# Transcriptome dynamics describe and predict state transition from health to leukemia

Russell C. Rockne[1]†*, Sergio Branciamore[2]†, Jing Qi[3]†, Guerry J. Cook[3], Wei-Kai Hua[3], Emily Carnahan[3], Ayelet Marom[3], Herman Wu[3], Davide Maestrini[1], Xiwei Wu[4], Chao Guo[4], Denis O'Meally[5], Yate-Ching Yuan[6], Zheng Liu[6], Nadia Carlesso[3], Leo D. Wang[7,8], Stephen J. Forman[3], Ya-Huei Kuo[3]††*, Guido Marcucci[3]††*

**Affiliations:**

[1]Department of Information Sciences, Division of Mathematical Oncology

[2]Department of Diabetes Complications & Metabolism

[3]Department of Hematology & Hematopoietic Cell Transplantation and the Gehr Family Center for Leukemia Research

[4]Department of Molecular and Cellular Biology; Integrative Genomics Core

[5]Center for Gene Therapy

[6]Department of Molecular Medicine; Bioinformatics Core

[7]Department of Immuno-Oncology

[8]Department of Pediatrics

Beckman Research Institute, City of Hope Medical Center

†Co-authors contributed equally.

††Co-senior authors.

*Co-corresponding authors

**Summary**

Time series gene expression (transcriptome) data provide information on global changes in expression patterns occurring through the course of cancer progression. The premise of this work is that cancer can be viewed as a state transition of the transcriptome, and that the transcriptome behaves as a particle in a force field obeying basic physical principles of motion. The implication of this concept is that cancer progression may be understood, and predicted, with mathematical models of motion of the transcriptome in a low-dimensional projection that can be constructed to retain a maximal amount of relevant information in the system. We use a genetic mouse model of acute myeloid leukemia (AML) to demonstrate the concepts of our mathematical model. We show that the transition of the transcriptome from a health state to a leukemia state can be understood in terms of mathematically-derived inflection points which characterize the dynamic probability of leukemia development.

**Keywords:** nonlinear dynamics; acute myeloid leukemia; state transition; dynamical systems; mathematical oncology; cancer theory; stochastic processes; critical transitions; time series; mathematical modeling; mathematical biology;

**Introduction**

Cancer evolves as a complex dynamic system wherein multilayer interconnected inputs, including genetic and epigenetic alterations, clonal expansion, evolutionary selection, and microenvironmental plasticity, collectively combine to produce a disease state corresponding to a clinical phenotype. Identification of gene mutations, epigenetic changes and gene expression profiles produced by high-throughput sequencing assays are becoming a part of the routine clinical assessment of cancer patients at diagnosis and subsequent time-points. While current

theories of cancer evolution conceptually consider changes in the genome as a whole, the predominant focus remains on a handful of specific oncogenic mutations and clonal dynamics (Graham and Sottoriva, 2016; Landau et al., 2014; Michor et al., 2005; Zhao et al., 2016). However, oncogenic mutations are often not mutually exclusive, and associate in a non-random fashion, thereby making it almost impossible to dissect their individual contributions to clinical phenotypes and prioritize them as therapeutic targets. Nevertheless, these combinatorial sets of mutations are often characterized by specific gene expression signatures representing aberrantly activated or inactivated homeostatic pathways that are ultimately responsible for the observed biological and clinical phenotypes of disease (Almassalha et al., 2016). Moreover, epigenetic alterations, clonal dynamics, and microenvironmental inputs are convergently manifested at the level of gene expression. Thus, global changes in gene expression signatures seem to represent ideal genomic biomarker for a complete assessment of disease evolution.

From a clinical perspective of predicting cancer progression, the most pragmatic use of genomic data is to identify subsets of genes or transcripts (a “signature” or “panel”) that are associated with predetermined biological or clinical features (Nabhan et al., 2015; Ng et al., 2016). Unfortunately, the predictive value of these signatures drops off precipitously if all elements of the signature are not apparent, nor do they provide a prediction of how the signatures might change over time (Jo et al., 2014; Sanavia et al., 2015). The exponential number of variations that could be observed from simple changes in a small number of genes in a signature points to a central challenge facing the precision medicine effort, which aims to target individual or sets of genes (Hyman et al., 2017; Lipinski et al., 2016; Tannock and Hickman, 2016). This leads us to agree with the contention that the era of “big data” is in need of “big theory” to guide interpretations of high-dimensional data sets, whose structures and intricate biological meanings are exceedingly complex and change dynamically over time (Coveney et al., 2016). We therefore desire a more holistic, conceptual, and mathematical

approach to study the smooth, continual change in configurations of the system that does not suffer from a combinatorial explosion of complexity, while retaining the maximum amount of information in the system. The central hypothesis of our theory is that cancer can be viewed as a state transition of the transcriptome, and that the transcriptome behaves as a particle in a force field obeying basic physical principles of motion.

State transition models have been used to characterize cellular differentiation states (Furchtgott et al., 2017; Hormoz et al., 2016; Li et al., 2016; Mojtahedi et al., 2016; Moris et al., 2016; Rizvi et al., 2017; Zhou et al., 2012), to characterize phenotypes of cells and gene regulatory networks (Wooten and Quaranta, 2017), or to define fitness landscapes in evolutionary models of therapeutic resistance (Nichol et al., 2016). Dynamical systems approaches have been used to study changes in gene expression (Kuwahara and Schwartz, 2012; Mackey et al., 2015), however, these and similar systems biology approaches, are focused on the scale of individual genes and regulatory networks (Zheng et al., 2012). Statistical tools have been developed to analyze time series transcriptomic data (Jo et al., 2014; Spies and Ciaudo, 2015), but these approaches are not predictive theories that hypothesize how the system should- or should-not evolve. To our knowledge, none of these or similar dynamical systems approaches has been applied to define state transition dynamics at a genome wide scale. Here we argue, and demonstrate, that complex processes occurring at the scale of individual or sets of genes, are manifested and represented by simple macroscopic dynamics that may be predictable.

Instead of tracking hundreds to tens of thousands of changes in specific sets of genes, we consider the state of the genome as a configuration and study the smooth continual change of gene expression (i.e., transcriptome) as a whole over the course of cancer progression. Here we propose a mathematical model, and more generally a theory, of cancer as a state transition of the transcriptome. We present this theory and model as a method to describe and predict the

evolution of cancer with a long-term goal of application in clinical settings where whole genome gene expression data over time is available. We use acute myeloid leukemia (AML) (Dohner et al., 2015) as a model system for this endeavor because of the close resemblance of mouse models to human disease and the easy access to blood samples *in vivo* over time. Herein we report on the dynamics of state transition from health to leukemia in a 2-dimensional "state-space" constructed with time series gene expression profiles derived from peripheral blood mononuclear cells (PBMC) from a genetic mouse model of AML. Our mathematical model provides a tool to interpret and predict changes in gene expression over time as well as introduces the novel concept of an leukemogenic (oncogenic) force which determines the dynamics of state transition from health to leukemia.

## Methods

### Inv(16) driven acute myeloid leukemia

**Acute myeloid leukemia (AML)** is a genetically and epigenetically heterogeneous malignant disease of the bone marrow (BM) comprising multiple entities characterized by gene mutations and chromosomal abnormalities. Recurrent chromosomal abnormalities in AML frequently and often result in leukemogenic fusion genes encoding chimeric proteins that contribute to the unique etiology and prognosis of distinct subsets of AML. One of the most common recurrent cytogenetic aberrations found in approximately 5-12% of AML patients is inv(16)(p13.1q22) or t(16;16)(p13.1;q22) [henceforth inv(16)]. At the molecular level, inv(16) disrupts the core-binding factor (CBF) transcription complex fusing *CBFB* with the *MYH11* gene and creates a leukemogenic fusion gene *CBFB-MYH11* (CM), which encodes a leukemogenic chimeric protein. We have established a *CM* knock-in mouse model (*Cbfb*$^{+/56M}$*/Mx1-Cre)*, in which conditional expression of *CM* fusion transcript drives AML development (Cai et al., 2016; Kuo et al., 2006).

To test our state-space mathematical model in which changes of the transcriptome can be modeled over time, we conducted a longitudinal study of conditional *CM* knock-in mice (*Cbfb*$^{+/56M}$*/Mx1-Cre;* C57BL/6; n=9) and littermate controls lacking the transgene (n=7). All mice were injected with a synthetic double-stranded RNA polyinosinic–polycytidylic acid [poly (I:C)]. In *CM* mice, poly(I:C) induced Cre-mediated recombination and in turn *CM* expression. Peripheral blood (PB) samples were collected before induction ("0" time point) and at 1, 2, 3, 4.5, 5.5 and 6.5 months after induction, or when *CM* mice became moribund (Fig. 2a). By 6.5 months ("end" time point), six *CM* mice developed leukemia, defined by the presence of 20% or more leukemic blasts in the PB and BM ("*CM* sick") while three *CM* mice had not ("*CM* not sick") but showed expansion of phenotypic hematopoietic stem and progenitor cell subpopulations in the bone marrow (BM) (Fig. S1), consistent with previous descriptions of *CM* expression in this mouse strain (Cai et al., 2016). PBMC whole genome RNA-Seq was performed in both *CM* and control mice. In *CM* mice, the *CM* fusion transcript levels were measured based on human *MYH11* sequencing reads originating from the *Cbfb*$^{+/56M}$ allele and reported in counts per million reads (cpm) at each sequential time point (Fig. 2b).

We constructed a matrix $\boldsymbol{X}$, consisting of all time series gene expression data in units TPM (see supplementary methods) with dimension $[M \times N]$ where *M* is the number of genes, and $N = n\tau$ is the number of mice ($n$) multiplied by the number of observations ($\tau$) across all mice (Fig. 2c). To simplify the complexity of the state-space and the state transition analysis, we elected to project the high-dimensional space of gene expression onto a 2-dimensional (2D) plane, which we defined to be the health-leukemia state-space, by factoring the data matrix **X** with the singular value decomposition (SVD) (Fig. 2d). The SVD is a matrix factorization method previously used to decompose gene expression data (Aiello and Alter, 2016; Ponnapalli et al., 2011). We selected the first two components of the SVD to define a 2D state-space, consisting of two basis vectors $v_0, v_1$ in the canonical Cartesian coordinate system which we label $(s, s_1)$,

so that each point in the state-space is a linear combination of all genes in the transcriptome. Of note, the 2D state-space constructed with gene expression data using the SVD approach results in a coordinate space without physical units.

**Mathematical model of gene expression state transition**

An essential concept of our state transition theory is that of a "state-space:" a low dimensional projection of the transcriptomic space, where states are defined geometrically and energetically by potential energy functions. To develop a mathematical model of cancer–and here we show for AML–as a state transition of the transcriptome, we postulate that the macroscopic dynamics of the transcriptome may be represented as a particle with unit mass moving in a potential energy field in a transcriptomic state-space. In healthy conditions, we suggest the transcriptome behaves as a stochastically varying homeostatic process where relatively large fluctuations in gene expression are prevented by feedback mechanisms. Thus, we represent the net force acting on a healthy transcriptome as a gradient of a homeostatic potential ($U_H$), $\vec{F} = -\nabla U_H + \sigma\vec{\eta}(t)$, where $\sigma\vec{\eta}(t)$ is a Brownian stochastic process that is uncorrelated in time $< \vec{\eta}(t), \vec{\eta}(t_i) > = \delta(t - t_i)$ with zero mean and variance $\sigma > 0$ (Fig. 1A). The Brownian process represents stochastic fluctuations in gene expression and the internal energy of the system to move in the space. Homeostasis is modeled as an Ornstein-Uhlenbeck process (Rohlfs et al., 2014; Uhlenbeck and Ornstein, 1930) which is a stochastic process that reverts to a mean state located at $M = [\mu_0, \mu_1]$ given by $\vec{F} = -(\Theta(S - M)) + \sigma\vec{\eta}(t)$ where in general, $S$ is an ordered n-tuple in Cartesian coordinates, with n=2 we have $S = (s_0, s_1)$, with homeostatic reversion rate constants given by the elements in the parameter matrix $\Theta = [\theta_{0,0}\ \theta_{0,1};\ \theta_{1,0}\ \theta_{1,1}]$.

Next, we hypothesize that the expression of an oncogene (here we will use the AML transgene *CM* as an example) can create a singularity in the health potential and generate a

"oncogenic" force that drives the system towards the singularity (Fig. 1B). Thus, we hypothesize the net force exerted on the particle (i.e. transcriptome) upon expression of *CM* would consist of a normal homeostatic health potential $U_H$, and a leukemic potential $U_L$ as follows, $\vec{F} = -\nabla(U_H + U_L) + \sigma\vec{\eta}(t)$ with $\nabla U_L$ and $\nabla U_H$ given by

$$\vec{F} = -(\overbrace{\Theta(S-M)}^{\substack{mean-reverting\\ homeostasis}} + \overbrace{A\,S^{-N}}^{\substack{singularity\ created\\ by\ CM\ transgene}}) + \overbrace{\sigma\vec{\eta}(t)}^{\substack{internal\ energy/\\ biological\ noise}}, \text{ with } S > 0. \quad (1)$$

Here $A = [\alpha_{0,0}\ \alpha_{0,1};\ \alpha_{1,0}\ \alpha_{1,1}]$ is a 2×2 matrix of parameters which determine the rate of transition to leukemia and *N* is a parameter that determines the magnitude of the *CM* leukemogenic force. Due to the stochastic component of the net force, the creation of the singularity and presence of the leukemogenic force does not guarantee or predetermine the transition to a leukemic state. We postulate that the equation of motion in the potential energy space is given by the Langevin equation $\vec{F} = \lambda\frac{d\boldsymbol{S}}{dt}$ where $\lambda$ is the damping constant in an over-damped motion on the potential and may be taken to be unity (Bezuglyy et al., 2006; Zhou et al., 2012).

**Structure and construction of a transcriptome state-space**

A key concept of our theory is that the transcriptome contains essential information necessary to describe the dynamical process of leukemia development. However, given the high dimensional nature of the transcriptome (order of $10^6$) and the fact that not all the information present in the transcriptome is equally important, we desire a projection of the transcriptome information into a lower dimensional space that can preserve a maximum amount of information in the system relevant to the process of interest. Ideally, in a 2-dimensional representation, the leukemic and non-leukemic components should also be orthogonal, and therefore their contributions to the overall dynamics should be separable. The orthogonality

property will be particularly useful in the modeling of the two processes (i.e., health and leukemia) independently.

We show that it is possible using information content measured by mutual information with respect to state variables, to construct an orthogonalized 2-dimensional state-space coordinate system ($S = (s_0, s_1)$) that maximizes the amount of information relevant to leukemia development along the ordinate axis (Fig. S2). Therefore, we contend the dynamics of state transition from health to leukemia in the orthogonalized 2-dimensional state-space are determined by stationary points of the equation of motion in one dimension along the leukemia ($s_1$) axis. In particular, the dynamics in the ordinate and absicca directions of the state-space are sufficiently decoupled so that we may take off-diagonal elements in the matrices $\Theta$ and $A$ to be zero, so that all entries in the matrix $A$ are zero except $\alpha_{1,1} > 0$ and the singularity is located at $(\mu_0, 0)$. This simplification may be relaxed to study the degree of coupling between the homeostatic and leukemogenic dynamics.

**Critical points in the dynamics of state transition in the state-space**

The critical points are positions in the state-space where the velocity of movement is zero $\frac{ds_1}{dt} = 0$ and are calculated by considering the deterministic terms in the equation of motion in one dimension along the leukemia axis, $\frac{dS_1}{dt} = -\theta_{1,1}(s_1 - \mu_1) - \alpha_{1,1}s_1^{-N}$. There can be zero, one, or two critical points in the system for $s_1 > 0$ (see Supplementary methods). In the case of two critical points, the critical point $c_1$ is an attractor point and represents the perturbation of the transcriptome away from the health equilibrium at $\mu_1$. The movement of point $c_1$ away from $\mu_1$ is due to a leukemogenic force directed towards the singularity at zero. The critical point $c_2$ is an unstable critical point and represents a transition point in the dynamics from health to leukemia. When the particle in the state-space crosses the critical point $c_2$, the velocity of the particle

increases unboundedly, and asymptotically approaches the singularity (Fig. 1C). These dynamics can be interpreted **biologically** as an acceleration of leukemia development after the threshold $c_2$ is crossed. The boundary of the health and leukemia states in the state-space is located where the magnitude of the leukemogenic potential equals the magnitude of the health potential, $U_L = U_H$ denoted $c_3$ which lies beyond the transition point (Fig. 1D). The parameters $(\alpha_{1,1}, N)$ can be calculated based solely on estimates of the critical points $c_1$, and $c_2$ in the state-space (see Supplementary methods). The values of the parameters $\theta_{1,1}$, $\mu_1$ may be taken to be unity without a change in the qualitative dynamics of the system.

## Results

### *CM* leukemogenic force correlates with *CM* expression

Because *CM* fusion transcript is a principle biological determinant in driving AML development, we examined the correlation between the expression level of CM ([*CM*]) and the force exerted on the particle in the state-space by evaluating the force function with the parameters estimated from the data (see Supplementary methods). At each point in the state-space, a force vector is defined by Equation (1), which represents the deterministic dynamics of the movement of the particle representing the transcriptome in the state-space at that position (Fig. 2F,G). At an arbitrary position in the state-space $S^0 = (s_0^0, s_1^0)$, the magnitude of the leukemic force is computed along the leukemic axis as $|\overrightarrow{F_L}| = |-\nabla U_L(s_1^0)|$ and the magnitude of the health force is given by $|\overrightarrow{F_H}| = |-\nabla U_H(s_1^0)|$ (Fig. 3A). Induced CM mice in general develop leukemia between 3 to 6 months within the observation time of our experiment (7 months); we defined these mice as "*CM* sick". A subset of mice may either take longer time or never develop leukemia; we defined these mice as "*CM* not sick". The median leukemogenic force in the *CM* sick mice was an order of magnitude higher than that in the *CM* not sick (0.20 vs. 0.08, $p < 0.001$), with the range of values extending several orders of magnitude higher in *CM* sick mice

than *CM* not sick mice (Fig. 3B). The magnitude of the leukemogenic force was able to discriminate *CM* sick from *CM* not sick mice as early as 1 month post induction; however the change in [CM] was not able to distinguish the two groups (Fig. S4A,B). The leukemogenic force and [CM] were not correlated at 1 month post induction (Fig. S4C), however the differences in the expression level of [*CM*] become more pronounced between the *CM* sick and *CM* not sick mice over time (Fig. S4D). We observed a strong correlation between [*CM*] and the leukemia force exerted on the transcriptome particle as it moved in the state-space, with a stronger correlation for the *CM* sick mice ($r^2$=0.81, p<0.001) than the *CM* not sick mice ($r^2$=0.21, p=0.117) and a weaker correlation between [*CM*] and the healthy force (*CM* sick $r^2$=0.60; *CM* not sick $r^2$=0.35; Fig. 3C).

**Energy states**

As a novel measure of the state of the transcriptome and its association with either a health or leukemia state, we calculated the potential energy in the state-space for the control and *CM* mice along the leukemic axis. At an arbitrary position in the space, $S^0 = (s_0^0, s_1^0)$, the potential energy for control mice (denoted $E^{Ctl}$) is given by evaluating the health potential at that position, $E^{Ctl}(S^0) = U_H(s_1^0)$. For the *CM* mice, the potential energy (denoted $E^{CM}$) at the position $S^0$ is given by evaluating the sum of the health and leukemic potentials, $E^{CM}(S^0) = U_H(s_1^0) + U_L(s_1^0)$. The potential energy at the boundary of the leukemia (denoted $E_L$) and the health states ($c_3$) is given by $E_L = E^{CM}(c_3) \approx -0.7021$. The shift in potential energy at $c_1$ caused by *CM* expression is $E^{CM}(c_1) \approx -0.5178$ vs. that of health equilibrium $E^{Ctl}(\mu_1) = -0.5$. Accordingly, the change in potential energy associated with the leukemogenic alteration is immediately seen at 1 month post induction, with a mean energy of *CM* sick mice of -0.517±4.6 x$10^{-4}$, a mean energy of control mice of -0.4999±-9.4x$10^{-5}$ (*CM* sick vs. control p<0.01), and a mean energy of *CM* not sick mice of -0.516±-5.5x$10^{-4}$ (*CM* sick vs. *CM* not sick p=0.55). The differences in potential

energy between the *CM* sick, *CM* not sick, and control mice becomes more pronounced as they transition from health to leukemia states (Fig. 3D).

**Dynamics of leukemia transition in the state-space**

The mathematical model predicts the effects of *CM* induction to be immediately observed as a shift in the stable equilibrium $\mu_1$ (translated so that $\mu_1 = 1$) to the critical point $c_1$ along the leukemia axis as a result of the singularity in the transcriptomic space and the leukemogenic force introduced by the *CM* fusion gene (Fig. 3E). As early as one month post induction, the *CM* sick mice are observed to exhibit a pronounced shift downward in their position in the state-space with *CM* sick mice having a mean position ($s_1$) of 0.83±0.04, while the *CM* not sick mice had a mean position of 0.96±0.02 ($p<0.01$) and controls 0.99±0.01 (*CM* sick vs. control, $p<0.01$). To use the model to predict the dynamics of transition from health to leukemia, 20,000 Monte Carlo simulations were initialized with the initial (t = 0) positions in the state-space taken from the data using the parameters estimated from the data with control and leukemia trajectories (Fig. 3F). To calculate the probability of transition to leukemia over time, we calculated the mean time to reach the leukemia state ($c_3$) from the Monte Carlo simulations and represented the time to transition as a survival curve. The time to leukemia for simulated and observed data were compared using the logrank test in a survival curve analysis and were found to be statistically indistinguishable from each other (p=ns).

Trajectories for control mice are confined to a specific region in the state-space, defined to be $\mu_1$, compared to *CM* mice. The trajectories of *CM* not sick mice fluctuate around the stationary point $c_1$ and do not approach the transition point $c_2$. In contrast, the trajectories of the *CM* sick mice are near $c_2$ and transition to the leukemia state at $c_3$ (Fig. 4A). Indeed, immunophenotypic analysis by fluorescence activated cell-sorting (FACS) demonstrates that

leukemia progression accelerates after crossing the critical point $c_2$ as evidenced by a sharp emergence of leukemia blast (cKit+) (Fig.4C, D).

## Discussion

The study of state transitions is common in physics, but is virtually nonexistent in biology. A simple example of a state transition is the transition of water from a solid, to a liquid based on changes in temperature and pressure. The theory and mathematics of state transitions is a powerful tool for interpreting and predicting the behavior of complex systems that has yet to be effectively leveraged with genomic data in cancer. Here we report a theory and mathematical model of state transition of the transcriptome from health to leukemia, and demonstrate the mathematical and biological implications of this theory in a mouse model of AML.

The construction of a transcriptomic state-space which defines health and leukemia states geometrically allows a privileged perspective to study leukemia dynamics by isolating as much as possible the processes directly affecting the transition to a leukemia state. Using Langevin dynamics in a state-space defined geometrically and energetically with a mathematical model, we demonstrate our ability to predict state transition from a health state to a leukemia state based on changes in genome-wide gene expression (i.e., transcriptome) over time. Using mathematical methods to project the transcriptome into a 2-dimensional “state-space” representation which maximally preserves relevant information in the system (SVD), we used simple mathematical equations to provide a physical perspective through which to understand progression to leukemia as a continuous change in the configuration of the transcriptome. Movement through this space could be understood in terms of critical points, mathematically-derived inflection points which provides a computational framework to predict

the development of leukemia for any point in the space, at any time point in leukemia progression.

This approach permitted us to differentiate, at very early time points (i.e., 1 month), mice that eventually developed leukemia from those that did not upon induction of the leukemogenic transgene, whose expression alone was not able yet to predict which mouse would develop overt leukemia within the observation time (Fig. S4). In addition, we show that the critical points, positions and trajectories in the state-space derived from our model remain valid and apply to an independent gene expression dataset from an independent cohort of mice (Fig. S5). Consistent with the mathematically-derived inflection points predicted by our theory, after crossing the critical point $c_2$ we observed accelerated trajectories towards a leukemia state based on FACS-based immunophenotypic analysis. Notably, at this point leukemia development is not yet evident based on conventional FACS assays. Thus, a predictive mathematical model of the dynamics of transition to leukemia that may anticipate critical transitions (Scheffer et al., 2012) has many implications and potential uses, including when to monitor disease changes, and as a novel tool to quantify and predict response to therapy.

The concept of a disease space in the context of infection (Torres et al., 2016) and information theoretic potential energy landscapes in the epigenome (Jenkinson et al., 2017), and the Ornstien-Uhlenbeck model has recently been proposed to describe the evolution of gene expression patterns in the mammalian genome (Chen et al., 2017). Although these approaches do not include a mathematical treatment of dynamics or changes over time. The mathematically-derived theory of state transition presented here intuitively contextualizes complex time series gene expression data into a simple conceptual framework that associates a state of disease with an increased state of energy of the gene expression space, and introduces the concept of a leukemogenic force associated with a particular genomic alteration. Our model is an innovative departure from conventional mathematical modeling approaches of cancer that

focus on the accumulation of oncogenic mutations and clonal dynamics, and introduces the novel concept of genetic alterations as forces that drive state transitions in gene expression that accompany the evolution of cancer.

Our theory leads to the hypothesis that there exists a necessary and sufficient force exerted on the genome required to initiate state transition, as well as to the conjecture that an equal and opposite therapeutic force could be required to return to a health equilibrium. Furthermore, this work suggests the homeostatic relaxation rate and noise parameters should be measured, perhaps through non-oncogenic perturbations of the system. This approach provides a novel framework to quantify the effects and magnitude of oncogenic alterations on the transcriptome, and ultimately on a disease process.

Analysis of trajectories and dynamics in the state-space of complex time series genes expression data are given context by our mathematical model and theory. We provide a new perspective on AML development at a level of abstraction which to our knowledge, has not been previously demonstrated in a controlled experiment. In principle, the state-space and mathematical model may be applied to other systems (e.g., other subtypes of leukemias or cancers), and to other forms of omics data (e.g., epigenomics, proteomics). Therefore, this work may have a broad impact on understanding cancer development, improving molecular-risk stratification and prognostication, guiding treatment selection, and more generally, studying the evolution of multidimensional changes in biological entities as a state-based dynamical system.

**References:**


Aiello, K.A., and Alter, O. (2016). Platform-Independent Genome-Wide Pattern of DNA Copy-Number Alterations Predicting Astrocytoma Survival and Response to Treatment Revealed by the GSVD Formulated as a Comparative Spectral Decomposition. PLoS One *11*, e0164546.

Almassalha, L.M., Bauer, G.M., Chandler, J.E., Gladstein, S., Szleifer, I., Roy, H.K., and Backman, V. (2016). The Greater Genomic Landscape: The Heterogeneous Evolution of Cancer. Cancer Res. *76*, 5605–5609.

Bezuglyy, V., Mehlig, B., Wilkinson, M., Nakamura, K., and Arvedson, E. (2006). Generalized Ornstein-Uhlenbeck processes. J. Math. Phys. *47*.

Cai, Q., Jeannet, R., Hua, W.K., Cook, G.J., Zhang, B., Qi, J., Liu, H., Li, L., Chen, C.C., Marcucci, G., et al. (2016). CBFβ-SMMHC creates aberrant megakaryocyte-erythroid progenitors prone to leukemia initiation in mice. Blood *128*, 1503–1515.

Chen, J., Swofford, R., Johnson, J., Cummings, B.B., Rogel, N., Lindblad-Toh, K., Haerty, W., Di Palma, F., and Regev, A. (2017). A quantitative model for characterizing the evolutionary history of mammalian gene expression. bioRxiv 229096.

Coveney, P. V, Dougherty, E.R., Highfield, R.R., Er, D., and Coveney, P. V (2016). Big data need big theory too. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. *374*.

Dohner, H., Weisdorf, D.J., and Bloomfield, C.D. (2015). Acute Myeloid Leukemia. N Engl J Med *373*, 1136–1152.

Furchtgott, L.A., Melton, S., Menon, V., and Ramanathan, S. (2017). Discovering sparse transcription factor codes for cell states and state transitions during development. Elife *6*, 1–33.

Graham, T.A., and Sottoriva, A. (2016). Measuring cancer evolution from the genome. J. Pathol.

de Hoon, M.J.L., Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. Bioinformatics *20*, 1453–1454.

Hormoz, S., Singer, Z.S., Linton, J.M., Antebi, Y.E., Shraiman, B.I., and Elowitz, M.B. (2016). Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements. Cell Syst. *3*, 419–433.e8.

Hyman, D.M., Taylor, B.S., and Baselga, J. (2017). Implementing Genome-Driven Oncology. Cell *168*, 584–599.

Jenkinson, G., Pujadas, E., Goutsias, J., and Feinberg, A.P. (2017). Potential energy

landscapes reveal the information-theoretic nature of the epigenome. Nat. Genet. 59766.

Jo, K., Kwon, H.-B., and Kim, S. (2014). Time-series RNA-seq analysis package (TRAP) and its application to the analysis of rice, Oryza sativa L. ssp. Japonica, upon drought stress. Methods *67*, 364–372.

Kuo, Y.H., Landrette, S.F., Heilman, S.A., Perrat, P.N., Garrett, L., Liu, P.P., Le Beau, M.M., Kogan, S.C., and Castilla, L.H. (2006). Cbfb-SMMHC induces distinct abnormal myeloid progenitors able to develop acute myeloid leukemia. Cancer Cell *9*, 57–68.

Kuwahara, H., and Schwartz, R. (2012). Stochastic steady state gain in a gene expression process with mRNA degradation control. J. R. Soc. Interface *9*, 1589–1598.

Landau, D. a, Carter, S.L., Getz, G., and Wu, C.J. (2014). Clonal evolution in hematological malignancies and therapeutic implications. Leukemia *28*, 34–43.

Li, Q., Wennborg, A., Aurell, E., Dekel, E., Zou, J., Xu, Y., Huang, S., and Ernberg, I. (2016). Dynamics inside the cancer cell attractor reveal cell heterogeneity , limits of stability , and escape. Proc. Natl. Acad. Sci. 1–6.

Liao, Y., Smyth, G.K., and Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Linsker, R. (1988). Self-Organization in a Perceptual Network. Computer (Long. Beach. Calif). *21*, 105–117.

Lipinski, K.A., Barber, L.J., Davies, M.N., Ashenden, M., Sottoriva, A., and Gerlinger, M. (2016). Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. Trends in Cancer *2*, 49–63.

Mackey, M.C., Santillán, M., Tyran-Kamińska, M., and Zeron, E.S. (2015). The utility of simple mathematical models in understanding gene regulatory dynamics. In Silico Biol. *12*, 23–53.

Michor, F., Hughes, T.P., Iwasa, Y., Branford, S., Shah, N.P., Sawyers, C.L., and Nowak, M.A. (2005). Dynamics of chronic myeloid leukaemia. Nature *435*, 1267–1270.

Mojtahedi, M., Skupin, A., Zhou, J., Castaño, I.G., Leong-Quong, R.Y.Y., Chang, H., Trachana, K., Giuliani, A., and Huang, S. (2016). Cell Fate Decision as High-Dimensional Critical State Transition. PLOS Biol. *14*, e2000640.

Moris, N., Pina, C., and Arias, A.M. (2016). Transition states and cell fate decisions in epigenetic landscapes. Nat. Rev. Genet.

Nabhan, C., Raca, G., and Wang, Y.L. (2015). Predicting Prognosis in Chronic Lymphocytic Leukemia in the Contemporary Era. JAMA Oncol. *60637*, 1–10.

Ng, S.W.K., Mitchell, A., Kennedy, J.A., Chen, W.C., McLeod, J., Ibrahimova, N., Arruda, A., Popescu, A., Gupta, V., Schimmer, A.D., et al. (2016). A 17-gene stemness score for rapid determination of risk in acute leukaemia. Nature *540*, 433–437.

Nichol, D., Robertson-Tessi, M., Jeavons, P., and Anderson, R.A.A. (2016). Stochasticicity in the genotype-phenotype map: implications for the robustness and persistence of bet-hedging. Genetics *XXX*, 1–23.

Ponnapalli, S.P., Saunders, M.A., van Loan, C.F., and Alter, O. (2011). A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms. PLoS One *6*.

Rizvi, A.H., Camara, P.G., Kandror, E.K., Roberts, T.J., Schieren, I., Maniatis, T., and Rabadan, R. (2017). Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. Nat. Biotechnol. *35*, 551–560.

Rohlfs, R. V., Harrigan, P., and Nielsen, R. (2014). Modeling gene expression evolution with an extended ornstein-uhlenbeck process accounting for within-species variation. Mol. Biol. Evol. *31*, 201–211.

Sanavia, T., Finotello, F., Di Camillo, B., Bar-Joseph, Z., Gitter, A., Simon, I., Oh, S., Song, S., Dasgupta, N., Grabowski, G., et al. (2015). FunPat: function-based pattern analysis on RNA-seq time series data. BMC Genomics *16*, S2.

Scheffer, M., Carpenter, S.R., Lenton, T.M., Bascompte, J., Brock, W., Dakos, V., van de Koppel, J., van de Leemput, I.A., Levin, S.A., van Nes, E.H., et al. (2012). Anticipating Critical Transitions. Science (80-. ). *338*.

Spies, D., and Ciaudo, C. (2015). Dynamics in Transcriptomics: Advancements in RNA-seq Time Course and Downstream Analysis. Comput. Struct. Biotechnol. J. *13*, 469–477.

Tannock, I.F., and Hickman, J.A. (2016). Limits to personalized cancer medicine. N. Engl. J. Med. *375*, 1289–31.

Torres, B.Y., Oliveira, J.H.M., Thomas Tate, A., Rath, P., Cumnock, K., and Schneider, D.S. (2016). Tracking Resilience to Infections by Mapping Disease Space. PLOS Biol. *14*, e1002436.

Uhlenbeck, G.E., and Ornstein, L.S. (1930). On the theory of the brownian motion. Phys. Rev. *36*, 823–841.

Wooten, D.J., and Quaranta, V. (2017). Mathematical Models of Cell Phenotype

Regulation and Reprogramming: Make Cancer Cells Sensitive Again! Biochim. Biophys. Acta - Rev. Cancer *1867*, 167–175.

Zhao, B., Sedlak, J.C., Srinivas, R., Creixell, P., Pritchard, J.R., Tidor, B., Lauffenburger, D.A., and Hemann, M.T. (2016). Exploiting Temporal Collateral Sensitivity in Tumor Clonal Evolution. Cell *165*, 1–13.

Zheng, M., Wu, J. nan, Huang, Y. xin, Liu, G. xia, Zhou, Y., and Zhou, C. guang (2012). Inferring Gene Regulatory Networks by Singular Value Decomposition and Gravitation Field Algorithm. PLoS One *7*, 1–6.

Zhou, J.X., Aliyu, M.D.S., Aurell, E., and Huang, S. (2012). Quasi-potential landscape in complex multi-stable systems. J. R. Soc. Interface *9*, 3539–3553.

**Acknowledgments:** This study was funded by the Gehr Family Center for Leukemia Research, and in part by the National Institutes of Health under award number R01CA178387 (to Y.-H.K.) and a St. Baldrick's Foundation Scholar Award (to L.D.W.). Research reported in this publication included work performed in the Analytical Cytometry Core and Animal Resource Center supported by the National Cancer Institute of the National Institutes of Health under award number P30CA33572. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank Art Riggs for his support of this project.

**Author contributions:** RCR, SB, Y-HK, GM, designed the experiments; RCR, SB designed the mathematical model; JQ, GJC, W-KH, EC, AM performed the experiments; HW, DM, DO'M, contributed data analysis and model simulations; XW, Y-CY, ZL, DO'M, contributed bioinformatics analysis and processing of raw data; NC, SJF, LDW, Y-HK, GM, RCR, SB, HW, DO'M, contributed to data interpretation and analysis; Y-HK, RCR, GM provided financial support. All authors contributed to manuscript editing. RCR, SB, NC, LDW, Y-HK, GM gave final approval of manuscript.

No competing financial interests.

Correspondence and requests for materials should be addressed to Russell Rockne (rrockne@coh.org); Ya-Huei Kuo (ykuo@coh.org); Guido Marcucci (gmarcucci@coh.org).

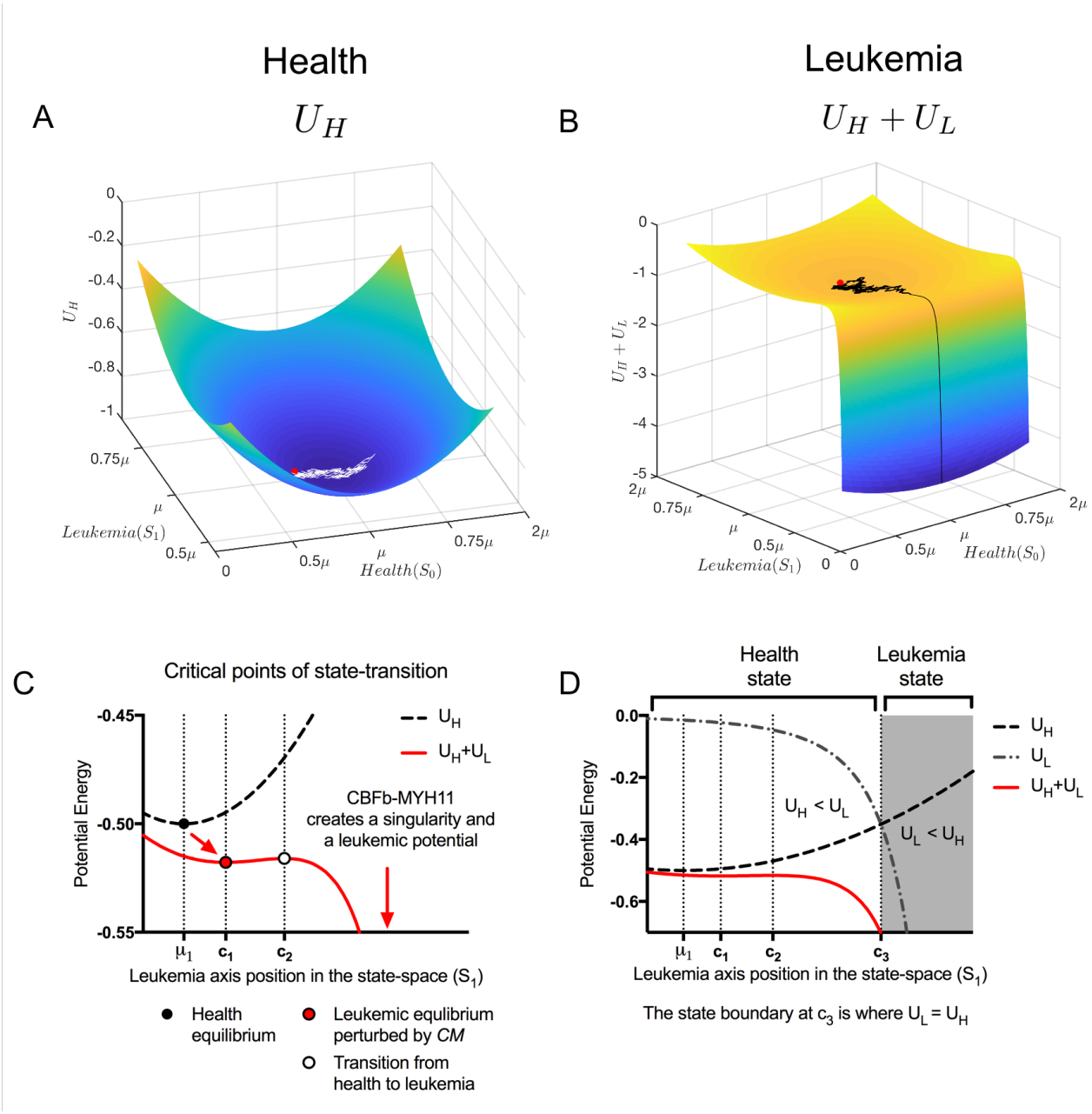


**Fig. 1**. **Cancer as a state transition of the transcriptome.** We represent the transcriptome acting under a directed force field in an "state-space" defined geometrically and energetically with potential energy functions. Our theory of cancer as a state transition of the transcriptome is general, however we demonstrate our approach with acute myeloid leukemia (AML). **(A)** In the absence of an oncogenic alteration, there exists only a health potential $U_H$ with a stable transcriptome configuration. A simulated trajectory of a particle undergoing stochastic motion in

the potential is shown in white. Red dot indicates initial position. **(B)** When an oncogenic alteration is introduced, a singularity is created in the state-space resulting in a shift of the health homeostatic equilibrium and the creation of an unstable critical point and an energetic landscape that characterizes the state of leukemia. Simulated trajectory shown in black with initial position indicated with red dot. An animation of this process is included as supplementary material. (**C**) In a one-dimensional representation, the dynamics of state transition are shown as critical points in the health and health-leukemia potential energy fields. In the health potential $U_H$, a single equilibrium exists at $\mu_1$ which defines the health state. In the case of acute myeloid leukemia, the *CM* transgene creates a singularity in the health potential which shifts the health equilibrium to the leukemia equilibrium perturbed by *CM* at $c_1$. The critical point $c_2$ defines the transition point from health to leukemia and precedes the trajectory towards the singularity. (**D**) The health and leukemia states are defined by energy boundaries. The health state is any point in the space such that $U_H < U_L$. The boundary of the leukemia state, denoted by $c_3$, is the point in the state-space where the health and leukemic potentials are equal, $U_H = U_L$. Any point in the state-space where the leukemic potential is less than the health potential is defined to be in a leukemia state $S_1 < c_3$, $U_L < U_H$. The boundary of the leukemia state lies beyond the transition point $c_2$. This definition is motivated by the stochasticity of the system and the internal energy of the particle in the state-space $\sigma\vec{\eta}(t)$ that allows the transcriptome to transiently cross $c_2$ without transitioning to the leukemia state.

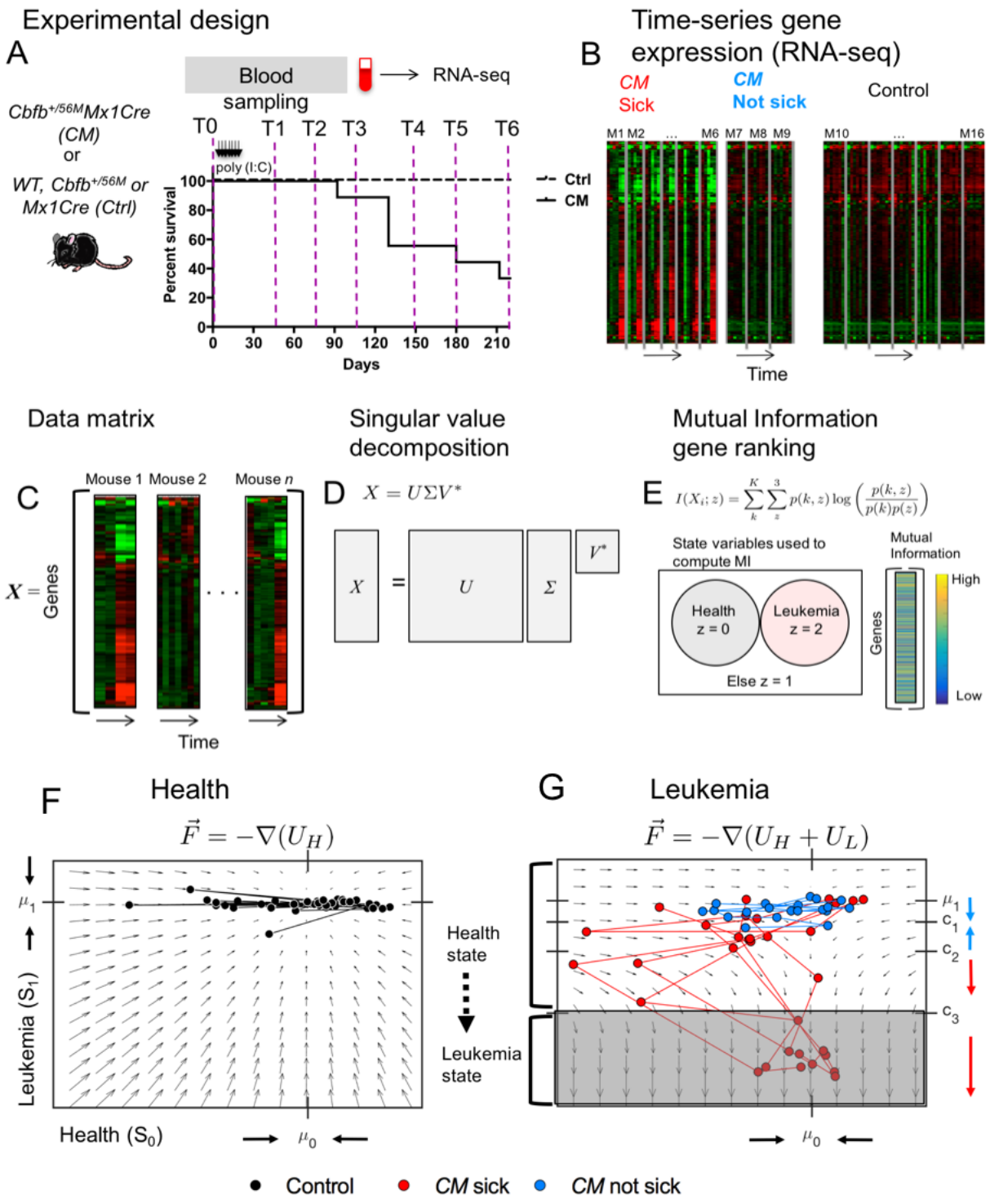


**Fig. 2. Construction of a health-leukemia state-space.** (**A**) A cohort of *CM* conditional knock-in mice (n=9) and littermate controls (n=7) were treated with poly (I:C) every other day for 7

doses (14 mg/kg/dose) to induce the expression of *CM*. Total RNA was isolated from peripheral blood samples collected prior to induction and at 1, 2, 3.5, 4.5, 5.5 and 6.5 months after induction and sequenced. Survival curve of all *CM* and control mice up to 6.5 months is shown. (**B**) Global expression patterns in mice harboring the *CM* fusion gene (columns 1-9) vs. controls (columns 10-16). By 6.5 months, 6 of 9 CM mice developed leukemia defined by the presence of over 20% leukemic blasts in the PB and BM ("*CM* sick", red) while 3 *CM* mice had not ("*CM* not sick", blue). Columns defining mice are separated by grey lines. Columns within each mouse group are time points. (**C**) Gene expression data for all mice at all sequential time points were combined into a data matrix **X**. (**D**) The singular value decomposition is used to reduce the complexity of whole-genome gene expression data and project the transcriptome space into a 2-dimensional plane. (**E**) Mutual information was computed for each gene with respect to state variables and used to rank the information content of each gene and select them to orthogonalize the health and leukemia states (see Supplementary methods). (**F**) Representation of the force field (arrows) and dynamics in a health potential energy landscape, predicting motion of the transcriptome towards a mean stable state located at $(\mu_0, \mu_1)$ (black arrows, left and bottom axis) and in a (**G**) *CBFB-MYH11* (*CM*) altered state-space, predicting dynamics of *CM not sick* mice towards the perturbed equilibrium at $(\mu_0, c_1)$ (blue arrows, right axis) and *CM sick* mice towards the leukemia state and singularity located at $(\mu_0, 0)$ (red arrows, right axis). The force vectors in the leukemia state-space (**G**) have been log transformed.

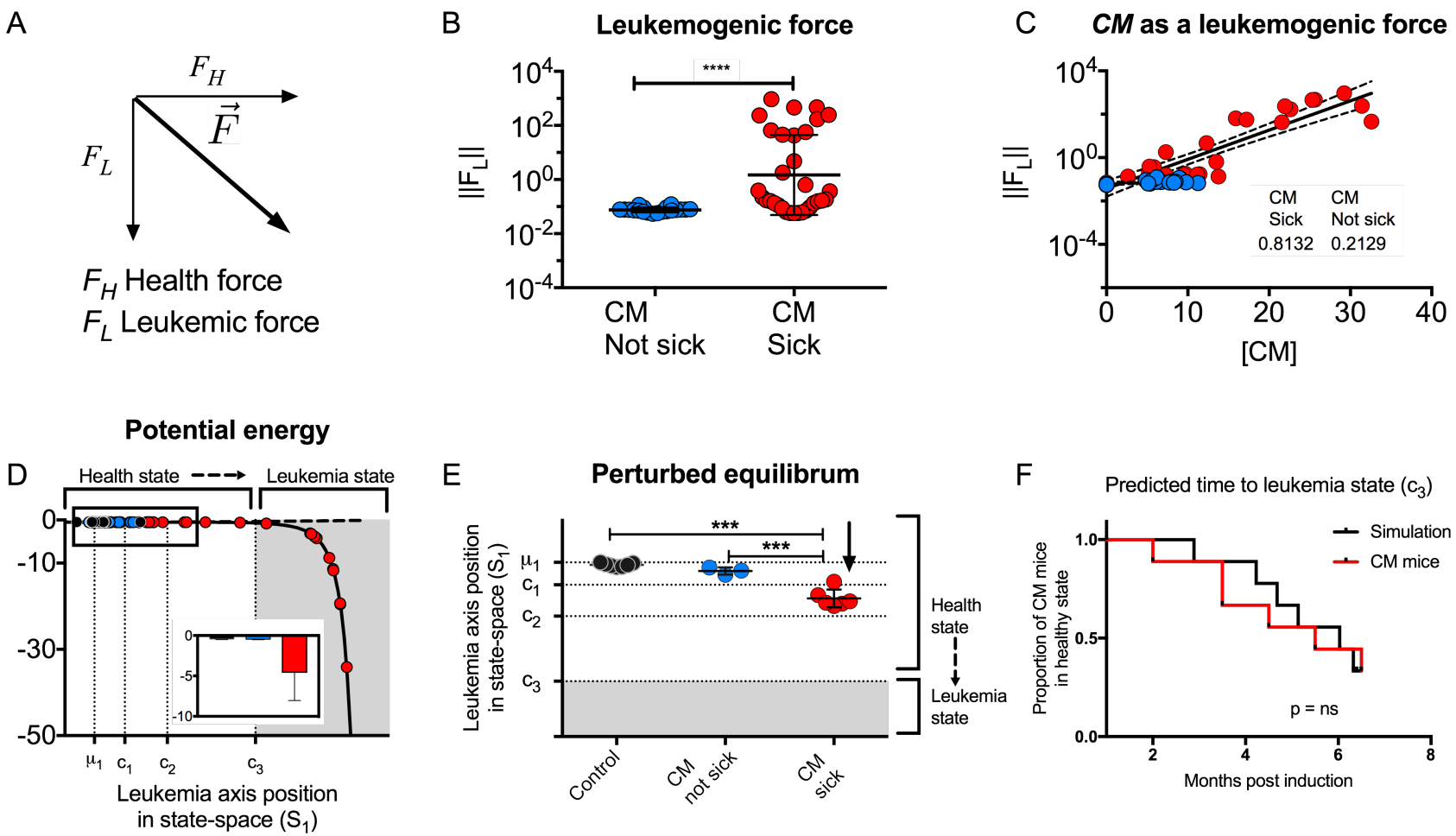
A
$F_H$
$\vec{F}$
$F_L$
$F_H$ Health force
$F_L$ Leukemic force
B
Leukemogenic force
****
$10^4$
$10^2$
$10^0$
$10^{-2}$
$10^{-4}$
$\|F_L\|$
CM Not sick
CM Sick
C
*CM* as a leukemogenic force
$10^4$
$10^0$
$10^{-4}$
$\|F_L\|$
CM Sick 0.8132
CM Not sick 0.2129
0
10
20
30
40
[CM]
Potential energy
D
Health state
Leukemia state
0
-10
-30
-50
$\mu_1$
$c_1$
$c_2$
$c_3$
Leukemia axis position in state-space ($S_1$)
E
Perturbed equilibrum
***
***
Leukemia axis position in state-space ($S_1$)
$\mu_1$
$c_1$
$c_2$
$c_3$
Health state
Leukemia state
Control
CM not sick
CM sick
F
Predicted time to leukemia state ($c_3$)
Simulation
CM mice
Proportion of CM mice in healthy state
1.0
0.5
0.0
p = ns
2
4
6
8
Months post induction

**Fig. 3. Interpreting the leukemogenic force, potential energy, and motion in the state-space .** (**A**) The deterministic components of the force vector acting on the transcriptome particle is composed of a leukemic and health component. (**B**) The magnitude of the leukemic force differentiates *CM* mice (*CM* sick, red) that develop leukemia from those that do not (*CM* not sick, blue). **CBFB-MYH11 fusion transcript level [*CM*].** Levels of *CM* over time (counts per million reads; cpm) (**C**) The expression level of *CM* is strongly correlated with the leukemic force exerted on the mouse transcriptome particle as predicted by the mathematical model, with a stronger correlation for the *CM* sick than the *CM* not sick mice. **Energy states.** (**D**) Potential energy for control and *CM* mice in the state-space over all time points (summarized in inset). (**E**) Position along the leukemic axis at 1 month post induction distinguishes *CM* mice from controls, as well as *CM* sick from *CM* not sick. The mathematical model predicts an immediate shift in equilibrium state from $\mu_1$ to $c_1$ following *CM* induction. (**F**) Proportion of *CM* mice in health state as a function of time, represented as a survival curve. Simulated data represents the mean time to leukemia state ($c_3$). The simulated and observed survival curves are not statistically different.

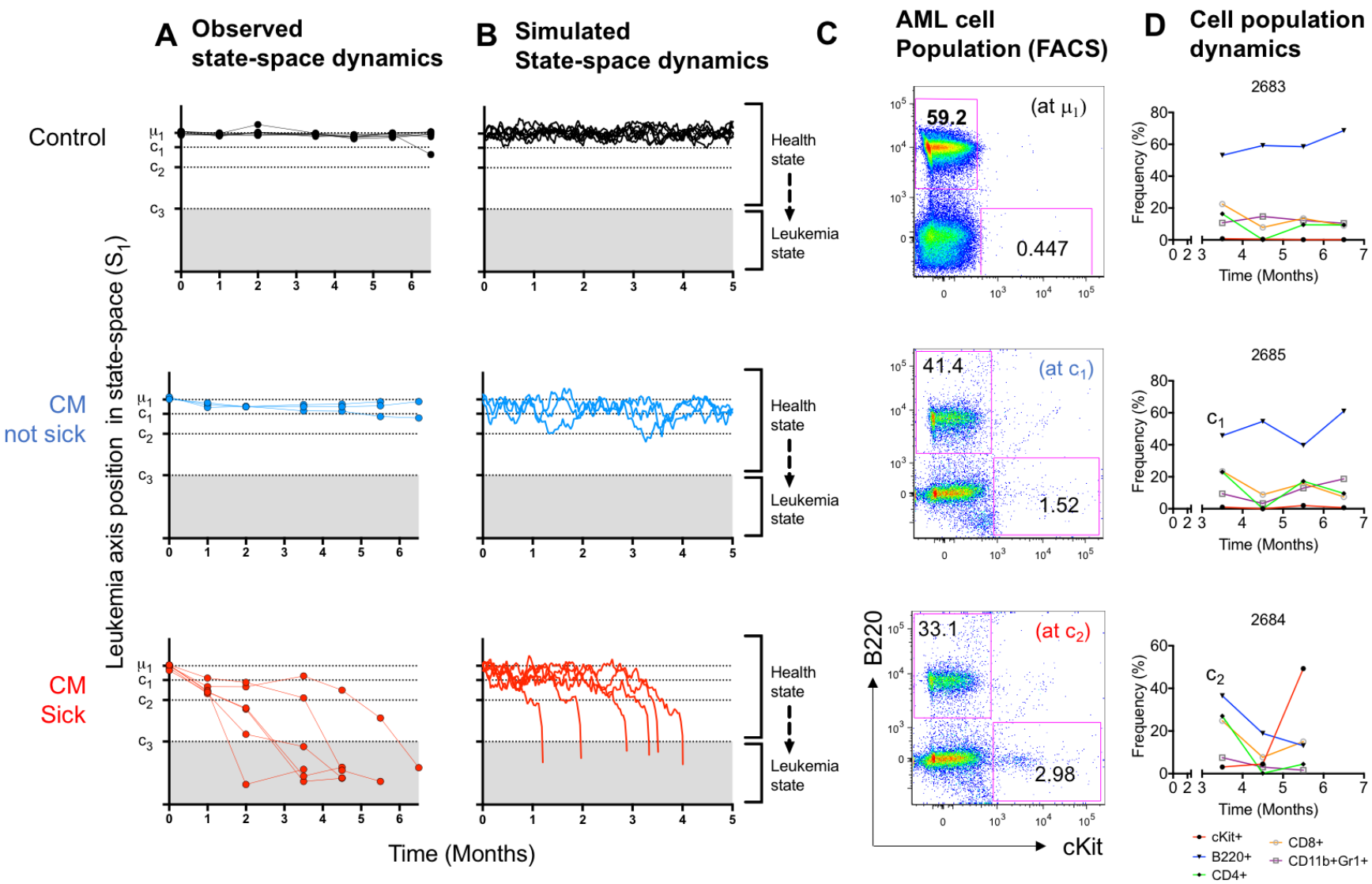
A Observed state-space dynamics
B Simulated State-space dynamics
C AML cell Population (FACS)
D Cell population dynamics
Control
CM not sick
CM Sick
Leukemia axis position in state-space ($S_1$)
$\mu_1$
$c_1$
$c_2$
$c_3$
Health state
Leukemia state
Time (Months)
59.2
(at $\mu_1$)
0.447
41.4
(at $c_1$)
1.52
33.1
(at $c_2$)
2.98
B220
cKit
2683
2685
2684
Frequency (%)
Time (Months)
cKit+
B220+
CD4+
CD8+
CD11b+Gr1+

**Fig. 4. Biological significance of state transition critical points**. (**A**) Observed trajectories along the leukemic axis in the state-space of control mice (n=7; black, top row), *CM* not sick (n=3; blue, middle row) and *CM* sick mice (n=6; red, bottom row). (**B**) Simulated dynamics initialized with the locations along the leukemic axis in the state-space from the data. (**C**) Representative FACS plots of cell surface markers for leukemia blasts (cKit) and normal B cells (B220) of control at $\mu_1$ (top), *CM* not sick at $c_1$ (middle) and *CM* sick mouse at $c_2$ (bottom) in the state-space. (**D**) Dynamics of blood cell population over time measured by FACS for one of the control (2683), CM not sick (2685) and CM sick (2684) mice. Cell populations shown are leukemia blasts (cKit+), B cells (B220+), T cells (CD4+ or CD8+) and myeloid cells (CD11b+/Gr1+).

## Supplementary methods

### Mice

*Cbfb*$^{+/56M}$*/Mx1-Cre* mice (Kuo et al., 2006) were backcrossed to C57BL/6 strain for more than 10 generations (Cai et al., 2016). To induce CM, 6-8 week old *Cbfb*$^{+/56M}$*/Mx1-Cre* mice were injected with 14 mg/kg/dose of poly (I:C) (InvivoGen) every other day for 7 doses. Similarly treated *Cbfb*$^{+/56M}$ *or Mx1-Cre* or WT littermates were used as control. Peripheral blood (PB) samples were collected periodically via retro orbital sinus. All mice were maintained in an AAALAC-accredited animal facility, and all experimental procedures were performed in accordance with federal and state government guidelines and established institutional guidelines and protocols approved by the Institutional Animal Care and Use Committee at the Beckman Research Institute of City of Hope.

### Flow cytometry analysis

Phenotypic analysis of the PB was performed by fluorescence activated cell-sorting (FACS) to monitor leukemia development. Cells were stained in PBS with 0.5% BSA for 15 min on ice with fluorescently labeled antibodies for cell surface markers including cKit, B220, CD4, CD8, CD11b, Gr1. A leukemia state was defined by presence of >20% cKit$^{+}$ blasts in PB. Bone marrow was analyzed when mice were moribund or at the experimental end point (6.5 months). Six out of 9 CM mice succumbed to lethal leukemia between 3.5 and 6.5 months. The remaining 3 CM mice that did not develop leukemia during the observation time (*CM* not sick) all showed a pre-leukemic phenotype in the bone marrow as previously described (Cai et al., 2016; Kuo et al., 2006). For FACS analysis, mononuclear cells were stained in PBS with 0.5% BSA for 15 min on ice with fluorescently labeled antibodies for cell surface markers, cKit, Sca1, CD16/32, CD105, CD150, CD48, CD41, CD71, Ter119. Lineage cocktail includes biotin-conjugated CD3, CD4, CD8, B220, CD19, IgM, NK1.1, CD11b, CD11c, and IL7Ra. All antibodies were purchased from BioLegend, BD Biosciences, or eBiosciences. Flow cytometry was performed using a 5-

laser, 15-detector BD LSRII and acquired data was analyzed by Flowjo software (Tree Star Inc.). Phenotypic hematopoietic stem and progenitor cell populations were defined as LSK ($Lin^-ckit^+Sca1^+$); MP ($Lin^-ckit^+Sca1^-$); Pre-GM ($Lin^-ckit^+Sca1^-CD41^-CD16/32^{-/lo}CD105^-CD150^-$); GMP ($Lin^-ckit^+Sca1^-CD41^-CD16/32^+CD150^-$); Pre-Meg/E ($Lin^-ckit^+Sca1^-CD41^-CD16/32^{-/lo}CD105^-CD150^+$); MkP ($Lin^-ckit^+Sca1^-CD41^+CD150^+$); Pre-CFU-E ($Lin^-ckit^+Sca1^-CD41^-CD16/32^{-/lo}CD105^+CD150^+$); CFU-E ($Lin^-ckit^+Sca1^-CD41^-CD16/32^{-/lo}CD105^+CD150^-Ter119^-$); ProEry ($Lin^-ckit^+Sca1^-CD41^-CD16/32^{-/lo}CD105^+CD150^-Ter119^+$); EP ($Lin^-ckit^+Sca1^-CD41^-CD16/32^{-/lo}CD105^+$).

**RNA extraction, sequencing and bioinformatics**

Total RNA was extracted from whole blood using the AllPrep DNA/RNA Mini Kit (Qiagen, Hilden, Germany); quality and quantity were estimated using a Nanodrop (Thermo Scientific, Waltham, Massachusetts). Sequencing libraries were constructed from time-point 0, 1 and endpoint (leukemic) samples using the Ribo-Zero™ rRNA Removal Kit (Human/Mouse/Rat) (Illumina, Madison, Wisconsin) and TruSeq RNA Sample Preparation Kit V2 (Illumina) following manufacturer's recommendations. For all other samples, the TruSeq Stranded Total RNA Human/Mouse/Rat kit (Illumina) was used. Libraries were loaded to cBot for cluster generation and sequencing was performed on a Hiseq 2500 (Illumina) with single end 51-bp for mRNA-seq to a nominal depth of 40 million reads. Image processing and base calling were conducted using Illumina's RTA pipeline. Reads were mapped to the UCSC mm10 reference with NovoAlign (NovoCraft, Selangor, Malaysia) and counted across genomic features using Subread featureCounts(Liao et al., 2014). Read counts were normalized by sequencing depth and transcript length (TMM) using the Bioconductor package "edgeR". Heatmaps were generated using Cluster v3.0 (de Hoon et al., 2004)(Fig 1b).

**State-space construction**

**Mutual information**

We ranked each gene *m* in the set of *M* genes ($m \in M$) by computing the mutual information (MI) value of that gene with respect to a state variable *Z* which defines states of health (*z* = 0), or leukemia (*z* = 1). The expression of gene *m* is given by $TPM_m = \frac{FPKM_m}{\sum_m^M FPKM_m} 10^6$. The state variable *S* describe the condition of the mice as healthy (*z*=0) if the position in the space is inside the health basin, defined by the smallest ellipses that includes all mice at time zero and all control mice, sick (*z*=2), defined by the smallest ellipse that contains all the mice with blast level greater than 60%, and intermediate (*z*=1) that describe all mice that do not fall in the previous two categories. We discretize the $TPM_m$ variable associated with each gene in the set ($m \in M$) into *K* values by partitioning the empirical cumulative distribution function of $TPM_m$ into *K* uniform intervals. We choose *K*=4 in this experiment so that each bin contained more than 20 elements. We than compute for each gene the mutual information ($MI_m$) between the discretized variable $X_m \in \{1,2,3,\ldots,M\}$ and the phenotype variable *z* with $z \in \{0,1,2\}$ as:

$$I(X_m; z) = \sum_k^K \sum_z^3 p(k,z) log(\frac{p(k,z)}{p(k)p(z)})$$

with $p(\mathrm{X_m}, z) = p(\mathrm{X_m} = x \wedge z = z)$ being the joint probability for the variables $X_m$and *z*.

**Orthogonolization of the state-space**

We note the SVD decomposition of the gene expression data matrix maximally preserves information when the data is a linear combination of signal and noise (Linsker, 1988). We then rotated and translated the space such that the principle basis vector ($v_1$) was parallel with the abscissa axis, $s_0$. We defined the transformed space as $S = a\ (RS' + T)$, where *R* is the standard 2-D rotation operator for the angle made between the principle (first) basis vector and the abscissa axis in a 2-D Cartesian coordinate space. The scale factor *a* and translation

operation *T* are chosen so that the values of *S* lie in the positive quadrant of the plane, $S \in \mathbb{R}_+^2$ and $\mu_1 = 1$.

We found there exists a distinct subset of 1047 genes (92% percentile of all genes, see supplemental data) that resulted in the maximal alignment of the center of mass of the health and leukemia states in the state-space (Fig. S2). This alignment makes orthogonal the health and leukemia components of the space and decouples the health and leukemia dynamics in the space. Of note, when a smaller subset of genes (top 99% percentile) are selected to build the state-space, the alignment of the health and leukemia center of mass was lost (Fig. S2), suggesting that there exists a subset of genes with an optimal amount of information to align and decouple the health and leukemia states in the space.

The data, when interpreted in this way, suggested that the transcriptome configuration of each mouse could be represented in the 2D state-space $S = (s_0, s_1)$ with the greatest variability in gene expression associated with health in the first component ($s_0$), and leukemia development in the second component ($s_1$), although we note that the alignment of the centers of mass and with the canonical axes is not required. These operations simplify our interpretation of the data and therefore the construction of the mathematical model. In support of our interpretation of a health (abscissa, $s_0$) and a leukemia (ordinate, $s_1$) axis, we observed a strong correlation between the *CM* transcript level and the leukemic axis position in the space ($r^2$=0.77, p<0.001), and no correlation between the level of the *CM* transcript and the health axis position in the space ($r^2$=0.04, p=0.137) (Fig. S3).

The direction of movement in the state-space is shown by the velocity vector, which differentiates *CM* sick and *CM* not sick mice as early as one month post induction. For *CM* sick mice, the velocity vector points towards the unstable transition point $c_2$; in contrast, in *CM* not sick mice the velocity vector points towards the perturbed stable equilibrium at $c_1$. In control mice, the velocity vector points towards the stable homeostatic equilibrium at $(\mu_0, \mu_1)$. The

directionality of the initial velocity towards the leukemia state measured in degrees was significantly different for *CM* sick vs *CM* not sick mice (-42.7°±21.7 vs. -17.4°±5.7, p=0.036) and *CM* sick vs control mice (-42.7°±21.7 vs. -12.6°±21.4 p=0.037). The signed magnitude of the velocity in the leukemia axis ($V_1$) of *CM* sick mice was significantly different from *CM* not sick ( -0.16±0.03 S/month vs. -0.05±0.03 S/month, p=0.002), and that of control mice (-0.01±0.01 S/month, p<0.001).

**State-space dynamics**

We chose the form of the leukemia potential $\nabla U_L$ both based on our hypothesized dynamics in the state-space that is a homeostatic fluctuation with a rapid acceleration towards a leukemia state, and on the principle of Occam's razor which suggests using the model with fewest assumptions possible to describe the phenomenon, and a model which provided clear biological interpretations. We also considered other functional forms for $\nabla U_L$ including bistable systems but were unable to justify the use of more complex models based on our data set. Of note, the potential functions vary only in space (*S)* and not in time, $\frac{\partial}{\partial t}(U_H) = \frac{\partial}{\partial t}(U_L) = 0$ so that Langevin dynamics can be applied.

**Critical points in the state-space**

The critical points that define the steady state solutions when $\frac{ds_1}{dt} = 0$ are given by the intersection of the two curves $G(s_1) = -\theta_{1,1}(s_1 - \mu_1)$ and $H(s_1) = \alpha_{1,1}s_1^{-N}$ for $s_1 > 0$. The curves may intersect at a single point, at two points, or not at all. For two critical points of the system (denoted $c_1, c_2$ where$c_1 > 0, c_2 > 0$), the following relation must hold: $\left(\frac{c_2}{c_1}\right)^N = \frac{(c_1-\mu_1)}{(c_2-\mu_1)}$. This can be seen by taking ratios of the equations $G(c_2) = H(c_2)$ and $G(c_1) = H(c_1)$. The value of $c_1 = 0.897$ was estimated from the data and taken to be the arithmetic mean position along the

leukemia axis for all time points mice induced to express the *CM* transgene mice that do not develop leukemia, defined by greater than 20% leukemic blasts in the peripheral blood, as well as *CM* sick mice at one month post induction. The critical value defining the state transition between health and leukemia, $c_2$ was estimated from the data as the least upper bound on the leukemia axis over all *CM* mice that eventually developed leukemia but that had not yet developed leukemia, $c_2 = 0.753$.

Due to the inherently stochastic nature of biological processes, including random biological fluctuations, the dynamics are not completely deterministic. The analysis of critical points and dynamics of transition to not explicitly account for these variations. In particular, because the transition point $c_2$ is an unstable stationary point of the dynamical system, from a purely deterministic perspective, leukemia would only develop if the system was initialized at some point past $c_2$. However, the stochasticity of the process allows the possibility of transition across $c_2$ for any initial state in the space, with sufficient variation in the noise function $\sigma\vec{\eta}(t)$. The stochastic dynamics of state transition are studied computationally with Monte Carlo simulations (Fig. 4D). The variance of the noise, $\sigma$ is estimated to be the variance about the mean of the control mice along the health axis.

**Model parameters**

In our proposed framework, the dynamics of leukemia development in the state-space along the leukemia axis are determined by the following parameters $(\mu_0, \mu_1, \theta_{0,0}, \theta_{1,1}\alpha_{1,1}, N)$. The parameters are: the position of the health state $(\mu_0, \mu_1)$, the homeostatic reversion rate ($\theta_{0,0}$ and $\theta_{1,1}$) which we assume to be symmetric so that $\theta_{0,0} = \theta_{1,1}$, the magnitude of the leukemia force (*N*), and the leukemogenic rate constant, $(\alpha_{1,1})$. We note the parameters are not explicitly fit to the data with a minimization procedure, rather, the critical points $c_1, c_2$ are estimated from the

observed dynamics in the data, from which the dynamical parameters may be calculated as follows. With $c_1$ and $c_2$ estimated from the data, the leukemia force constant can take only one value and may be calculated directly from the relation $\left(\frac{c_2}{c_1}\right)^N = \frac{(c_1-\mu_1)}{(c_2-\mu_1)}$. Without loss of generality, or change in the fundamental dynamics of the system, the homeostatic reversion rate constant may be taken to be unity. Finally, once $c_1, c_2$ have been estimated, we calculated $N = 4.999 \approx 5$. The homeostatic reversion rate constants and mean position may be taken to be unity $\theta_{0,0} = \theta_{1,1} = \mu_1 = 1$, the leukemia rate constant $\alpha_{1,1}$ can be computed directly by solving the one-dimensional dynamics equation using either $c_1$ or $c_2$ as critical value to give $\alpha_{1,1} = 0.0598\ S^2/month$. We suspect that the leukemogenic and homeostatic rate constants are stochastically increasing variables in time, however, we take as a first assumption that each of these variables are constant. The critical point defining the leukemia state, $c_3 = 0.4542$ is the position in the state-space where $U_H = U_L$ computed with the rate constant parameters.

**Testing model predictions with data from an independent experiment**

To determine whether the critical points, positions, and trajectories in the state-space remain valid and apply to independent gene expression datasets at different time points after induction, we analyzed PBMC RNA-Seq data from a second cohort of mice from an independent experiment including *CM* (n=3, *CM1, CM2, CM3*) and control (n=2) mice collected at 10 weeks or 20 weeks (t1) after induction but before the appearance of leukemia in the PB (data not shown). We collected PB for RNA-Seq one month later (t2) and when *CM* mice became moribund (t3). We used the SVD projection operator and the 1047 genes identified in the original state space to map the second cohort of *CM* mice into the space, with positions (Fig. S5a) and trajectories (Fig. S5b) consistent with the critical points in the space identified with the original cohort. The healthy (Fig. S5c) and leukemic force fields (Fig. S5d) predict the dynamics

of motion in the state-space for the second cohort of mice. In particular, we observed the control mice of the second cohort remaining at the stable stationary health state located at $\mu_1$ for all time points, consistent with the results obtained from the first cohort. Without knowledge of whether a mouse was *CM* or control, the positions, trajectories, and leukemic force identified all three *CM* mice in the second cohort with 100% probability of transition to leukemia at or before t2, and that 1 of the 3 *CM* mice had already crossed $c_2$ at t1 (Fig. S5e). Leukemic force computed at all time points for the second cohort clearly differentiates *CM* and control mice (Fig. S5f).

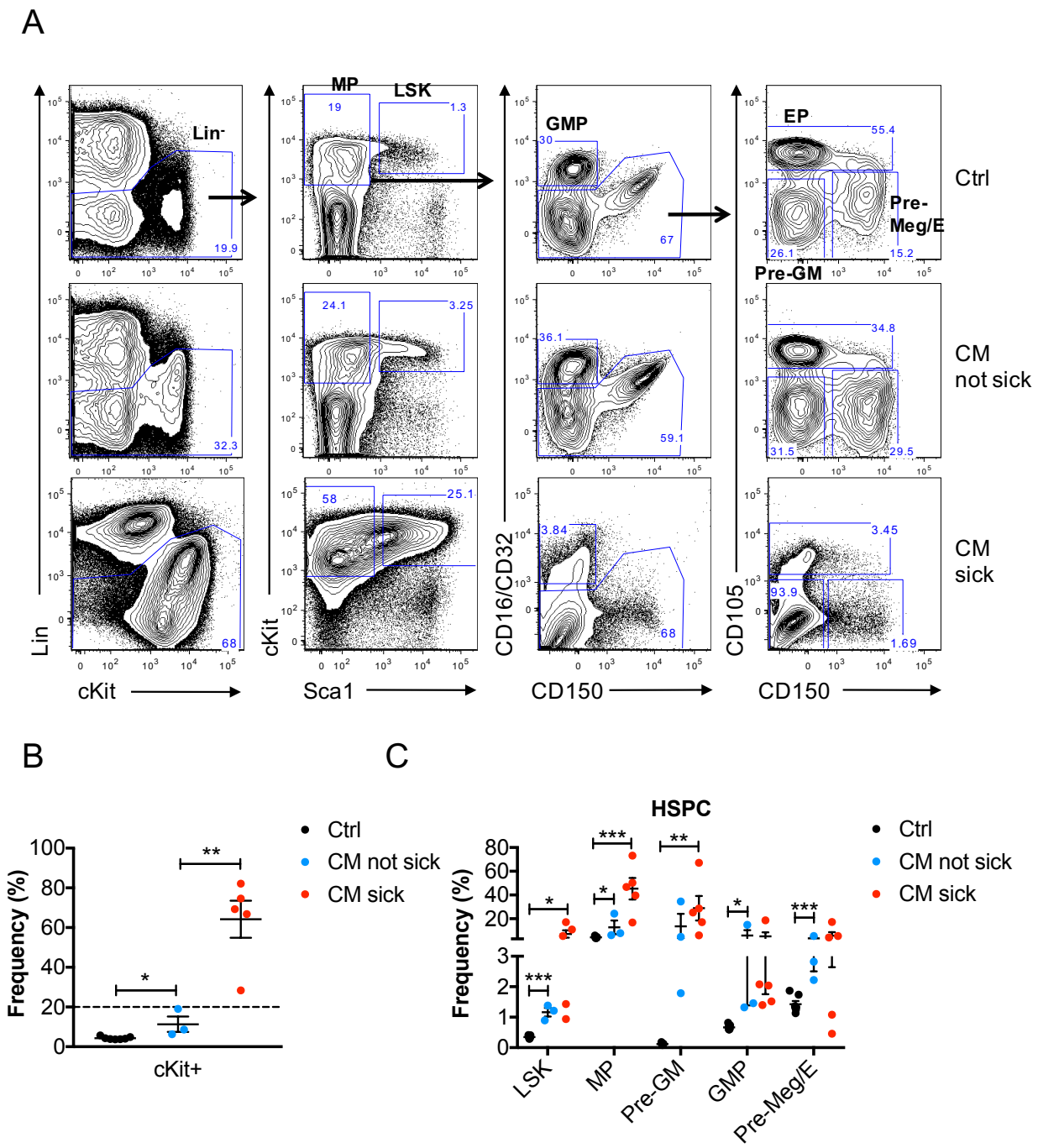


**Fig. S1. Phenotypic analysis of bone marrow (BM) at experimental end point.** (A) Representative fluorescence-activated cell sorting (FACS) plots showing the gating strategy of various phenotypic progenitor populations including lineage-negative ($Lin^-$), LSK ($Lin^-$ $ckit^+Sca1^+$); myeloid progenitor (MP; $Lin^-ckit^+Sca1^-$); pre-granulocyte-macrophage (Pre-GM; $Lin^-$ $cKit^+Sca1^-CD16/32^{-/lo}CD150^-CD105^-$), granulocyte-macrophage progenitors (GMP; $Lin^-$

cKit$^{+}$Sca1$^{-}$CD16/32$^{hi}$CD150$^{-}$), pre-megakaryocyte/erythrocyte (Pre-Meg/E: Lin$^{-}$cKit$^{+}$Sca1$^{-}$CD16/32$^{-/lo}$CD150$^{+}$CD105$^{-}$), and erythroid progenitors (EPs: Lin$^{-}$cKit$^{+}$Sca1$^{-}$CD16/32$^{-/lo}$CD105$^{hi}$) in the BM of moribund CM mice (CM sick), CM mice that had not developed leukemia at 6.5 months (CM not sick) compared to Ctrl mice analyzed at the experiment end point. The frequency within total BM is shown for Lin$^{-}$ (first plot from left), and the frequency in the respective parent gate is shown for the other populations. (B) The frequency of cKit$^{+}$ cells in the BM of moribund CM sick (n=6), CM not sick (n=3) compared to control (Ctrl) mice (n=7) at end point. Leukemia features expansion of immature cKit$^{+}$ blasts (>20%, indicated by dished line) in both PB and BM. (C) The frequency of phenotypic progenitor populations in the CM sick (n=5), CM not sick (n=3) compared to Ctrl mice (n=7) BM as defined above. Increase of stem and progenitor cell populations (including LSK, MP, Pre-GM, GMP, Pre-Meg/E) were observed in all 3 CM mice that had not developed leukemia at 6.5 months. Mean ± SEM is shown. *: $p<0.05$; **: $p< 0.01$; ***: $p<0.001$.

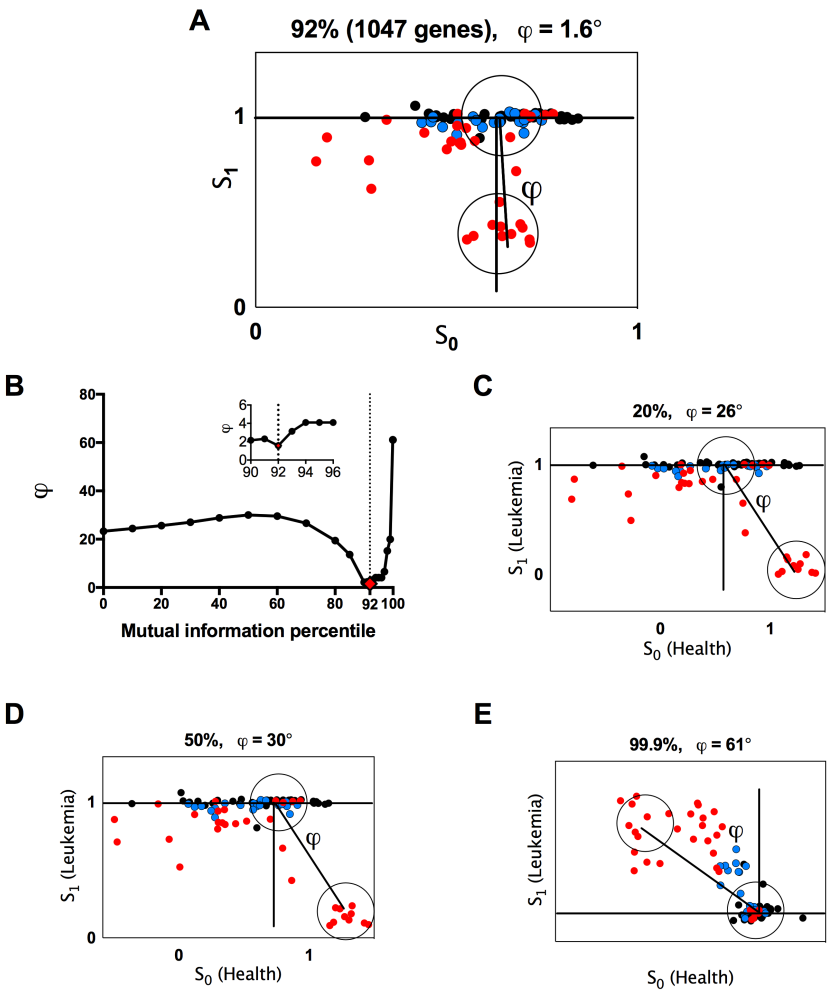


**Fig. S2 Orthogonalization of the state space: alignment of health and leukemia state center of mass.** (**A**) We used mutual information to rank the information content of genes relative to the center of mass of a health and leukemia state and calculated the angle between the health and leukemia states ($\varphi$). (**B**) We selected genes for the space to maximize the alignment between the center of mass of the data from mice in the healthy and leukemia states (angle $\varphi$ close to 0'), which occurs at a unique minimum of 1047 genes which have a MI score at or above 92% percentile (see inset). (**C**) Example of a non-aligned center of mass in the state-space using genes which rank 20%, (**D**) 50%, the optimal and (**E**) 99.9% percentile of genes.

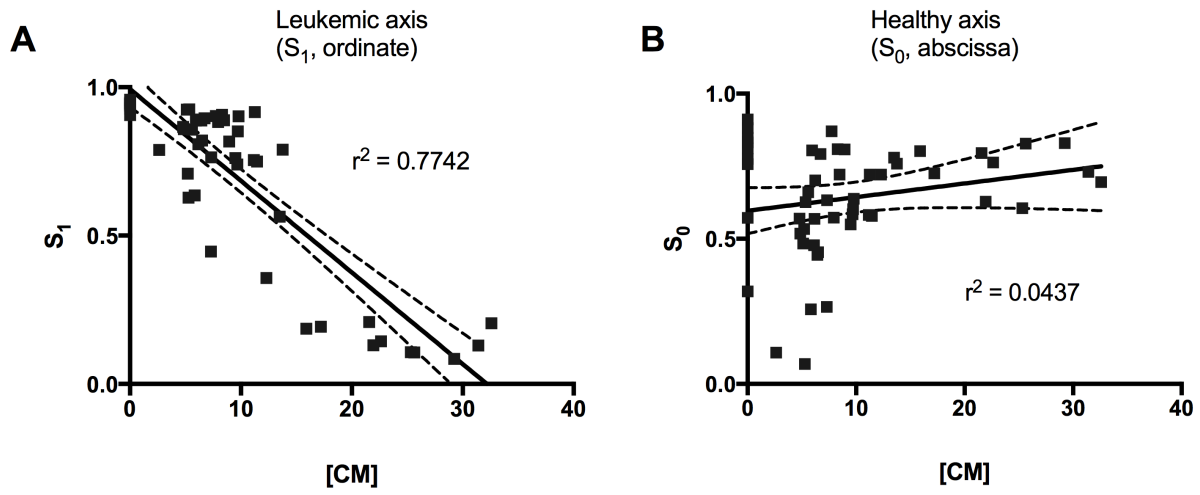


**Fig. S3 Correlation of the *CM* fusion transcript level with the axes of the state-space.** (**A**) The *CM* transcript level and the leukemic axis (ordinate) trajectories in the space ($r^2$=0.77, p<0.001) (**B**) and a weaker correlation with the healthy axis (abscissa) ($r^2$=0.04, p=0.137).

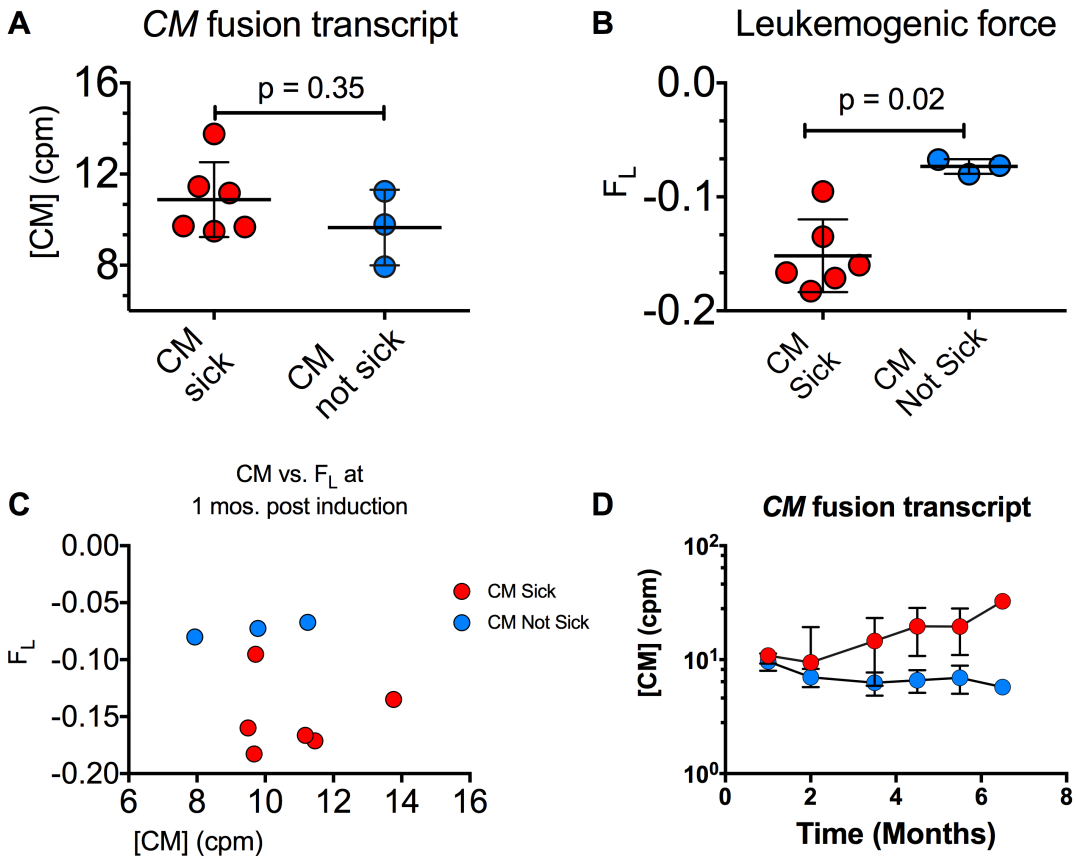


**Figure S4. Expression of fusion gene CM and leukemic force.** (**A**) The expression level of the *CM* fusion transcript between the *CM* sick and *CM* not sick mice is not different at one month post injection (p = 0.35). (**B**) The leukemic component of the force ($F_L = -\nabla U_L$) is significantly different between *CM* sick and *CM* not sick mice at one month post induction (p = 0.02). (**C**) The *CM* fusion transcript expression level vs. leukemic force. (**D**) Levels of *CM* over time (counts per million reads; cpm) over time.

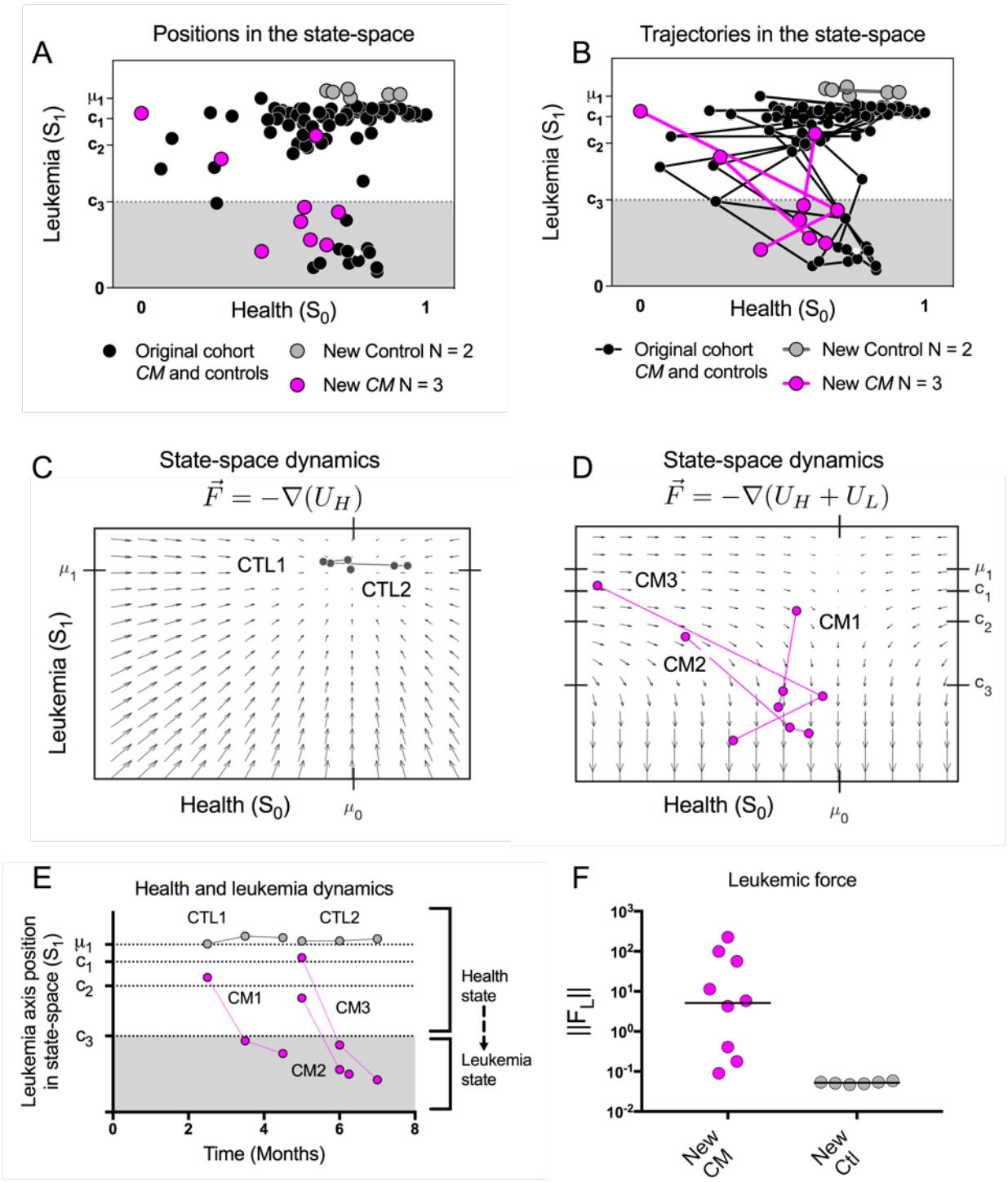


**Fig. S5. Application to data from an independent experiment.** RNA-seq data of five new mice (3 *CM,* magenta; 2 control, grey) from an independent experiment were mapped into the state-space using the SVD projection operator. (**A**) Positions and (**B**) trajectories in the state-

space of the new mice differentiate controls from *CM* induced mice and are consistent with those observed in the original cohort (black). Trajectories and dynamics in the state-space predicted by the mathematical model and force field for the (**C**) controls and (**D**) *CM* mice. (**E**) Trajectories along the leukemia axis for the new control and *CM* mice. (**F**) Leukemic force evaluated at all time points for new *CM* and control mice.