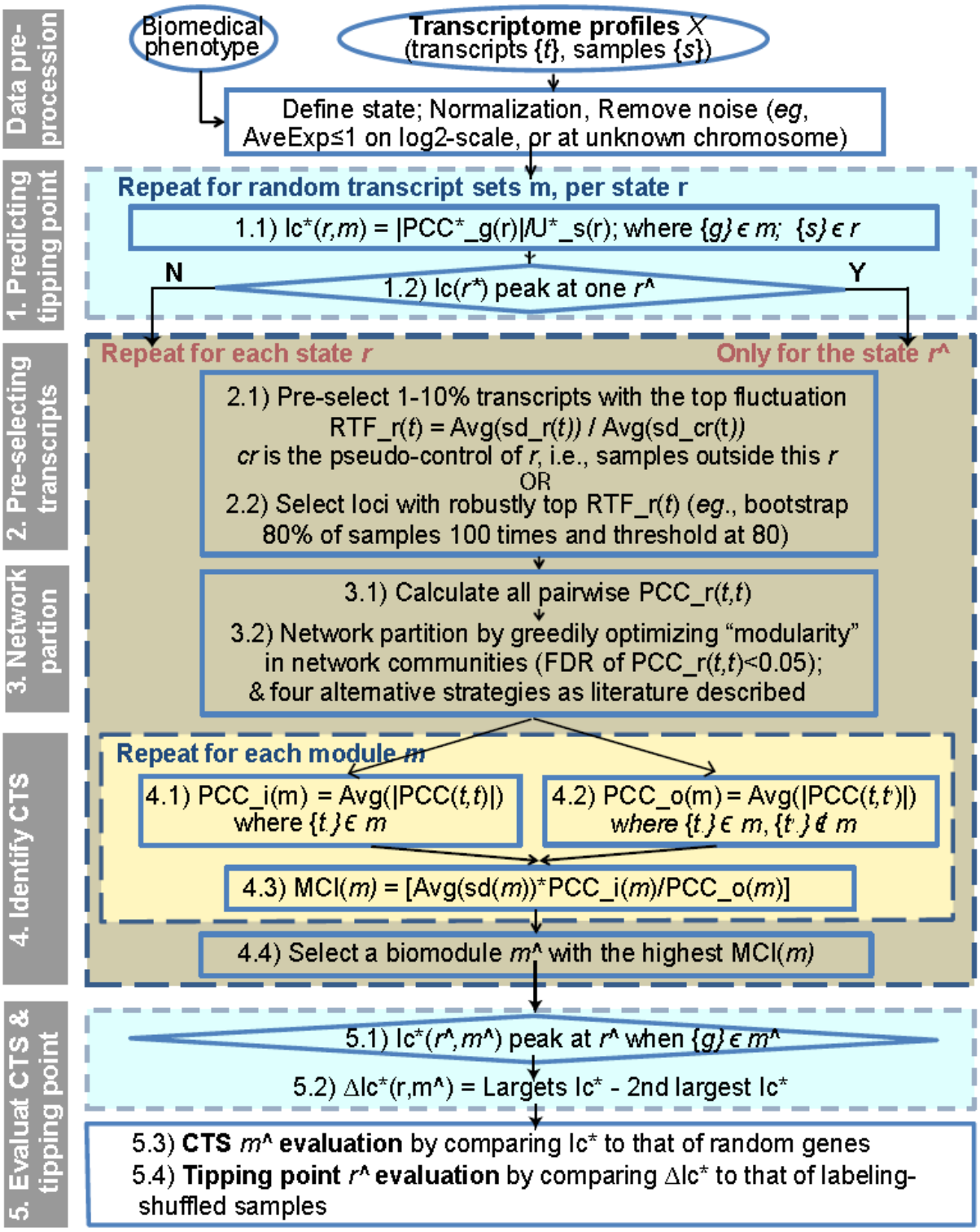

Figure EV1. BioTIP's five-step analytic workflow

Figure EV2

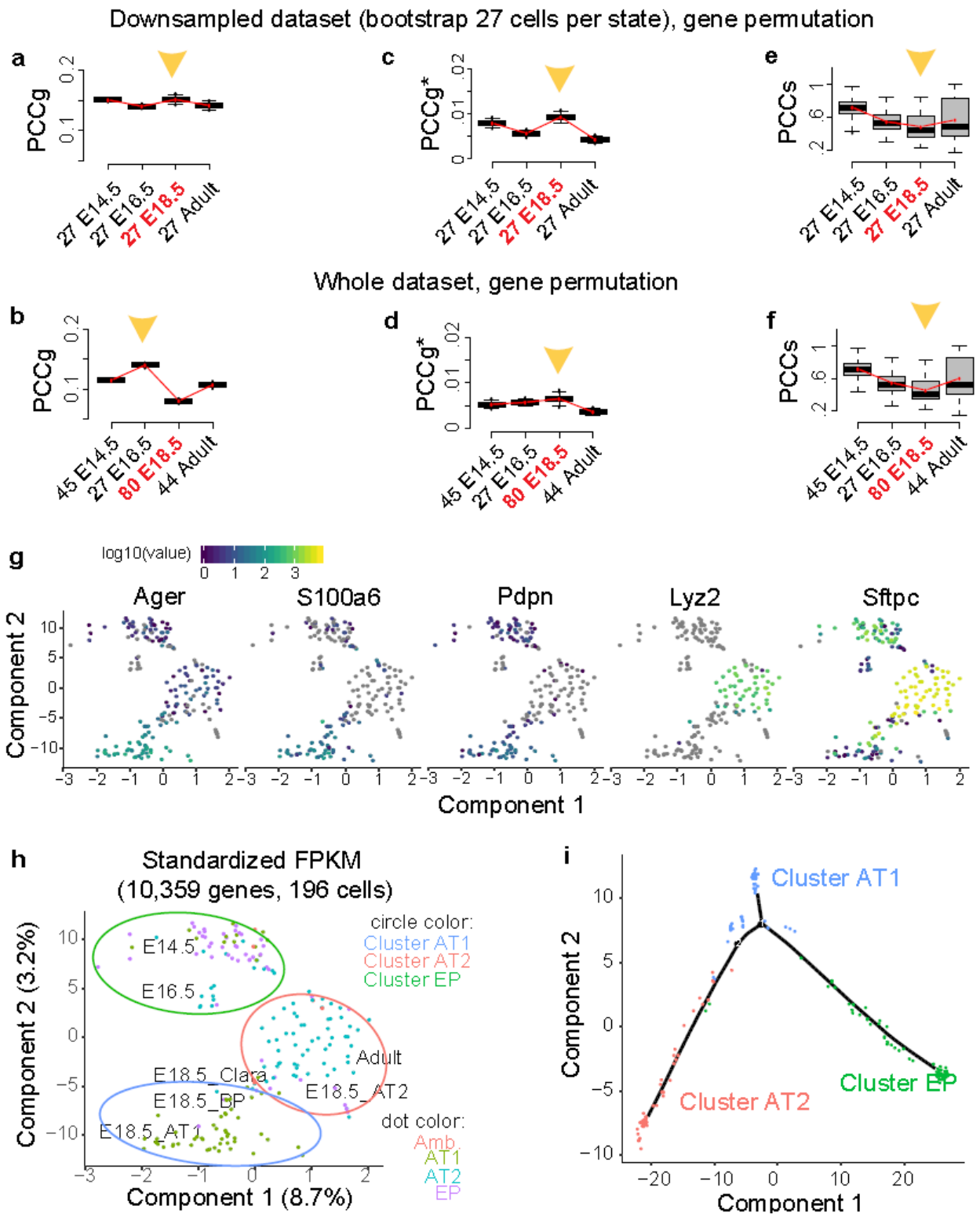
Downsampled dataset (bootstrap 27 cells per state), gene permutation
a
PCCg
27 E14.5
27 E16.5
27 E18.5
27 Adult
c
PCCg*
e
PCCs
Whole dataset, gene permutation
b
45 E14.5
27 E16.5
80 E18.5
44 Adult
d
f
g
log10(value)
Ager
S100a6
Pdpn
Lyz2
Sftpc
Component 2
Component 1
h
Standardized FPKM
(10,359 genes, 196 cells)
circle color:
Cluster AT1
Cluster AT2
Cluster EP
E14.5
E16.5
Adult
E18.5_Clara
E18.5_BP
E18.5_AT2
E18.5_AT1
dot color:
Amb.
AT1
AT2
EP
Component 2 (3.2%)
Component 1 (8.7%)
i
Cluster AT1
Cluster AT2
Cluster EP
Component 2
Component 1

Figure EV2. The Ic* scoring system corrects the bias of the original Ic, demonstrated by the analysis of single-cell transcriptomes of lung epithelia development (GSE52583).

a, Boxplot showing the average gene-gene correlation (PCCg), at four time points. Yellow triangle points to the state (E18.5) with summit correlations of random genes. The same number of cells (n=27) from each time point were extracted and 1000 random genes were sampled to calculate the random scores.

b, A different maximum (E16.5) of PCCg was identified from the same scRNA-seq data, when all detected cells were used to calculate the classic Ic, as opposed to the subset of 27 cells used in panel a.

c, Same as shown in panel a for an analysis of subsampled cells, but the Ic* scoring system is applied.

d, Same as shown in panel c for an analysis of all cells, but the Ic* scoring system is applied. Here, consistent highest at E18.5 were observed.

e, Boxplot showing the cell-cell correlation (PCCs), at four time points. Yellow triangle points to the state (E18.5) with the lowest correlations. The same number of cells (n=27) from each time-point were extracted and 1000 random genes were sampled to calculate the random scores.

f, Same as shown in panel e for an analysis of all cells. Here, correlations were lowest at E18.5 was observed.

In panels a-f, the red line links the median values of random scores.

g, Dot plots present the first two principal components of 196 single-cells reported in the dataset, grouped by expression of each marker gene. Each dot is a single cell, colored by the relative expression level of each marker gene. Top: Three markers for alveolar type 1 (AT1) (Ager, S100a6, Pdpn) and Bottom: two markers for AT2 (Lyz2, Sftpc). Principle component analysis (PCA) was performed on 10,359 genes expressed in at least 10 cells and selected by trajectory analysis using Monocle.

h, PCA plot of the same 196 single cells classified and colored according to the relative expression levels of the above five mark genes (Monocle analysis: AT1-lineage - *Ager* and *S100a6* higher but *Lyz2* low; AT2-lineage - *Sftpc* high but *Pdpn* low; amb: ambiguous). Circles identify three cellular clusters determined by the relative expression of global gene expression. These published ages and lineages of the cells are annotated. Circle color distinguishes the global gene expression-based cell clusters, and the dot color denotes the biomarker-based cell types.

i, Dot-line plot presenting the pseudotime of 196 single cells, branching into two lung epithelial developmental lineages: AT1 (top middle) and AT2 (bottom left). The pseudotime trajectory was built by the R package Monocle (v2.17.0) (Qiu, Mao et al., 2017). Two branching points are marked with filled black circles.

Figure EV3

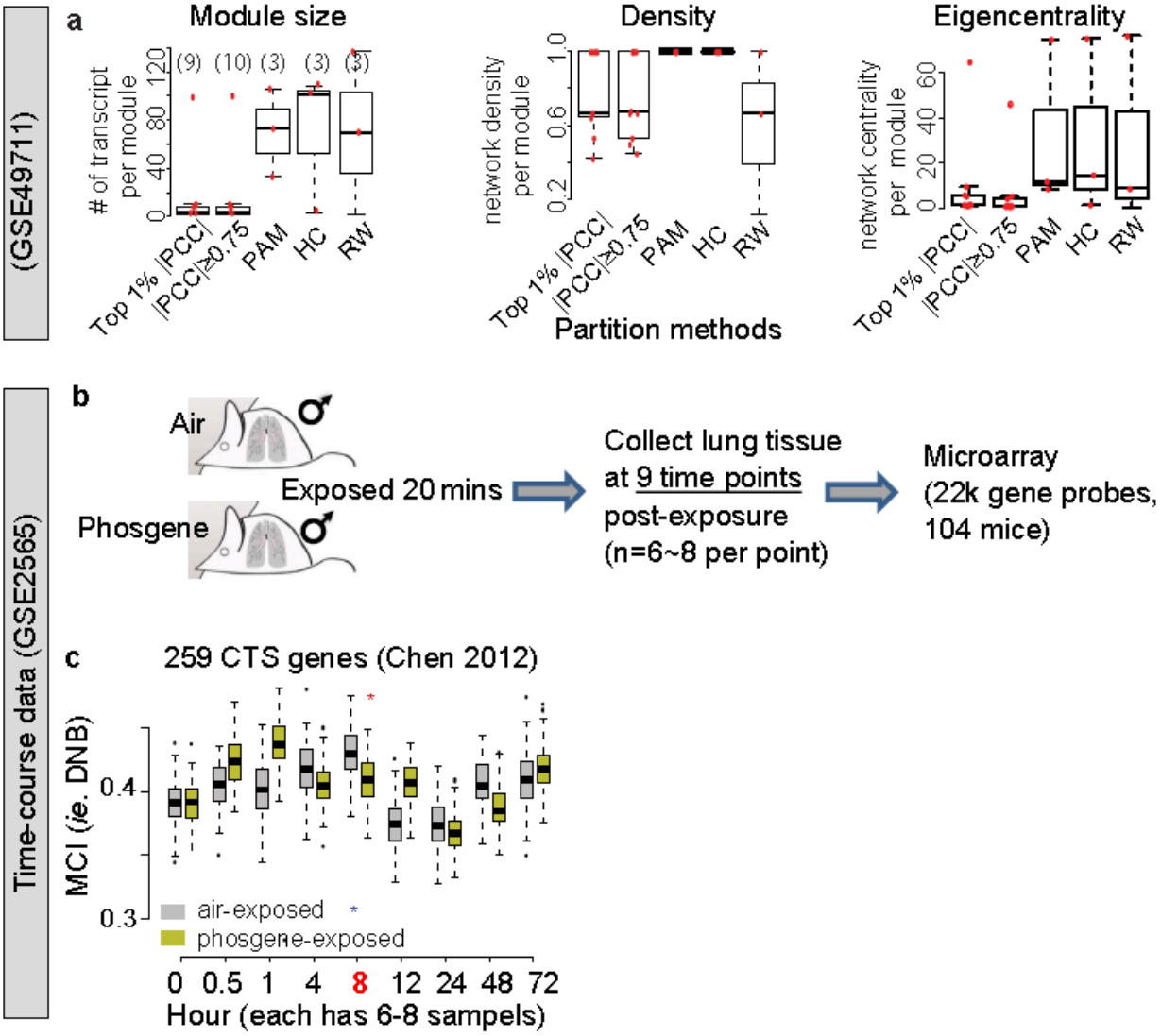
(GSE49711)
a
Module size
# of transcript per module
(9) (10) (3) (3) (3)
Top 1% |PCC|
|PCC|≥0.75
PAM
HC
RW
Density
network density per module
Partition methods
Eigencentrality
network centrality per module
Time-course data (GSE2565)
b
Air
Phosgene
Exposed 20 mins
Collect lung tissue at 9 time points post-exposure (n=6~8 per point)
Microarray (22k gene probes, 104 mice)
c
259 CTS genes (Chen 2012)
MCI (ie. DNB)
air-exposed
phosgene-exposed
0 0.5 1 4 8 12 24 48 72
Hour (each has 6-8 sampels)

Figure EV3. Comparison of network-partition methods and an example of insignificant CTSs identified by DNB.

a, Boxplots comparing network features generated by five network-partition methods, with a number of isolated modules in parentheses. The edge density (between two co-expressed genes) is the ratio of present edges to all possible edges per module. Of note, PAM and HC yielded the highest density of 1 because they considered all possible between-gene correlations without filtering out insignificant edges.

b, Top: Carton plot illustrating a benchmark dataset with a time-course disease model (GSE2565). This dataset profiles the gene expression of two groups of CD-1 male mice after exposure to air (as control) and phosgene, respectively. Bottom: MCI score of the DNB-identified 259 CTS genes (y-axis) in this model. A tipping-point identification at 8 hours was previously reported, when the observed MCI score in the phosgene-exposed mice (red star) is significantly higher than that of random genes ($P<0.01$, 100 runs). The observed score in the air-exposed mice (blue star), as a control, is significantly lower than that of random genes (grey box).

Figure EV4

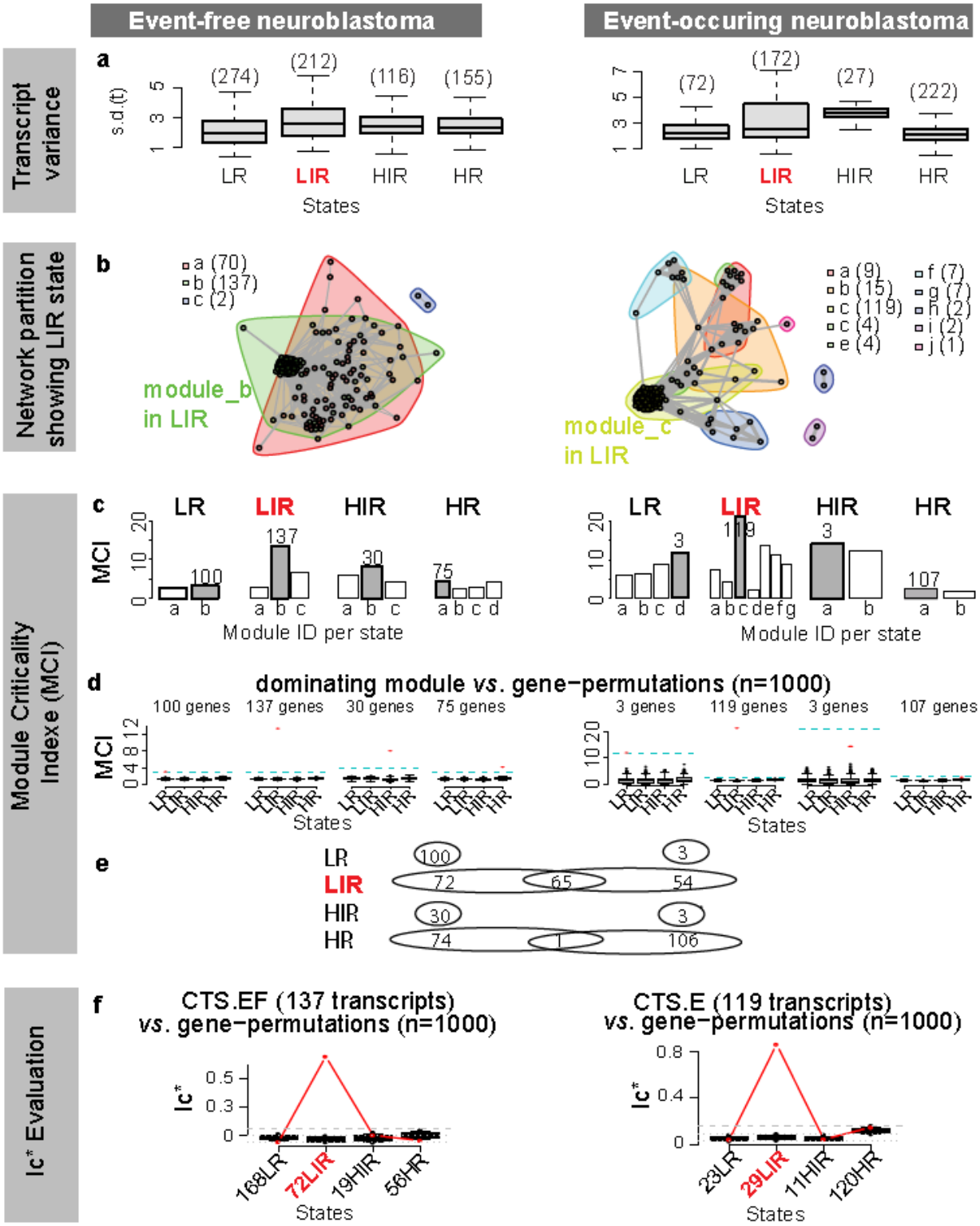
Event-free neuroblastoma
Event-occuring neuroblastoma
Transcript variance
a
s.d.(t)
(274)
(212)
(116)
(155)
(72)
(172)
(27)
(222)
LR
LIR
HIR
HR
States
Network partition showing LIR state
b
a (70)
b (137)
c (2)
module_b in LIR
a (9)
b (15)
c (119)
c (4)
e (4)
f (7)
g (7)
h (2)
i (2)
j (1)
module_c in LIR
Module Criticality Index (MCI)
c
MCI
100
137
30
75
3
119
3
107
Module ID per state
d
dominating module vs. gene−permutations (n=1000)
100 genes
137 genes
30 genes
75 genes
3 genes
119 genes
3 genes
107 genes
e
100
3
72
65
54
30
3
74
1
106
Ic* Evaluation
f
CTS.EF (137 transcripts) vs. gene−permutations (n=1000)
Ic*
168LR
72LIR
19HIR
56HR
CTS.E (119 transcripts) vs. gene−permutations (n=1000)
23LR
29LIR
11HIR
120HR

Figure EV4. Applying BioTIP to bulk-cell expression profiles of two NB populations (GSE49711)

a, Boxplot comparing the standard deviation (s.d.) of the preselected transcripts, with the transcript numbers at the top. In both populations, the highest variance presents at the LIR state.

b, A graph of network modulation determined by random walk for the preselected transcripts at the LIR state. Background colors represent different clusters. The module size (number of transcripts) is listed in parentheses.

c, Bar plots illustrate the MCI scores calculated for transcript modules per state. The dominating module per state is in grey with module size on top. In both patient populations, the MCI score peaked in the LIR state.

d, Comparison of the MCI score of the dominating module per state (red star) to that of randomly-selected genes (grey boxes), calculated throughout all four risk states. Green dashed line indicates the suggested significance cutoff. For the Event-occurring (E) population, three dominating CTSs are empirically significant, but in the event-free (EF) population, only the LIR CTS is significant.

e, Venn-diagram comparing the CTS identification in the E population (left) to that in the EF population (right), for each state, respectively. Only the CTS detected from the LIR states significantly overlap.

f, Plot comparing the observed Ic*-score (red line) to the random Ic* scores (grey boxes). The observed score was calculated for the 137 LIR-state-specific transcripts (herein the 'CTS') but at all four states; the random scores were calculated in the same way used 137 randomly chosen transcripts from the 48.5k-transcript background. In both populations, the observed Ic* score peaked in the LIR state.

Figure EV5

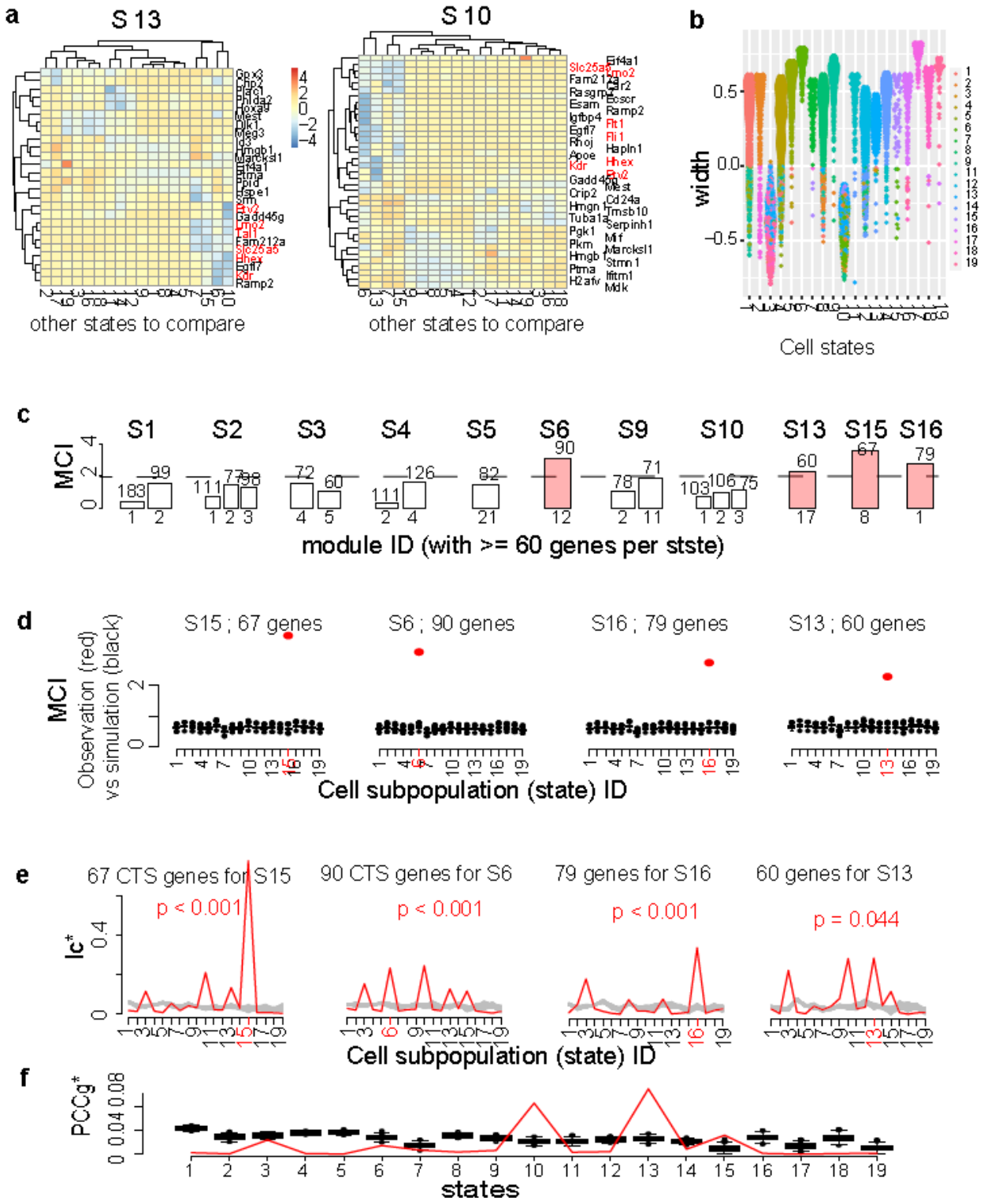
a
S 13
S 10
other states to compare
other states to compare
b
width
Cell states
c
S1
S2
S3
S4
S5
S6
S9
S10
S13
S15
S16
MCI
module ID (with >= 60 genes per stste)
d
S15 ; 67 genes
S6 ; 90 genes
S16 ; 79 genes
S13 ; 60 genes
MCI
Observation (red)
vs simulation (black)
Cell subpopulation (state) ID
e
67 CTS genes for S15
90 CTS genes for S6
79 genes for S16
60 genes for S13
p < 0.001
p < 0.001
p < 0.001
p = 0.044
Ic*
Cell subpopulation (state) ID
f
PCCg*
states

Figure EV5. Applying BioTIP to single-cell expression profiles of developing mesoderm (GSE87038)

a, Heatmap showing the top-ranked regulated genes for state 13 (left) and state 10 (right), respectively, over other states. Pairwise comparisons between cell states were run using the Wilcoxon rank-sum test. Euclidean distance was measured, and normalized log counts were centered and scaled in the row direction. The red color in the y-axis highlights the haemato-endothelial progenitor markers, the hematopoietically expressed homeobox gene Hhex, and the endothelial marker Slc25a5.

b, Distribution of the approximate silhouette width (y-axis) across 19 cell states (x-axis) of the dataset. Each point represents a cell and is colored with the identity of its own cluster if its silhouette width is positive, or that of the closest other cluster if the width is negative.

c, Bar plots illustrating the MCI scores calculated for transcript modules per state, with module size on top. Small modules with less than 60 genes were removed and the 11 cell states having module identification are shown. The dashed line indicates a threshold of score =2, above which 4 modules of interest are shown in red.

d, Comparison of the MCI score of the identified CTS (red dot) to that of randomly-selected genes (black boxes) within each cell state, for the 4 modules of interest. The simulation was run 1000 times, respectively.

e, Plot comparing the observed Ic* scores (red line) to the random Ic* scores of randomly-chosen genes among randomly grouped cells (1000 grey lines). For each CTS, the observed Ic* scores peaked in the predicted tipping-point state (labeled in red in the x-axis. For each tipping point, the P-value was estimated by Delta scoring.

f, Red line showing the average gene-gene correlation (PCCg) across states for the above identified 60 CTS genes. In contrast, the boxplots show the simulated PCCg scores (1000 runs).

Appendix

Other supporting Materials are available with this article online, including two S Tables and S Methods.