State-transition analysis of time-sequential gene expression identifies critical points that predict development of acute myeloid leukemia

Russell C. Rockne1†\*, Sergio Branciamore2†, Jing Qi3†, David E. Frankhouser2,4, Denis O’Meally5, Wei-Kai Hua3, Guerry Cook3, Emily Carnahan3, Lianjun Zhang3, Ayelet Marom3, Herman Wu3, Davide Maestrini1, Xiwei Wu4, Yate-Ching Yuan7, Zheng Liu6, Leo D. Wang8,9, Stephen Forman3, Nadia Carlesso3, Ya-Huei Kuo3††\*, Guido Marcucci3††

**Supplemental figures, tables, and methods**

In this supplement we provide additional data and methods to support our results. In particular, we show: details of our mouse model of AML; demonstrate the robustness of our state-space construction to normalization methods and gene selection criteria; compare dimension reduction methods (PCA, Diffusion Mapping, t-SNE); show bootstrap cross-validation of our predictions; compare hierarchical clustering with our state-space approach; and finally, provide full lists of differentially expressed genes and gene ontology (GO) term enrichment. Each of the concepts is illustrated with a figure and are organized as follows:

**Figures**

**Figure S1.** Details of animal model and fusion gene CM.

**Figure S2.** Ex vivo flow cytometry analysis of bone marrow.

**Figure S3.** Details of principal component analysis.

**Figure S4.** Comparison of state-space construction with different dimension reduction methods.

**Figure S5.** Hierarchical clustering of time-series RNA-seq data and relation to state-space.

**Figure S6.** Details of critical point estimation, construction of quasi-potential, and state-space dynamics.

**Figure S7.** Correlation of state-space geometry with Kit and CM gene expression. Sensitivity of state-space geometry to inclusion of Kit and CM.

**Figure S8.** Volcano plots of critical-point based differential gene expression analysis.

**Figure S9.** Heatmaps of selected GO pathways in early, transition, persistent, and leukemic events.

**Figure S10.** Computation of eigengene angle in state-space.

**Figure S11.** Details of principal component analysis of validation cohorts and state-space dynamics.

**Figure S12.** Mean-squared displacement analysis of particle trajectories in state-space and calibration of Fokker-Planck diffusion coefficient with training cohort.

**Figure S13.** Sensitivity analysis of state-space construction to sample and normalization thresholds.

**Figure S14.** Bootstrap cross validation of state-space construction.

**Tables**

**Table S1.** Differentially expressed genes for vs

**Table S2.** Differentially expressed genes for vs

**Table S3.** Differentially expressed genes for vs

**Table S4.** Differentially expressed genes for vs

**Table S5.** Differentially expressed genes for vs

**Table S6.** Differentially expressed genes for vs

**Tables S7-S10.** Differentially expressed genes for early, transition, persistent, and leukemia events.

**Tables S11-S14.** GO terms enriched for early, transition, persistent, and leukemia events.

**Table S15.** Top 1% of eigengenes.

**Methods**

**Mice**

To induce expression of CM, *Cbfb+/56M/Mx1-Cre* C57BL/6 mice (4-8 weeks old, including both females and males) were injected with 14 mg/kg poly (I:C) (InvivoGen, San Diego CA) every other day for 7 doses. Similarly treated, wild-type, *Cbfb+/56M or Mx1-Cre* littermates were used as controls. As previously described(1), induction of CM expression results in expansion of pre-leukemic hematopoietic stem and progenitor cell (HSPC) populations in the bone marrow and subsequent development of overt leukemia characterized by >20% cKit+ leukemia blasts, increased white blood cell counts, and splenomegaly. Peripheral blood samples were taken via retro-orbital venous sinus before induction and monthly thereafter for the duration of the experiment. 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**

PBMCs were stained in PBS with 0.5% BSA for 15 min on ice with fluorescently labeled antibodies for cell surface markers, including cKit, CD3, B220, CD11b, CD11c, CD71, Ter119. Phenotypic HSPC populations were defined as LSK (Lin-ckit+Sca1+); MP (Lin-ckit+Sca1-); Pre-GM (Lin-ckit+Sca1-CD41-CD16/32-/loCD105-CD150-); GMP (Lin-ckit+Sca1-CD41-CD16/32+CD150-); Pre-Meg/E (Lin-ckit+Sca1-CD41-CD16/32-/loCD105-CD150+). All antibodies were purchased from BioLegend, BD Biosciences, or eBiosciences. Flow cytometry was performed using a 5-laser, 15-detector BD LSRII or LSRFortessa. Data analysis was performed using FlowJo (Tree Star, Ashland OR).

**RNA extraction, sequencing, and bioinformatics**

Peripheral blood samples were accrued for all timepoints and allocated to randomized batches for RNA extraction. Total RNA was extracted from whole blood using the AllPrep DNA/RNA Mini Kit (Qiagen, Hilden, Germany); quality and quantity were estimated using a BioAnalyser (Agilent, Santa Clara, CA). Samples with a RIN > 4.0 were included. External RNA Controls Consortium (ERCC) Spike-In Control Mix (Ambion, Foster City, CA) was added to all samples per the manufacturer’s recommendations, although these were not used for downstream analyses. Samples were allocated to randomized batches for library preparation, such that samples from each timepoint were distributed evenly over all sequencing runs. Sequencing libraries were constructed using the KapaHyper with RiboErase (Kapa Biosystems, Wilmington, MA), loaded on to a cBot system for cluster generation, and sequenced on a Hiseq 2500 (Illumina) with single end 51-bp for mRNA-seq to a nominal depth of 40 million reads. To mitigate batch effects, samples were randomly assigned to one of eight flow cells such that each flow cell contained a sample from at least one mouse and one timepoint. Image processing and base calling were conducted using Illumina's Real-Time Analysis pipeline.

Raw sequencing reads were processed with the nf-core RNASeq pipeline version 1.0 (2). Briefly, trimmed reads were mapped using Spliced Transcripts Alignment to a Reference (STAR)(3) to the GRCm38 reference amended with ERCC sequences and the human *MYH11* fusion gene sequence. Each library was subjected to extensive quality control, including estimation of library complexity, gene body coverage, and duplication rates, among other metrics detailed in the pipeline repository(2). Reads were counted across genomic features using Subread featureCounts(4) and merged in to a matrix of counts per gene for each sampled timepoint. *CM* fusion transcript expression was measured by counting reads that spanned the *CM* fusion boundary. Surrogate variable analysis was used to check for confounding experimental effects(5). None were apparent (data not shown); however, library preparation batch was used as a covariate in differential expression analyses using the Bioconductor package edgeR(6) as implemented in SARTools (3). The RNA-seq dataset is submitted and assigned accession number GSE133642.

**Estimation of critical points in the transcriptome state-space**

The PC-constructed transcriptome state-space allowed us to identify the position of the critical points in the state-space of the double-well quasi-potential and to predict the dynamics of the transcriptome-particle. Because our state-transition model has 3 critical points, we used K-means clustering to identify 3 clusters (labeled K1,K2,K3) along the leukemia axis (PC2) to estimate the locations of the critical points in our data (Figure S6A). The centroid of all control samples was used as the point \*. The centroids of the clusters K1 and K3 of the CM samples were used as estimates of the leukemia critical points , respectively. Because the critical point is unstable, it is unlikely to be correctly identified with a centroid approach. To estimate , we constructed quasi-potentials with values of which ranged from to (Figure S6B, left), and computed the Boltzmann Ratio (BR) for each quasi-potential (Figure S6B, right; we assumed the temperature is constant). The BR gives the proportion of mice predicted to be in either health or leukemic states over a long period of time at the steady-state distribution. We took to be the “upper” boundary (maximum value) of the cluster, which resulted in a BR of 81.4, corresponding most closely to the observed ratio of mice in the leukemia state at the experimental endpoint in the training dataset.

We remark that other clustering approaches were investigated, including hierarchical clustering which did not cluster samples by time or by condition (Figure S5) and that there may be multiple, equivalent ways to estimate the critical points. We found K-means to be the most simple and objective approach which ensured the identification of 3 distinct critical points.

**Construction of the leukemia transcriptome quasi-potential**

Mathematically, a double-well quasi-potential can be represented as a 4th degree polynomial . We choose a polynomial representation of the potential because it is mathematically simple and the odd powered terms allow for an asymmetric potential. The equation of motion of the particle in the double-well quasi-potential then takes the form of a Langevin stochastic differential equation where denotes the location of the particle at time *t* and is a Brownian motion that is uncorrelated in time , with being the Dirac delta function and is the diffusion coefficient.

Given the positions of critical points, the double-well quasi-potential could be constructed up to a constant of integration by evaluating where where is a scaling parameter and *x* is the position of the particle in the quasi-potential. The coefficients can be expressed in terms of by expanding and integrating with respect to *x*. State-transition theory predicts that the energy barrier—defined as the energetic difference between the initial state and the transition state—will be lowered by the expression of the *CM* gene, resulting in significantly increased probability and rate of transition from normal hematopoiesis to leukemia in *CM*-induced mice compared to control mice. The parameter a is a scaling factor and can be chosen arbitrarily. Because the state-space coordinates are on the scale of 102 (PC1 ranges from [-200 – 400], and PC2 ranges from [-300 to 100]) and the quasi-potential function is a quadratic polynomial producing values on the scale of , in order for the potential to have unit scaling, (1=100), the factor a must be on the scale of . We calculated  = 4.85 x 10-8 by fitting the diffusion coefficient in the Fokker-Planck equation to the training cohort data (Figure S12). We observed that the energy barrier ( was 0.99 (arbitrary units, A.U.) for *CM*-induced mice, nearly an order of magnitude lower than the energy barrier for control mice [6.45 (A.U.)].

**Validation study with independent cohorts**

To validate our state-transition mathematical model, state-space, and analytical approach, we collected PBMC RNA-sequencing data from two additional independent cohorts of control and CM mice. We collected validation (v) cohort 1 samples monthly for up to 6 months; (vControl1-7; vCM1-9) and collected validation cohort 2 samples sparsely at 3 randomly selected timepoints; (vControl8-9; vCM10-12) during leukemia progression. We performed PCA analysis of the validation cohort 1 and 2 data via singular value decomposition. The singular value decomposition is computed on the mean-centered data where represents the column mean so that (\* denotes the conjugate transpose). The columns of the unitary matrix *U,* not to be confused with the potential function , form an orthonormal basis for the sample space (i.e., the temporal dynamics of the transcriptome), the diagonal matrix contains the singular values, and the columns of the matrix *V\** correspond to the eigengenes(7) (see Table 1 glossary of terms), or loadings, of each gene in the transcriptome per PC.

PCA of the validation cohort data again demonstrated that the majority of the variance was encoded in the first 4 PCs (Figure S11A) and the leukemia-related variance was encoded in PC2 (Figure S11B,C). We then evaluated our ability to map state-transition trajectories and predict leukemia development in the validation cohorts by projecting the data from the validation cohorts into the state-space constructed using the training cohort. We accomplished this by using the eigengenes from the singular value decomposition of the training data as follows: a matrix of new data, , was projected into the state-space by multiplying by the matrix *V* from the training data, so that the coordinates of the new data in the reference state-space were given by . Because the matrices and *V* must have the same dimension, and more importantly the weights of the genes in the matrix *V* must match one-to-one with the genes in , the projection may use only genes in the intersection of and . Thus, we mapped the trajectory of each mouse in the validation cohorts in the PC2 space (Figure S11D). State-transition critical points were estimated using k-means clustering (k=3), using the same method that was used on the training cohort.

**Robustness of the leukemia transcriptome state-space construction**

The construction of the state-space using PCA was not sensitive to variations in data-normalization method, sample number, or gene selection criteria, as shown by bootstrap cross-validation (Figure S14). In fact, the geometry of the state-space could be inferred from time-series RNA-sequencing data derived from just 1 control mouse and 1 CM mouse (Figure S14B,C) and was not changed by the exclusion of the known leukemia genes *Kit* or *CM* (Figure S7C,D). There was no difference in state-space positions when the state-space was constructed without *Kit* or CM, up to machine precision, 2.2x10-16.These results demonstrate that PCA-based state-space construction is robust and reproducible regardless of variation in data-processing methods.

To evaluate the sensitivity of our state-space construction to variations in sample frequency, number of reads per gene, and number of timepoints, each of these quantities was varied independently (Figure S13). Sample frequency was assessed by varying the gene inclusion criteria from 5 counts in at least 2 samples, to 5 counts in each of 5, 10, 30, 50, 70, 90, 110, 120, and all 132 samples, with 5 counts in 2 samples being the most permissive criteria, resulting in 21,482 genes, to 5 in 132 samples being the most restrictive criteria, resulting in 8,995 genes. The number of reads per gene was assessed by varying the log of the counts per million reads (cpm) log2(cpm) threshold in increments of 0.01, 0.05, 0.5, 5, 1, 3, 5, 7, 10, 15, and 20 for each of the sampling frequencies, resulting in 100 combinations. For a subset of combinations, the effect of normalization methods [e.g., trimmed means of M values (TMM), relative log expression (RLE), upper quartile (UQ), and RLE with ComBat regularization] on the state-space was examined. Sampling frequency during leukemia evolution was assessed by performing a leave-“x”-out cross-validation technique, in which x=70 samples were randomly identified and removed from the data set. The remaining 62 samples were used to predict the positions in the state-space for the 70 removed samples. This cross-validation procedure was performed 100 times and the absolute difference between the true state-space position and the predicted position was computed.

**Leukemia eigengene selection**

Due to the large number of differentially expressed genes at the leukemic endpoint (c3), genes were filtered based on the angle subtended () by the gene in the 2-dimensional state-space. The range of angles () identified as being associated with leukemia evolution was defined as the minimal sector of a circle centered at (0,0) that included all leukemic samples as well as the mirror image of this sector along the x-axis of symmetry (Figure S10).

**References**

1. Cai Q, Jeannet R, Hua WK, Cook GJ, Zhang B, Qi J, et al. CBFβ-SMMHC creates aberrant megakaryocyte-erythroid progenitors prone to leukemia initiation in mice. Blood [Internet]. 2016;128:1503–15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27443289

2. Ewels P, Peltzer A, Moreno D, rfenouil, Garcia M, Panneerselvam S, et al. nf-core/rnaseq: nf-core/rnaseq version 1.0. 2018 [cited 2018 Dec 6]; Available from: https://zenodo.org/record/1400711

3. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

4. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.

5. Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Storey JD, et al. sva: Surrogate Variable Analysis. 2018.

6. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26:139–40.

7. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-Wide expression data processing and modeling. Proc Natl Acad Sci U S A. 2000;97:10101–6.