Open access • Posted Content • DOI:10.1101/668442

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

Xinan Yang, Zhezhen Wang, Andrew S. Goldstein, Yuxi Sun ...+5 more authors

Institutions: University of Chicago, Stanford University

Published on: 03 Mar 2021 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Tipping point (climatology)

Adopting tipping-point theory to patient transcriptomes unravels gene regulatory network dynamics

Running title: biological tipping-point characterization

Xinan H Yang[1,*], Zhezhen Wang[1], Antonio Feliciano[1], Dannie Griggs[1], Fangming Tang[1], Qier An[1,2], John M Cunningham[1]

[1] Department of Pediatrics, [2] Graham School of Continuing Liberal and Professional Studies, The University of Chicago, Chicago, IL 60637, USA

Key words: transcriptional oscillation; co-expression; critical transition; gene regulatory network; long noncoding RNA (lncRNA)

Highlights:

- Adopting tipping-point theory to transcriptomes
- Robust framework for the identification of critical-transition signal at tipping points
- Statistics on a high probability space enables the tipping-point analysis of cross-sectional data
- Application to gene expression of neuroblastoma reveals a gene regulatory network transition and underlying mechanisms

Summary

Abrupt and irreversible changes (or tipping points) are decisive in disease and normal phenotypic progress. Often, however, computational approaches detect critical-transition signals (CTSs) indicating tipping points from longitudinal data – which often are not available for patient transcriptomes. Here we adopt historical tipping-point approaches to cross-sectional data by modeling high probability spaces of phenotypes. We formulate this task as a generalized CTS-searching problem and derive a robust algorithm to solve it. We construct a comprehensive scoring scheme and successfully apply the scheme to lymphoma, lung-injury, heart-development, and neuroblastoma systems. Thus, we identify a spatial gene-expression feature for systematic dynamics at phenotypic tipping points, which can be exploited to infer functional genetic variations and transcription factors. Our framework ('BioTIP') can analyze not only time-course but also cross-sectional transcriptomes and is compatible with noncoding RNA profiles. Additional knowledge discovery that explores the critical transition of a system can be tested using our approach.

Introduction

'Tipping-point' theory has historically been used to characterize critical transitions and

been widely applied to ecosystems, climates, and other complex systems in various fields (*1*). In this theory, complex system dynamics exhibit abrupt shifts between two distinct, stable states. Each stable state is characterized by a quick recovery rate from perturbations within a large basin of attraction maintained by specific attractors (e.g., steady gene interactions). Therefore, monitoring a system for increased oscillations and correlations can indicate a competition between two controlling factors of adjacent stable states, thus identifying critical-transition signals (**CTSs**) (*2*).

Accurately modelling biological CTSs is of crucial importance, as they are associated with disease-phenotype transitions (**Fig 1a**). Phenotypic transitions are prevalent in disease progress, such as epithelial-mesenchymal transition, metabolic transition, and tumor-immune interactions (*3-5*). Because the first order for cellular function is to enable transcription, transcriptomic CTSs can shed light on disease mechanisms (*6-8*).

However, there exists a gap between tipping-point theory and transcriptomic application (**Fig 1b**). Practically, the only available tool, the R package *earlywarning* (*9*), analyzes univariate data, while biological data is frequently multivariate. Biologically, all existing transcriptomic CTS studies ignored the effect of long-noncoding RNAs (**lncRNA**) whose expression has been recognized to be of importance in determining tissue and phenotype specificity (*7, 8, 10-13*). Computationally, existing methods for identifying CTSs have major limitations: one previous method can detect tipping points only if given pre-defined genes (*8*), and while other methods, such as DNB (dynamic network biomarker) and its variations (*6*), support both tipping-point and CTS identification, they require the user to specify the parameter to cluster genes. Furthermore, existing methods are designed for longitudinal cohort studies and thus cannot be efficiently applied to cross-sectional transcriptional data.

Here, we introduce an advanced biological tipping-point (**BioTIP**) scoring scheme to characterize transcriptional CTS in a high probability space defined by phenotype, rather than an individual time-course space. The assumption is that slight perturbations of a hidden process (i.e., expression of a master regulator) could trigger a critical phenotypic transition through interrupting a gene regulatory network (**GRN**). As proof of concept, BioTIP was applied to three benchmarking datasets and compared with existing methods. Additionally, two CTSs identified from neuroblastoma (**NB**) transcriptomes repeatedly exposed potentially causative factors and epigenetic mechanisms for GRN transitions. This study demonstrates the broad utility of tipping-point theory for a wide range of complex disease systems.

Results

The BioTIP scoring system

To fill the gap between theory and transcriptomic application, particularly for the CTS identification in widely available transcriptomes of complex disease, we designed a detailed five-step analytical scheme (**Fig 1c**).

i) For data pre-procession, BioTIP inherits a concept of 'phenotype-defined states,' assuming the GRN within a state represents the probability distribution of the expression-configuration effects that maintain this state (*6, 10, 14*).

ii) BioTIP can find increased variance without normal-control samples. When selected, this feature employs a relative transcript fluctuation (**RTF**)-score that assesses the variance of transcripts t_r among patients in a state *r*, relative to its complement set t_{cr} of patients outside the state (i.e., a relative 'control state'). This gives

$$RTF(t)|r = \frac{\mathrm{sd}(\mathrm{t}_r)}{\mathrm{sd}(\mathrm{t}_{cr})} \qquad \textbf{(Formula 1)}$$

Given a large sample size, the empirical optimization of RTF by bootstrapping samples can pre-select robustly variable features (genes), therefore minimizing the error due to mislabeled samples or heterogeneity.

iii) It is essential to cluster pre-selected features into 'functional' modules that will serve as the inputs of downstream scoring calculation. BioTIP constructs densely connected subnetworks (or modules) based on random walks on graphs (**RW**) (*15*). In contrast, hierarchical clustering (**HC**), requires prior knowledge to set either a partitioning threshold or a limit for the largest sub-cluster (*6, 7*). Alternative strategies, including semi-supervised Partitioning Around Medoids (**PAM**), prove challenging when trying to optimize a number of intended clusters -- the parameter of *K* (*11*). A simple partitional clustering by splitting (*16*) is too limited as it returns only two modules. RW avoids these pitfalls while generating results that generally agreed with four other clustering methods (**Fig S1a**) and constructing relatively concentrated modules with modest module size (**Fig S1b**). Therefore, RW is as efficient as conventional methods without requiring prior knowledge of the resultant clusters.

iv) The two most essential indicators of an impending tipping point are increased transcript correlation and variance (*2*). BioTIP quantifies these indicators by first employing the DNB score (*6*), which we renamed as a module-criticality index (**MCI**). For a state *r* and its pre-selected transcript module *m*, comparing transcript deviation and correlation in this module t_m relative to its complement set t_{cm}, gives

$$MCI(m)|_r = Avg(sd(\mathrm{t}_{m.})) \times \frac{Avg(|PCC(\mathrm{t}_{mi},\mathrm{t}_{mj})|)}{Avg(PCC(\mathrm{t}_m,\mathrm{t}_{cm}))}|r \qquad \textbf{(Formula 2)}$$

Taking the absolute value (|.|) is meaningful because both positive and negative feedbacks are indicative of a looped regulatory network. The module with the highest MCI is termed the dominating "biomodule" for each state.

v) To further predict critical transitions and determine their significance, BioTIP employs a composite index of critical transitions (**Ic**) (*8*). Given a biomodule *m^* defined from a state *r* with samples $s_{r.}$, the Ic compares pair-wise correlations between samples (*k, l*) to that between transcripts (*i, j*) :

$$Ic(r)|_{m^{\wedge}} = \frac{Ave(|PCC(t_{m^{\wedge}i}, t_{m^{\wedge}j})|)}{Ave(PCC(s_{rk}, s_{rl}))} \qquad \textbf{(Formula 3)}$$

Following the theoretical induction, Ic(r) reaches its maximum value at the estimated tipping point r^ when given a successful biomodule m^ (*8*). Therefore, BioTIP can assess empirical significance for any CTS identification.

Lastly, an exploration in the biological relevance of CTS transcripts is introduced (**Fig 1d**). Because co-regulated GRN can amplify the effect of an upstream transcription factor (**TF**), we reason that CTS identification will allow for novel predictions of master regulators at or near the top of the regulatory hierarchy including noncoding variants. These relevance analyses include (but are not limited to) DNA-binding enrichment analysis, NB-susceptibility loci scan from genome-wide association study (**GWAS**), and additional knowledge discovery from CTS on disease progress.

Reproducibility, robustness, and empirical significance in transcriptomic data analysis

In transcriptome analysis, feature-preselection leverages between dimension and information. We first examined whether BioTIP performs robustly to feature-preselection choices, using a benchmarking dataset (GSE6136). In this dataset, murine lymphomas were originally grouped into five states according to their clinical presentation, pathology and flow cytometry (*17*), and the normal-tumor-transitional 'activated B cell' (P2) state was previously defined as a tipping point (*6*). BioTIP recaptured this critical transition, regardless of its parameter of feature-preselection with empirical significance (**Fig 2, a-b**).

Using the same dataset, we then tested whether BioTIP is sensitive to the way phenotypic states are defined. Note that global gene expression patterns of aggressive lymphomas (P5) are more similar to activated B cells (P2) than resting B cells (P1), and one P5 and one P3 sample showed the most similarity to P2 samples (**Fig 2c**) (*17*). Therefore, we re-assigned this margining P3 sample into P2 to generate a new but relative state definition; and instead re-assigned the clinically distinct P5 sample into P2 to simulate a misclassified state definition. As expected, BioTIP recaptured the P2 state

from the relative but not the misclassified sample definition (**Fig 2, d-e**). The enriched TF-binding motifs were consistently stable among the promoters ([-500,+200]nt) of these different CTS genes (all with the P2-identification) (**Fig 2f**). A similar stability was observed with a larger promoter-window size ([-1k,+500] nt). Additionally, two simplified phenotypic definitions of samples indicated P5 as a tipping point with the same CTS identification (**Fig S2, a-c**). We conclude that BioTIP is stable when given clinically and transcriptionally related definitions of states, particularly with repeatable TF predictions.

Next, we showed that BioTIP produces more significant CTS predictions than existing methods, having applications to both time-course and cross-sectional data. While DNB lacked empirical significance (**Fig S2c**), the BioTIP identification showed a global maximum specifically at the normal-disease transitional 'activated B cell' state (**Fig 2a**). We repeatedly observed this improvement in a time-course data analysis, in which temporal gene expression data of lung tissues collected from air-exposed (as control) or phosgene-exposed mice at nine time points (GSE2565, **Fig S2d**). Although CTSs at 8h had been reported by the DNB methods (*6, 18*), previous identifications lacked the global maximum (red dot vs. yellow boxes, **Fig S2e**) or focal empirical significance (**Fig S2f**). In contrast, BioTIP identification showed a global maximum significantly at 8h for the phosgene-exposed mice (solid red line, **Figs 2g, S2g**). We conclude that BioTIP outperforms three existing tipping-point scoring approaches for empirical significance.

Biological applications

We have demonstrated the CTS identification from gene expression profiles measured by microarray, although lncRNA transcriptome contributes significantly to the dynamics and diversity of cell specificity. Therefore, it is important to identify CTS from total RNA-seq profiles that cover lncRNA transcriptome. We first tested BioTIP on available time-course data of human embryonic stem cell differentiation to cardiomyocytes (GSE115575, **S Method**) (*19*). We identified 237 CTS transcripts containing important cardiac or neural developmental regulating genes, e.g., *CACNA1D*, *CDH13* (30, 31). Although more functional evaluations are required, the identification of the cardiac progenitor stage could be a meaningful tipping point (**Fig 2h**) because, in a similar cardiac differentiation system, early beating cardiomyocytes were observed only afterwards (*20*).

Then, a focused application to cross-sectional NB transcriptomes demonstrates three impacts on computational genomics. First, NB represents complex disease systems that require a data-driven model because there is limited knowledge on its heterogeneous mechanisms (e.g., few common variants with modest effect size) (*21, 22*). Second, regardless of distinct risk-stratifications between which no transition has been reported on individual level, high Myc-pathway activity and low neuronal

differentiation associate with poor outcome (*23*), suggesting an underexplored GRN trajectory. The availability of multi-omics data (*24*) (**Table S1**) allows for a systematic integration to understand GRN transitions. Lastly, transcriptomic profiles of NB had been collected not only from large populations (n=498, GSE49711, **Table S2**) (*25*) but also been sequenced for not only coding but also noncoding RNAs.

We identified a critical transition appearing at the Low-or-Intermediate-Risk (**LIR**) state in both event-free (**EF**) and event-occurring (**E**) populations (**Fig S3**). The 'tipping-point' character was verified by an increased variability and transcript correlation among the CTS transcripts at the LIR state (**Fig 3a**). Derived from independent cohorts, the two CTSs at the LIR-state significantly overlapped, sharing 65 transcripts including lncRNAs (**Fig 3b**, P<2e-16), thus indicating a common transcriptional configuration committing to phenotype-defined states in NB.

To understand how lncRNAs impact GRN, we evaluated the network-degrees of lncRNA-gene links. In both identified CTSs, lncRNAs and genes exhibited a significantly higher proportion of connections than gene-gene or lncRNA-lncRNA connections did (P<1.4e-7, **Fig S4, a-b, S Method**). Over 90% of these co-expressed lncRNA-gene pairs resided on different chromosomes (**Fig S4c**), agreeing with the fact that most heritability of gene expression is due to remote trans-regulation (*26*).

It's notable that both CTSs were significantly depleted with differentially expressed transcripts (**Fig 3c, yellow box**), confirming that a CTS cannot be identified by conventional statistics on group difference in mean (*6*). We therefore asked whether CTS transcripts could disclose underappreciated genetic variants, given that most noncoding variants had modest effect size and thus need to define function with transcriptomic information (*22*). Overall, four lncRNAs and five genes (**Figs 3c**, **S4d**, red and blue dots, respectively) are CTS-marked with promising functional significance for future investigation (**Table S3**). One notable finding was gene *DCDC2*, whose copy-number-gain presented in 6 out of 29 NB tumors studied (*27*) and lied adjacent to NB suppressive lncRNA pairs *CASC14/15* (*28*). Note that eight suggestively (8e-5<P<5e-5) susceptibility loci were overlooked in the previous GWAS (*29*), which all appear to have some connection with NB in the literature (**Table S4**).

To infer the biological mechanisms underpinning a phenotypic transition through the identified CTS transcripts, we performed upstream regulator analysis (**Fig 3d**). TF-binding motif analysis of the identified two sets of LIR-CTS promoters revealed known TFs implicating neurological development and disease, such as YY1 (*30*).

In summary, BioTIP captures unique CTSs based on correlations and deviations among transcriptional dynamics, which are uncharacterized by traditional statistics for state averages, thus shedding light on potential causative mechanisms that tip the phenotypic

transitions alongside time-course or GRN transitions between stratified non-time-course states.

A model of CTS-recognized GRN suggesting metabolic reprogramming

We subsequently investigated the mechanisms underlying the GRN transition in NB by focusing on the 13 recurrently identified TFs (**Fig 4a**). In GRN, little malfunction of master regulators may have enlarged consequence, and network hubs often indicate master regulators. Using Ingenuity pathway analysis (IPA) (*31*), 11 TFs were involved in a disease-associated network in which NHF4A stood out as a hub (**Fig 4b**). *HNF4A* gene encodes HNF4α, a nuclear transcription factor that binds DNA as a homodimer, controls the expression of many genes, and interacts with over thousands of proteins (*32*).

We tested the presumption that HNF4α may cooperate with the well-known NB-associated myc-family to implicate NB tumor progression. In line with the significant number of shared target genes between HNF4α and N-myc (OR=3.4, **Fig 4c**), we further examined the transcriptional dependence (i.e., global fold changes between high *HNF4A*-expressing NBs versus low *HNF4A*-expressing NBs). This analysis presented a pattern strongest among the LIR samples -- a positive correlation between *HNF4A*- and *MYCN*-dependence, reverse correlation between *HNF4A*- and *MYC*-dependence, and reverse correlation between *MYCN*- and *MYC*-dependence (**Fig S5a**). Surprisingly, we recaptured this pattern across all samples, and also in three independent cohorts (**Fig S5b**). Furthermore, there were reports supporting transcriptomic *MYC-HNF4A* co-regulation (*33, 34*). In NB, a recent study predicted HNF4α working together with N-myc as activated upstream regulators of proteins overexpressed in *MYCN*-amplified compared to non-amplified tumors (*35*). We conclude that HNF4α may involve a temporal transcriptional coordination underlying the GRN transition in NB.

We then searched for the potential role of HNF4α in NB tumorigenesis. Confirming its leading enrichment among CTS promoters (**Fig 3d**), there were 39 HNF4α-targeted CTS genes (OR=12) in contrast to only 7 N-myc-targeted CTS genes (OR=0.2) (**Fig 4c, Table S5**). These 39 genes significantly over-represented the Reactome 'lipoprotein metabolism' pathway (GO:0042157, P<2e-16) and the biological process 'protein-lipid complex' (GO:0032994, **Fig 4d**). These results supported a notion of how metabolic stress makes trade-off between the endogenous synthesis and exogenous uptake of fatty acids that are essential for tumor energy production and protein modification (*36*).

We next tested whether HNF4α-targeted lipid genes spotlight any metabolic features that derive the NB ontogenesis from its neuroectodermal origin. This was motivated by both the above observation and the report that lipids were particularly elevated in NB and associated with adverse clinical outcomes (*37*). Among human NB cell lines

(GSE89413, **Table S6, Fig S5c**), 8 of 9 HNF4α-targeted, lipid-related CTS genes were expressed more in tumorigenic N-type cells than in non-tumorigenic S-type cells (Wilcox t-test P=7e-6, **Fig 4e**). These results support a therapeutic potential of lipotoxicity for children with NB by influencing cancer metabolism (*38*).

Since LR- and HR-NB have distinct phenotypes and no 'transit' has been reported on individuals, to explain our results, we propose a new model regulating the CTS-derived GRN (**Fig 4f**). In the model, a subtle but fine-tuning CTS network reshapes tumor microenvironmental potentials, which lead to distinct phenotypes that are irreversible and thus present strong, conventional core network signals.

In summary, BioTIP provides a mathematical basis to identify the critical GRN transition from cross-sectional data by seeking a maximum score of probability distributions. On the one hand, the CTS-GRN identified from time-course data could be involved in pushing the system over the saddle-node tipping point towards stable points. On the other hand, the CTS-GRN identified from cross-sectional data could reprogram the shape of the epigenetic landscape. In either case, a critical-transition state reflects the vulnerable regime of the GRN that could be reversible (*39*). Therefore, CTS derives novel microenvironment insights in the context of phenotypic transitions beyond the specific context of neuroblastoma. Additional knowledge discovery that explores the critical transition of a system can be tested using our approach.

Discussion

This work introduced an advanced computational framework, BioTIP, which adopts tipping-point theory to identify CTSs among phenotypic states. Using BioTIP, we recaptured the known normal-lymphoma transitional state, the time-point just before critical lung injury, the stem-cell differentiating day just before heart beating, and the LIR state in NB. These signals were previously overlooked due to their insignificant expression change in group-mean tests.

A broader significance of this study lies in the fact that applying pharmacological intervention during the fragile and reversible window of state transition has the most potential for halting disease progression (*5, 12*). For example, intermediate-risk NBs have been shown to have high cure rates and are susceptible to chemotherapy (reversible) (*40*). Thus, this study serves as proof-of-concept that transcriptional signatures can be indicators of a reversible state (or GRN) in disease progression and may help improve treatment strategies.

It is worthwhile to point out three methodological contributions of BioTIP to aid new biological discoveries, thus having broad applications.

1) Given the underlying assumptions that critical transitions exist alongside a phenotypically defined progressive trajectory, BioTIP is robust, reproducible, and sensitive in CTS identification.

2) BioTIP performs transcriptomic analysis on phenotypic units (sharing regulatory mechanism) by identifying the 'probability-distribution transition' and demonstrating its ability to reveal master regulators of disease phenotypes. Its application to cross-sectional data has the advantage of removing random noise that may also be present in CTS (*2*). Distribution-transition pinpoints transcriptomic changes whose individual effects are insignificant in static statistics, but whose coherent oscillation makes a unique dynamic signature of GRN.

3) Transcriptomic architecture of complex traits may be dominated by lncRNA interactions where lncRNAs are controlled by upstream TFs and which in turn affect gene expression (*26*). However, the small effects of lncRNA expression are hard to detect. The CTS transcripts propose a new theoretical path for linking weak lncRNA signals to trans-regulatory GRNs and disease phenotypes. In this way, BioTIP provides immense value by spotlighting gene-lncRNA-shared TFs that may mark disease-state transitions.

Certain limitations of the BioTIP approach must be addressed. This approach relies on sample-grouping, which is not always practical. When normal controls are unavailable, BioTIP uses a 'relative control' by comparing one state against the other complement states in the same disease. This requires a large sample size and enrollment of most states of a disease. Additionally, it remains unclear whether the current framework can correctly identify multiple critical-transition states that are postulated in some complex biological systems *(41, 42).*

After applying BioTIP to NB transcriptomes, we reported four systems biology outlooks:

1) The CTS analysis could facilitate the discovery of functional loci from GWAS-findings, providing versatile perspectives for future biomarker research and the development of treatment. In this way, this model goes beyond the previous 'omnigenic' hypotheses (*26*) by considering GRN dynamics. Because small-scale variants are able to escape natural selection's strong constraints against large-effect variants (*43, 44*), grouping individuals into phenotype-defined groups allows us to unveil the 'distribution transition' that is seemingly trivial at the individual level but significant alongside the population-wide windows.
2) HNF4α may be a master regulator of the transcriptional regulatory hierarchy. The HNF4A GRN may engage in repressive histone modification and be affected by other genes' expression, such as CTS member *HNF4A-AS* and known NB core pathway members *MYC* and *MYCN* (**Figs 4f, S4d**). A complex if not

contradictory role of HNF4α in tumor progression has recently emerged in cancer of multiple organs and embryonal cells (*45-50*). The role of HNF4α in tumor progression is complex because transcriptional regulation of HNF4A is likely to be temporal (*51*).

3) Epigenetic regulation plays a role in phenotypical CTS. The presence of histone deacetylases (HDACs, **Fig 4b**) is of particular interest given that the HNF4α1 (isoform 1) recruits HDAC activity to target promoters and thus mediates multiple interactions (*52*). In multiple cancer cell lines, a HDAC inhibitor has been shown to partially recover p53-mediated repression and target HNF4α modifications, thus seeming to serve as an effective treatment for advanced carcinomas (*53, 54*). Furthermore, HDACs are known to control MYCN function and are upregulated in chemotherapy-resistant NB cells (*55*). Therefore, a functional interaction seems to exist among the key genetic regulators, epigenetic regulator HDAC, and the identified transcriptional CTS in NB tumorigenesis.
4) Our prediction brings the lipoprotein-metabolism-associated neural cell microenvironment to the forefront of attention. Dynamic CTS-GRN may involve lipid metabolism that helps cancer cells thrive in a harsh tumor microenvironment (*36*). Indeed, increased lipid accumulation has been associated with NB suppression and adverse clinical outcome (*37, 38*). In this way, this application pokes a novel computational inference for understanding the epigenetic microenvironment impacted by temporal GRN dynamics.

In conclusion, studying the perturbed connectivity in a transcriptional network between two equilibrium states can identify important CTS biomarkers (*56, 57*). The identified CTS and their upstream regulators are novel treatment targets for making use of a fragile and reversible state in disease progression. We expect future functional follow-up on CTS genes to underscore many underappreciated mechanisms such as lipid-metabolism alteration in cancer (*58*). We expect the CTS concept will have broad utility for computational analysis of gene-regulation networks and single-cell transcriptomes in a wide range of biological systems.

STAR Methods

CTS identification from NB transcriptomes

NB samples were categorized into four states based on the phonotypic Children's Oncology Group (**COG**) risk group criteria (*59*). Given the lower expression levels of lncRNA relative to coding genes (*60*), we applied a mild 'expression' cutoff that leveraged noise for lncRNAs' transcriptional insights. 48.5k (80%) of transcripts remained after this cutoff (the cross-sample average expression level ≥1 on log2 scale). 20.6k (57.5%) of these transcripts were lncRNAs. A robust RTF-score was calculated by bootstraping 80% of samples (n=100) in each state to decrease the impact of

individual heterogeneity. This bootstrapping pre-selected a variety of transcripts ranging from 27 to 274 at each state, and the LIR state showed increased and scattered transcript variance (**Fig S3a**). After employing a network partition strategy, each pre-selected transcript set was further divided into 2-10 modules (**Fig S3b**). Then, a dominating module, herein the 'biomodule,' was identified by ranking the MCI for each state (**Fig S3c**). Again, the biomodule of the LIR stage showed the highest MCI-value, in both populations. This critical transition was confirmed by the pattern of Ic -- the Ic of the LIR-derived biomodule peaked at the LIR-state. Additionally, the observed Ic was significantly higher than the scores (Ic*) of any randomly combined transcripts at the LIR-state (black line versus grey lines in **Fig S3d**). Sample-size-controlled bootstrap analysis further confirmed a higher value for the observed Ic for the LIR patients than any randomly grouped patients (empirical P<0.001).

Integration of GWAS findings with CTS identification in NB

Seeing as GWAS rarely identifies single variant-trait associations but instead finds blocks of associated variants in linkage-disequilibrium (**LD**) blocks (*61*), we scanned 163 NB-associated blocks hosting significant or suggestive NB-susceptibility SNP(s) (GWAS P<5e-5, **S Method**) (*62*). Because regulatory interactions are frequently observed within topologically associating domains (**TADs**) (*63, 64*), it was assumed that most NB-associated cis-actions would happen in the 151 TADs defined by the 163 LDs. Querying these 151 NB-associated TADs with the identified CTSs allowed for pinpointing nine TAD regions with susceptible SNPs and CTS transcripts.

Motif-enrichment analysis

A promoter was defined as the 2k-nt window ([-1,500, +500]) (NB study) and/or 700-nt window ([-500, +200]) (mouse lymphoma study) around the CTS transcription start site (**TSS**). The blacklist, an identification of problematic regions of the human and mouse genome, was removed for the downstream analysis (*65*). Enrichment was evaluated for 1,571 known DNA binding motifs (**S Method**). The significance was considered with the criteria: FDR <0.05, average binding-affinity score >1; and the breadth of enrichment (using a 10% ranking threshold) >15%.

Define TF-target genes in NB

673 'HNF4α-target genes in NB' (**Fig 4c**) are the intersection of 4,576 HNF4(α)-binding targets and 3,434 *HNF4A*-dependently expressed genes. HNF4(α)-binding targets were the union from two resources. In the first, over 1.1k targets were defined in the MSigDB database (v3.1) (*66*). In the second, 3.9k targets presented *in vitro* ChIP-seq signal at promoters, which was defined as [-1000, 500]nt around TSS, in both the GEO and ENCODE databases (**Table S1**). In parallel, the *HNF4A*-dependent genes in NB were detected from four independent cohorts of NB patients (E-MTAB-1781, E-MTAB-

179, GSE3960, and GSE49711). A gene with significant dependence (FC>1.5, FDR<0.05) occurring in any one of the four cohorts was counted as '*HNF4A*-dependent in NB.' Using the same methods, we defined 2,845 c-myc targets in NB and 2,381 N-myc targets in NB.

N- and S-type NB expression

The Fragments Per Kilobase of transcript per Million mapped reads (FPKM) of 26.8k transcripts in 41 human cell lines were downloaded from GEO (GSE89413), among which 39 were NB cell lines. From the literature (**Table S6**), we annotated ten neuroblastic (N-type) and three substrate-adherent (S-type) NB cells. Lowly expressed, and noncoding transcripts were excluded, leaving 16k 'expressed' coding genes with an average FPKM across all cells > 0.01. Normalized rank values were assigned to these expressed coding transcripts per cell, $R = \frac{\mathrm{rank(FPKM+a)}}{\mathrm{N}}$, where N is the number of all expressed transcripts and a=1e-6 is a parameter to adjust zero value for log-transformation. 8 out of the 9 HNF4A-targeted lipid genes in the CTS identification were expressed. Therefore, their *R* values were compared between N-type and S-type cells using the Wilcox test (**Fig 4e**).

Acknowledgements

This work was supported by NIH grant 5R21LM012619 (XY, ZW, FT, JC), the University of Chicago, Biological Sciences Collegiate Division Quantitative Biology Fellowship (DG), and Graham School of Continuing Liberal and Professional Studies, Master of Science in Biomedical Informatics Capstone Project (QA). We thank Drs. Charles Y Lin and Brian J. Abraham for providing global ATAC signature and H3K27ac signature, respectively, in NB cells. We thank Helen Salwen for reviewing and categorized the neuroblastoma cell lines, and Linsin Smith for discussion on figure legends. We thank UChicago Center for Research Informatics for supporting the Ingenuity IPA analysis and high performance computing services. We thank Susan Cohn for her insightful discussion when initiating the project.

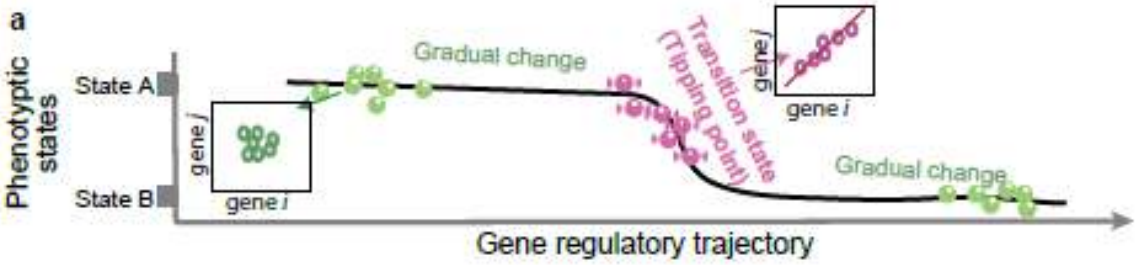
Author contributions

XHY designed the study, helped data analysis, interpreted the results, and wrote the manuscript, with comments from JMC. ZW collected data, performed data analysis, and contributed to result interpretation. AF analyzed the data (Fig 2h) and drafted its method. DA wrote the Table S2 and edited the manuscript. FT contributed to the biological discussion and model evaluation. QA committed to literature review (Fig 1b). JMC directed the project design and result interpretation and revised the abstract.

Competing interests

The authors declare no competing interests.

Figure Legends



b

Function	BioTIP	Tesc 2012	Chen 2012	Teschendorff 2014	Richard 2016	Mojtahedi 2016	Yang 2018	Zhang 2019
Gene expression feature to be modeled / Publication	■ ■ ■	■	■ ■	■ ■	■ ■	■ ■	■ ■	■ ■
Sample-sample Correlation								
Gene-gene Correlation								
Variance								
1. Applying to phenotypic series (cross-sectional transcriptomes)	✓	✓	✓	✓	x	x	x	x
2. Control-free feature selection	✓	a	a	✓	a	NA	a	b
3. Parameter-free network partition	✓	NA	HC x	PAM x	NA	NA	|PCC| x	#
4. Identifying biomodule	✓	x	✓	✓	✓	x	✓	✓
Robust to noisy data	✓	x	x	x	x	x	x	✓
Biological (clinical) relevance	✓	✓	x	✓	x	x	✓	✓
Cohort-independence	✓	✓	x	✓	x	NA	x	x
Incl. lncRNA (mRNA-lncRNA pair)	✓	x	x	x**	x	x	x	x
5. Determining "tipping-point"	✓	✓	✓	✓	✓	✓	✓	✓
Biomarker-free	✓	x	✓	✓	✓	x	✓	✓
Robustness (empirical significance)	✓	c	x	c	x	✓	x	c

c

BioTIP
1) Pre-process gene & lncRNA expression data matrix (remove noise)
2) Pre-select transcripts with highest fluctuation
3) Partition transcript-interaction network into clusters
4) Identify biomarkers (i.e., the dominating cluster)
5) Determine tipping-point

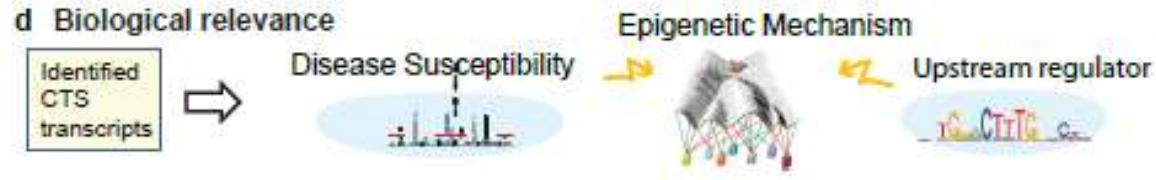


Figure 1. Tipping-point theory and adaption to gene expression data analysis.
a, Schematic illustration of samples (balls) in a 'tipping point" (red), where a tiny perturbation can trigger a critical transition from one stable state to another (green). In a stable state, the samples show lower gene-gene variation across their homogenous population (embedded panel, green balls). Therefore, increased gene expression variance and correlation are the detectable critical transition signal (CTS)-features (embedded panel, red balls).
b, Comparing the functions of BioTIP to that of seven previous publications. BioTIP and these publications study biological critical transition signals (CTSs), and the modeled gene expression feature(s) are denoted by colored square(s) on top. Abbreviations are as follows. a: Feature selection requires a normal (or, time zero) control group; b: Used the previous-time point as a relative control; c: Empirical evaluation was conducted for gene biomarkers but not for the identified 'tipping-point;' #: Network partitioning requires prior knowledge on gene-gene interaction; **: Analysis was conducted on coding-gene methylation profiles. HC: Hierarchical clustering; PAM: Partitioning Around Medoids; PCC: Pearson Correlation Coefficient.
c, A five-step analytics pipeline of the BioTIP method for identifying tipping points and the corresponding biomarkers (multiple CTS transcripts).
d, Example of potential downstream implications of BioTIP to understand gene regulatory network dynamics. Abnormally expressed CTS genes might reprogram a normal epigenetic landscape by an amplified effect of local disease susceptibility or transcription factor activity.

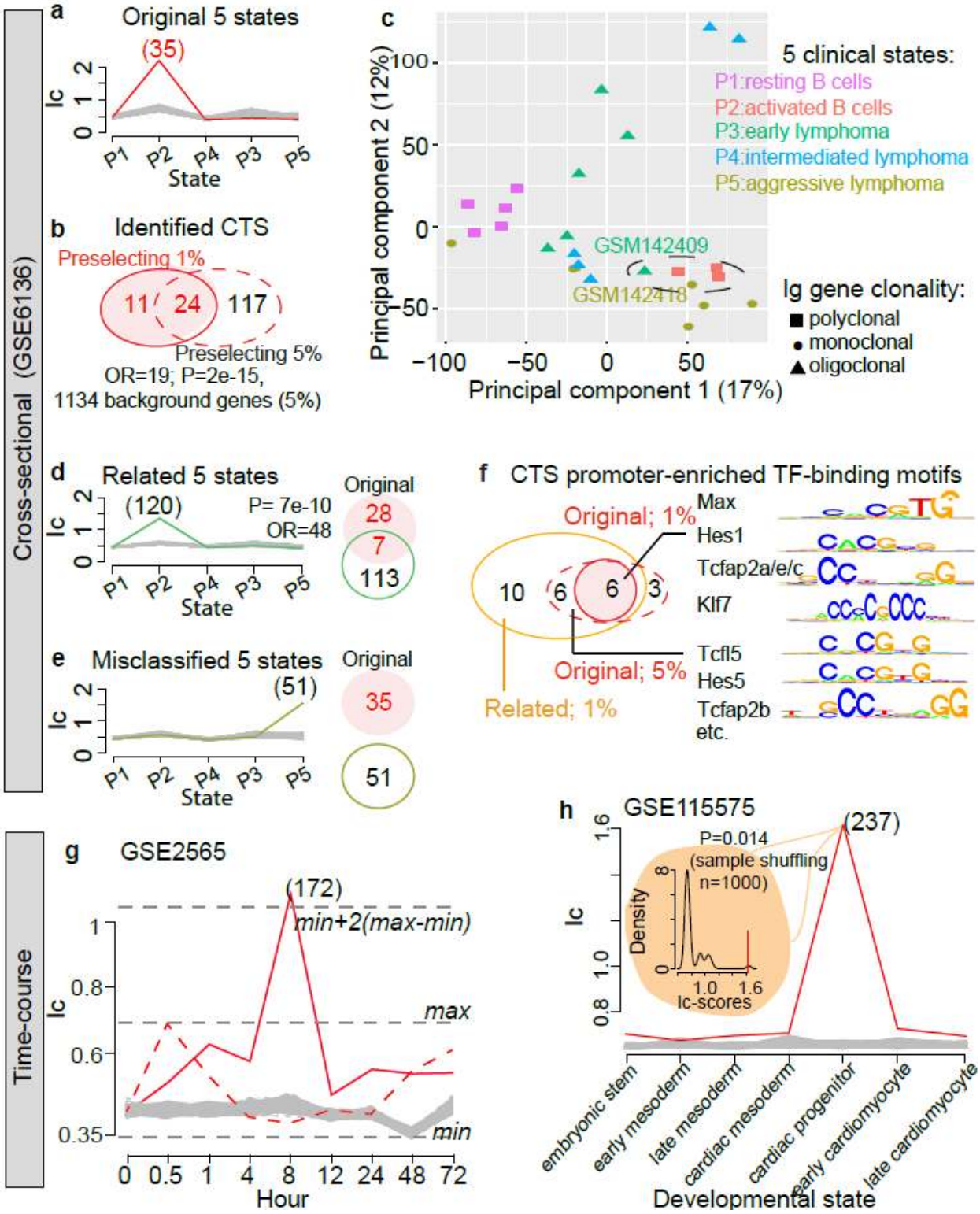
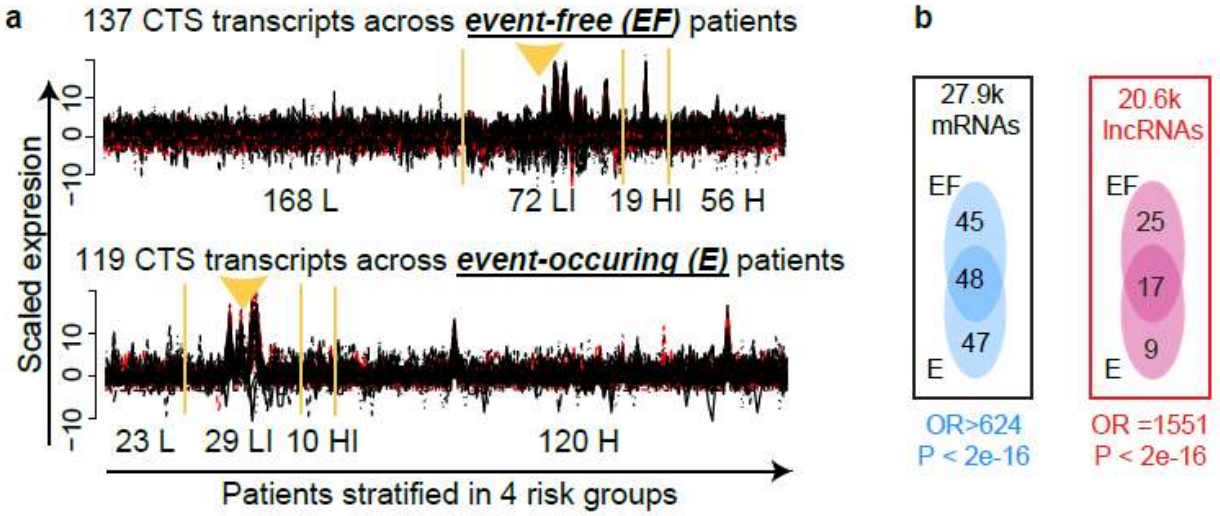


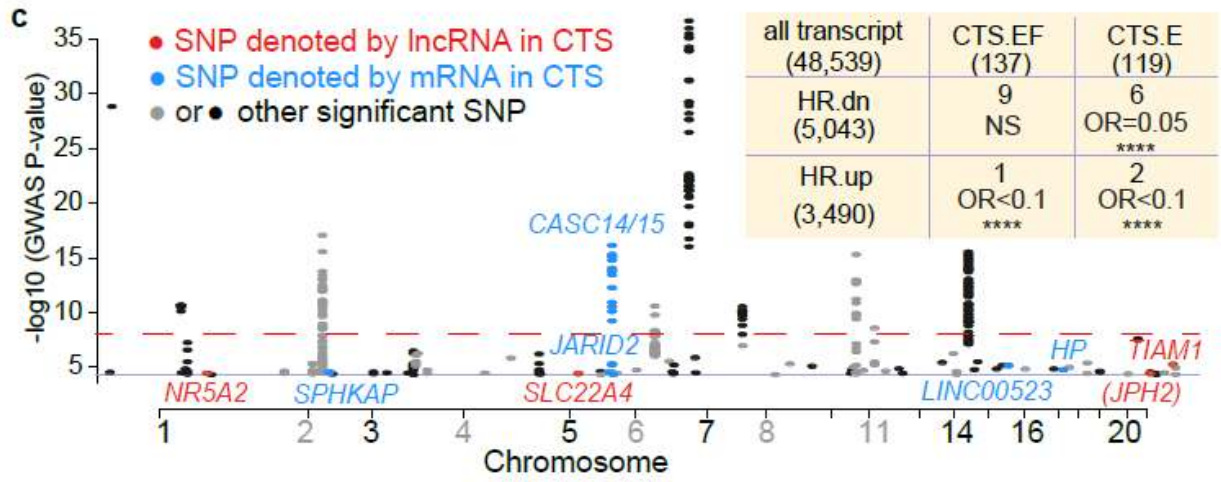
Figure 2. Robustness of the BioTIP identifications and exemplified applications to benchmark datasets.

a, BioTIP results given an originally reported 5-state sample-definition, in which P2, the tipping point between normal B cells and lymphoma, is identified with a set of 35 genes presenting a significantly higher score at P2. Plot shows index of critical transition (Ic, y-axis) against state (x-axis). Red line indicates score of observation against random genes (grey).
b, Venn-diagram depicting two-thirds of the above original CTS identification was recaptured after setting an alternative parameter, the percentage of pre-selected genes. The P-value of enrichment was estimated using Fisher's Exact test (FET) against 1134 commonly selected probes. OR: Odds ratio.
c, Principle component plot visualizing global gene expression patterns, based on all 18k expressed mouse gene probes in this dataset (S Methods). Each node is one sample, and color denotes the phenotypical states. Dashed-circle spotlights two samples (GSM142409 in green and GSM142418 in yellow) whose gene expression patterns are similar to the transitional P2 samples.
d, Left: BioTIP results given 1% probe selection and a related 5-state definition (i.e., grouping an early lymphoma sample (GSM142409) into the activated B cells)*. Right: Venn-diagram showing one-fifth of the original CTS identification was recaptured by this "related" CTS-identification (n=120). FET was used to estimate the P-value against 22k measured probes in this dataset.
e, BioTIP results given 1% probe selection and a misclassified 5-state sample-definition (i.e., grouping an aggressive lymphoma sample (GSM142418) into the activated B cells)*. No overlap between the 35 original CTS probes and this misclassified CTS-identification (n=51).
f, Left: Venn-diagram showing overlap among the enriched transcription factor (TF)-binding motifs among CTS-promoters of P2 in each category (original 5-state with 1% probes preselected; original 5-state with 5% probes preselected, and related 5-state with 1% probes preselected). Promoters ([-500, 200] around TSSs) were tested using the R-function *PWMEnrich*, and significance was considered with criteria: P<0.005, (average binding-affinity) score >1.1; and the breadth of enrichment (using a 10% ranking threshold) >15%.
g, BioTIP results of time-course data of disease model (GSE2565). This dataset profiles the gene expression of two groups of CD-1 male mice after exposing to air (as control) and phosgene, respectively. The Ic-score of 172 identified genes significantly peaked at the 8 hour with phosgene-exposure (solid red line*), but not the air-exposure (dashed red line). The top horizontal dashed line indicates an empirical significance threshold which is 2-fold above the range of random scores.
h, BioTIP results of time-course data of normal tissue development (GSE115575)*. This dataset profiles 7 different time-points (in triplicates), covering key stages of human embryonic stem cell differentiation to cardiomyocytes. Embedded histogram shows the empirical significance of identifying the cardiac progenitor as a tipping-point, the

observation score (red) sitting far higher than scores of random sample grouping (black).

*: Plot shows lc versus state, identifications (yellow), and random genes (grey); and the number of identified CTS transcripts (or probes) is given in parentheses on top.





all transcript (48,539)	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 ****

d

Top 12 of 23 enriched motifs

Target	PWM	Score	P-value	top %
OLIG1		1.3	2.3e-07	19
HNF4A		5.0	1.5e-06	17
NFE2L1		2.6	2.6e-06	23
HNF4G		4.2	4.3e-06	19
MYF6		1.4	2.0e-05	17
EHF		2.8	2.0e-05	25
KIF22		1.1	3.2e-05	17
GPAM		1.1	3.5e-05	22
FIP1L1		1.1	7.8e-05	19
NR4A1		1.0	0.00012	17
PICK1		1.2	0.00016	23
HNF1B		7.2	0.00016	16

Top 12 of 29 enriched motifs

Target	PWM	Score	P-value	top %
YY1		1.1	6.7e-10	20
HNF4A		8.5	9.1e-10	24
HNF4G		6.6	2.5e-09	23
NR2F1		5.6	8.7e-09	21
ATF6		1.4	1.6e-06	24
DUS3L		1.3	2.4e-06	25
RXRA		2.7	1.3e-05	19
PTCD1		1.2	1.7e-05	17
NR2F6		5.5	2.7e-05	17
EHF		2.8	2.7e-05	24
NFIX		1.3	4.7e-05	20
GPAM		1.1	4.9e-05	21

Figure 3. Applying BioTIP to two neuroblastoma transcriptomes identifies two sets of CTS transcripts, both predicting a risk-transitional regulation at the Low-or-Intermediate-risk (LI) state. L: low-risk; HI: High-or-Intermediate-risk; H: High-risk.
a, Scaled expression levels of the CTS transcripts (lncRNAs in red and mRNAs in black) showing increased fluctuation at the identified tipping-point (yellow triangle). Vertical lines separate the samples into four risk states, with sample sizes listed below.
b, Venn diagram comparing the CTS transcripts identified between two populations (EF and E). Between the two identified CTSs, there are 48 shared mRNAs and 17 shared lncRNAs. Enrichment was assessed using FET against its corresponding transcript background given on the bottom. OR: Odds Ratio.
c, Manhattan plot showing the 379 suggestive NB-susceptibility SNPs extracted from publication (via the GRASP database) with a cutoff of P<5e-5. The dashed red line represents the conventional significance threshold of P<1e-8. Each SNP is marked by one dot, and the CTS transcript-localized SNPs are denoted in color. Embedded box shows the depletion (OR<0.5) between CTS transcripts and historically differentially expressed transcripts (FC>2, FDR<0.05, comparing high-risk patients to non-HR patients) in the same transcriptomic dataset.
d, DNA-binding motifs overrepresented at the CTS-transcription promoters ([1500, 500] around TSS) in the event-free population (left) and the event-occurring population (right), respectively. TF-binding was tested using the R-function *PWMEnrich*, and significance was considered with criteria: FDR <0.05, Score >1, and the breadth of enrichment (i.e. top %) >15% using a 10% ranking threshold. TF motifs recaptured in both shown in red. PWM: position weight matrix.

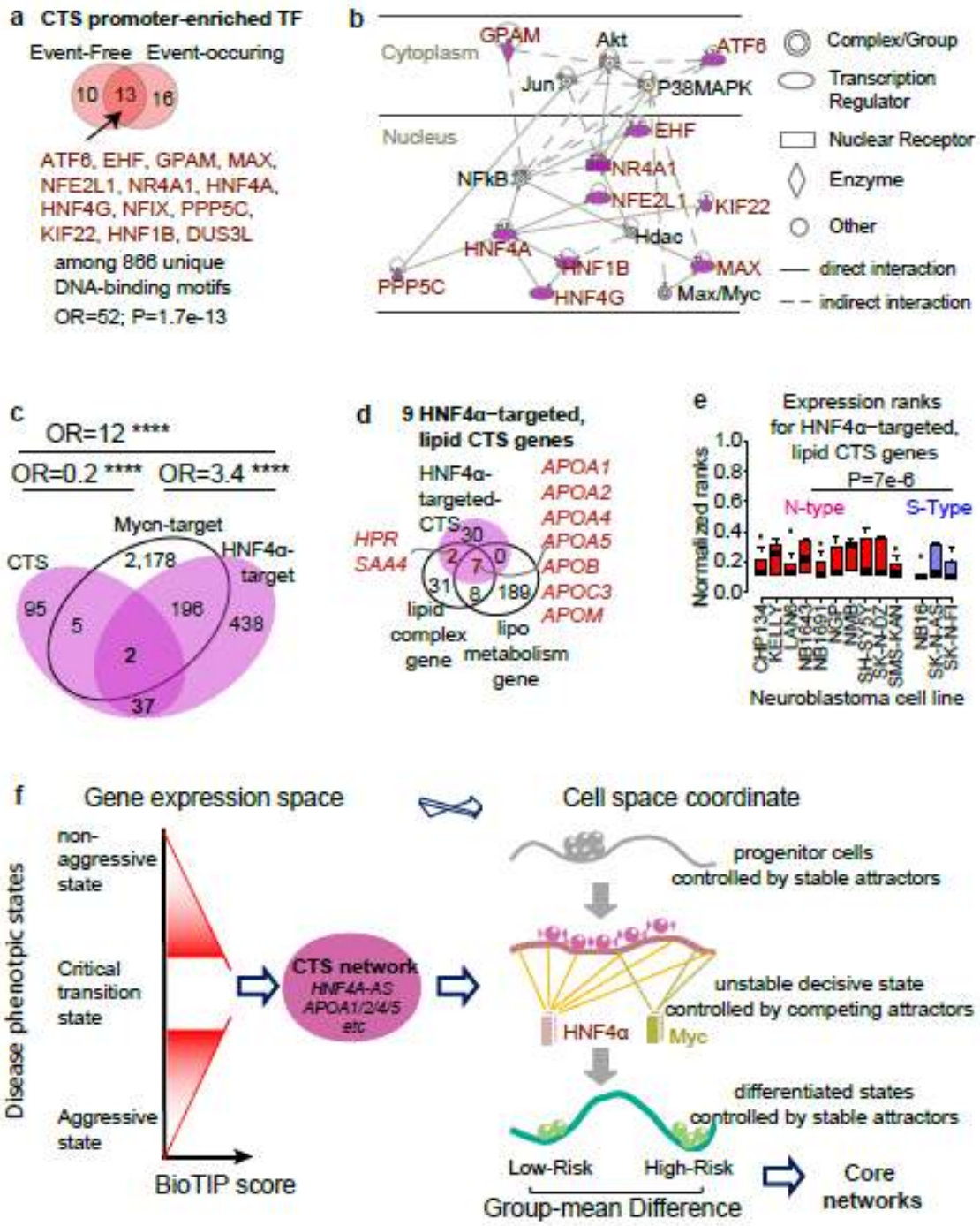


Figure 4. HNF4A represents a master regulator of CTS genes that reveal metabolic reprogramming in the underlying GRN transitions.

a, Venn-diagram comparing significant promoter-enriched DNA-binding motifs for the two CTS identifications. Gene symbols encoding the 13 commonly enriched transcription factors (TFs) are listed. The P-value of overlapping was tested using FET.

b, Protein-protein-interaction network view of the 13 reproducibly CTS-overrepresented TFs (red nodes). Shown are 11 out of the 13 TFs that interact in a network related to endocrine system disorders in the Ingenuity Pathway Analysis.

c, Venn-diagram showing the overlap of all 139 identified CTS genes with the target genes of the transcription factors Mycn, and with the targets of HNF4α. These 139 genes were encoded by 140 CTS mRNA transcripts, in which 39 (=2+37) are HNF4α-targeted CTS genes. ****: FET P-value<2e-16, given 21k human gene symbols.

d, HNF4A-targeted CTS genes are over-represented in the Reactome 'lipoprotein metabolism' pathway (GO:0042157) and the 'protein-lipid complex' (GO:0032994). The gene symbols of the 9 HNF4A-targeted lipid genes are shown.

e, Boxplot comparing the gene expression pattern of these HNF4α-targeted lipid genes between tumorigenic N-type and non-tumorigenic S-type cells (GSE89413, **Table S6**). The P-value of the Wilcoxon rank-sum test is indicated. One gene was not reported in the dataset and thus ignored.

f, A model of CTS tuner. Left: The BioTIP-scoring system identifies the critical transition state and a CTS-mediated gene regulatory network (GRN) from gene expression features. Right: The GRN reveals a critical epigenetic landscape reprogramming moment in which the flat platform visualizes a relatively unstable state representing a vulnerable phenotype. In contrast, the core network of disease was historically identified by group-mean difference between later stable states (bottom right).

Appendix

Supporting Materials are available with this article online, which are a file of five Supplementary Figures and a file of other Supplementary Materials including Abbreviation, Supplementary Methods, and six Supplementary Tables. More details on benchmark data analyses are given at https://github.com/xyang2uchicago/BioTIP_application.

Reference

1. C. F. Clements, M. A. McCarthy, J. L. Blanchard, Early warning signals of recovery in complex systems. *Nat Commun* **10**, 1681 (2019).
2. M. Scheffer *et al.*, Anticipating critical transitions. *Science* **338**, 344-348 (2012).
3. X. Tian *et al.*, Polo-like kinase 4 mediates epithelial-mesenchymal transition in neuroblastoma via PI3K/Akt signaling pathway. *Cell Death Dis* **9**, 54 (2018).

4. A. Kianercy, R. Veltri, K. J. Pienta, Critical transitions in a game theoretic model of tumour metabolism. *Interface Focus* **4**, 20140014 (2014).
5. W. J. Lesterhuis *et al.*, Dynamic versus static biomarkers in cancer immune checkpoint blockade: unravelling complexity. *Nat Rev Drug Discov* **16**, 264-272 (2017).
6. L. Chen, R. Liu, Z. P. Liu, M. Li, K. Aihara, Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Scientific reports* **2**, 342 (2012).
7. B. Yang *et al.*, Dynamic network biomarker indicates pulmonary metastasis at the tipping point of hepatocellular carcinoma. *Nat Commun* **9**, 678 (2018).
8. M. Mojtahedi *et al.*, Cell Fate Decision as High-Dimensional Critical State Transition. *PLoS Biol* **14**, e2000640 (2016).
9. V. Dakos *et al.*, Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. *PLoS One* **7**, e41010 (2012).
10. A. E. Teschendorff, M. Widschwendter, Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics* **28**, 1487-1494 (2012).
11. A. E. Teschendorff *et al.*, The dynamics of DNA methylation covariation patterns in carcinogenesis. *PLoS Comput Biol* **10**, e1003709 (2014).
12. F. Zhang *et al.*, Genome-wide dynamic network analysis reveals a critical transition state of flower development in Arabidopsis. *BMC Plant Biol* **19**, 11 (2019).
13. A. Richard *et al.*, Single-Cell-Based Analysis Highlights a Surge in Cell-to-Cell Molecular Variability Preceding Irreversible Commitment in a Differentiation Process. *PLoS Biol* **14**, e1002585 (2016).
14. M. Honda, S. Kaneko, H. Kawai, Y. Shirota, K. Kobayashi, Differential gene expression between chronic hepatitis B and C hepatic lesion. *Gastroenterology* **120**, 955-966 (2001).
15. P. Pons, M. Latapy, Computing communities in large networks using random walks. *Lect Notes Comput Sc* **3733**, 284-293 (2005).
16. M. Li, T. Zeng, R. Liu, L. Chen, Detecting tissue-specific early warning signals for complex diseases based on dynamical network biomarkers: study of type 2 diabetes by cross-tissue analysis. *Briefings in bioinformatics* **15**, 229-243 (2014).
17. M. E. Lenburg, A. Sinha, D. V. Faller, G. V. Denis, Tumor-specific and proliferation-specific gene expression typifies murine transgenic B cell lymphomagenesis. *J Biol Chem* **282**, 4803-4811 (2007).
18. R. Liu, P. Chen, K. Aihara, L. Chen, Identifying early-warning signals of critical transitions with strong noise by dynamical network markers. *Scientific reports* **5**, 17501 (2015).
19. S. Frank *et al.*, yylncT Defines a Class of Divergently Transcribed lncRNAs and Safeguards the T-mediated Mesodermal Commitment of Human PSCs. *Cell Stem Cell* **24**, 318-327 e318 (2019).
20. Q. Liu *et al.*, Disruption of mesoderm formation during cardiac differentiation due to developmental exposure to 13-cis-retinoic acid. *Scientific reports* **8**, (2018).
21. J. M. Maris, M. D. Hogarty, R. Bagatell, S. L. Cohn, Neuroblastoma. *Lancet* **369**, 2106-2120 (2007).

22. L. E. Ritenour, M. P. Randall, K. R. Bosse, S. J. Diskin, Genetic susceptibility to neuroblastoma: current knowledge and future directions. *Cell Tissue Res* **372**, 287-307 (2018).
23. E. Fredlund, M. Ringner, J. M. Maris, S. Pahlman, High Myc pathway activity and low stage of neuronal differentiation associate with poor outcome in neuroblastoma. *Proc Natl Acad Sci U S A* **105**, 14094-14099 (2008).
24. R. Zeid *et al.*, Enhancer invasion shapes MYCN-dependent transcriptional amplification in neuroblastoma. *Nat Genet* **50**, 515-523 (2018).
25. W. Zhang *et al.*, Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biol* **16**, 133 (2015).
26. X. Liu, Y. I. Li, J. K. Pritchard, Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* **177**, 1022-1034 e1026 (2019).
27. R. A. Costa, H. N. Seuanez, Investigation of major genetic alterations in neuroblastoma. *Mol Biol Rep* **45**, 287-295 (2018).
28. T. Mondal *et al.*, Sense-Antisense lncRNA Pair Encoded by Locus 6p22.3 Determines Neuroblastoma Susceptibility via the USP36-CHD7-SOX9 Regulatory Axis. *Cancer Cell* **33**, 417-434 e417 (2018).
29. K. Wang *et al.*, Integrative genomics identifies LMO1 as a neuroblastoma oncogene. *Nature* **469**, 216-220 (2011).
30. Y. He, P. Casaccia-Bonnefil, The Yin and Yang of YY1 in the nervous system. *J Neurochem* **106**, 1493-1502 (2008).
31. A. Kramer, J. Green, J. Pollard, Jr., S. Tugendreich, Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30**, 523-530 (2014).
32. A. D. Rouillard *et al.*, The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database (Oxford)* **2016**, (2016).
33. J. Wang *et al.*, Regulation of neural stem cell differentiation by transcription factors HNF4-1 and MAZ-1. *Mol Neurobiol* **47**, 228-240 (2013).
34. T. Ando, R. Kato, H. Honda, Differential variability and correlation of gene expression identifies key genes involved in neuronal differentiation. *BMC Syst Biol* **9**, 82 (2015).
35. H. Zaatiti *et al.*, Tumorigenic proteins upregulated in the MYCN-amplified IMR-32 human neuroblastoma cells promote proliferation and migration. *Int J Oncol* **52**, 787-803 (2018).
36. R. Munir, J. Lisec, J. V. Swinnen, N. Zaidi, Lipid metabolism in cancer cells under metabolic stress. *Br J Cancer* **120**, 1090-1098 (2019).
37. S. E. Kohe *et al.*, Metabolic profiling of the three neural derived embryonal pediatric tumors retinoblastoma, neuroblastoma and medulloblastoma, identifies distinct metabolic profiles. *Oncotarget* **9**, 11336-11351 (2018).
38. A. Chlenski *et al.*, Secreted protein acidic and rich in cysteine (SPARC) induces lipotoxicity in neuroblastoma by regulating transport of albumin complexed with fatty acids. *Oncotarget* **7**, 77696-77706 (2016).
39. C. Trefois, P. M. Antony, J. Goncalves, A. Skupin, R. Balling, Critical transitions in chronic disease: transferring concepts from ecology to systems medicine. *Curr Opin Biotechnol* **34**, 48-55 (2015).

40. D. L. Baker *et al.*, Outcome after reduced chemotherapy for intermediate-risk neuroblastoma. *N Engl J Med* **363**, 1313-1323 (2010).
41. A. Majdandzic *et al.*, Multiple tipping points and optimal repairing in interacting networks. *Nat Commun* **7**, 10850 (2016).
42. M. Scheffer *et al.*, Early-warning signals for critical transitions. *Nature* **461**, 53-59 (2009).
43. Y. B. Simons, K. Bullaughey, R. R. Hudson, G. Sella, A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS Biol* **16**, e2002985 (2018).
44. H. Chen, C. I. Wu, X. He, The genotype-phenotype relationships in the light of natural selection. *Mol Biol Evol*, (2017).
45. B. Lucas *et al.*, HNF4alpha reduces proliferation of kidney cells and affects genes deregulated in renal cell carcinoma. *Oncogene* **24**, 6418-6431 (2005).
46. H. Chiba *et al.*, Activation of p21CIP1/WAF1 gene expression and inhibition of cell proliferation by overexpression of hepatocyte nuclear factor-4alpha. *Exp Cell Res* **302**, 11-21 (2005).
47. T. Tanaka *et al.*, Dysregulated expression of P1 and P2 promoter-driven hepatocyte nuclear factor-4alpha in the pathogenesis of human cancer. *J Pathol* **208**, 662-672 (2006).
48. Q. Sun *et al.*, Role of hepatocyte nuclear factor 4 alpha in cell proliferation and gemcitabine resistance in pancreatic adenocarcinoma. *Cancer Cell Int* **19**, 49 (2019).
49. L. Wei *et al.*, Oroxylin A activates PKM1/HNF4 alpha to induce hepatoma differentiation and block cancer progression. *Cell Death Dis* **8**, e2944 (2017).
50. S. S. Marable, E. Chung, M. Adam, S. S. Potter, J. S. Park, Hnf4a deletion in the mouse kidney phenocopies Fanconi renotubular syndrome. *JCI Insight* **3**, (2018).
51. M. Qu, T. Duffy, T. Hirota, S. A. Kay, Nuclear receptor HNF4A transrepresses CLOCK:BMAL1 and modulates tissue-specific circadian networks. *Proc Natl Acad Sci U S A* **115**, E12305-E12312 (2018).
52. M. E. Torres-Padilla, F. M. Sladek, M. C. Weiss, Developmentally regulated N-terminal variants of the nuclear receptor hepatocyte nuclear factor 4alpha mediate multiple interactions through coactivator and corepressor-histone deacetylase complexes. *J Biol Chem* **277**, 44677-44687 (2002).
53. Y. Maeda *et al.*, Tumour suppressor p53 down-regulates the expression of the human hepatocyte nuclear factor 4alpha (HNF4alpha) gene. *Biochem J* **400**, 303-313 (2006).
54. Y. Maeda, S. D. Seidel, G. Wei, X. Liu, F. M. Sladek, Repression of hepatocyte nuclear factor 4alpha tumor suppressor p53: involvement of the ligand-binding domain and histone deacetylase activity. *Mol Endocrinol* **16**, 402-410 (2002).
55. O. Witt, H. E. Deubzer, M. Lodrini, T. Milde, I. Oehme, Targeting histone deacetylases in neuroblastoma. *Curr Pharm Des* **15**, 436-447 (2009).
56. M. Grechkin, B. A. Logsdon, A. J. Gentles, S. I. Lee, Identifying Network Perturbation in Cancer. *PLoS Comput Biol* **12**, e1004888 (2016).

57. S. Mukherjee, A. Carignano, G. Seelig, S. I. Lee, Identifying progressive gene network perturbation from single-cell RNA-seq data. *Conf Proc IEEE Eng Med Biol Soc* **2018**, 5034-5040 (2018).
58. M. Yi *et al.*, Emerging role of lipid metabolism alterations in Cancer stem cells. *J Exp Clin Cancer Res* **37**, 118 (2018).
59. K. K. Matthay *et al.*, Neuroblastoma. *Nat Rev Dis Primers* **2**, 16078 (2016).
60. T. Derrien *et al.*, The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research* **22**, 1775-1789 (2012).
61. M. Wainberg *et al.*, Opportunities and challenges for transcriptome-wide association studies. *Nat Genet* **51**, 592-599 (2019).
62. Z. Wang, J. M. Cunningham, X. H. Yang, CisPi: a transcriptomic score for disclosing cis-acting disease-associated lincRNAs. *Bioinformatics* **34**, i664-i670 (2018).
63. J. R. Dixon *et al.*, Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012).
64. R. Jager *et al.*, Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat Commun* **6**, 6178 (2015).
65. H. M. Amemiya, A. Kundaje, A. P. Boyle, The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific reports* **9**, 9354 (2019).
66. A. Liberzon, A Description of the Molecular Signatures Database (MSigDB) Web Site. *Methods in molecular biology* **1150**, 153-160 (2014).

Supplementary Figures

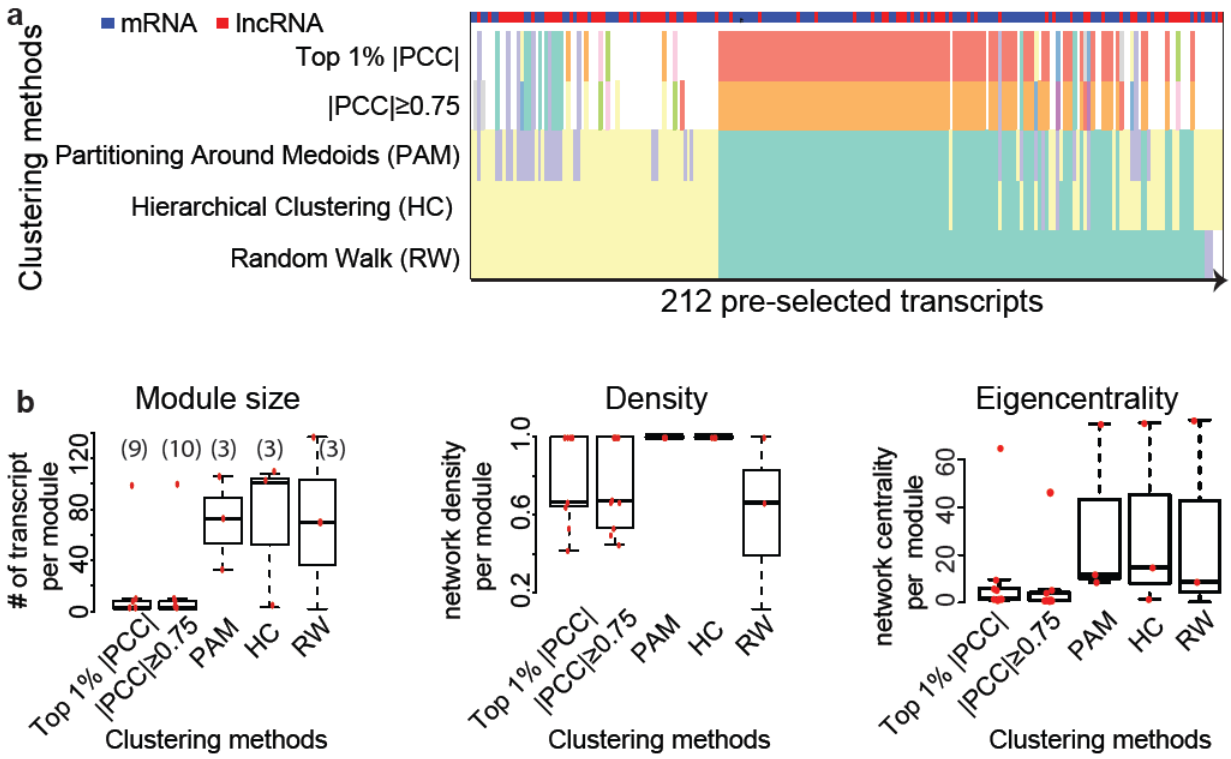


Figure S1. Random-walk and four transcriptional network partitioning methods in the existing CTS-scoring systems.

a, Heatmap showing transcripts representing different subnetworks (modules, denoted in colors), using five different partitioning methods. Shown are 212 transcripts preselected in the gene expression variance analysis of neuroblastoma profiles (GSE49711, Event-free subpopulation, Low-or-Intermediate-risk state). The top red-blue bar indicates whether a transcript represents lncRNA (red) or mRNA (blue).

b, Boxplots comparing isolated subnetworks derived from different approaches in three features: module size, density, and eigencentrality. Left: The number of isolated modules by each method is given in parentheses at the top. Middle: The edge density of a subnetwork is calculated from the proportion of present edges over the number of all possible edges in the subnetwork using the R-function *edge_density* in the *igraph* package. Note that PAM and HC received the highest density value of 1 because they considered all possible gene-gene correlations without filtering out insignificant edges. Right: To measure the influence of network nodes in a module, the average vertex centrality of each module is calculated using the R-function *eigen_centrality*.

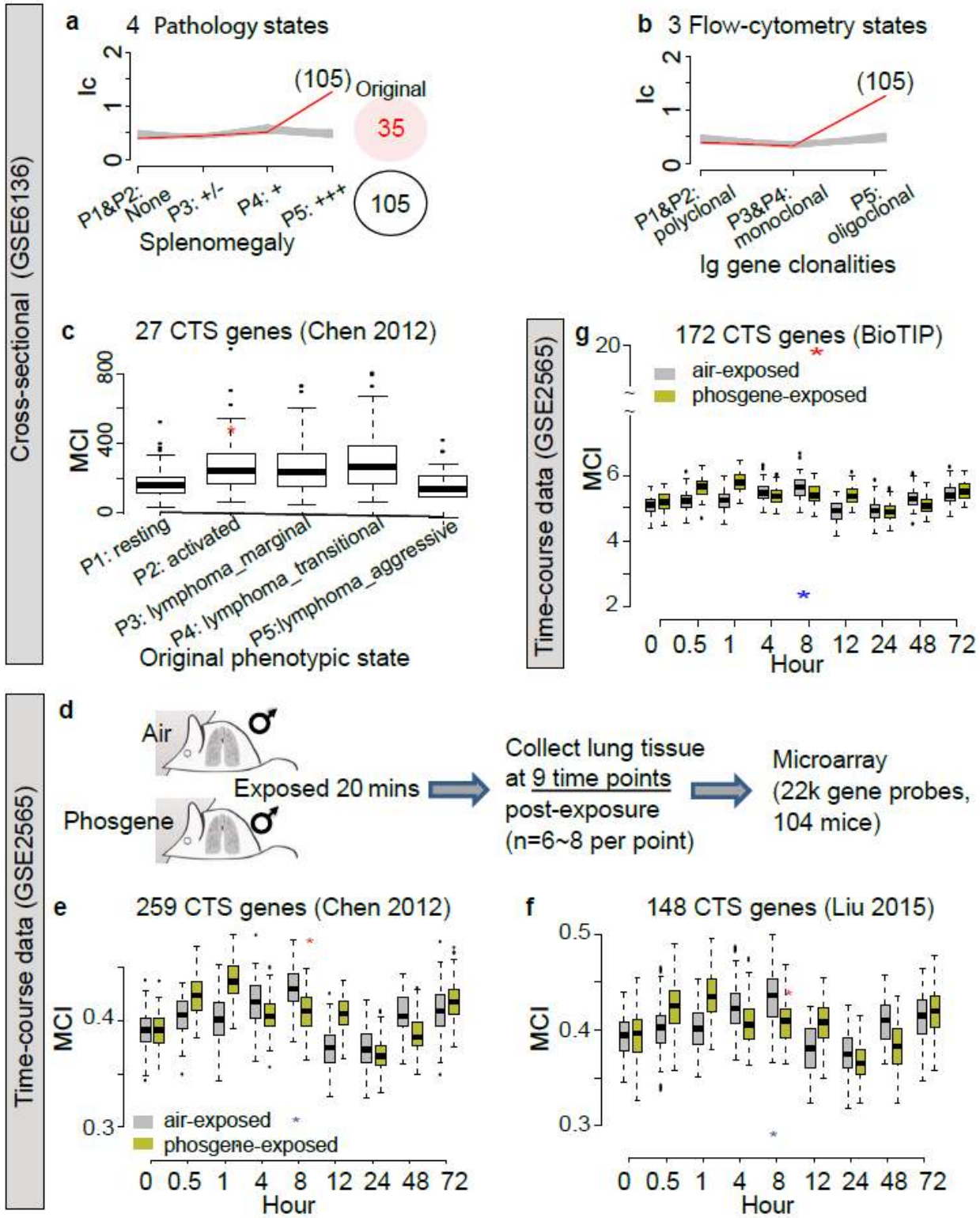


Figure S2. Applying BioTIP to available benchmark gene expression datasets.
a-b, BioTIP identifications given 1% probe selection and a simplified 4-state (a) or 3-state (b) sample-definition. Both indicate a tipping point at P5 with the same set of 105 identified CTS probes. In each panel, plot shows index of critical transition (Ic, y-axis)

against state (x-axis). Red line indicates score of observation against random genes (grey). No overlap of this misclassified identification with the 35 CTS probes identified based on the original 5-state definition (Fig 2a).

c, Plot showing module-criticality index (y-axis, MCI) against state (x-axis). Red dot indicates MCI of 27 published CTS genes, previously identified by the MCI-scoring system, against random genes (black box). Although P2 was identified, this signature lacks empirical significance because it falls below the third quartile of random scores.

d, Scheme of the experiment design of a publicly available time-course dataset (GSE2565). This dataset is unique because it measured the time-series of both case and control in parallel.

e-f, Plot showing MCI of two sets of previously published CTS genes (y-axis) against state (x-axis). The observed MCI scores (red dot) is compared to its size-controlled random genes (yellow box) in phosgene-exposed; or in control mice (blue dot *vs*. grey box) per state. Both CTSs were published for a tipping-point identification at 8 hours. However, one observed score (right, red) falls into the range of random scores at 8 hour, and both are lower than the maxim of random scores at other time points.

g, Plot showing MCI (y-axis) of BioTIP-identified CTS genes against state (x-axis). The observed MCI-score (red dot) indicates a transitional state at 8 hour with empirical significance not only locally at 8 hour but also globally higher than the maximum random score of any time-points. In contrast, the MCI score calculated for control mice (blue dot) is the lowest.

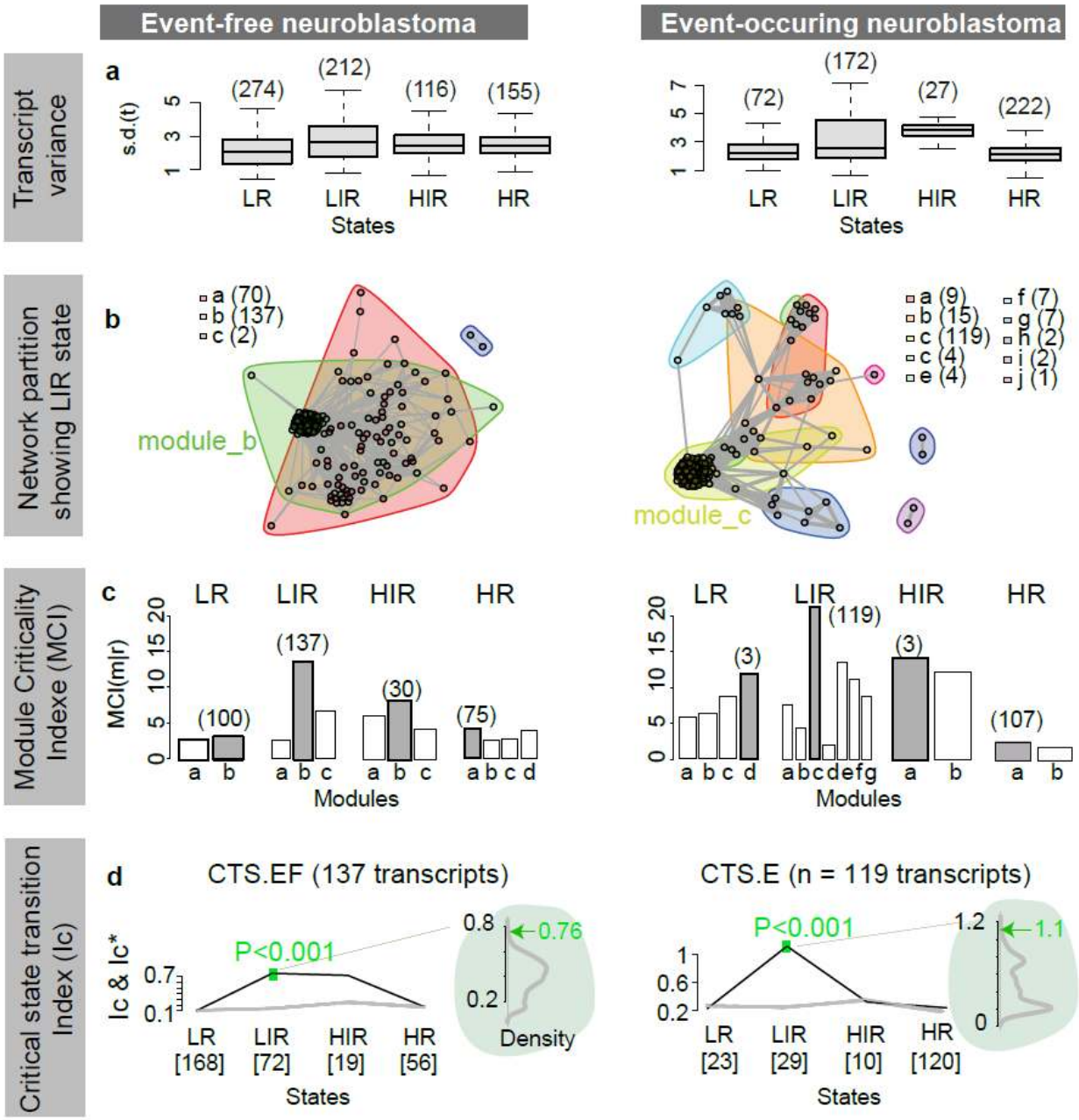


Figure S3. Applying BioTIP to the gene expression profiles of two neuroblastoma populations: (Event-free, EF) and Event-occurring (E)

a, Boxplot showing the standard deviation (s.d) of preselected transcripts, with transcript numbers in parenthesis on the top. In both populations, the LIR state presents the highest standard deviation.

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

c, Bar plots illustrating the MCI scores (y-axis) calculated for each module per state (x-axis). For each state, the dominating cluster (biomodule) is denoted in grey with module size in parentheses. In both populations, the MCI score peaked at the LIR state.
d, Line plot compares the observed Ic-score (black) to 100 random scores (grey). The observed score was calculated for the LIR-state specific biomodule (herein the 'CTS') at every four states to be compared; the random scores (Ic*) were calculated in the same way used randomly-chosen transcripts (listed in parentheses) from the 48.5k-transcript background. In both populations, the observed Ic score peaked in the LIR state. Additionally, density histogram compares each observed Ic-score to the random results obtained from 1000 times of sample bootstraps.

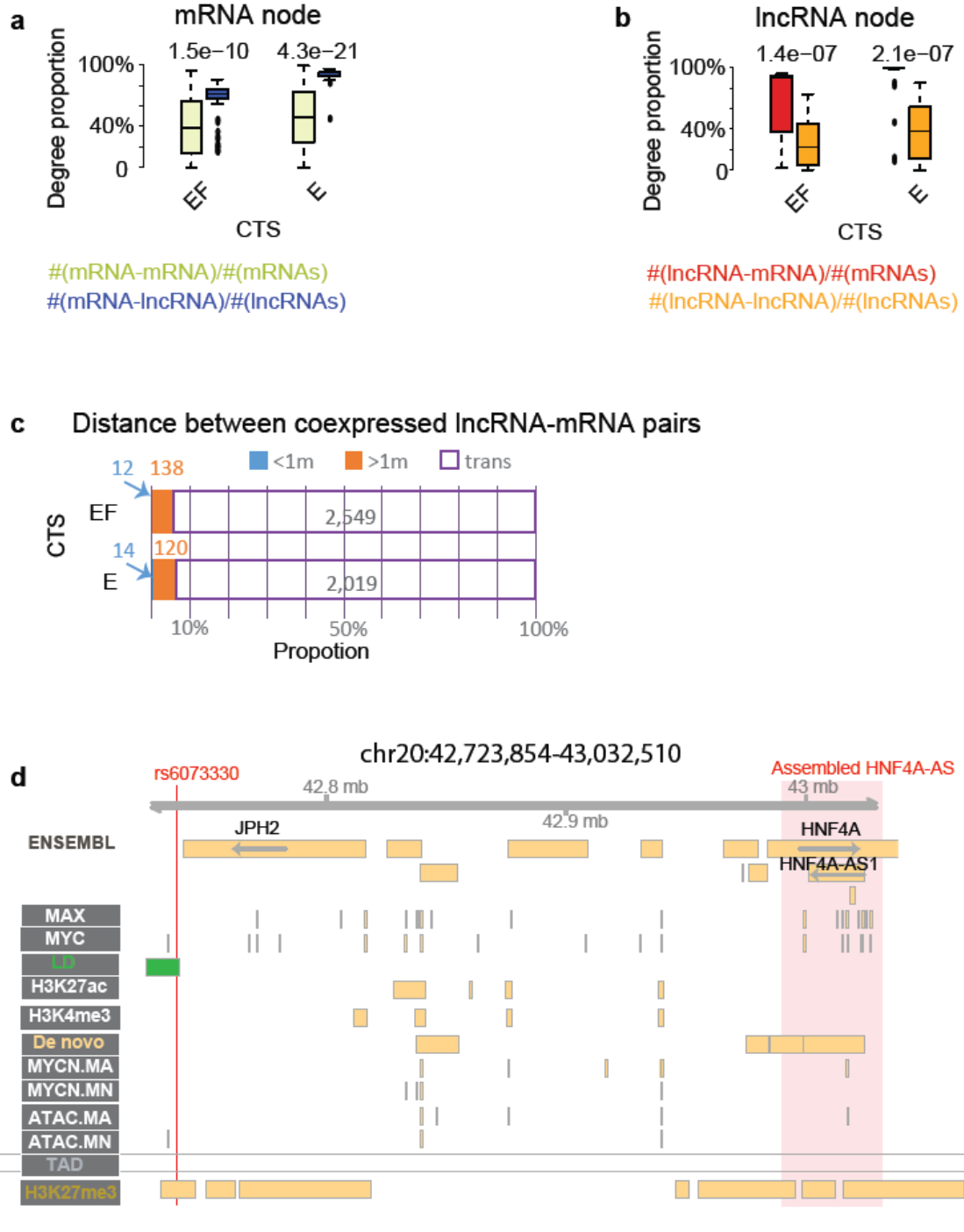


Figure S4. Genomic features of the identified noncoding CTS transcripts.
a, Boxplot comparing network-degree proportion in the two identified CTS from neuroblastoma subpopulations, event-free (EF) and event-occurring (E). In each network, nodes are CTS transcripts and links are the significant correlation (PCC, FDR<0.05). Comparison is between the proportion of only mRNAs (yellow) and the proportion of potential ci-regulation (i.e., the number of mRNA-lncRNA pairs over the

number of CTS mRNAs, blue). The P-value is calculated using the Wilcoxon rank-sum test.

b, Boxplot comparing network-degree proportion in the two identified CTS from neuroblastoma subpopulations, EF and E respectively. Comparison is between the proportion of only lncRNAs (i.e., the number of mRNA-mRNA pairs over the number of CTS mRNAs, red) and the proportion of potential ci-regulation (i.e., the number of mRNA-lncRNA pairs over the number of CTS lncRNAs, orange). The P-value is calculated using the Wilcoxon rank-sum test.

c, Stacked bars comparing all co-expressed (FDR<0.05) lncRNA-mRNA pairs for two sets of CTSs whose edge-to-edge distances fall within a certain nucleotide range of each other (denoted in colors). The cross-chromosomal pairs take the largest fraction in both CTSs.

d, Genomic view of the CTS transcript *HNF4A-AS* identified (pink box) overlapping with the ENSEMBL-annotation *HNF4A-AS1*. The neuroblastoma-suggestive SNP (red vertical line) is 250k nt away from *HNF4A-AS*. The occupancy of repressive histone marker H3K27me and lack of enhancer marks at both loci suggests their inactivity in neuroblastoma cell lines. A shared cis-regulatory region TAD (topological associating domain, transparent bar) was obtained from the YUE lab at http://promoter.bx.psu.edu/hi-c/.

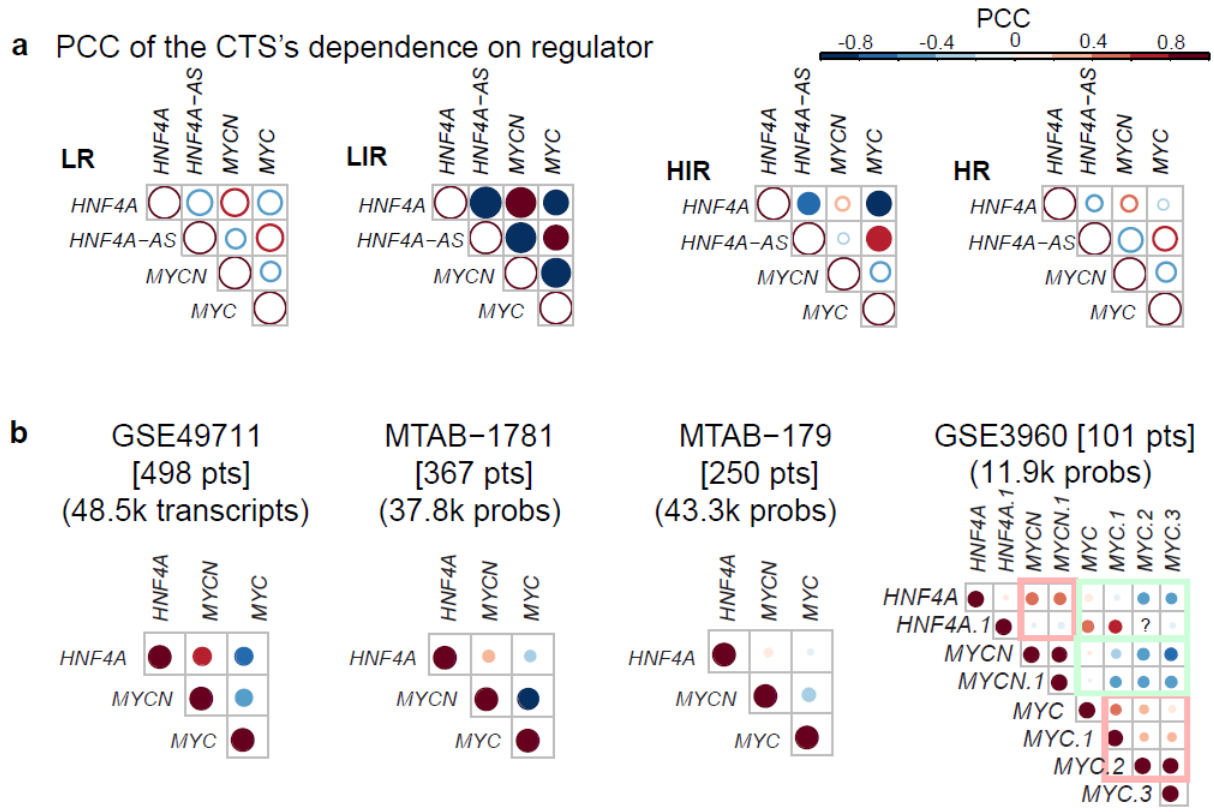


Figure S5. CTS transcripts disclosing underappreciated HNF4α and its potential collaboration with MYC proteins N-myc and c-myc in NB progression.

a, For each state, visualization of the pair-wise transcription co-regulation among five regulators after pooling the EF and E subpopulations to enlarge the sample size for a Pearson Correlation Coefficient (PCC) calculation. PCC was calculated for pairs of log2 fold changes by comparing the transcription of the tumors within the upper-quarter of a gene *(e.g. HNF4A)* to the tumors within the lower-quarter levels. Circle size and color denote the strength of the correlation, with larger, more saturated colors signifying stronger correlations versus subdued, smaller circles. The filled circle indicates an empirical significance ($P < 0.05$, 100 bootstraps of transcripts).

b, Visualization of the transcription co-regulation among all expressed transcripts for all patients in four independent datasets, set to display only significant correlations (Pearson correlation p-value < 0.05). Circle size and color denote the strength of the correlation, with larger, more saturated colors signifying stronger correlations versus subdued, smaller circles. A question mark denotes insignificant PCC. When multiple probes measuring the same gene, each probe was independently calculated and shown in a colored square.