State-transition Modeling of Blood Transcriptome Predicts Disease Evolution and Treatment Response in Chronic Myeloid Leukemia (CML)

David E. Frankhouser[1^*], Russell C. Rockne[1^*], Lisa Uechi[1], Dandan Zhao[2], Sergio Branciamore[1], Denis O'Meally[3], Jihyun Izarriy[2], Lucy Ghoda[2], Haris Ali[2], Jeffery M. Trent[4], Stephen Forman[2], Yu-Hsuan Fu[2], Ya-Huei Kuo[2], Bin Zhang[2#], Guido Marcucci[2 #,^]

[1]Department of Computational and Quantitative Medicine, Beckman Research Institute, City of Hope National Medical Center, Duarte, California, 91010, USA

[2]Department of Hematologic Malignancies Translational Science, Beckman Research Institute and Division of Leukemia, City of Hope National Medical Center, Duarte, California, 91010, USA

[3]Department of Diabetes and & Cancer Discovery Science, Beckman Research Institute, City of Hope National Medical Center, Duarte, California, 91010, USA

[4]Translational Genomics Institute, Phoenix, Arizona, 85004, USA

^These first authors equally contributed

[#]These senior authors equally contributed

*Corresponding authors

David Frankhouser (dfrankhouser@coh.org)

Russell Rockne (rrockne@coh.org)

Guido Marcucci (gmarcucci@coh.org)

Keywords: CML; BCR-ABL; state-transition; critical points; transcriptome; nilotinib; TKI

Abstract

Chronic myeloid leukemia (CML) is initiated and initially maintained solely by the fusion gene BCR-ABL, encoding a multifaceted chimeric kinase targeted in the clinic with tyrosine kinase inhibitors (TKIs) TKIs are effective in inducing long-term remission, but are also frequently not curative. Thus, CML is an ideal system to test our hypothesis that transcriptome-based state-transition models accurately predict cancer evolution and treatment response. To test our hypothesis, we collected time-sequential blood samples from tetracycline-off (Tet-Off) BCR-ABL-inducible transgenic mice and wild-type controls. Using the time-series bulk RNA-seq analysis to capture a system-wide view of distinct disease states, we identified a single principal component that constructed a CML state-space with a three-well BCR-ABL leukemogenic potential landscape. The potential stable critical points defined observable disease states. Early states were characterized by anti-CML genes opposing leukemic transformation; late states were characterized by pro-CML genes. Genes with expression patterns shaped similarly to the potential landscape were identified as disease transition drivers. Re-introduction of tetracyclines to silence the BCR/ABL gene returned diseased mice transcriptomes to a stable state near to health, without reaching it, suggesting partly irreversible transformation changes. TKI treatment only reverted the diseased mice transcriptomes to an earlier disease state, without approaching health; disease relapse occurred soon after treatment completion. Using only the earliest time-point as initial conditions, our parametrized state-transition models accurately predicted both disease progression and treatment response, supporting this as a potentially valuable approach to time clinical intervention even before phenotypic changes become detectable.

Introduction

Chronic myeloid leukemia (CML) is characterized at the cytogenetic level by t(9;22), the Philadelphia chromosome that creates BCR-ABL[1]. This fusion gene encodes a ligand-free

activated mutant tyrosine kinase that transforms normal hematopoietic stem cells into leukemia stem cells (LSCs), primitive leukemic cells capable of initiating and maintaining the disease. While the advent of tyrosine kinase inhibitors (TKIs) has revolutionized the treatment of CML by inducing long-term remissions and largely preventing disease evolution from a chronic phase (CP) into blast crisis (BC), these drugs often fail to fully eradicate LSCs[2,3]. Lack of treatment compliance, drug intolerance or acquired BCR-ABL mutations or additional genomic "hits", therefore may result in disease relapse or progression to BC[4–6]. Thus, while for a subset of patients is possible to eventually discontinue TKIs, most of them remain on lifetime treatment[7,8].

Quantification of BCR/ABL fusion transcript levels is diagnostic and effective in assess ongoing treatment response but has been less useful to predict timely disease evolution at the earliest time points. In recent years, the transcriptome has emerged as a promising tool for assessing treatment response and predicting outcome in cancer and leukemia[9,10]. Gene expression profiling using high-throughput technologies has enabled comprehensive analysis of the transcriptome in CML patients and identified gene signatures that reveal molecular mechanisms of transformation and treatment response[11,12]. Recently, Giustacchini et al. reported on a distinct molecular signature of LSCs that selectively persisted after TKI treatment[13]. Ross et al. investigated the transcriptomic landscape of CML patients at diagnosis and during TKI therapy and identified differentially expressed genes and pathways associated with cell proliferation, apoptosis, and drug resistance, thereby predicting treatment response and long-term outcomes[14]. Radich et al. explored transcriptomic changes in CML patients who discontinued TKI therapy and found that specific gene expression patterns, particularly involving immune-related genes, could distinguish patients who successfully maintained treatment-free remission from those who relapsed[15,16]. However, these studies are mainly population-based, and are therefore difficult to apply to predict disease evolution and treatment outcome for individual patients. To our knowledge, no recent studies have examined how transcriptomic changes detected at the earliest time points of disease

or at the start of treatment can accurately predict outcome in individual CML mouse or human, even before phenotypic changes had occurred.

Here, we applied state-transition theory to model temporal dynamics of the CML transcriptome and predict CML evolution and treatment response[9,10]. CML is an ideal model to test this hypothesis as the disease is initiated and initially maintained solely by BCR-ABL, targeted in the clinic by TKIs, a treatment effective in inducing disease remission, but often not curative. We showed that the transcriptome of peripheral blood mononuclear cells (PBMC) from a transgenic CML mouse can be modeled as a particle undergoing state-transition in a CML-transformed potential landscape. Using this model, we accurately predicted the trajectories of disease evolution, treatment response, and relapse of individual mice at the earliest time-point. Furthermore, the state-transition framework allowed for identification of the patterns of gene expression that drove leukemogenesis, thereby connecting transcriptome state-transition to molecular mechanisms of disease growth and treatment resistance.

Results

Peripheral blood transcriptome state-transition during CML development

To investigate the molecular mechanisms of CML initiation and evolution, we collected PBMC samples from the CP CML mice at sequential time points after BCR-ABL induction by tetracycline discontinuation (Tet-off; Fig. 1A). Samples were also collected from control mice that continued to receive tetracycline (Tet-on) in order to maintain BCR-ABL suppression. Bulk RNA-sequencing (RNA-seq) was performed on each PBMC sample. To study the transcriptomic changes that characterized disease evolution, we applied a state-transition approach that mapped the time-series transcriptomes of each mouse into a CML state-space. The resulting trajectories of each

mouse in the state-space were modeled using a stochastic equation of motion to determine whether the disease evolution of each mouse could be predicted based on its initial condition.

To start, the CML state-space was built using Principal Component Analysis (PCA). Briefly, we performed singular value decomposition[9,10] (SVD; see methods) on all expressed genes measured with RNA-seq of the sequentially collected samples of both control and CML mice. By including all the time-sequential samples collected from the CML and control mice, we were able to capture the continuum of transcriptional states that spanned from health to terminal states but also included intermediate states of the disease. Among all PCs, PC2 (hereafter called CML state-space) provided the greatest separation between samples collected from the healthy control mice and those collected from CML moribund mice. We obtained transcriptome trajectories of each individual mouse by plotting the time sequential samples in the CML state-space (Fig. 1B, left panel).

Supporting the ability of the CML state-space to identify distinct disease states, we demonstrated that among all the PCs, PC2 had the best correlation with BCR-ABL expression levels. Nevertheless, the correlation was relatively weak (R^2 = 0.48; Fig. 1B, right panel; Table S1), suggesting that BCR-ABL was not the only driver of transcriptome state-transition. In fact, immediately after tetracycline discontinuation (Tet-off), we observed that the CML mice began to move through the CML state-space, from the health state toward disease. Despite BCR-ABL expression not yet being detectable, it likely influenced disease state changes even at those lowest, initial levels (Fig. 1C).

CML transcriptome state-transition model predicts disease evolution

As the CML state-space contained the continuum of the transcriptomic states from health to disease, we used this information to build a state-transition model that predicted the disease

trajectories of individual mice. These predictions were made by setting initial conditions for each mouse using only the data from the first collected time point. The theoretical framework of this model postulated that changes in the transcriptome could be modeled as a particle undergoing Brownian motion in a potential energy landscape using the Langevin equation [9,10]. Thus, the Fokker-Planck (FP) solution of the equation of motion was used to predict disease state trajectories across the CML state-space.

To understand the system dynamics, we hypothesized that the CML potential energy landscape could be modeled using a three-node network that resulted in a tri-stable system (Fig. 2A).[17] In this system, BCR-ABL was the input signal that interacted with the PBMC cell population. The changing PBMCs and the consequent transcriptome variation created a double negative feedback loop representing how leukemogenesis dysregulated the transcriptome and how the dysregulated altered transcriptome affected the composition (normal vs leukemic) of the cell population. The transcriptome, which was represented by a CML state-space variable, was modeled as self-activating progressive dysregulation of gene networks during leukemia progression[18]. The potential landscape resulting from this system ranges from a single well centered at the health state (BCR-ABL off) to three wells representing the most energetically favorable, CML steady states (BCR-ABL maximum; Fig. 2B). The equations describing the dynamic configuration of the CML potential resulting from this three-node network are described in the Methods.

To test if the theoretically-derived three-node network leukemogenic potential reliably recapitulated the experimental observations, we considered the empirical distribution of the CML mouse samples in the state-space and derived an experimental CML potential (Fig. 2C; see Methods). The derived CML potential was three-welled and with five critical points identified from the kernel density estimator of the empirical distribution (denoted $c_1 - c_5$). We labeled the stable (minima) critical points as Early state (Es) at c_1, Transition state (Ts) at c_3, and Late state (Ls) at

c_5 (Fig. 2C). The unstable (maxima) critical points were labeled as Early-Transition state (T-Es) at c_2 and Transition-Late state (T-Ls) at c_4; these unstable points represent the boundaries between two stable states. We also used the sample density of the control mice to determine the location of a Health state (Hs) labeled c_h as the single stationary state and critical point. The samples from the control mice remained confined to a region of the CML state-space corresponding to the Health state (Hs) c_h. Importantly, the three-wells of the experimentally derived potential directly mapped into those of the theoretically predicted potential (Fig. 2B-C).

To predict time-to-disease [i.e., the amount of time that elapsed before the mouse crossed the boundary of the Ls (c_5)], we modeled the location and trajectory of each sample's transcriptome in the state-space using the CML potential and the Langevin equation of motion. We were able to generate predicted transcriptome trajectories that matched the observed ones using only the data from samples collected at the first time point (Fig. 2D). The Fokker-Planck (FP) solution of the Langevin equation, when solved forward in time, accurately predicted time-to-disease (Fig. 2E). As all the CML mice died after passing Ls (c_5), we also compared the time-to-disease (defined as the first crossing of the Ls (c_5) critical point) between the observed and the predicted trajectories. The predicted and observed time-to-disease curves overlapped and were not statistically different, demonstrating accuracy of the state-transition model's prediction ($p = 0.8$, Log-rank; Fig. 2F). Of note, one CML mouse never developed disease after Tet-off and remained healthy (Fig. 2D). Our model predicted that this mouse would not cross Ls (c_5), thereby accurately reproducing the experimental observation.

State-transition critical points of BCR-ABL potential landscape defined distinct disease states

In order to identify the transcriptomic drivers of CML state-transition, we reasoned that samples from different mice must be compared at equivalent disease states. Since mice developed CML at different times, we could not simply rely on grouping transcriptome samples by their collection time point. Thus, we utilized the samples' positions in the state-space to group them into distinct disease states. We then compared gene expression at each disease critical point disease [i.e., Es (c_1), Ts (c_3) and Ls (c_5)] with that of control samples at health critical point [i.e., Hs (c_h); Fig. 3A; Fig. S3A; Table S2].

A total of 78 differentially expressed genes (DEGs) were identified at Es (c_1) as compared to Hs (c_h). A large proportion of DEGs (55 out of 78 total DEGs) observed at the Es (c_1) state were unique to this state and were not identified in the other disease states (Fig. 3B). Among the most upregulated genes, were Ins1 and Ins2 that encode insulin, in addition to Fam177a1 and several SnRNAs (Table S2). Fam177a1 encodes a protein localized to the Golgi complex and endoplasmic reticulum but of unknown function, and when mutated, is associated with developmental and neurological disorders[19]. Of note, BCR-ABL was not among the upregulated DEGs at the Es (c_1). Among the most downregulated DEGs, there were several genes encoding for collagen and elastin. Compared to Hs (c_h), the transition state Ts (c_3) presented with 366 DEGs (Fig. 3B; Table S2). In addition to BCR-ABL, among the most upregulated DEGs, was Prrt4, encoding for a protein predicted to be an integral component of the cell membrane, and the lncRNA Gm6093, predicted to play a role in cell differentiation[20,21]. Among the most downregulated DEGs, we observed collagen and elastin encoding genes, in addition to Il33. Compared to Hs (c_h), Ls (c_5) presented with 1860 DEGs (Fig. 3B; Table S2). In addition to BCR-ABL, among the most upregulated DEGs, there were Pgk1-rs7, involved in glycolysis, and Prrt4. Among the most downregulated DEG, there were several pseudogenes and other genes coding

proteins with unknown function. Of note, 94% of DEGs identified at Ts (c_3) were also DEGs and expressed at higher levels at Ls (c_5) (Fig. 3B).

Using gene set enrichment analysis (GSEA), we identified Hallmark gene sets enriched at distinct disease states (Fig. 3C). Compared with Hs (c_h), we observed that all significantly enriched gene sets were downregulated at Es (c_1). These included Epithelial Mesenchymal Transition (EMT), Apical Junction, and Myogenesis, that have been previously associated with malignant transformation[22–24]. In contrast, compared with Hs (c_h), all significantly enriched gene sets were upregulated in the Ts (c_3) and Ls (c_5) (Fig. 3C). Among the most upregulated gene sets, we found those related to metabolism, angiogenesis, and inflammation. Of note, while all enriched gene sets at the Ts (c_3) and Ls (c_5) were upregulated with respect to Hs (c_h), we observed that these genes sets had higher expression at Ls (c_5) compared with Es (c_1), with the exception of Pancreas Beta Cells which was downregulated (Fig. S3B).

Eigengene analysis quantifies DEG contribution to CML development.

To quantify the contribution of each DEG to CML development, we used the CML state-space loading value of each gene which we defined as the eigengenes[25]. Located in the second column of the matrix V^* from the SVD procedure that we used to construct the CML state-space (Table S2), the eigengenes described the direction and the magnitude of the effect that each gene had on the sample's location in the state-space. Thus, by combining the eigengenes with the expression changes at each disease critical point with Hs (c_h), we quantified the impact of DEGs on disease onset and evolution and categorized them as pro- or anti-CM eigengenes. A pro-CML eigengene had either a positive loading value with downregulated expression or a negative loading value with upregulated expression; an anti-CML eigengene had either a positive loading value with upregulated expression or a negative loading value with downregulated expression

(Fig. 3D; S4A). The eigengenes were then visualized in a 2D space constructed by plotting the CML contribution vs the PC1 loading value (Fig. S4A). Of note, since PC1 was not directly associated with BCR-ABL expression or CML development, its selection was made simply to map the CML contribution; any other PC could have been used instead.

The net contribution to CML of the Es (c_1), Ts (c_3) or Ls (c_5) DEGs was calculated as the vector mean of the vectors representing individual DEGs (Fig. 3D; Fig. S4B). The DEGs at Es (c_1) had a net anti-CML contribution, while those at Ts (c_3) and Ls (c_5) had a net pro-CML contribution. We also calculated the CML contribution of the significantly enriched gene sets as a mean vector of all the eigengenes in a given enriched gene set. All the Es (c_1) enriched gene sets had a net anti-CML contribution (n=13), whereas all the Ts (c_3) and Ls (c_5) enriched gene sets had a net pro-CML contribution (n=17 and 26 respectively; Fig. 3E; Fig. S4C).

To assess whether the eigengene-based contributions corresponded to phenotypic differences between disease states, we compared myeloid counts at Es (c_1), Ts (c_3), and Ls (c_5) states. We observed that no myeloid cell growth at Es (c_1); this was followed by an initial expansion at Ts (c_3) and further growth at Ls (c_5) (Fig. S3D). The growth of the myeloid population between Ts (c_3), and Ls (c_5) was associated with an expansion of the DEGs at Ls (c_5) and an increase in enrichment in many of the shared processes (Fig. 3B; Fig. S3B). GSEA indicated these genes to be members of metabolism- and inflammation-related gene sets. In addition, 405 downregulated DEGs were unique to the Ls (c_5) state (Fig. S3C; Table S2). However, over half of the downregulated DEGs were unnamed genes and failed to be significantly enriched in any pathways (265 of 405). Nevertheless, the eigengene analysis showed that all but one of the downregulated Ls (c_5) DEGs had a pro-CML contribution.

Gene expression dynamics identify drivers of CML state-transition

While we used the state-space critical points as a classifier of distinct disease states, the continuous state-space coordinates could also be used to align transcriptomes by their progress toward CML. By doing so, we created a pseudotime ordering of the disease, which allowed us to identify the patterns of gene expression involved in CML transformation. To this end, we used a correlation-based analysis to identify subsets of genes with similar patterns of expression during disease evolution (hereafter called gene "modules"; Fig. S5). At Es (c_1), we found no gene modules, meaning that the DEGs identified at Es (c_1) were not co-regulated at subsequent states (Fig. 4A, left; Fig. S5A). At Ts (c_3), we identified an "increasing module", meaning that these DEGs were co-regulated and continued to increase during disease progression (Fig. 4A, middle; Fig. S5B). At Ls (c_5), we identified two gene "modules", one increasing and one decreasing (Fig. 4A, right; Fig. S5C).

Different from the increasing gene module observed at Ts (c_3), the increasing gene module at Ls (c_5) had non-linear dynamics as it presented an inflection point near T-Ls (c_4) (Fig. 4A). The inflection represented the point where the rapid increase of DEG expression started to slow down and became constant. The "decreasing" Ls (c_5) gene module also had a non-linear expression curve, with an inflection occurring at approximately T-Ls (c_4) and continued through Ls (c_5) (Fig. 4A). It is important to emphasize that the expression dynamics of the gene modules were not derived from any critical point information; rather were derived independently, using only the mean expression of the DEGs and the CML state-space location of each transcriptome sample. Thus, it was striking to us that the inflections of the gene modules occurred almost exactly at T-Ls (c_4), an unstable critical point between Ts (c_3), and Ls (c_5). The observation of co-regulation at the unstable critical points led us to hypothesize that these types of coordinated expression changes

could break the equilibrium of a stable critical point and drive the disease transition from that stable state of the disease to the next one.

To identify the driver genes of disease transition between two otherwise stable states, we correlated the expression of individual DEGs with the shape of the CML potential (Fig. 4B-C; Fig. S6A). For this analysis, we included all DEGs (n=3,304) resulting from the comparisons of the stable disease states with the health state (Hs). We identified 194 DEGs at the T-Es (c_2) state and 497 at the T-Ls (c_4) state whose expression followed the CML potential (Fig. S6B). All but one of the T-Es (c_2) genes and all T-Ls (c_4) genes had a pro-CML contribution, which supports the hypothesis that these genes drove the CML state-transition. Interestingly, only 23 DEGs were identified as being involved in both T-Es (c_2) and T-Ls (c_4) transitions, suggesting that unique and distinct processes likely drove disease evolution from across the distinct disease states (Fig. S6B).

To investigate which cellular processes were involved in the transition between two stable states [i.e., Es (c_1) to Ts (c_3) and Ts (c_3) to Ls (c_5)], we constructed protein-protein interaction networks and performed pathway enrichment analysis (Fig. S6C). Both T-Es (c_2) and T-Ls (c_4) were significantly enriched for protein interactions ($p<0.0001$; 31 and 232 interacting proteins respectively). The T-Es (c_2) transition point was characterized by inflammation and immune related processes and involved multiple collagen genes (Fig. 4B; Table S3). T-Ls (c_4) transition point was characterized by metabolism-related processes in addition to inflammation- and immune-related processes also enriched in T-Es (c_2) (Fig. 4C, S6C; Table S3). All enriched gene sets for both T-Es (c_2) and T-Ls (c_4) had a pro-CML contribution, supporting their role in driving disease evolution.

State-space transition predicts treatment response

To investigate how treatment impacted CML state-transition, we performed time-series experiments using two approaches (Fig. 1A, 5A, S7A). The first approach made use of the CML mouse model's ability to induce and eventually suppress BCR-ABL expression upon tetracycline discontinuation (Tet-off) and subsequent re-administration (Tet-on); hence we referred to this treatment as Tet-off-Tet-on (TOTO)]. The TOTO approach represented a hypothetical best scenario where a treatment is able to completely suppress BCR-ABL expression (Fig. S7A). The second approach consisted of treatment with nilotinib, a TKI used in clinical practice (hereafter referred to as "TKI"). TKI was assumed to be less effective than TOTO as, rather than suppressing BCR-ABL expression, it inhibited BCR-ABL at the protein level. Both Tet-on and TKI treatments were initiated six weeks after Tet-off. In the TOTO group, Tet-on was continued for 12-weeks. In the TKI cohort, TKI was administered for 4 weeks, and the mice were followed for the subsequent 12 weeks, for a total of 18-weeks or until they succumbed to disease, whichever occurred first.

To determine the effect of each therapy on CML state-transition, the transcriptome dynamics of mice receiving one of the two treatments were compared with those of the untreated diseased and healthy mice at each of the disease [(Es (c_1), Ts (c_3), Ls (c_5)] or health Hs (c_h) states respectively; Fig. 5B, S7B; Table S4). In order to assess post-treatment CML disease states, we used the eigengenes to project the TOTO and TKI samples into the CML state-space (see Methods). Supported by an observed relationship between the number of DEGs and the mean location of sample groups in the state-space, we used both the number of DEGs and the state-space to assess transcriptomic similarities or differences (Fig. S7C). We defined treatment response as movement of the transcriptome in the state-space away from the CML LS (c_5) state and toward Hs (c_h), with a complete response being a return to health [Hs (c_h)].

When disease states of TOTO and TKI treated mice were compared with those of controls, we observed significantly different results. In the TOTO cohort, we observed a rapid movement from LS (c_5) toward Hs (c_h) in the state-space soon after tetracycline was re-administered. Post-treatment TOTO samples were tightly distributed around the c_2 critical point, which represented a near-health, newly created stable state in the treatment transformed potential landscape (Fig. 5A). When treated mice at c2 were compared to the controls, few DEGs were identified (Fig. 5A-B). The observed DEGs and enriched gene sets that were detected by comparing TOTO mice to healthy controls were similar to those observed as time-related effect in the control mice alone and therefore, were possibly associated with aging (Fig. S7D). However, after subtracting out the time-related DEGs, 38 DEGs remained and had a small pro-CML contribution (Fig. 5C), indicating that once BCR-ABL transformation had occurred, the transcriptome maintained some leukemic fingerprints even if BCR-ABL expression is completely suppressed.

Different from TOTO, TKI treatment moved the mouse transcriptomes more slowly from the initial Ls (c_5) state. These transcriptomes reverted only to Ts (c_3) and none of these mice achieved the response of the TOTO mice (Fig. 5A). When compared to controls, the TKI treated mice had 1269 DEGs (Fig. 5B), and a net pro-CML contribution with BCR-ABL the top pro-CML DEG (Fig. 5D). Using GSEA to compare TKI treatment with healthy controls, gene sets that were upregulated during CML remained upregulated during TKI treatment (Fig. 5E) and had a overall pro-CML contribution (Fig. S7F). Of note, BCR-ABL was the top ranked pro-CML DEG.

Modeling the effects of TOTO and TKI on the BCR-ABL potential

We next tested whether treatment could alter the potential landscape and, in turn, whether the state-transition model could predict treatment response at the earliest time point after treatment initiation. We started by identifying which parameters should be perturbed to recapitulate the state-transition observed with each of the treatments. To model TOTO treatment, we set the BCR-

ABL signal to zero ($S = 0$) and introduced a treatment response parameter γ_{TOTO} which modulated the action of the signal on the transcriptome as $g^*_{ab} = \gamma_{TOTO}\ g_{A,S-B+}$ with $\gamma_{TOT0} = 7.5$ (Fig. 6A left). This resulted in a potential with a new steady state occurring near the T-Es (c_2) state, the exact location in the state-space where the TOTO mice mapped post-treatment (Fig 6A right). We observed that both a decrease in BCR-ABL expression and an increase in the transcriptome sensitivity to BCR-ABL were necessary to derive a post-treatment potential that matched the observed TOTO trajectories. Once BCR-ABL signal was set to zero, the modulation of the single parameter g^*_{ab} was sufficient to transform a pre-treatment tri-stable CML potential ($V_{CML}(x)$) into a system with only one stable state corresponding to c_2 following TOTO treatment ($V_{TOTO}(x)$). Using the TOTO potential ($V_{TOTO}(x)$), we then produced simulated TOTO trajectories that matched those observed (Fig. 6B; Fig. S8).

To model TKI treatment, we reduced the BCR-ABL signal (BCR-ABL > 0) and introduced a treatment response parameter γ_{TKI} (Fig. 6C left). In contrast to TOTO, the post TKI potential was produced only when both the BCR-ABL signal was reduced ($S < S_{max}$) and the response rate between the PBMCs and the BCR-ABL signal was increased, $g^*_{as} = \gamma_{TKI}\ g_{A,S+B-}$, with $\gamma_{TKI} = 1.4$. The change in this single parameter was sufficient to transform the tri-stable CML potential ($V_{CML}(x)$) into a bi-stable potential with new stable states at c_1 and c_3 following TKI treatment ($V_{TKI}(x)$). Importantly, the TKI model predicted that the only way TKI treated mice could return to Hs (c_h) was if the BCR-ABL level approached zero (Fig. 6C right). When we used this potential to simulate TKI treatment, we again observed an accurate prediction of the observed transcriptome trajectories (Fig. 6D; Fig. S8).

DISCUSSION

We have previously reported that state-transition models can be used to predict time-sequential transcriptome dynamics and disease evolution in biological systems such as murine acute myeloid leukemia (AML).[9,10,18] Constructing a state-space to model biological transitions can capture changes produced by genetic, epigenetic, microenvironmental perturbations on the transcriptome, which simultaneously deregulate multiple biological processes in order to generate a leukemic phenotype. Using a transgenic CP CML mouse model that recapitulated the human disease, we report here the application of state-transition theory to interpret the temporal transcriptomic changes involved both in CML development and in treatment (where CML is "reversed") to accurately predict outcome and treatment response.

We demonstrate the feasibility of predicting state-transition dynamics and time to disease using the time-sequential transcriptomic data collected from a murine model of CML. Our results show that the movement of the blood transcriptome in the state-space could be modeled as that of a particle undergoing Brownian motion in a leukemogenic potential defined by stable and unstable critical points. We hypothesized that BCR-ABL creates a leukemogenic potential that arises due to the action of BCR-ABL on both the PBMCs and the transcriptome. We also derived a CML potential landscape based on the position of the critical points derived from the observed distribution of the CML mouse transcriptomes in the state-space. The theoretically and experimentally derived landscape potentials were qualitatively identical and could be used to predict the trajectory of disease development over time. We showed that at the earliest time points, the CML state-space was more sensitive at detecting the disease state transition than the BCR-ABL expression itself. Furthermore, we showed that our state-transition model predicted transcriptome trajectories that matched well the observed ones utilizing only the transcriptomic state derived from data collected at the first time point after Tef-off BCR/ABL induction.

The biological features that characterized the distinct disease states were inferred from the analysis of DEGs obtained by comparing gene expression at each of the CML critical points with the control health state. Because the SVD decomposed the gene expression into linear combinations of orthogonal basis vectors, we were able to assess the contribution of individual genes, DEGs, or entire Hallmark gene sets to CML initiation and growth. We observed that the Es (c_1) state was largely characterized by anti-CML DEGs that resisted CML development whereas the Ts (c_3) and Ls (c_5) disease states were characterized by pro-CML DEGs associated with the movement of state-transition toward late disease state. In other words, the anti-CML contributions present only at Es (c_1), were suggestive that at this early state of the disease, the transcriptome changes opposed BCR-ABL-driven disease development, whereas the pro-CML contribution as observed at Ts (c_3) and subsequently at Ls (c_5) suggested that transcriptome changes at these points overwhelmed any initial anti-leukemic force and contribute to disease evolution.

Interestingly, while we noted that genes involved in inflammation and angiogenesis were enriched at both Ts (c_3) and Ls (c_5), we also observed upregulation of genes involved both in glycolysis and oxidative metabolisms at Ls (c_5), suggesting an increase in the bioenergetic demand in the state of highest growth[5,26,27]. Studying the changes in the expression of individual DEGs, we also discovered that groups of co-regulated genes (i.e., gene modules) that increased or decreased concurrently. Some of these gene modules displayed non-linear expression trends with inflection points that mirrored the shape of the CML potential at the unstable critical points Es (c_2) and Ts (c_4). Therefore, we hypothesized that they were likely the major drivers of transition between two stable disease states. Among others, metabolic processes were involved in the transition between Ts (c_3) and Ls (c_5), suggesting that the increased metabolic demand was one of the possible

causes of the transition to the final state of the disease, when maximum leukemic growth was observed.

Importantly, we also applied the state-transition model to predict response to therapy. To this end, first we treated CML mice in Ls (c_5) with TOTO, which resulted in complete suppression of BCR-ABL. We showed that TOTO brought the system rapidly back close to Es (c_1), without reaching it. In fact, TOTO-treated mice remained at a new stable critical point TOTO (c_2), with a small number of pro-CML DEGs. Thus, it is likely that once the system is altered, some fingerprints of BCR-ABL transformation remain, despite complete suppression of the fusion gene. Importantly, using only the initial time point and treatment start date, our potential model accurately predicted treatment response and post-treatment dynamics.

In contrast, TKI treatment returned the transcriptome only to Ts (c_3). Once TKI treatment was completed, the transcriptome rapidly returned to Ls (c_5) and mice had disease relapse. This mimics a possible trend observed in those patients that are either not compliant to the treatment and need to stop due to intolerable side-effects or have developed TKI resistant disease that evolves to a blastic phase. While the model also accurately predicted TKI trajectories, it also predicted that with administration of higher dose or more effective TKIs, the system may potentially return to the healthy state, Hs (c_h). However, the optimal dosage of TKI compatible with the return to health may not be attainable in vivo[28]. Although the treatment response parameters between TOTO and TKI may not be directly compared, it is interesting that the treatment response effect for TOTO (γ_{TOTO}) was higher than that for TKI (γ_{TKI}) suggesting that the suppression of BCR-ABL gene with tetracycline had a greater positive impact on the transcriptome than BCR/ABL protein inhibition by TKI.

In summary, we show here that state-transition dynamical models can predict disease evolution and treatment response in a murine model that recapitulates human disease. It is possible therefore that this approach could be scaled up and translated into the clinic as a assay to predict disease evolution and treatment response at the earliest time points, even before phenotypic changes will occur.

METHODS

Mouse model

Inducible transgenic SCLtTA/BCR-ABL mice of B6 background were maintained on tetracycline (tet) water at 0.5 g/L. Tet withdrawal (tet-off) results in expression of BCR-ABL and generation of a CML-like disease in these mice with a median survival of approximately 10 weeks after induction of BCR-ABL, which recapitulates human CP CML. Six to eight weeks old male and female SCLtTA/BCR-ABL mice were randomly divided into 4 groups: control (Tet on, n=3 mice), CML (Tet off, n=6 mice), Tet Off-On (TOTO; Tet off 6 weeks then Tet on for 12 weeks, n=4 mice), TKI [Tet off 6 weeks then TKI (Nilotinib, 50mg/kg, oral gavage, daily) treatment 4 weeks, n=7 mice]. Blood was collected from these mice weekly, i.e., before BCR-ABL induction (t = 0) and weekly after induction up to 18 weeks (t = 1 to 18) or when the mouse developed leukemia and became moribund, whichever event occurred first.

RNA-seq sequencing, data processing and analysis

For each mouse, peripheral blood samples were accrued for all timepoints before 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. 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 Novaseq 6000 (Illumina) with paired end 100-bp for mRNA-seq to a nominal depth of 40 million reads. To mitigate batch effects, samples were assigned to flow cells such that treatments were approximately evenly represented across runs and samples from the same mouse shared a run. 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 3.7[29]. Briefly, trimmed reads were mapped using STAR to the GRCm39 reference (Genbank accession GCA_000001635.9) amended with the human *BCR-ABL1* fusion gene sequence (Genbank accession EF158045.1) using GENCODE annotation Release m28[30]. 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[29]. Transcript abundance was estimated across genomic features using Salmon and merged into a matrix of pseudocounts per million transcripts (TPM) per gene for each sample[31]. *BCR-ABL1* transcript abundance was measured by counting reads that spanned the fusion boundary. Surrogate variable analysis was used to check for confounding experimental effects[32]. None was apparent (data not shown); however, for all comparisons, sex was used as a covariate in differential expression analyses using the default setting of DESeq2 and counts generated by tximport using the Salmon abundance estimation[33]. The RNA-seq dataset is submitted to the Gene Expression Omnibus and assigned accession number GSE244990.

State-space construction

To construct the CML state-space, we performed singular value decomposition (SVD) on all expressed genes in the transcriptome. By including all time points of the time-series peripheral blood samples in both the control and the BCR-ABL expressing CML mice, we both captured the

full transcriptional system and spanned all intermediate disease states (Fig. 1B). The SVD is a dimensionality reduction technique that produces low rank approximations of the input data, called principal components (PCs), that capture sources of variation in the original data set (Fig. S1B). Among the principal components, we found an axis on the PC1 vs PC2 plane to be the component with the best separation between the control and CML endpoint samples. To align this axis with the y-axis, we found the rotation angle that made the linear fit line of all control samples parallel to the x-axis. After applying a 37.8° rotation to the PC1 vs PC2 plane, we identified the rotated PC2 to be the principal component that best encoded CML development. We determined this based on two criteria: 1) it had the largest separation between the control samples and the endpoint samples for the CML mice; and 2) it demonstrated the best correlation with both BCR-ABL expression (abundance and log-normalized) and the myeloid population in the CML mice (Fig. S1A; Table S1). Therefore, we defined PC2 as the CML state-space. When we plotted PC2 against time for all samples, the CML state-space provided disease trajectories for each mouse (Fig. 1B). Although the correlation with BCR-ABL was by far the best for the CML state-space compared to all other PCs, the correlation was low (R^2 = 0.48; Table S1).

Using the expression of the same genes used when performing SVD to construct the state-space, the projection of the TOTO and TKI treatment cohorts was performed as follows: the treatment expression values X_T, were multiplied by the state-space loading values, V, and the inverse of the eigenvalues, Σ^{-1} (Fig. 3B). The second column of the resulting treatment cohort PC matrix, U_T, represented the coordinates of the treatment samples in the CML state-space.

Mathematical model

The three compartment mechanistic model used to generate the CML, TOTO, and TKI potentials is adapted from Dey and Barik 2021.[17] Using the double negative feedback loop B (DNFL-B)

circuit framework, BCR-ABL concentration is modeled as the signal (S) acting on PBMC cells (A) and the corresponding observable variation in the transcriptome (B), as

$$\frac{dA}{dt} = G_A(S,B) - k_A A$$

$$\frac{dB}{dt} = G_B(A,B) - k_B B$$

where $G_A(S,B)$ and $G_B(A,B)$ are rate functions relating the concentration of the signal to the cell and transcriptional states as

$$G_A(S,B) = g_{A,S-B-}H^-(S)H^-(B) + g_{A,S+B-}H^+(S)H^-(B)$$

$$+g_{A,S-B+}H^-(S)H^+(B) + g_{A,S+B+}H^+(S)H^+(B)$$

and similarly for $G_B(A,B)$, see Dey and Barik 2021. The functions H^- and H^+ are switching Hill functions given by

$$H^-(x) = \frac{1}{1 + \left(\frac{x}{x_0}\right)^n}$$

and $H^+(x) = 1 - H^-(x)$. The parameters are as stated in Dey and Barik Table S2 for the DNFL-B circuit model.

The effective potential $V(x)$ for the system is given by integrating the effective force of the signal (S) on the cells (A) and observable transcriptome (B) as

$$F(B,S) = G_B(G'_A(S,B),B) - k_B B$$

Where G'_A is the steady state rate of A $\frac{dA}{dt} = 0$, so that

$$V(B,S) = -\int_0^B [G_B(G'_A(S,x),x) - k_B x]dx$$

is a function of the state variable B, which in our case is the state-space coordinate corresponding to PC2, which parametrically depends on the signal, or BCR-ABL expression level S.

Mean squared displacement analysis

Mean squared displacement (MSD) was calculated for each mouse from their CML state-space trajectories vs time (Fig. S2A). The MSDs of each experimental group was fit to a linear model and the slope of the line was used to estimate the diffusion coefficient (β) used in the state-transition model (Fig. S2B).

The MSD at a time t is given as $MSD = \langle |x(t) - x_m|^2 \rangle$, where $x(t)$ is the position of trajectory from the simulation by a treatment model and x_m is the CML state-space trajectory of the mouse model. The MSD was calculated for each mouse for each time point t, and the average MSD for all time points was evaluated by changing the coefficient of treatment force γ and post-treatment diffusion coefficients β. The MSDs were examined within the range of γ for $0.1 \leq \gamma \leq 2.0$ and β for $0.001 \leq \beta \leq 0.1$, and the combination of values of γ and β that has the minimum MSD was selected. The search ranges of parameter, β, was empirically estimated from the average slope of linear regression to CML and control mouse data. The upper bound of γ was determined by the convergence of trajectories of the simulation.

DEG expression dynamics

Expression dynamics were investigated by performing hierarchical clustering on the correlation matrixes produced using all CML mouse time points for each of the unique Early, Transition, and Late DEGs. Genes with similar expression dynamics (gene modules) were determined by selecting cutting the dendrogram to obtain clusters and then selecting for clusters that had a median correlation coefficient greater than 0.25 (Fig. S5A-C). For each gene module, the expression dynamic plots were created by calculating the mean of the log-normalized mean-centered expression for each CML sample and then plotting that value vs the CML state-space coordinate of the sample (Fig. 4A).

Transition point driver genes

To identify which genes had expression dynamics that were most similar to the inflection points of the CML potential, we used the following approach (Fig. S6A). First, we included all DEGs obtained by comparing CML disease state vs both healthy (c_h) and other CML disease states in the analysis. Second, we defined each transition point by selected the CML samples that were located near the T-Es (c_2) (CML state-space coordinates = $[c_1, c_3]$) or the T-Ls (c_4) (CML state-space coordinates = $[c_3, c_5]$). Finally, we performed both a correlation analysis and linear regression separately for the T-Es (c_2) and the T-Ls (c_4) transition points between the log-normalized mean-centered expression and the value of the potential at the corresponding CML state-space of each sample. The driver genes for each transition point were those genes with a correlation coefficient greater than 0.5 and a linear regression significance of $p < 0.05$ (Fig. S6B). To identify the processes that were involved at each transition point, we used STRINGdb protein-protein interaction database to extend the driver genes to all high-confident (interaction score > 900) interaction partners (Fig. 4B-C).[34] Using the drivers plus their interaction partners, EnrichR gene set enrichment analysis was performed on the Hallmark gene sets to identify the significantly enriched transition point processes (Fig. 4B-C).[35,36] Since the T-Ls network was quite large and included a number of enriched gene sets, we identified the comprising subnetworks and performed EnrichR on each subnetwork to better resolve the involved processes (Fig. S6C)

TOTO and TKI treatment simulations

We integrated treatment into the state-transition treatment model by modeling the treatment as a force that affects both the potential ($\vec{F}$) and internal energy of the particle, or the diffusion rate ($\vec{R}$). The equation of motion for the transcriptome particle is then given by the Langevin equation as

$$dX_t = \left(-\nabla U_p + \vec{F}\right)dt + \sqrt{2\vec{R}}dB_t,$$

where the position of the particle in the quasi-potential is denoted by X_t, and B_t is a Brownian stochastic process that is uncorrelated in time $\langle B_i, B_j \rangle = \delta_{i,j}$.

The corresponding probability distribution $p(x,t)$ at time t is given by a Fokker-Plank (FP) equation including the treatment force $\vec{F}$ and $\vec{R}$ as follows:

$$\frac{\partial}{\partial t} p = \frac{\partial}{\partial x}\left(-\nabla U_p + \vec{F}\right)p + \vec{R}\frac{\partial^2}{\partial x^2} p.$$

We modeled mRNA transcriptome for Tet-off CML as $\nabla U_p = a(x-c_1)(x-c_2)(x-c_3)(x-c_4)(x-c_5)$, with critical points c_1, c_2, c_3, c_4 and c_5 as identified in the state-space.

The state-transition potential U_p was mapped to the potential $V_{CML}(x)$ using the critical points in the state-space as follows.

For TOTO treatment, the CML and normal state-transition potentials approximated those generated by the network model as

$$\nabla U_{p-TOTO-CML} = a(x-c_1)(x-c_2)(x-c_5) \approx V_{TOTO}(x),$$
$$\nabla U_{p-TOTO-normal} = a(x-c_2) \approx V(x).$$

The transformation from CML to normal was therefore given as
$$\nabla U_{p-TOTO} = \nabla U_{p-TOTO-CML} - \nabla U_{p-TOTO-normal} = a\,(x-c_2)\,(x^2 + c_1(c_5 - x) - c_5 x - 1).$$

The treatment force started t=6 until the study endpoint therefore, the treatment force, $\vec{F}_{TOTO}$, and $\vec{R}_{TOTO}$ were given as follows:

$$\vec{F}_{TOTO} = \gamma_{TOTO} H(t - t_{Tet-on})\,\nabla U_{p-TOTO},$$
$$\vec{R}_{TOTO} = \beta^{-1} + (\beta_{Tet-on}^{-1} - \beta^{-1})H(t - t_{Tet-on}).$$

The parameter γ_{TOTO} represents the treatment strength, H is the Heaviside step function and t_{Tet-on} is the timing when Tet treatment started.

For TKI, the CML and normal potentials are approximated using critical points and midpoints between c_1 and c_5 based on the observations from the network model. It was given as

$$\nabla U_{p-TKI-CML} = a(x-c_1)(x-c_2)(x-c_5) \approx V_{TKI}(x),$$
$$\nabla U_{p-TKI-normal} = a(x-c_1)(x-c_2)(x-c_3) \approx V(x).$$

The transformation from CML to normal was therefore given as

$$\nabla U_{p-TKI} = \nabla U_{p-TKI-CML} - \nabla U_{p-TKI-normal} = a(c_3-c_5)(x-c_1)(x-c_2)$$

The TKI treatment was introduced during the t=6 and t=9, and the drug half-life, λ, reduced the effect of the treatment force.[37] The treatment force for TKI system, $\vec{F}_{TKI}$, and $\vec{R}_{TKI}$, were defined as the following,

$$\vec{F}_{TKI} = \gamma_{TKI}\, exp\{-\lambda\big(t-t_{TKI-off}\big)\, H\big(t-t_{TKI-off}\big)\}\, H(t-t_{TKI-on})\nabla U_{p-TKI,}$$
$$\vec{R}_{TKI} = \beta_{TKI}^{-1} + (\beta_{TKI-on}^{-1} - \beta_{TKI}^{-1})(H(t-t_{TKI-on}) - H\big(t-t_{TKI-off}\big)) + \big(\beta_{TKI-off}^{-1} - \beta_{TKI}^{-1}\big)H\big(t-t_{TKI-off}\big).$$

In the model, the transition from normal potential to CML potential was approximated using three critical points using c_1, c_2 and c_5 as $\nabla U_{p-CML} = a(x-c_1)(x-c_2)(x-c_5)$, and $\nabla U_{p-normal} = a(x-c_1)(x-c_5)$.

FP equation was solved to generate the trajectories in the state space for each treatment, TOTO, TKI, and CML system using the state transition potentials. The FP equation for the TOTO system is defined as $\frac{\partial}{\partial t}p = \frac{\partial}{\partial x}\big(-\nabla U_{p-TOTO-CML} + \vec{F}_{TOTO}\big)p + \vec{R}_{TOTO}\frac{\partial^2}{\partial x^2}p$. For TKI, the FP solution is $\frac{\partial}{\partial t}p = \frac{\partial}{\partial x}\big(-\nabla U_{p-TKI-CML} + \vec{F}_{TKI}\big)p + \vec{R}_{TKI}\frac{\partial^2}{\partial x^2}p$, and $\frac{\partial}{\partial t}p = -\frac{\partial}{\partial x}\nabla U_{p-CML}\, p + \beta_{CML}^{-1}\frac{\partial^2}{\partial x^2}p$ for the CML system. The probability density and trajectories at a given time were quantified by solving each FP solution with the initial condition, c_1, which is the healthy state as the initial time point.

Table 1 Values of parameters used in the modeling.

	γ	β	λ	a	c_1	c_2	c_3	c_4	c_5
TOTO	γ_{TKI} =1.1	β_{Tet}^{-1} =0.088 β_{Tet-on}^{-1} =0.001		1	1	0.9231	0.4744	0.2393	0
TKI	γ_{TKI} =1.3	β_{TKI}^{-1} =0.037 $\beta_{TKI-on=}^{-1}$0.001 $\beta_{TKI-off}^{-1}$ =0.001	6.9315	1	1	0.9231	0.4744	0.2393	0
CML		β_{CML}^{-1}=0.005		100	1	0.9231	0.4744	0.2393	0

Time to disease prediction

To compute the probability of disease progression, we integrated the solution of the Fokker-Planck equation beyond the critical point c_5 over time as

$$\mathrm{P}[c_5|t] = 1 - \int_{c_5}^{\infty} \int_0^t p(x,\tau) dx d\tau$$

where $p(x,t)$ is the solution to $\frac{\partial}{\partial t} p = \frac{\partial}{\partial x}\left(-\nabla U_p + \vec{F}\right)p + \vec{R}\frac{\partial^2}{\partial x^2} p$ with initial conditions given by the PC2 values at time t=0 of the CML (tet-off) samples. The continuous curve given by $\mathrm{P}[c_5|t]$ was then evaluated at the sample observation timepoints, corresponding to 1-week intervals. The observed time to disease was given by the timepoint at which the CML state-space trajectory crossed the c_5 critical point and was computed for each mouse. The predicted and observed time to disease curves were compared with a log-rank test and were found to not be significantly different (p=0.8).

Early vs late control analysis

An analysis comparing the early (first five time points) vs late (last five time points) control samples was performed to determine how the samples changed over time since they were observed to

move downward in the CML state-space (Fig. S7D; *top*). The total CML contribution of the 118 DEGs identified showed a small pro-CML contribution in the late controls (Fig. S7E; *bottom left*). The significantly enriched Hallmark gene sets from GSEA were very similar to those identified when the TOTO post-Rx were compared to all Hs samples (Fig. S7E; *bottom right*). Because of the small observed pro-CML effects observed in this analysis that could be due to aging processes in the mice over the course of the experiment, the TOTO post-Rx DEGs were controlled for aging by removing the DEGs observed from the early vs late control analysis.

References

1. Cortes, J., Pavlovsky, C. & Saußele, S. Chronic myeloid leukaemia. *Lancet (London, England)* **398**, 1914–1926 (2021).
2. Goldman, J. M. & Melo, J. V. Targeting the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.* **344**, 1084–1086 (2001).
3. Druker, B. J., Tamura, S., Buchdunger, E., Ohno, S., Segal, G. M., Fanning, S., *et al.* Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr–Abl positive cells. *Nat. Med. 1996 25* **2**, 561–566 (1996).
4. Zhang, B., Zhao, D., Chen, F., Frankhouser, D., Wang, H., Pathak, K. V., *et al.* Acquired miR-142 deficit in leukemic stem cells suffices to drive chronic myeloid leukemia into blast crisis. *Nat. Commun. 2023 141* **14**, 1–21 (2023).
5. Kuntz, E. M., Baquero, P., Michie, A. M., Dunn, K., Tardito, S., Holyoake, T. L., *et al.* Targeting mitochondrial oxidative phosphorylation eradicates therapy-resistant chronic myeloid leukemia stem cells. *Nat. Med. 2017 2310* **23**, 1234–1240 (2017).
6. Mojtahedi, H., Yazdanpanah, N. & Rezaei, N. Chronic myeloid leukemia stem cells: targeting therapeutic implications. *Stem Cell Res. Ther.* **12**, 1–27 (2021).
7. Mahon, F. X., Réa, D., Guilhot, J., Guilhot, F., Huguet, F., Nicolini, F., *et al.* Discontinuation of imatinib in patients with chronic myeloid leukaemia who have maintained complete molecular remission for at least 2 years: the prospective, multicentre Stop Imatinib (STIM) trial. *Lancet. Oncol.* **11**, 1029–1035 (2010).
8. Ross, D. M., Branford, S., Seymour, J. F., Schwarer, A. P., Arthur, C., Yeung, D. T., *et al.* Safety and efficacy of imatinib cessation for CML patients with stable undetectable minimal residual disease: results from the TWISTER study. *Blood* **122**, 515–522 (2013).
9. Rockne, R. C., Branciamore, S., Qi, J., Frankhouser, D. E., O'Meally, D., Hua, W. K., *et al.* State-Transition Analysis of Time-Sequential Gene Expression Identifies Critical Points That Predict Development of Acute Myeloid Leukemia. *Cancer research* vol. 80

3157–3169 at https://doi.org/10.1158/0008-5472.CAN-20-0354 (2020).

10. Frankhouser, D. E., O'Meally, D., Branciamore, S., Uechi, L., Zhang, L., Chen, Y. C., *et al.* Dynamic patterns of microRNA expression during acute myeloid leukemia state-transition. *Sci. Adv.* **8**, 1664 (2022).

11. Janowski, M., Ulańczyk, Z., Łuczkowska, K., Sobuś, A., Rogińska, D., Pius-Sadowska, E., *et al.* Molecular Changes in Chronic Myeloid Leukemia During Tyrosine Kinase Inhibitors Treatment. Focus on Immunological Pathways. *Onco. Targets. Ther.* **15**, 1123 (2022).

12. Iezza, M., Cortesi, S., Ottaviani, E., Mancini, M., Venturi, C., Monaldi, C., *et al.* Prognosis in Chronic Myeloid Leukemia: Baseline Factors, Dynamic Risk Assessment and Novel Insights. *Cells* **12**, 1703 (2023).

13. Giustacchini, A., Thongjuea, S., Barkas, N., Woll, P. S., Povinelli, B. J., Booth, C. A. G., *et al.* Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat. Med. 2017 236* **23**, 692–702 (2017).

14. Ross, D. M., Pagani, I. S., Irani, Y. D., Clarson, J., Leclercq, T., Dang, P., *et al.* Lenalidomide maintenance treatment after imatinib discontinuation: results of a phase 1 clinical trial in chronic myeloid leukaemia. *Br. J. Haematol.* **186**, e56–e60 (2019).

15. Radich, J. P., Dai, H., Mao, M., Oehler, V., Schelter, J., Druker, B., *et al.* Gene expression changes associated with progression and response in chronic myeloid leukemia. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 2794–2799 (2006).

16. Radich, J. P., Hochhaus, A., Masszi, T., Hellmann, A., Stentoft, J., Casares, M. T. G., *et al.* Treatment-free remission following frontline nilotinib in patients with chronic phase chronic myeloid leukemia: 5-year update of the ENESTfreedom trial. *Leuk. 2021 355* **35**, 1344–1355 (2021).

17. Dey, A. & Barik, D. Potential Landscapes, Bifurcations, and Robustness of Tristable Networks. *ACS Synth. Biol.* **10**, 391–401 (2021).

18. Hari, K., Harlapur, P., Saxena, A., Haldar, K., Girish, A., Malpani, T., *et al.* Low dimensionality of phenotypic space as an emergent property of coordinated teams in biological regulatory networks. *bioRxiv* 2023.02.03.526930 (2023) doi:10.1101/2023.02.03.526930.

19. Alazami, A. M., Patel, N., Shamseldin, H. E., Anazi, S., Al-Dosari, M. S., Alzahrani, F., *et al.* Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families. *Cell Rep.* **10**, 148–161 (2015).

20. Shen, R., Soeder, R. A., Ophardt, H. D., Giangrasso, P. J., Cook, L. B. & Author, C. Identification of Long Non-coding RNAs Expressed During Early Adipogenesis. *OnLine J. Biol. Sci. Orig. Res. Pap.* (2019) doi:10.3844/ojbsci.2019.245.259.

21. Zhou, J. dong, Zhang, T. juan, Xu, Z. jun, Deng, Z. qun, Gu, Y., Ma, J. chun, *et al.* Genome-wide methylation sequencing identifies progression-related epigenetic drivers in myelodysplastic syndromes. *Cell Death Dis.* **11**, (2020).

22. Oshi, M., Yan, L., Kunisaki, C., Endo, I. & Takabe, K. Association of enhanced epithelial-mesenchymal transition signature with tumor microenvironment, angiogenesis, and survival in gastric cancer. *https://doi.org/10.1200/JCO.2023.41.16_suppl.e16002* **41**, e16002–e16002 (2023).

23. Takahashi, H., Oshi, M., Yan, L., Endo, I. & Takabe, K. Gastric cancer with enhanced apical junction pathway has increased metastatic potential and worse clinical outcomes. *Am. J. Cancer Res.* **12**, 2146 (2022).

24. Greaves, D. & Calle, Y. Epithelial Mesenchymal Transition (EMT) and Associated Invasive Adhesions in Solid and Haematological Tumours. *Cells* **11**, (2022).

25. Alter, O., Brown, P. O. & Botstein, D. Singular value decomposition for genome-Wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 10101–10106 (2000).

26. Qiu, S., Sheth, V., Yan, C., Liu, J., Chacko, B. K., Li, H., *et al.* Metabolic adaptation to tyrosine kinase inhibition in leukemia stem cells. *Blood* **142**, 574–588 (2023).

27. Zhang, B., Zhao, D., Chen, F., Frankhouser, D., Wang, H., Pathak, K. V., *et al.* Acquired miR-142 deficit in leukemic stem cells suffices to drive chronic myeloid leukemia into blast crisis. *Nat. Commun.* **14**, (2023).

28. Wang, Z., Jiang, L., Yan, H., Xu, Z. & Luo, P. Adverse events associated with nilotinib in chronic myeloid leukemia: mechanisms and management strategies. *Expert Rev. Clin. Pharmacol.* **14**, 445–456 (2021).

29. Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology* vol. 38 276–278 at https://doi.org/10.1038/s41587-020-0439-x (2020).

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

31. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods 2017 144* **14**, 417–419 (2017).

32. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).

33. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4**, (2015).

34. Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).

35. Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., *et*

al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90 (2016).

36. Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D. J. B., Evangelista, J. E., Jenkins, S. L., *et al.* Gene Set Knowledge Discovery with Enrichr. *Curr. Protoc.* **1**, e90 (2021).

37. Pushpam, D. & Bakhshi, S. Pharmacology of tyrosine kinase inhibitors in chronic myeloid leukemia; a clinician's perspective. *DARU, J. Pharm. Sci.* **28**, 371–385 (2020).

Acknowledgement

Research reported in this publication was supported in part by Robert and Lynda Altman Family Foundation and included work performed in the integrative genomics core and biostatistics and mathematical oncology Shared Resource supported by the National Cancer Institute of the National Institutes of Health under grant number P30CA033572 and grant U01CA250067. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Figures

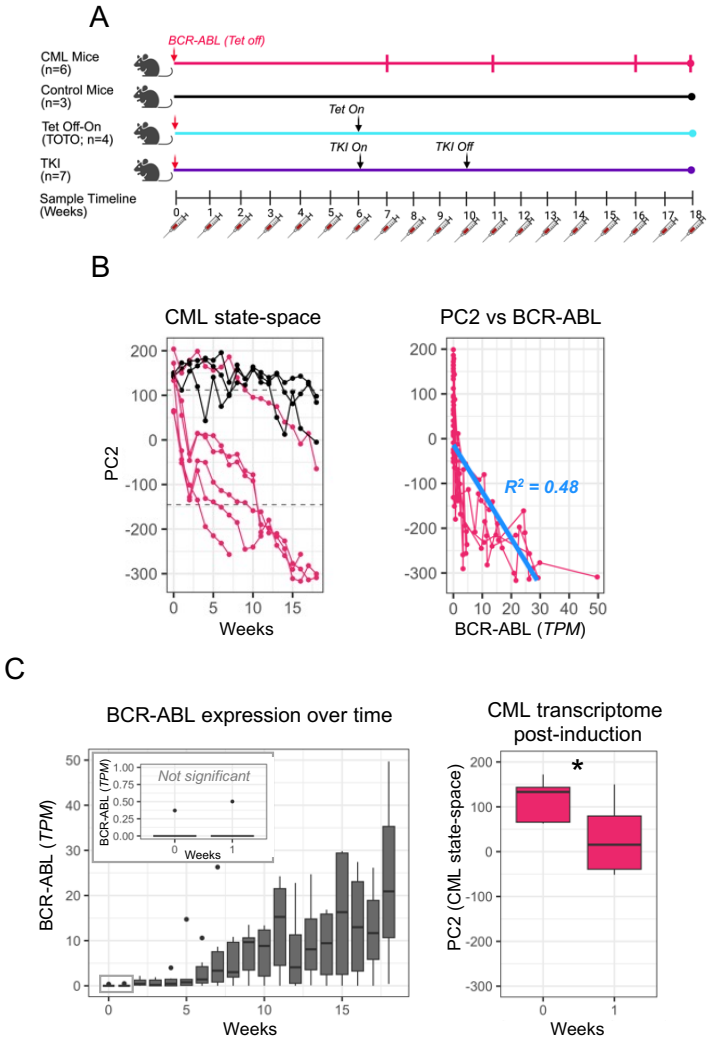


Figure 1: Time-series peripheral blood transcriptome encodes the CML state-space.

A) The experimental design collected time-series sampling of peripheral blood from four cohorts of mice over 19 weeks. All mice had Tet-inducible expression of the BCR-ABL fusion gene which results in CML when expressed. **B)** Singular value decomposition (SVD) was performed on the full transcriptome using all time points from both the CML and control mouse cohorts to identify the CML state-space (PC2). When plotted against time, the CML state-space shows sample disease trajectories of each sample over time (*left*). PC2 was selected as the CML state-space because it had the largest separation between control mice and the endpoint CML samples, and because it had the best linear fit (in *blue*) with BCR-ABL expression (*right*; Table S1). **C)** BCR-

ABL expression only began to increase 4-5 weeks after Tet was turned off (*left*). Whereas BCR-ABL expression remained undetectable at week 1 (*left inset*), the CML state-space already shows a significant movement toward the CML disease state (*right*; p-value=0.006).

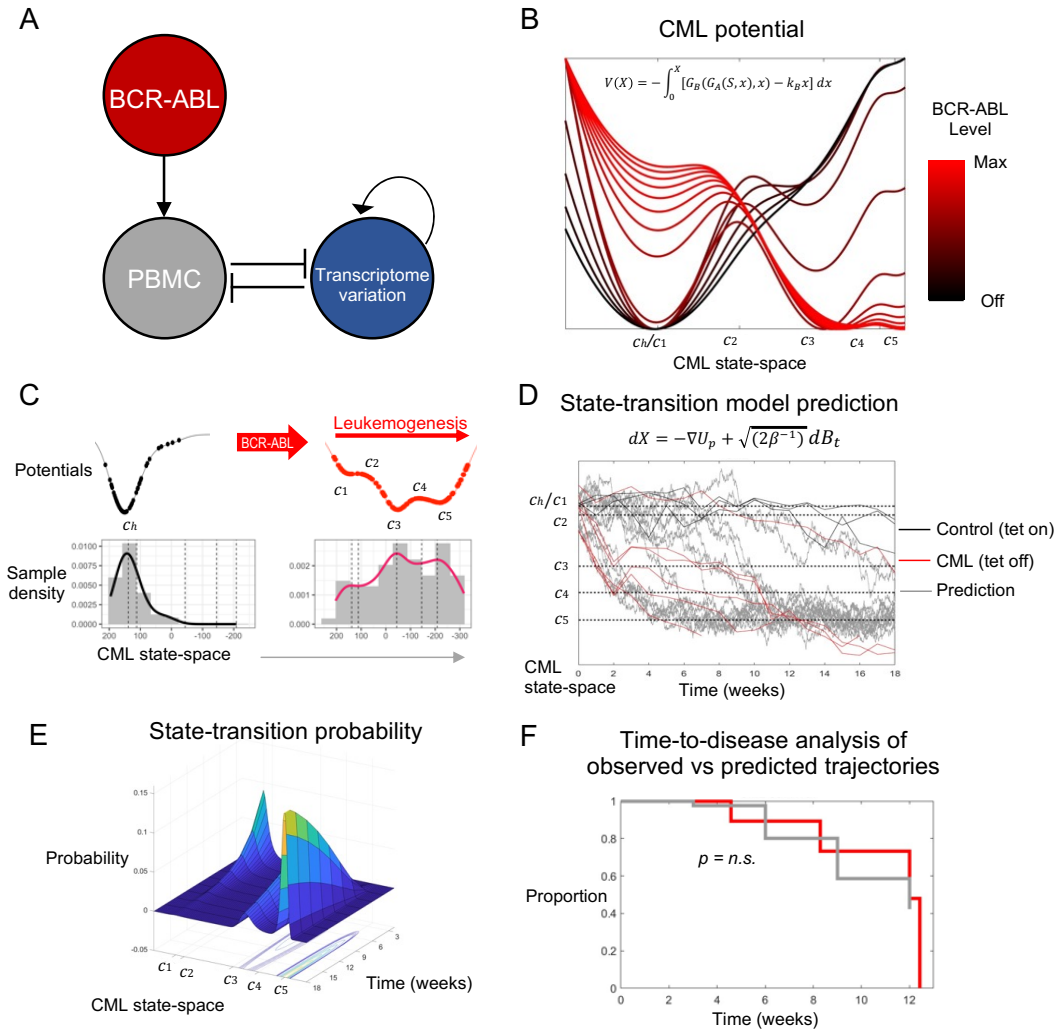


Figure 2: State-transition model accurately predicts time-to-disease.

A) The three node CML network model that produced a tri-stable systems and captures the features of the transcriptomic CML potential. **B)** BCR-ABL input signal determines the shape of the CML potential. When BCR-ABL is off (*black*), the system has one steady state at the healthy (c_h) state. As the BCR-ABL signal increases, the potential gains additional steady states. When

BCR-ABL signal reaches its maximum, which we use to model CML development in this experiment, the potential has five critical point which we labeled $c_1 - c_5$. Three of these critical points are steady states (c_1, c_3, c_5), and the most energetically favorable state (c_5) is the CML disease state. **C)** Using the distribution of samples in the CML state-space, we empirically constructed the control (*left*) and CML (*right*) potentials. CML potentials were created by fitting histograms of samples in the CML state-space with kernel densities (bottom). The polynomials that defined potentials (*top*) were then constructed by finding the critical points of the density curves and using those as the zeros of the derivative of the potential. **D)** Using only the first (week 0) time point of each sample as an initial condition, the stochastic equation of motion was solved forward in time for the CML and control potentials to predict the trajectory of each sample in the CML state-space (*grey lines*). **E)** The solution of the state-transition model is given as a probability density over time which represents the likelihood of finding a sample in the CML state-space as time evolves. **F)** To test the accuracy on the state-transition model predictions, a time-to-disease analysis was performed showing that there was no significant difference between the predicted and observed trajectories (log-rank p-value = 0.8).

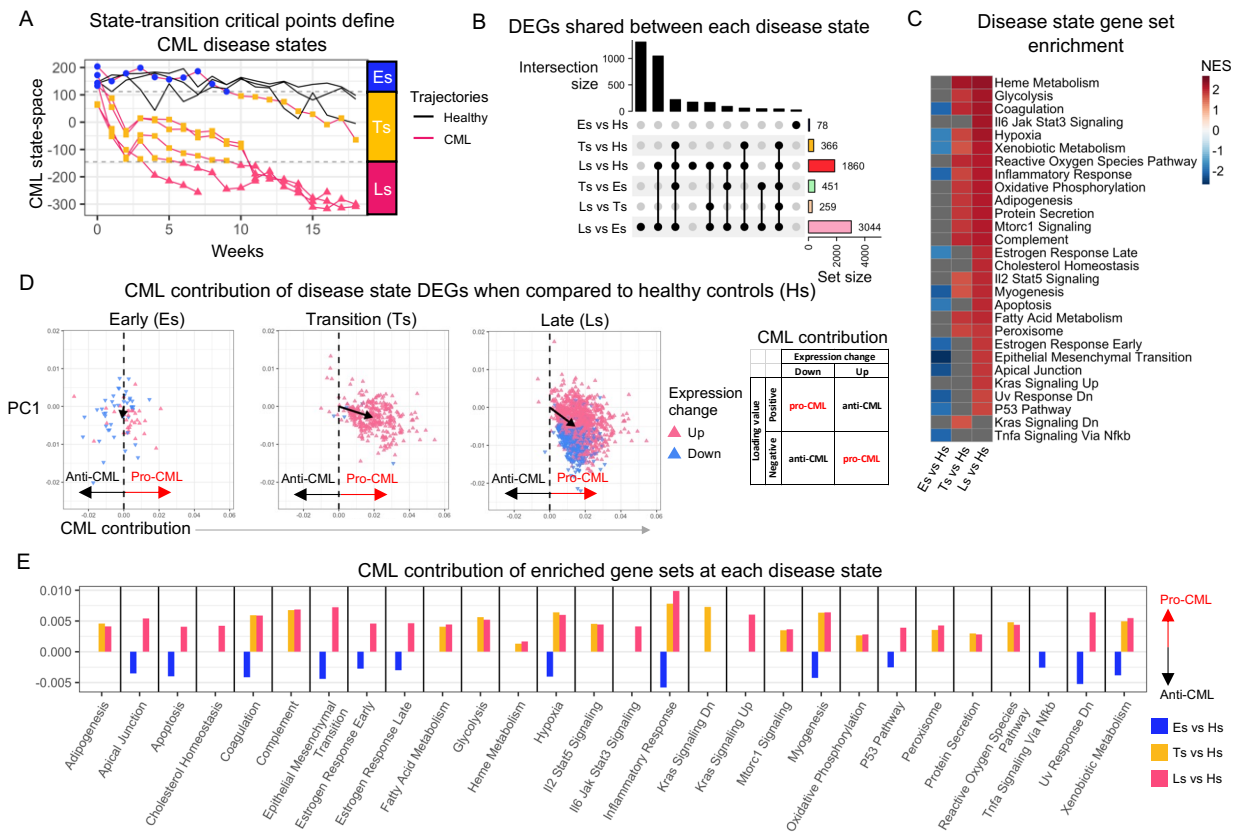


Figure 3: Differential expression analysis performed on CML state-space disease states. **A)** The three stable critical points of the CML state-space (c_1, c_3, c_5) were used to group the CML samples into Early (Es), Transition (Ts), and Late (Ls) disease states based on their location in the CML state-space. The unstable critical points (c_2, c_4; *dashed lines*) were used as the boundaries to define the three disease states. **B)** Differential gene expression was performed both between the disease states and healthy control state (Hs) and between the disease states. The number of differentially expressed genes (DEGs) contained in the intersection between the comparisons was used to determine how similar the disease states were to each other. **C)** Gene set enrichment analysis was performed on the $\log_2$ fold change of the healthy vs disease states to determine which Hallmark gene sets were significantly enriched (between disease states shown in Fig. S3B). The results were summarized by showing a heatmap of the normalized enrichment score (NES) for all significant pathways (non-significant pathways are in *grey*) across the three disease states. The NES score shows the direction of expression change for

each gene set. **D)** Using the eigengenes (PC2 loading values) used to construct the CML states-space, we determined the contribution to CML of each gene that was a DEG in the disease state vs control comparisons. The CML contribution is determined by combining the eigengene value with the observed expression change for that comparison (*right box*). The total contribution for each DEG comparison was indicated by the mean vector (*black arrow*) of all DEGs in each comparison. **E)** To visualize the contribution of each significantly enriched gene set, the total CML contribution (mean vector) for each comparison was plotted for each comparison (missing bars indicate non-significant gene set in that comparison).

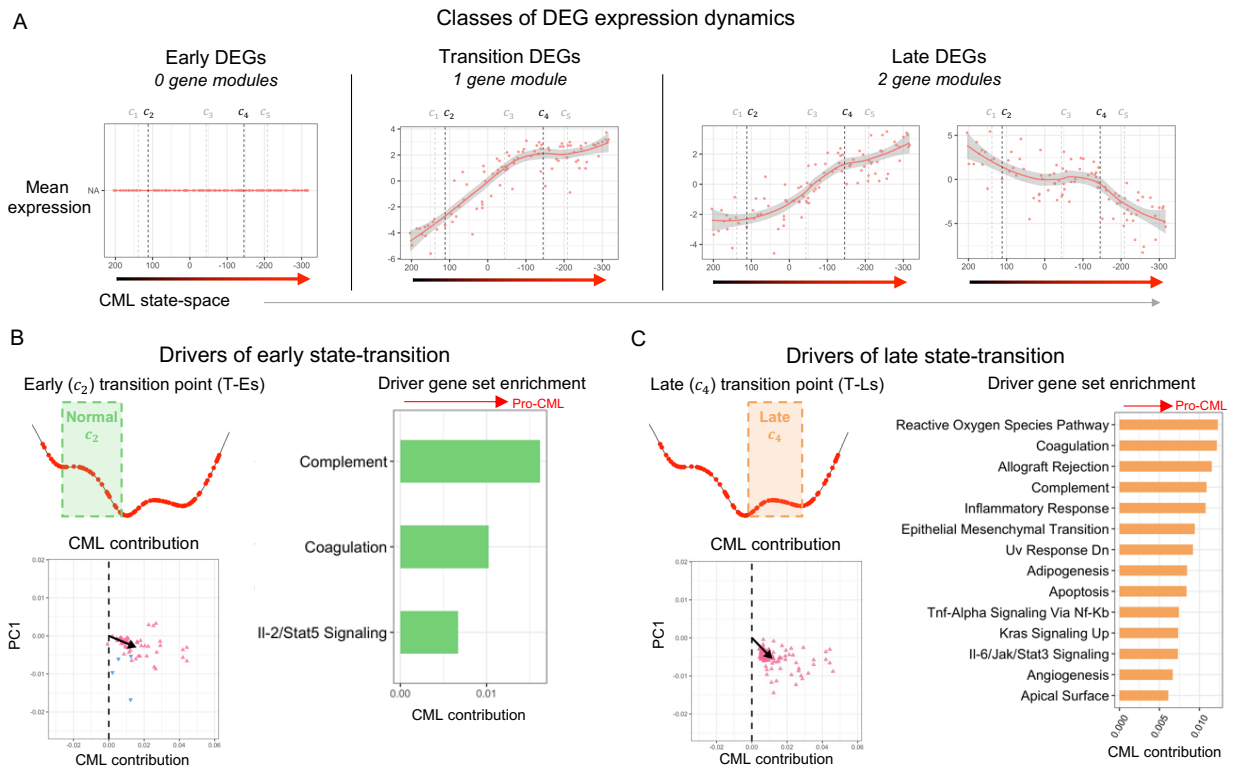


Figure 4: CML state-space aligns samples and identifies transition driver genes.

A) Using a correlation analysis to identify the DEGs with similar dynamics during CML development, gene modules were identified for the unique DEGs of the three disease states (Es, Ts, Ls; Fig. S3C, S4A). To visualize the expression of each detected gene module that had

a mean correlation coefficient > 0.25, the mean log-normalized mean-centered expression of all genes in the module was plotted as function of the CML state-space coordinate for each CML sample. **B-C)** For the early transition point (T-Es; **B**) and the late transition point (T-Ls; **C**) DEGs that were involved in state-transition were identified by correlating the expression of each DEG with the shape of the potential around each of the transition points (*left top*; Fig. S6*B*). To identify the driver gene processes, the driver genes for each transition point were then extended to their high-confidence interaction partners using STRINGdb (Fig. S6C). The CML contribution were summarized for the resulting protein-protein interaction networks (*left bottom*). The processes enriched by the STRINGdb extended driver genes involved were identified and their CML contribution was calculated (*right*).

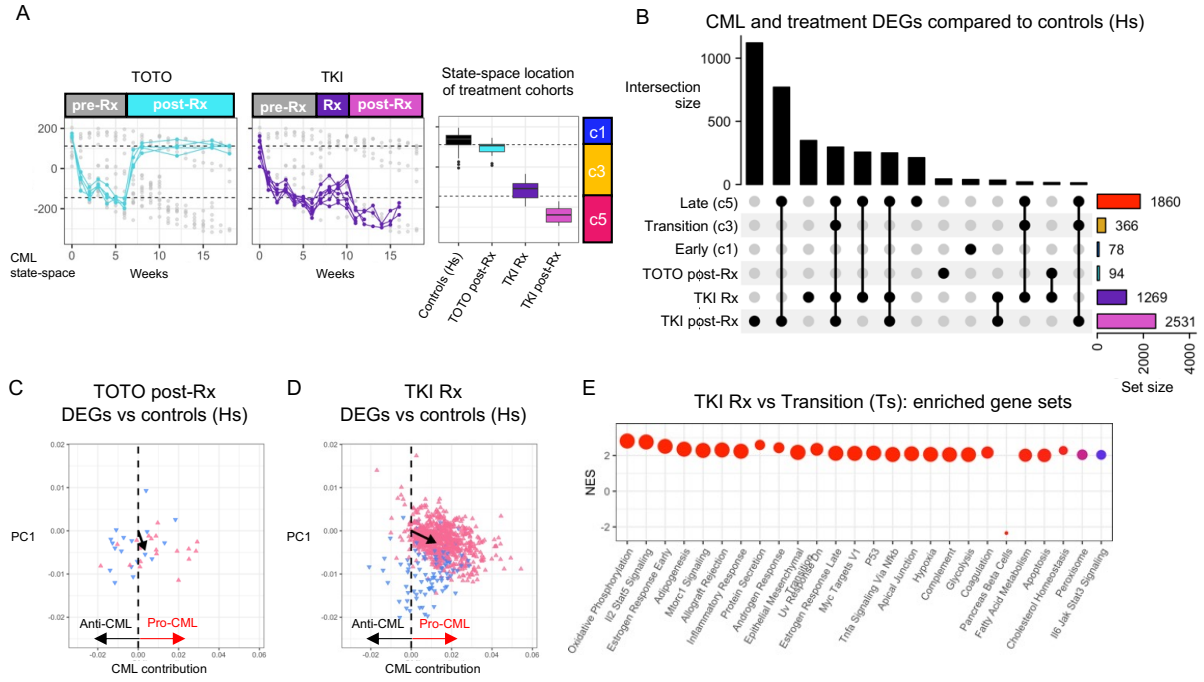


Figure 5: CML treatment effects determined using CML state-space disease states. **A)** Samples from the TOTO (*left*) and TKI (center) were projected into the CML state-space to allow the treatment effects to be compared to the CML disease states (*right*). **B)** DEGs were identified by comparing the TOTO post-Rx and both the TKI Rx and post-Rx samples to the

controls (Hs). All intersections with 10 or more DEGs are shown (all intersections shown in Fig. S7B). **C)** After removing the age-related DEGs that were detected when comparing the early vs late control samples, the CML contribution of the 38 TOTO post-Rx vs controls DEGs were plotted and showed a small overall pro-CML contribution (*black arrow*). **D)** The CML contribution of each of the TKI Rx vs control (Hs) DEGs are shown along with the overall CML contribution for all DEGs (*black arrow*). **E)** Since the TKI Rx samples returned to a similar CML state-space location, the significantly enriched (adjusted p-value < 0.0001) GSEA Hallmark gene sets for the TKI Rx vs transition (Ts) DEGs were determined to identify what processes were affected by TKI treatment.

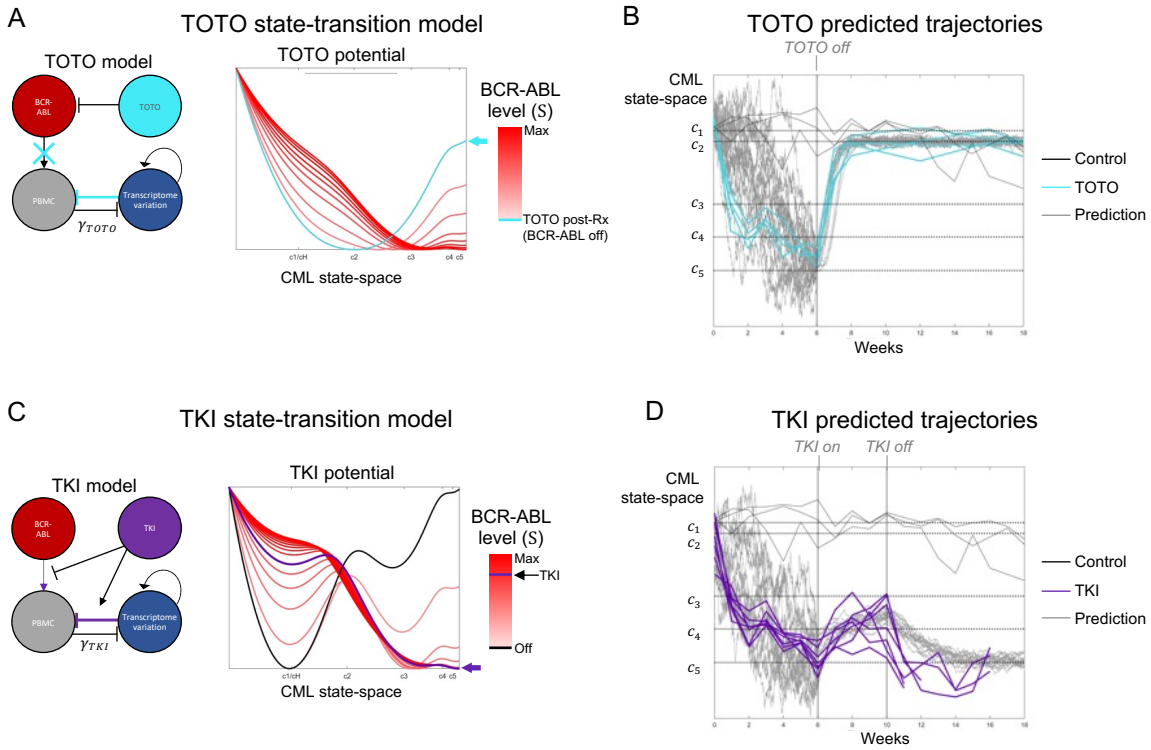


Figure 6: State-transition model predicts CML treatment dynamics.

A) The model for the TOTO cohort is modified from the CML model to include turning off the BCR-ABL input signal and increasing the γ_{TOTO} term in the model (*left*). As the input BCR-ABL signal is decreased in this model, the potential transforms from a tri-stable CML potential (*red*

curve) to a potential with one state that has minimum energy at the unstable transition critical point c_2 (*teal curve*; *right*). **B)** Using the TOTO potential, sample trajectories were predicted by solving the stochastic equation of motion forward in time using the first time point as an initial condition. **C)** The model for the TKI cohort was modified to reduce the BCR-ABL signal and increase the γ_{TKI} term in the model (*left*). The TKI potential (*purple*) lies between the CML (BCR-ABL max) and the BCR-ABL off potentials. The model showed that the healthy state (Hs) could only be achieved as an energetically favorable state while on TKI was for the BCR-ABL signal approach zero. **D)** Using the TKI potential, sample trajectories were predicted by solving the stochastic equation of motion forward in time using the first time point as an initial condition.

Supplemental Figures and Table

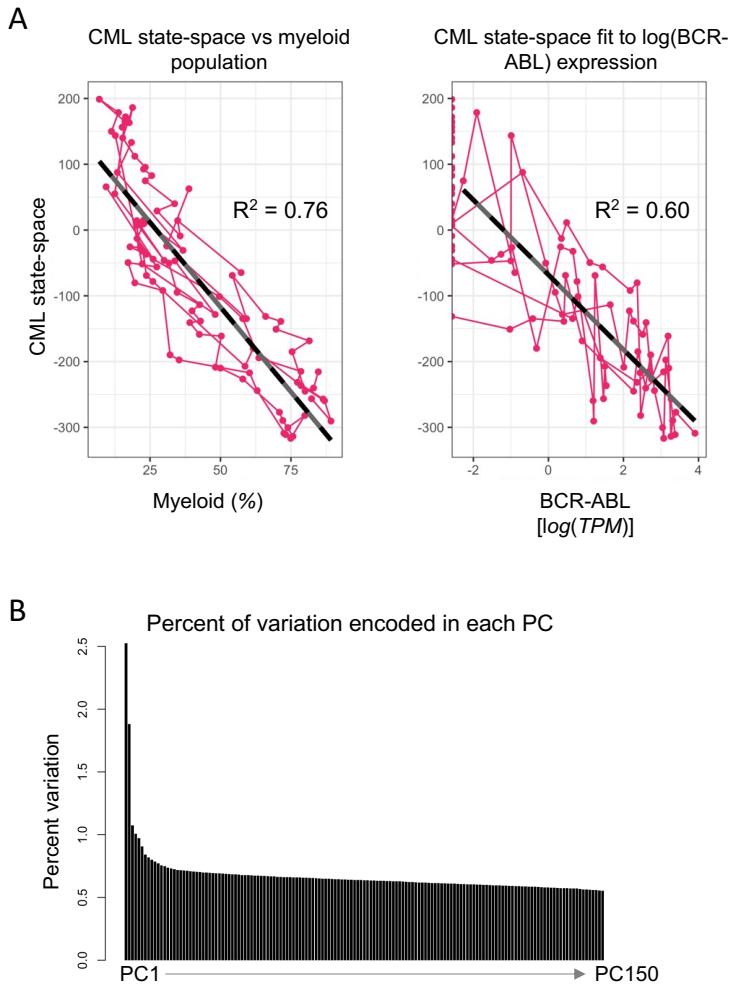


Figure S1: Features of the CML state-space.

A) The CML state-space was correlated with clinical markers of CML the myeloid population in the peripheral blood (*left*) and the log-expression of BCR-ABL (*right*). Linear fits (*black dashed*) were used to show that the CML state-space had the best correlation with both myeloid and BCR-ABL (Table S1). **B)** The percent of variation encoded by each PC that resulted from the SVD operation was plotted and showed that the variation encoded by PCs decreased and leveled off after PC2

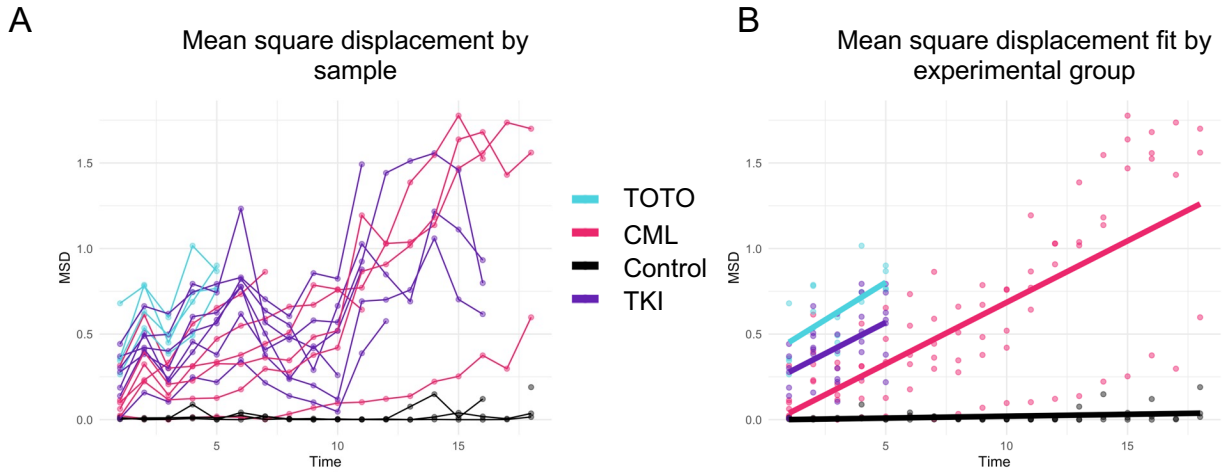


Figure S2: Mean squared displacement estimates for the diffusion coefficients.

A) A mean square displacement (MSD) was performed for each mouse using their trajectories in the CML state-space as a function of time. **B)** For each cohort of mice, a linear fit of all trajectories in the cohort was calculated and the slope of the line was used to estimate the diffusion coefficient (β). Treatment cohorts were shown only until treatment was initiated.

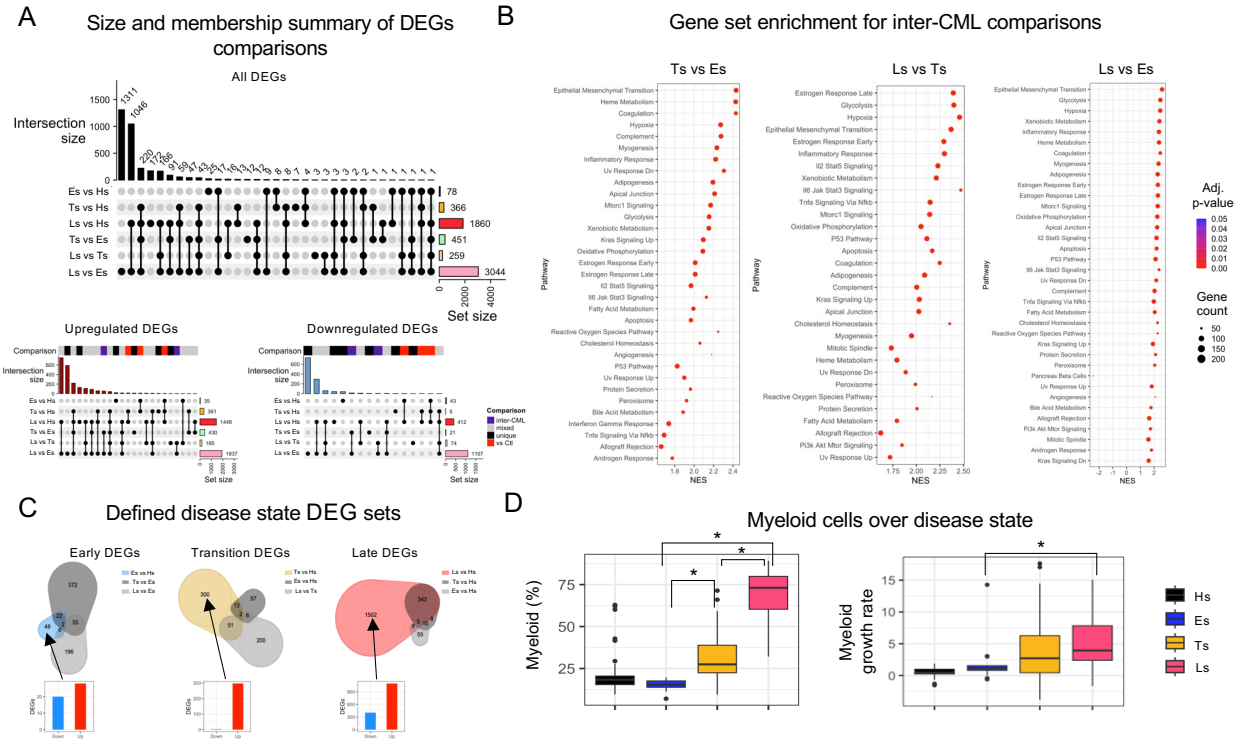


Figure S3: Analysis of CML disease states.

A) Differentially expressed genes (DEGs) were identified between all pairwise comparisons of the CML disease states (Es, Ts, Ls) and healthy controls (Hs). The results for the intersection between each comparison were plotted separately for all genes (*top*), the upregulated genes in each comparison (*left*), and the downregulated genes in each comparison (*right*). **B)** Gene set enrichment analysis (GSEA) was also performed on all comparisons using the Hallmark gene sets. Here, the inter-CML disease state comparisons (pairwise comparisons of Es, Ts, and Ls) were plotted using the normalized enrichment score (NES) to show the direction of the expression in all significantly enriched gene sets (adjusted p-value < 0.001). **C)** The DEGs unique to each disease state were defined by making intersections between the relevant comparisons to determine which gene expression changes only occurred in each disease state. **D)** The percent of myeloid cells in the peripheral blood was compared in each disease state using a t-test (*left*). The myeloid growth rate was calculated as the derivative of the spline fits for each mouse's myeloid percentage (*right*) and a t-test was again used to compare growth rate between disease states (significant p-value < 0.05).

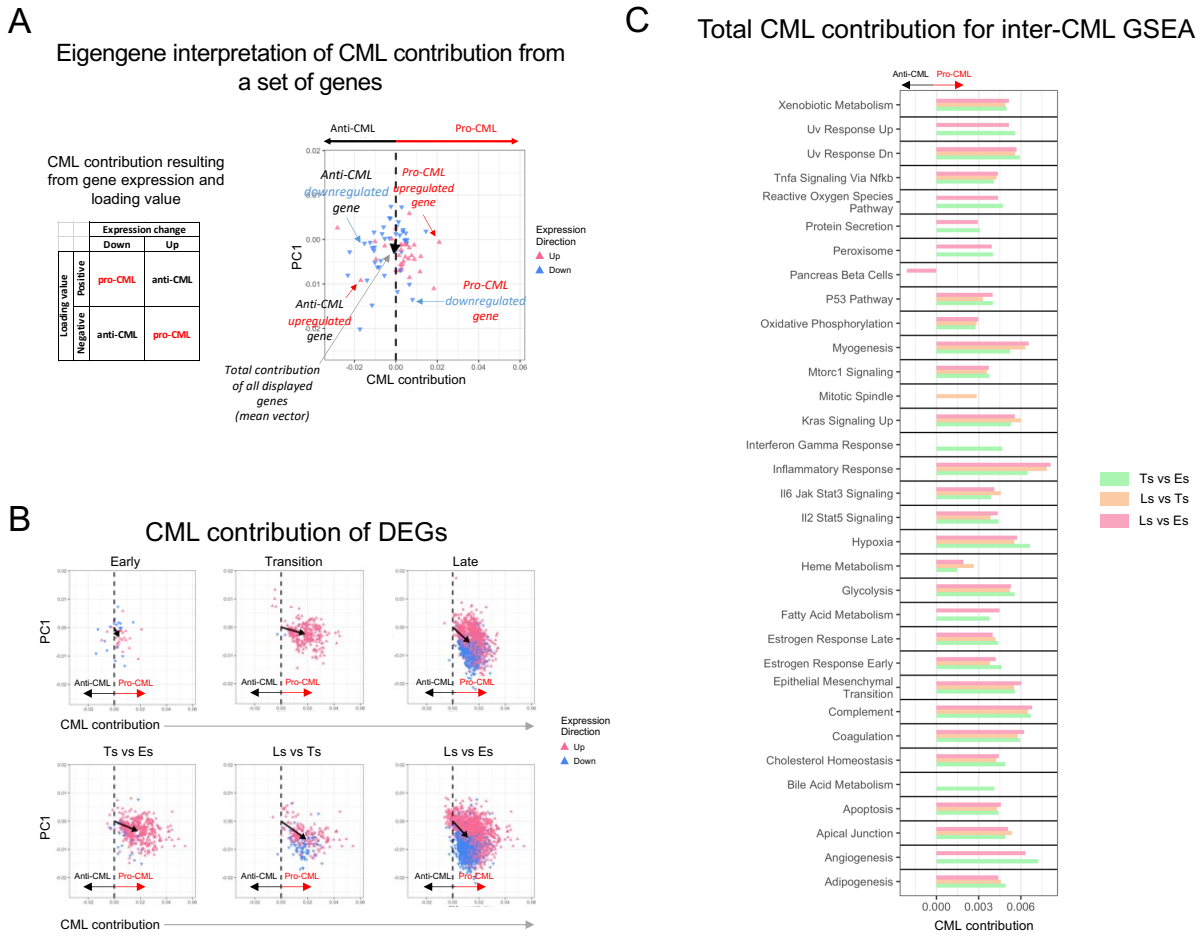


Figure S4: Eigengenes quantify CML contribution of genes and groups of genes.

A) Using the eigengenes, which were defined as the CML state-space (PC2) loading values, we determined a gene's quantitative contribution to CML state-transition. For each gene, the CML contribution was determined by combining the eigengene with the observed gene expression change using the table (*left*) to determine whether the contribution was pro- or anti-CML. The CML contribution was illustrated for a set of genes (*right*) by plotting the CML contribution vs the PC1 value for all genes in the set. The total contribution of the set was represented by the mean vector of all genes in the set (*black arrow*). **B)** The CML contribution was used to summarize the overall effect of each DEG comparison on CML state-transition. **C)** The CML contribution was summarized for the three inter-CML comparison's GSEA results by plotting the total CML

contribution for the genes in leading edge of each significantly enriched gene set (adjusted p-value < 0.0001). Missing bars represent gene sets that were not significant in that comparison.

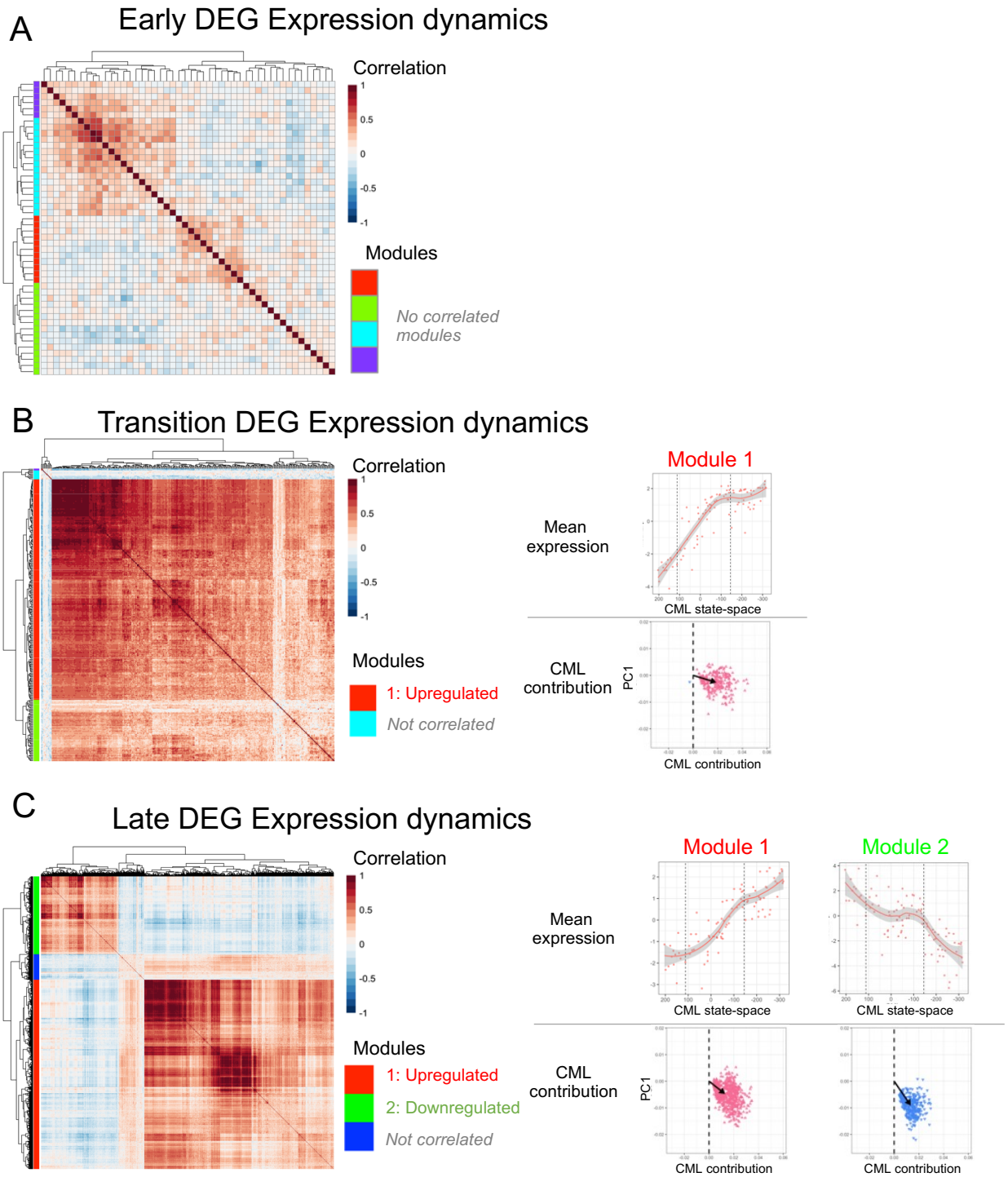
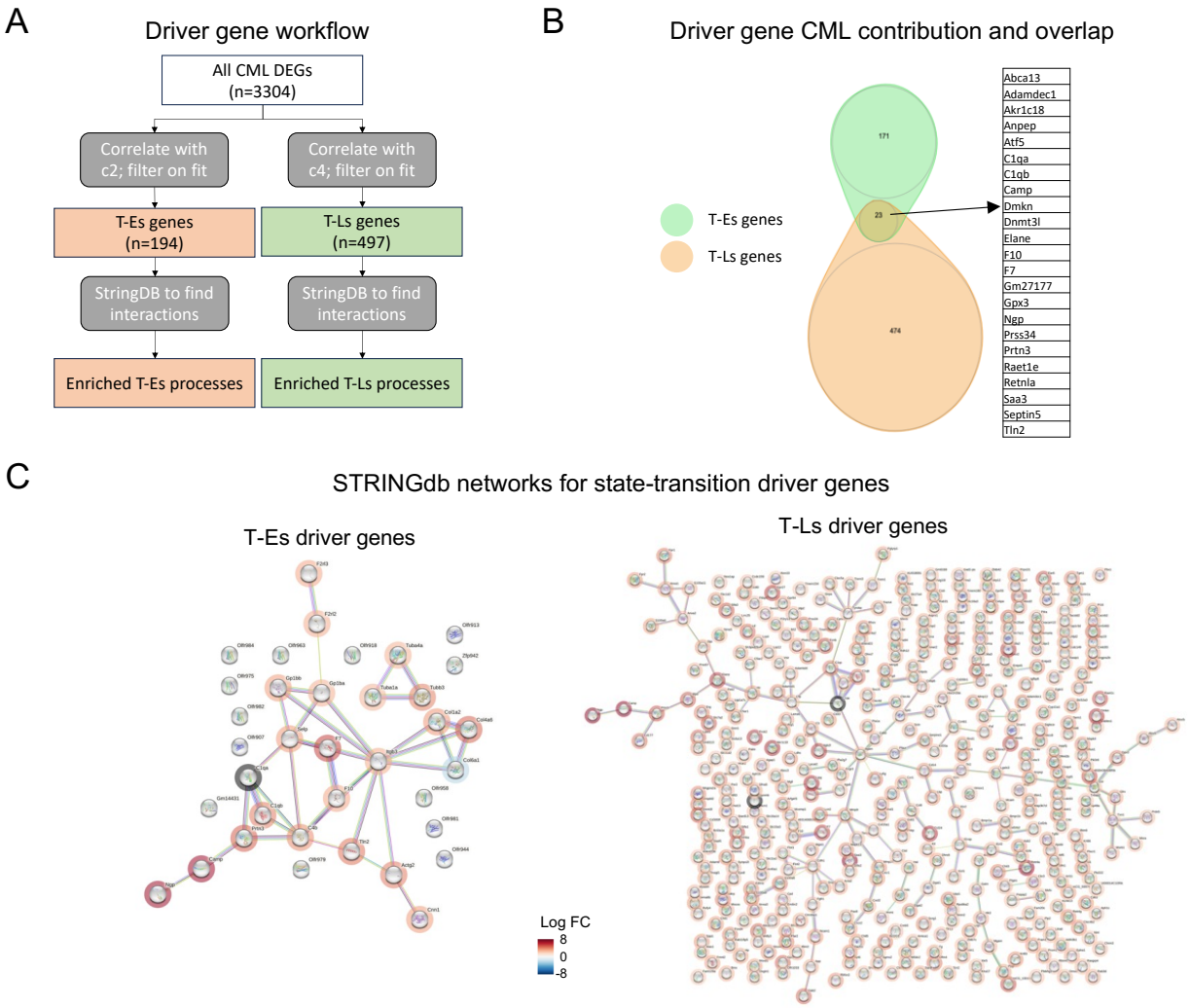
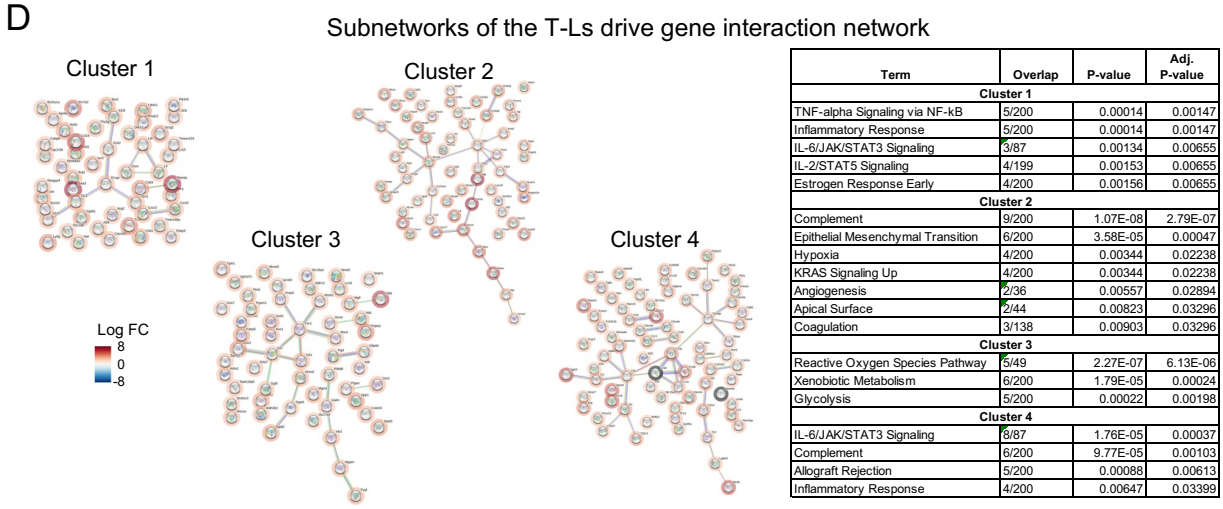


Figure S5: Expression dynamics of the unique disease-state DEGs.

A correlation analysis using all mice from the CML cohort was used to determine which genes have similar expression dynamics over CML development. Using hierarchical clustering on the correlation matrix, gene modules were defined as a cluster of genes with a mean correlation coefficient greater than 0.25. To visualize how the expression changed over CML, the average expression of each gene module was plotted for each sample as a function of the CML state-space of the sample and summarized using a loess fit (*red line*). Finally, the overall CML contribution (*black arrow*) of each gene module was determined. This process was performed on the **A)** DEGs unique to Es, **B)** the DEGs unique to Ts, and **C)** the DEGs unique to Ls.





Term	Overlap	P-value	Adj. P-value
Cluster 1			
TNF-alpha Signaling via NF-kB	5/200	0.00014	0.00147
Inflammatory Response	5/200	0.00014	0.00147
IL-6/JAK/STAT3 Signaling	3/87	0.00134	0.00655
IL-2/STAT5 Signaling	4/199	0.00153	0.00655
Estrogen Response Early	4/200	0.00156	0.00655
Cluster 2			
Complement	9/200	1.07E-08	2.79E-07
Epithelial Mesenchymal Transition	6/200	3.58E-05	0.00047
Hypoxia	4/200	0.00344	0.02238
KRAS Signaling Up	4/200	0.00344	0.02238
Angiogenesis	2/36	0.00557	0.02894
Apical Surface	2/44	0.00823	0.03296
Coagulation	3/138	0.00903	0.03296
Cluster 3			
Reactive Oxygen Species Pathway	5/49	2.27E-07	6.13E-06
Xenobiotic Metabolism	6/200	1.79E-05	0.00024
Glycolysis	5/200	0.00022	0.00198
Cluster 4			
IL-6/JAK/STAT3 Signaling	8/87	1.76E-05	0.00037
Complement	6/200	9.77E-05	0.00103
Allograft Rejection	5/200	0.00088	0.00613
Inflammatory Response	4/200	0.00647	0.03399

Figure S6: Driver gene of CML state-transition.

A) Process for identifying the T-Es and the T-Ls drive genes. **B)** Venn diagram of the T-Es and T-Ls driver genes. **C)** Protein-protein interaction networks for the T-Es (*left*) and the T-Ls (*right*) driver genes. STRINGdb was used to identify high confidence interactions (interaction score > 900). Each gene is colored based on the observed log foldchange between Ts vs Hs for the T-Es driver genes or Ls vs Hs for the T-Ls driver genes. **D)** To further refine the processes involved in the large network of the T-Ls driver genes, four subnetworks were identified, and the significantly enriched Hallmark gene sets for each cluster were reported for each cluster (adjusted p-value < 0.01).

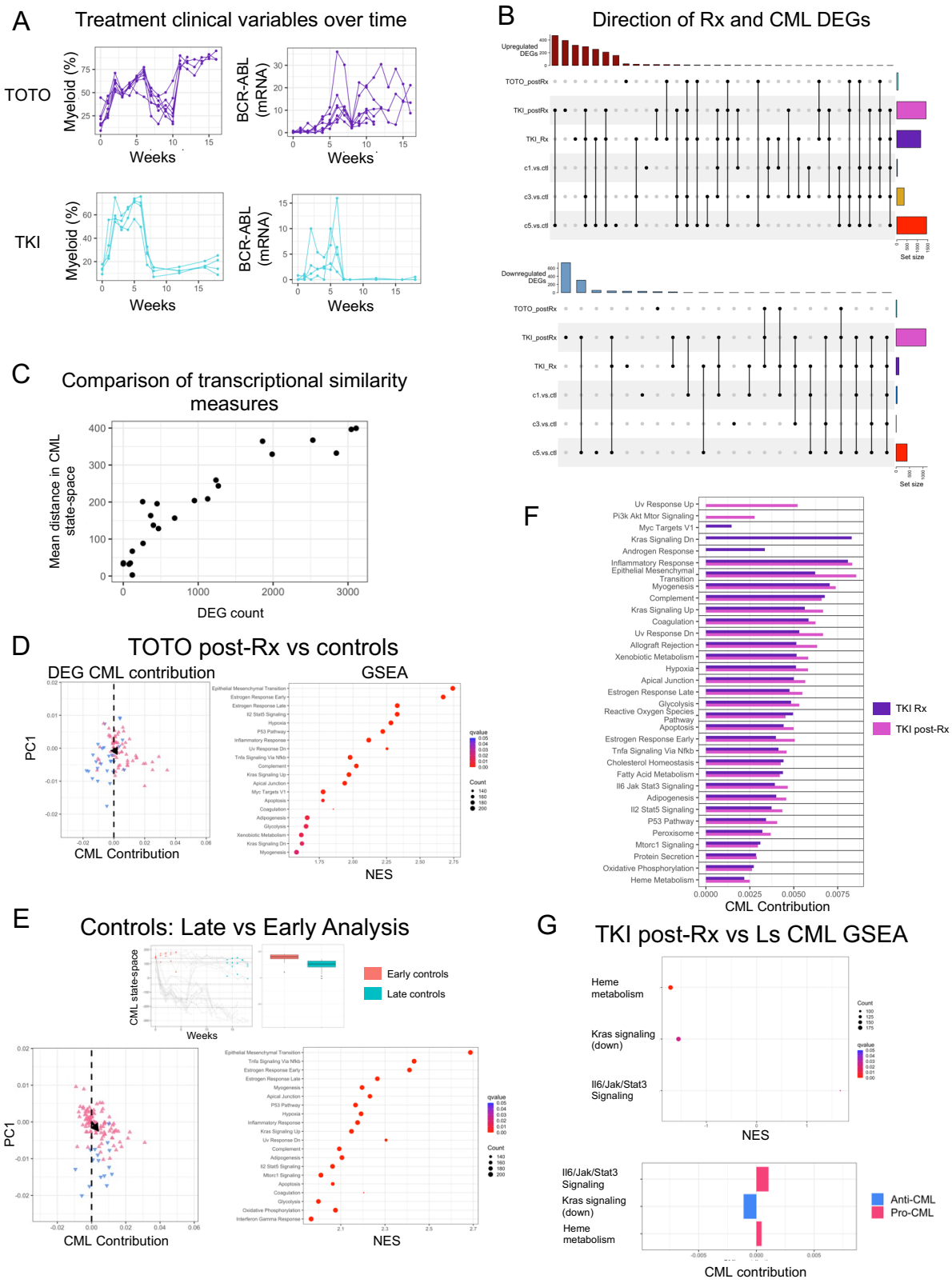


Figure S7: Analysis of treatment cohorts.

A) Clinical variables plotted vs time for Tet-on Tet-off (TOTO) mice (*top*) and TKI mice (*bottom*). **B)** Intersection up- and down-regulated treatment cohorts with the disease states (Es, Ts, and Ls) compared with the control (Hs). **C)** To compare two measures of transcriptomic similarity between two groups of samples, the number of DEGs was compared to the average distance between the samples of each group in the CML state-space for all DEG comparisons. **D)** The CML contribution of the DEG (*left*) and the significantly enriched Hallmark gene sets from GSEA (adjusted p-value < 0.001; *right*) that resulted from the TOTO post-Rx samples vs healthy control (Hs) samples. **E)** An analysis comparing the early vs late control CML samples due to the observed change in CML state-space location over the course of the experiment (*top*). The total CML contribution was calculated for all DEGs that resulted from comparing the early vs late control samples (*bottom left*) which showed a small pro-CML effect in the late control mice. GSEA was also performed, and all significantly enriched Hallmark gene set pathways were upregulated in the late control mice (adjusted p-value < 0.001; *bottom right*). **F)** Total CML contribution for all significantly enriched Hallmark gene sets after GSEA for the TKI Rx and TKI post-Rx samples when compared to controls (adjusted p-value < 0.001). Missing bars indicate a gene set that was not significant in that comparison. **G)** GSEA results from the TKI post-Rx samples compared to the Ls CML samples. Significantly enriched Hallmark gene sets were shown with enrichment in expression with respect to the TKI post-Rx samples (*top*). The total CML contribution of the TKI post-Rx samples was summarized for each gene set (*bottom*).

Table S1: Correlation analysis of BCR-ABL expression and myeloid population with each principal component.

Table S2: DEG and GSEA results for both CML disease state comparisons (Es, Ts, Ls) vs healthy controls (Hs) and inter-CML disease state comparisons.

Table S3: Hallmark gene set enrichment analysis for state-transition driver genes at T-Es and T-Ls.

Table S4: DEG and GSEA results for treatment groups (TOTO post-Rx, TKI Rx, TKI post-Rx) compared to both CML disease states (Es, Ts, Ls) and healthy controls (Hs)