Deep generative model deciphers derailed trajectories in acute myeloid leukemia

Achille Nazaret[1,2*], Joy Linyue Fan[2,3*], Vincent-Philippe Lavallée[4,5,6*], Andrew E. Cornish[4,7,8], Vaidotas Kiseliovas[4,9], Ignas Masilionis[4,9], Jaeyoung Chun[4,9], Robert L. Bowman[10,11], Shira E. Eisman[10], James Wang[1], Lingting Shi[2], Ross L. Levine[10], Linas Mazutis[4,9,12], David Blei[1,13,14], Dana Pe'er[4,15#], Elham Azizi[1-3,14,16#]

[1]Department of Computer Science, Columbia University, New York, NY 10027, USA
[2]Irving Institute for Cancer Dynamics, Columbia University, New York, NY 10027, USA
[3]Department of Biomedical Engineering, Columbia University, New York, NY 10027, USA
[4]Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA.
[5]Centre Hospitalier Universitaire Sainte-Justine Research Center, Montréal, QC, Canada
[6]Department of Pediatrics, Université de Montréal, Montréal, QC, Canada
[7]Immunology Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[8]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA.
[9]Alan and Sandra Gerry Metastasis and Tumor Ecosystems Center, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[10]Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[11]Department of Cancer Biology, University of Pennsylvania, Philadelphia, PA 19104, USA
[12]Institute of Biotechnology Vilnius University, Life Sciences Centre, Vilnius 02158, Lithuania
[13]Department of Statistics, Columbia University, New York, NY 10027, USA
[14]Data Science Institute, Columbia University, New York, NY 10027, USA
[15]Howard Hughes Medical Institute, Memorial Sloan Kettering Cancer Center, New York 10027, NY 10065, USA
[16]Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY 10032, USA
*equal contribution
[#]correspondence: dpeer@mskcc.org, ea2690@columbia.edu

ABSTRACT

Single-cell genomics has the potential to map cell states and their dynamics in an unbiased way in response to perturbations like disease. However, elucidating the cell-state transitions from healthy to disease requires analyzing data from perturbed samples jointly with unperturbed reference samples. Existing methods for integrating and jointly visualizing single-cell datasets from distinct contexts tend to remove key biological differences or do not correctly harmonize shared mechanisms. We present Decipher, a model that combines variational autoencoders with deep exponential families to reconstruct derailed trajectories (https://github.com/azizilab/decipher). Decipher jointly represents normal and perturbed single-cell RNA-seq datasets, revealing shared and disrupted dynamics. It further introduces a novel approach to visualize data, without the need for methods such as UMAP or TSNE. We demonstrate Decipher on data from acute myeloid leukemia patient bone marrow specimens, showing that it successfully characterizes the divergence from normal hematopoiesis and identifies transcriptional programs that become disrupted in each patient when they acquire *NPM1* driver mutations.

INTRODUCTION

Single-cell genomic technologies have empowered the comprehensive mapping of cellular states in health and disease, including cancer[1–4], inflammatory bowel disease[5], and COVID-19[6,7]. A powerful feature of this data is the ability to derive cell-state transitions and dynamics from a single snapshot using

pseudotime inference[8–10]. Reconstructing the order of events by which cells derail from their normal trajectories to assume pathological states can improve the characterization of initial disease stages, identify drivers underlying derailment from normal states, and nominate strategies for early detection and prevention. However, we lack methods to rigorously align trajectories and compare conditions, including health to disease and normal states to perturbations such as gene knockout or chemical treatment.

A prime example of derailed development occurs in acute myeloid leukemia (AML), a lethal cancer of the hematopoietic system. AML cells originate from specific subsets of bone marrow (BM) hematopoietic stem and progenitor cells (HSPCs) that acquire genetic and epigenetic abnormalities. These events lead to the accumulation of HSPC-like leukemic cells called “blasts" that fail to terminally differentiate into mature cells. However, the originating pre-leukemic and leukemic stem cell states are not well characterized in AML[11,12], and our inability to target and eradicate these cells leads to disease recurrence[13–15].

AML comprises many genetic subtypes, such that the deviation from normal hematopoiesis can differ across patients[16]. While the genetic events driving leukemic transformation in AML have been exhaustively cataloged, we do not know how specific mutations engender distinct pathological trajectories, which can initiate from healthy states or pre- and early-malignant states such as clonal hematopoiesis (CH) and myelodysplastic syndromes (MDS)[17–19,20,21]. Numerous other contexts, including disorders of embryonic development, neurodegenerative diseases, and T-cell exhaustion[22,23,24,25], face a similar challenge of aberrant trajectory inference.

To visualize and compare trajectories, it is critical to obtain a faithful joint embedding. Whereas many methods visualize data by projecting it onto two dimensions[26,27,28,29], this often distorts topology and obscures functional relationships[30]. Embedding multiple samples, especially for heterogeneous cancers, is also sensitive to minor gene program differences between samples, such that they typically fail to co-embed in a biologically meaningful way. Existing integration methods[31,32,33,34,35,36] are primarily designed for batch correction; they assume that samples share similar cell states and attempt to eliminate differences as technical effects. This can also result in removing genuine biological differences—particularly for trajectories and other continuous contexts.

Moreover, most approaches compress information from thousands of genes into 10–50 independent factors, constituting a latent space that neglects dependencies between related biological processes. As a result, latent spaces are often useful for annotating discrete cell types but do not preserve cell order in healthy trajectories or disease progression[37]. These limitations highlight the need for approaches that address interpretability, preserve global geometry, and capture confounding mechanisms to better model trajectories perturbed by mutation, genetic manipulation, drugs, or disease.

In this work, we present Decipher (deep characterization of phenotypic derailment), an interpretable deep generative model for the simultaneous integration and visualization of single-cell data from two disparate conditions. A key principle behind Decipher is that perturbed and unperturbed trajectories include both shared and unique components. Our method re-examines a common assumption of single-cell dimensionality reduction methods[38]: that latent factors shaping the low-dimensional structure of cell states are independent. Decipher boosts the degree of nonlinearity over standard deep probabilistic models by adding a second latent layer to the generative model. Thus, Decipher hierarchically learns the correlation structure of the latent factors by further factorizing them into a second latent space of only two dimensions. By allowing some latent factors to be correlated, the resulting joint representation has sufficient flexibility to encompass both biological differences and

commonalities. Decipher's high-level two-dimensional latent space also provides a direct visualization of the single-cell representations without any modification to the geometry learned by the model.

To demonstrate Decipher's performance, we used single-cell RNA-seq (scRNA-seq) to profile bone marrow specimens from AML patients and healthy donors. Despite significant genomic and transcriptomic heterogeneity across patients, Decipher successfully characterized divergence from normal myeloid differentiation in each patient, as confirmed by genotyping, whereas alternative methods distorted the global geometry of trajectories. We show that these disrupted transcriptional programs are supported by single-cell epigenetic data. In addition, we use Decipher to jointly map transitions from premalignant to malignant cell states in two subtypes of gastric cancer, demonstrating its broad applicability to different systems. Indeed, Decipher represents a comparative approach to resolve trajectories across any pair of normal and perturbed conditions.

RESULTS

The Decipher framework

Aligning trajectories from normal and perturbed contexts requires a joint representation of cells that preserves their topology and order along both trajectories without forcing overlap. To achieve a more faithful and flexible joint representation that captures both biological differences and shared mechanisms, Decipher allows some latent factors to be correlated. Our key assumption is that, though a deviated trajectory can lead to a substantially different terminal cell state from normal, it can still maintain a subset of shared transcriptional programs such as normal cell maturation. Dependent factors can capture shared mechanisms between trajectories as well as divergent gene programs.

To enable correlated latent factors, Decipher extends the successful single-cell variational auto-encoder (VAE) architecture[35,39–41] into a deeper generative model inspired by the deep exponential family[42]. In Decipher, each cell has not one but two complementary latent representations (**Fig. 1a**). We first embed cells in a two-dimensional representation that encodes global cell-state dynamics. We refer to this high-level embedding as the *Decipher space* and its two dimensions as *Decipher components*. The Decipher components represent the dominant axes of variation in the data, typically maturation and disease derailment. We then generate a higher-dimensional representation conditional on the two Decipher components, which is designed to capture more refined cell state information such as transitional intermediates. We call the space of refined representations the *latent space* and refer to its dimensions as *latent factors*. The latent space is similar in principle to that of previous VAEs[35,39,40], but in Decipher, latent factors can have dependencies between them. Finally, Decipher generates the denoised gene expression of each cell from its inferred cell state (**Fig. 1b** and Methods).

With this three-step generative process, Decipher interpolates between different degrees of non-linearity and generates latent factors whose dependencies are shaped by high-level Decipher components. A major advantage is that Decipher components can be used for direct 2D visualization of the cell representations, eliminating the need for further dimensionality reduction, such as with tSNE or UMAP, whose usage has been a subject of debate[43]. Another key feature is the ability to reconstruct derailed trajectories and a joint pseudotime with direct use of the Decipher components and latent factors (**Fig. 1c**). This is accomplished by clustering cells in the latent space and constructing a minimum spanning tree to link clusters then interpolating between clusters in Decipher space to create a *Decipher time* (see Methods).

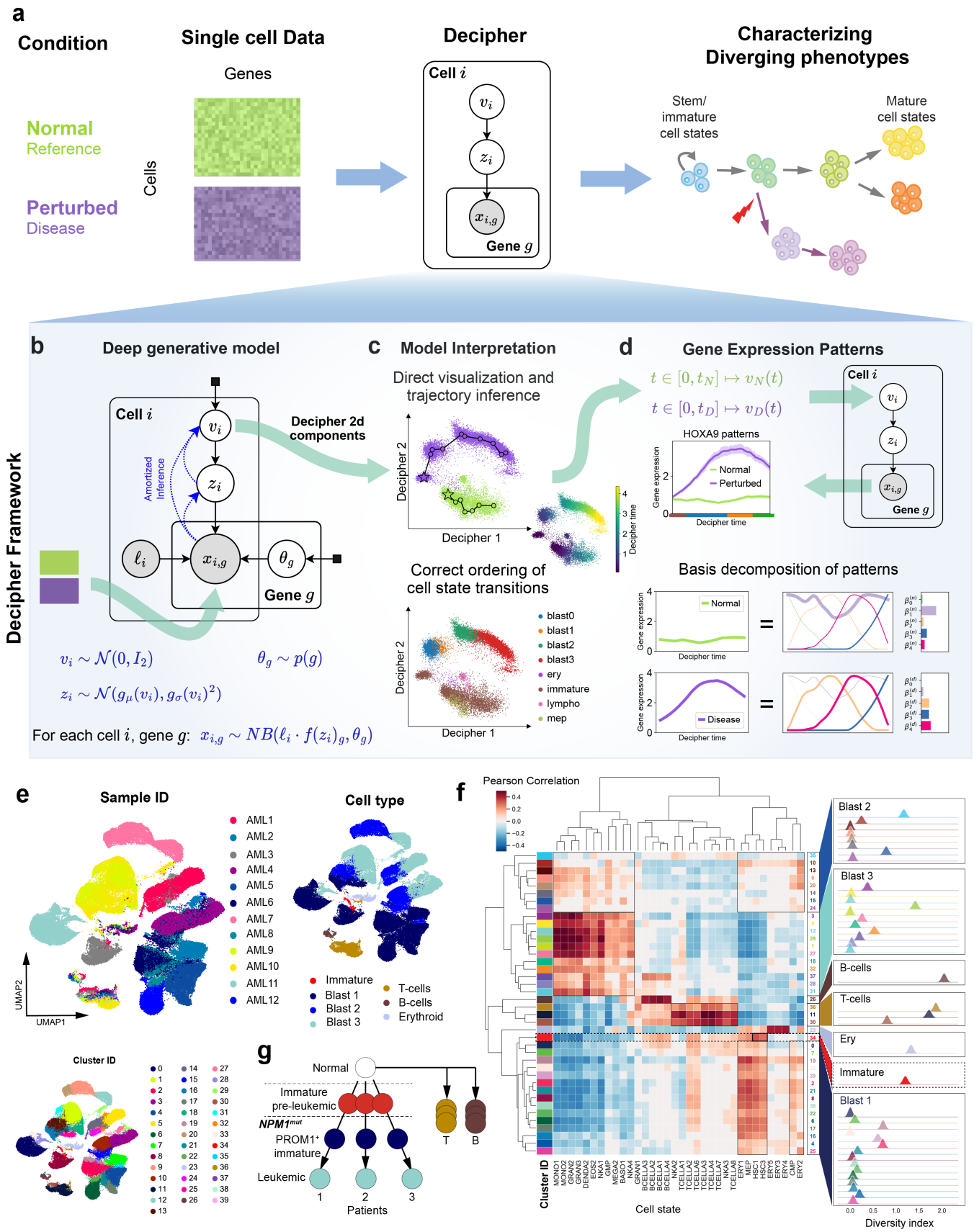


Figure 1. Overview of the Decipher framework and AML data. (a) Decipher input: two single-cell datasets from a normal condition (reference) and a perturbed (e.g. disease) condition; Decipher graphical model: The probabilistic model and inference learn to represent the two datasets in a shared latent space without removing biological differences; Decipher's interpretation:

The latent space reveals shared cell-state transitions and characterizes diverging phenotypes. **(b)** Detailed graphical representation of Decipher's deep generative model with the corresponding distributions. Decipher has three levels of cell representations: the 2D Decipher components *v*, the Decipher latent factors *z*, and the gene expression *x*. The Decipher components summarize heterogeneous cell states and are used for direct visualization of the latent space. **(c)** Example visualization of the Decipher space, colored by the dataset of origin (normal or disease) and by manually annotated cell states. Two distinct trajectories are identified and computed in the Decipher space. **(d)** Computation of the gene expression patterns along each trajectory using Decipher's mapping from *v* to *x* and decomposed into representative patterns (basis). The corresponding weights are used to compare patterns between the two contexts to pinpoint disrupted and conserved genes. **(e)** 2D UMAP projection of unsorted BM transcriptomes from 12 samples from 11 AML patients (AML 9 and AML10 are biopsies from the same patient). Each dot represents a cell colored by patient ID, cluster ID, and annotated cell type. **(f)** Pearson correlation between cluster centroid expression and bulk gene expression data from sorted subsets of healthy HSPCs[44] (left; Methods). Shannon Diversity index computed for distribution of patients in each cluster with controlling for cluster size (right). Higher Diversity showing more mixing of patients (Methods) **(g)** Illustration of AML patient-specific trajectories with shared initial states, similar immature NPM1 WT cells, and divergent terminal states after NPM1 mutation. Patient-specific trajectories are shown in **Extended Data Fig. 2a**.

In addition, Decipher's generative model and inference are formulated to give explicit mapping functions between gene space and Decipher space, enabling a straightforward reconstruction of gene expression anywhere in Decipher space (Methods) and the imputation of gene trends along the entire trajectory (**Fig. 1d**). It is especially useful to know the gene expression levels in locations with few observed cells. It also enables the reciprocal computation of Decipher components for any cell with measured gene expression for a straightforward visualization. In contrast, there is no explicit mapping between the UMAP space and the gene expression space. Decipher offers a unique framework for dimensionality reduction, 2D visualization, trajectory alignment, and characterization.

Enrichment of rare immature populations reveals *PROM1* as a marker of immature *NPM1*^{mut} cells

To demonstrate Decipher's capabilities on a complex disease, we applied it to understand how leukemic cells in *NPM1*^{mut} AML derail from healthy HSPCs. *NPM1* is among the most commonly mutated genes in AML, occurring in 20-30% of cases, but how it leads to leukemogenesis is not completely known[45], and we lack effective treatment options for eliminating persistent leukemic stem cells that lead to relapse in a large proportion of patients[46]. *NPM1*^{mut} AML patients often have additional mutations in epigenetic modifiers (e.g., *TET2* and *DNMT3A*), which are also among the most frequent somatic mutations in apparently healthy elderly (>65 years) individuals, a condition known as clonal hematopoiesis associated with an elevated relative risk of progression to myeloid malignancy[47]. These epigenetic mutations likely occur earlier in pre-leukemic hematopoietic stem cells[48] and can facilitate AML development following subsequent acquisition of genetic lesions, such as *NPM1*^{mut}, in a process of clonal evolution[49]. The transcriptomic consequences of *NPM1* mutation and how it is influenced by pre-existing epigenetic abnormalities remain unclear.

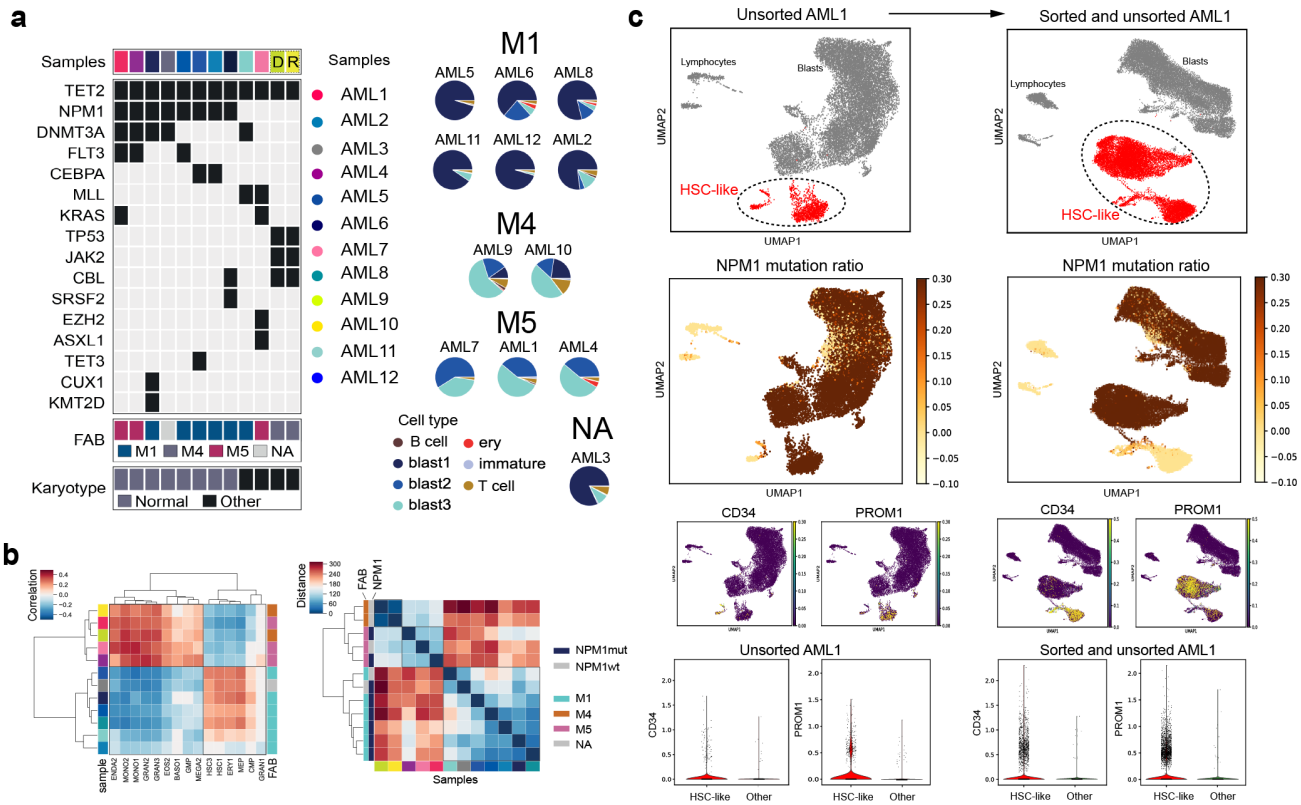
To characterize the impact of *NPM1*^{mut} and reduce confounding effects, we focused on patients bearing *TET2* epigenetic mutations (n = 12), with and without *NPM1* mutations, and used scRNA-seq to profile 104,116 AML bone marrow (BM) cells (**Fig. 1e**, **Extended Data Fig. 1a** and **Supplementary Table 1**). Assuming that leukemic cells preserve a subset of normal hematopoietic differentiation programs, we used well-characterized normal references[44] to achieve cluster-level annotation of maturation stage (**Fig. 1f**; Methods). This analysis revealed a wide spectrum of leukemic cell states that partially recapitulates microscopic assessment of the diverse morphological subtypes in our cohort[50,51] (**Extended Data Fig. 1a**): while undifferentiated (M1) samples are mostly populated by immature blasts and myelomonocytic (M4) and monoblastic (M5) samples by cells with greater maturation, a variety of cell types characterized most samples (**Extended Data Fig. 1b**).

Similar to previous studies[16,52], we found significant inter-patient heterogeneity in leukemic blast cells, which are unlikely to be technical effects given that lymphocytes are well mixed (**Fig. 1f**). Surprisingly, we found that the most immature cluster (448 cells; 0.4% of total) characterized by a marked HSC-like phenotype (**Fig. 1f**), was the top non-lymphoid cluster shared by most patients. The vast majority of *NPM1* mutations occur at the 3' end of the gene[45], which, combined with high *NPM1* expression in AML, makes it possible to detect these mutations directly from scRNA-seq data[53]. Interestingly, scanning for *NPM1* mutations (Methods) identified a subset of cells in this HSC-like population that are NPM1^{wt} (**Extended Data Fig. 1c**). These results suggest that while leukemic cells display marked heterogeneity, immature leukemic or pre-leukemic cells are phenotypically more similar across patients, motivating reconstruction of trajectories diverging from normal HSPCs upon *NPM1* mutation.

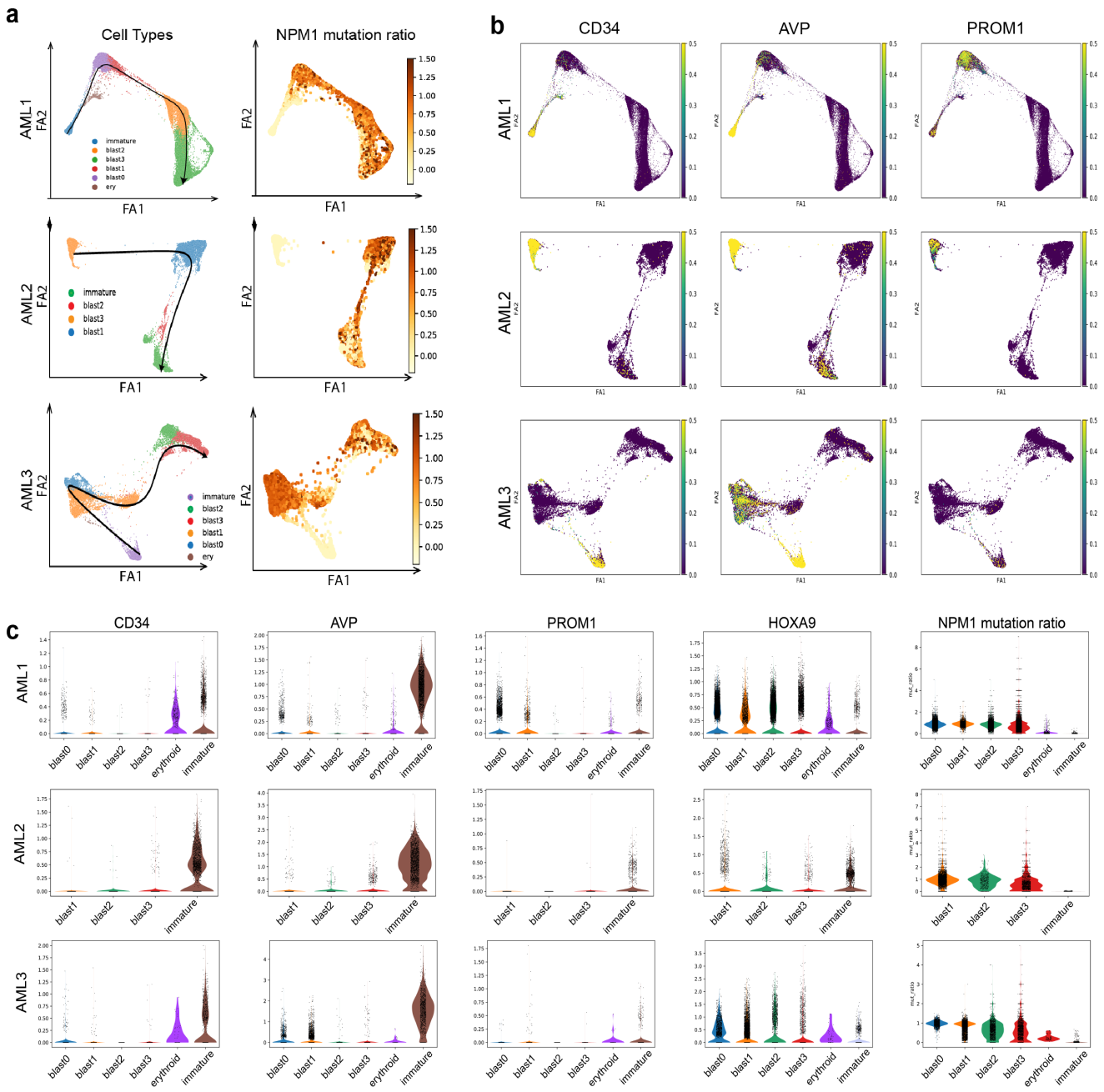
There are too few HSC-like cells to effectively characterize *NPM1*-mediated derailment, so we enriched the immature population in *NPM1*^{mut} patients. Specifically, we identified two differentially expressed genes in this population that code for surface proteins, namely CD34 (log fold change: 6.67, adjusted $p < 1e-6$) and a less-conventional marker, PROM1 (CD133[54]; log fold change = 7.23, adjusted $p < 1e-6$), and sorted for cells expressing either CD34 or PROM1. This approach successfully expanded the target population in patient AML1 from 160 to 37,395 cells (**Extended Data Fig. 1c**), for a total of 72,945 sorted cells from AML1–AML4 (Methods). Interestingly, we found that these rare cells clustered into two broad populations defined by the presence or absence of *NPM1* driver mutations—a pre-leukemic immature (*NPM1*^{wt}) population closest to HSCs and an *NPM1*^{mut} population most resembling progenitor populations (labeled blast0,1; **Extended Data Fig. 1c; Extended Data Fig. 2a,b**). Both expanded populations have markedly different expression profiles than leukemic cell states (labeled blast2,3 and most resembling myeloid committed cells) in AML1–AML3 (**Extended Data Fig. 2a,b**) suggesting diverging leukemic states across patients (**Fig. 1g)**.

Importantly, comparison of pre-leukemic cell and normal HSC phenotypes consistently returned *PROM1* as a marker for the immature population in human *NPM1*^{mut} AML samples. This analysis also identified consistent underexpression of HSC genes such as *CD34* and *AVP* and overexpression of *HOX* cluster genes in *NPM1*^{mut} cells (**Extended Data Fig. 2c**), concordant with the demonstrated ability for *NPM1* mutations to induce *HOX* genes *in vitro*[55] and confer self-renewal properties to myeloid progenitors in mice[48]. Overall, this approach provides an expanded map of both frequent and rare cellular populations that exist in each patient, and unambiguously delineates *NPM1* mutation status.

The shared initial state and divergent terminal states across patients suggest patient-specific trajectories (**Fig. 1g**), motivating a comparison of immature populations in each patient to a normal HSPC reference trajectory. Specifically, we aimed to elucidate shared and distinct transcriptional shifts at different stages of leukemic initiation, *NPM1*^{mut} acquisition and leukemic blast development. We thus integrated single-cell data from each AML patient with data from a healthy individual sorted for CD34^{+} marrow cells. Given the disparity between the two samples, their joint visualization using UMAP or force-directed layouts (FDL) does not show overlapping cell states or distorts global transitions and the order of blast maturation stages (**Extended Data Fig. 3a**). In addition, harmonization methods tend to force all cell states, including leukemia cells and normal HSPCs, to overlap, thus failing to achieve an interpretable characterization of derailed differentiation in leukemia (**Extended Data Fig. 3b**). These observations demand frameworks capable of obtaining improved representation of leukemia and normal cells and a joint pseudotime between AML and normal that is faithful to cell-state order.



Extended Data Figure 1. Details of AML samples and enrichment of immature cells (a) Targeted mutation, karyotype, and morphological (FAB) assessments of patient BM samples (**Supplementary Table 1**). D, R denote diagnosis and relapse paired samples from the same patient. Pie charts indicate composition of cell types according to annotations in **Fig. 1f. (b)** Left: Pearson correlation between sample centroid expression and bulk gene expression data from sorted subsets of healthy HSPCs (Methods); unsupervised hierarchical clustering matches FAB classification. Right: Euclidean distance between sample centroid expressions confirming stratification by FAB. **(c)** UMAP projection of AML1 single-cell transcriptomes before sorting (left) and with the inclusion of sorted CD34^{+}/PROM1^{+} cells (right). HSC-like cells are marked in red showing enrichment of this population with sorting (top). Cells are colored by the proportion of NPM1 mutation in the neighborhood of each cell (Methods) and markers (middle). Violin plots show the differential expression of CD34 and PROM1 in the HSC-like population enriched with sorting (bottom).

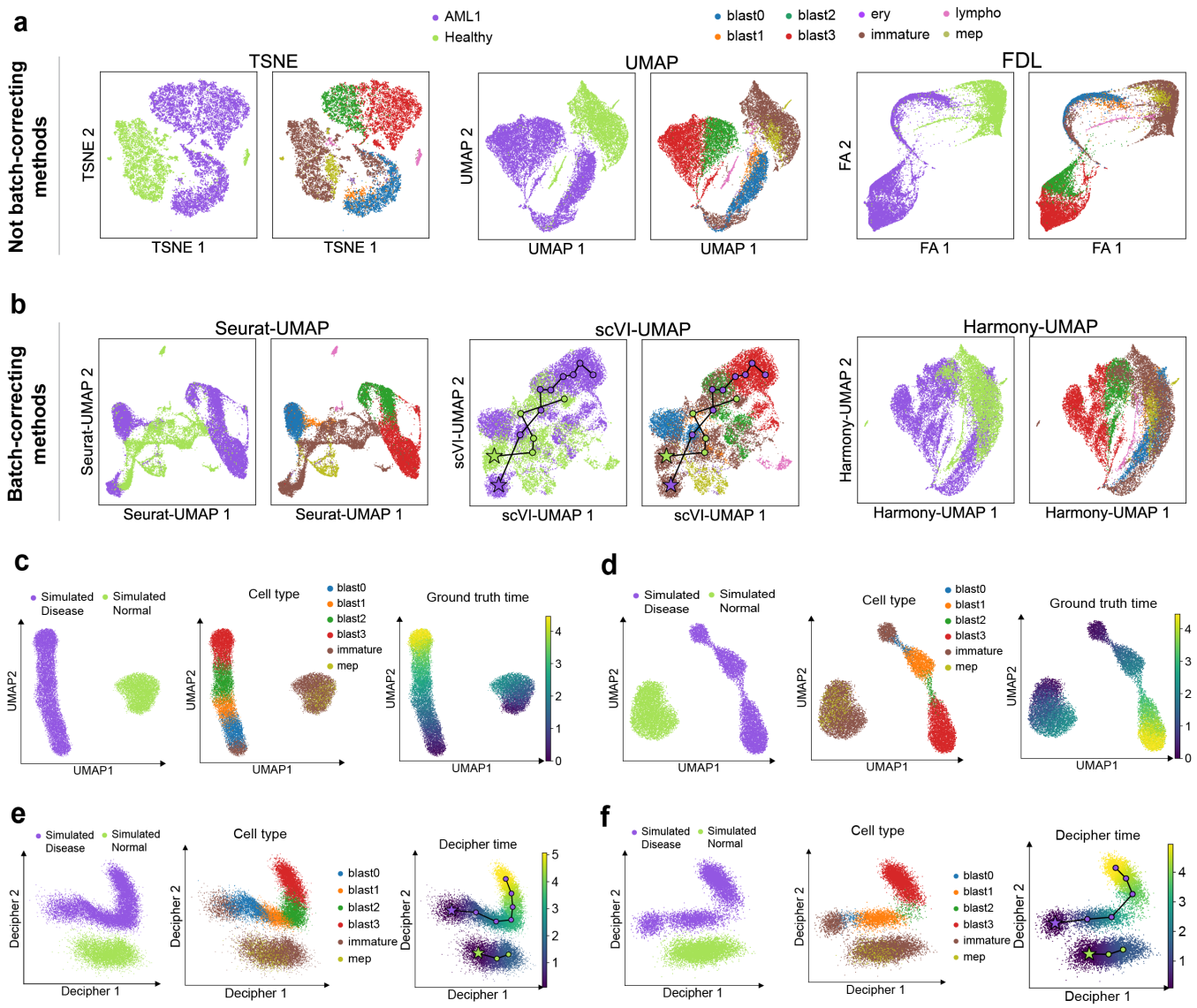


Extended Data Figure 2. (a) Force-directed layouts (FDL) of single-cell transcriptomes in three AML patients colored by cell type (left); proportion of NPM1 mutation (right; Methods). **(b)** Expression of key marker genes across three AML patients shown in the FDL. **(c)** Violin plots showing expression of key marker genes (*CD34*, *AVP*, *PROM1*, *HOXA9*) in identified cell types.

Performance of Decipher on simulated data

We first assessed Decipher's performance on data simulated in a semi-synthetic fashion using the generative model of Decipher (**Extended Data Fig. 3c-f**). We started by fitting Decipher to real data (normal HSPCs and sorted cells from AML1) in order to obtain realistic values for the probabilistic model. We then simulated new data by (i) forming two divergent ground-truth trajectories in the Decipher component space (Methods), (ii) generating synthetic cells by sampling their Decipher components uniformly along these trajectories, and (iii) generating synthetic gene expression values for these simulated cells using our probabilistic model (**Extended Data Fig. 3c**). We subsequently fitted another

Decipher instance to this simulated dataset. Decipher successfully learned a latent space in which cell-state transitions from immature to blast0–3 and both normal and AML trajectories are correctly aligned along the Decipher 1 component (**Extended Data Fig. 3e**). The derailment in the AML condition is captured by the Decipher 2 component, which precisely locates the bifurcation point. We assessed the recovery of cell order by evaluating the correlation between the inferred Decipher time and the ground-truth trajectory. To confirm the robustness of these results, we repeated the entire Decipher pipeline across 5 simulated datasets and consistently observed a correlation above 0.94. Finally, we examined robustness in the context of rare or poorly represented cell states, which commonly occur in real samples. To do so, we removed 90% of cells from two intermediary cell states (**Extended Data Fig. 3d**) and fitted Decipher on this new simulated data. Even with non-uniform sampling of cell states, Decipher successfully preserved the order of cell states (**Extended Data Fig. 3f**).



Extended Data Figure 3. Integration of AML and normal samples with existing methods and Decipher performance on simulated data (a,b) Projection of 37,395 cells from patient AML1 and from a healthy donor using different tools (Methods). Each dot represents a cell colored by origin (left) and type (right). Non-batch-correcting visualization methods TSNE, UMAP, and Force Directed Layout (FDL) (a), and batch-correcting methods – UMAP of the Seurat space, scVI space, and Harmony space (b) – all fail to integrate AML and healthy cells. Lines indicate trajectories inferred on the scVI space (Methods). **(c)** UMAP visualization of semi-synthetic data (18,000 cells) simulating an AML patient and a healthy reference (Methods) along simulated

trajectories. Each dot represents a cell colored by sample (left), cell type (middle), and ground truth time (right). **(d)** UMAP visualization of the data simulating rare cell types. Similar to (c) but removing 90% of blast0 (blue) and blast2 (green). Each dot represents a cell colored by sample (left), cell type (middle), and ground truth time (right). **(e)** Example of Decipher applied to the data in (c). It preserves the order of cell states and shows the shared maturation (Decipher1) as well as the divergence of mature leukemic blast cells (blast2,3) in AML (Decipher2). Each dot represents a cell colored by sample (left), cell type (middle), and inferred Decipher time (right). **(f)** Decipher applied to data shown in (d) is still able to learn meaningful representation and trajectories.

Decipher characterizes derailed development and inflammation in AML

To characterize personalized trajectories, we used Decipher to generate joint representations of sorted cells from a healthy donor and just one AML patient at a time. Although cells from patient AML1 and healthy samples do not overlap in Decipher space, we found that they are spatially organized to reflect their similarities and differences. We further rotated the axes of the latent space to align the Decipher components with interpretable biological mechanisms (Methods). Specifically, the Decipher 1 component is associated with cell maturation in both conditions; cell types (which we annotated independently) are coherent with maturation stages, stem markers (*CD34, AVP*) decrease and myeloid differentiation genes (*MPO, LYZ*) increase along Decipher 1, and PROM1 is specifically upregulated in *NPM1*^{mut} blast 0 and blast 1 cell states (**Fig. 2a**). Decipher 2 distinguishes leukemic progression from normal cell states and, remarkably, captures the rise of *NPM1* mutation and leukemic initiation, placing *NPM1*^{wt} cells closest to the normal states (**Fig. 2a**). We performed gene set enrichment analysis (GSEA) on cells projected onto Decipher 2, and identified TNFa signaling, inflammatory response, IL6/JAK/STAT3 signaling, IFNg response, and KRAS signaling pathways (**Supplementary Table 4**). These findings comport with the well-elucidated functional role of *Tet2* in repressing IL6 transcription[56], as well as loss-of-function mutations in *Tet2* associating with an increased abundance of inflammatory myeloid populations in conditions ranging from CH-related atherosclerosis[57] to most relevantly AML[58]. While our data does not contain significant non-leukemic myeloid populations and cannot resolve the cellular source of the IL6 and other cytokines responsible for the induction of these programs, their stark upregulation along the Decipher 2 component supports work in *Tet2* murine models[59] suggesting this inflammation drives cellular plasticity enabling leukemogenesis, rather than merely being coincident to it.[58–62].

Decipher also automatically computes a trajectory in Decipher space for each sample (**Extended Data Fig. 4a** and Methods). Both trajectories originate from immature populations before branching (**Fig. 2a**) – the AML trajectory follows cancer progression as cells transition from *NPM1*^{wt} immature to *NPM1*^{mut} blast0, blast1, blast2, and blast3, while the healthy trajectory follows normal myeloid differentiation, as reflected by increasing *MPO* expression.

In addition to Decipher space, our model learns a higher-dimensional latent space (10 dimensions here) that captures cell states in greater detail. This hierarchical dimensionality reduction obtains more expressive factors that are not attainable by simply increasing the number of latent factors in other methods[42]. Since Decipher's latent factors can be correlated (**Fig. 2b**), it allows for complex dependencies. Importantly, correlated factors are capable of capturing overlapping transcriptional programs between trajectories as well as shared mechanisms underlying consecutive cell state transitions. For example, z7 is mainly encoded with Decipher 1 (**Extended Data Fig. 4b**) and captures common trends along both normal and AML trajectories (**Extended Data Fig. 4c**).

These latent factors are also capable of highlighting features that distinguish normal and perturbed states. For example the first two factors (z1, z2) are correlated in immature cells but not in blasts (**Fig. 2b**). These factors are mainly encoded by Decipher 2 and capture different cell-state transitions along AML derailment (**Extended Data Fig. 4b, Fig. 2c**). Factor z1 expression is highest early

in blast formation (marking blast0, 1 and 2), z10 marks an intermediate (from blast1 to 2), and z2 marks the final stage of leukemic maturation (blast3) (**Fig. 2c**). Applying GSEA, we found that z1 is enriched in reduction in oxidative phosphorylation, a pathway that is altered in leukemic stem cells[63], and z10 and z2 are enriched for TNFα, IFNg and inflammation[59,64] (**Fig. 2c** and **Supplementary Table 2**). Indeed, the enrichment of IFNg and inflammation in z2 (**Supplementary Table 4**) is expected as it captures most mature myelo/monoblastic cells (**Extended Data Fig. 4c**)[59,64]. Decipher thus enables a more comprehensive and nuanced understanding of cellular dynamics and lineage relationships, especially when integrating normal and perturbed conditions.

Decipher reconstructs continuous and divergent trajectories and highlights patient-specific features in AML

The ability to capture dependencies in the latent space is a unique feature of Decipher, whereas the independent factors retrieved by other methods such as scVI[35] can remove biological differences between cell states, thereby collapsing the two conditions onto each other in every latent dimension (**Fig. 2b**), deforming the geometry and disrupting continuous trajectories.

To systematically and quantitatively benchmark the ability of different methods to characterize derailed trajectories, we defined two metrics that evaluate biological faithfulness (**Fig. 2d** and Methods). The *ordering score* evaluates whether the latent space of each method orders cell states correctly, according to known maturation stages; for instance, blast1 should be located between blast0 and blast2. The *divergence score* assesses the preservation of biological differences between normal and disease - it rewards the proximity of early immature cells between normal and perturbed samples, and penalizes mixing between terminal healthy and progressed disease states.

We applied these metrics to a range of commonly used visualization (FDL[65], UMAP[27,65], tSNE[66], Phate[67]), dimension reduction (PCA), and integration methods (Seurat[68], scVI[35], Harmony[69]) (**Fig. 2e** and **Extended Data Fig. 3a,b**). Some methods, such as FDL, fail to produce a latent space with correct cell-state ordering. Common 2D visualization methods such as UMAP or FDL fail to recover proximity of immature leukemic cells and normal cells, despite their similarities (possibly due to insufficient batch correction). Other methods such as scVI or Seurat tend to mix leukemic blasts and normal HSPCs despite their distinct characteristics, possibly due to overcorrection during integration. Decipher components as well as latent factors scored highest in both metrics, demonstrating that the method best balances the trade-off between integrating across conditions and preserving their unique geometries and cell states (**Fig. 2e**).

Decipher also successfully captured global geometry and identified similar structures and biological processes in patients AML2 and AML3, outperforming other methods (**Fig. 2e** and **Extended Data Fig. 4d,e**). The Decipher space of both patients has a similar structure; AML and healthy immature cells are embedded close to each other, and the two trajectories branch out, correctly ordered by maturation. The Decipher 2 axis shows significant enrichment for TNFα, inflammatory response, and hypoxia in all three patients (**Supplementary Table 4**), suggesting that similar mechanisms are behind the leukemic initiation captured by this axis.

The Decipher results also highlight key differences between patients. There is a larger gap between *NPM1*^{wt} and *NPM1*^{mut} cells in the Decipher space for AML2, reinforced by the absence of detected blast0 cells in this patient (**Extended Data Fig. 4d**), demonstrating Decipher's ability to preserve global geometry and distances to reconstruct derailed trajectories despite nonuniform sampling of cells. In AML3, branching occurs during blast1 rather than before blast0 (**Extended Data Fig. 4e**), suggesting

that derailment occurred at a later stage in this patient than the other two. Such differences highlight the significant heterogeneity among AML patients and argue forcefully for the development of personalized models that can capture shared and unique features of each disease trajectory.

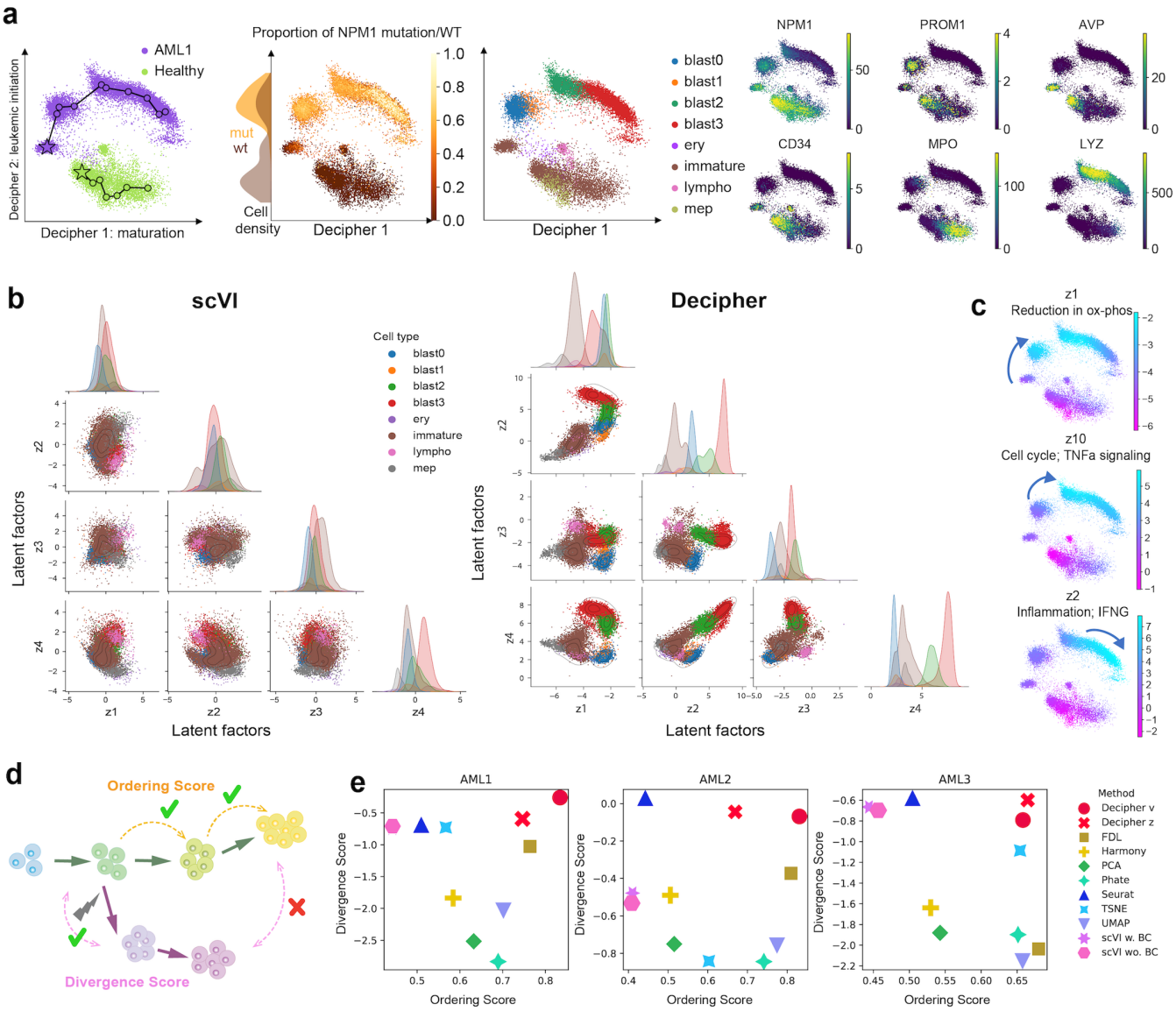
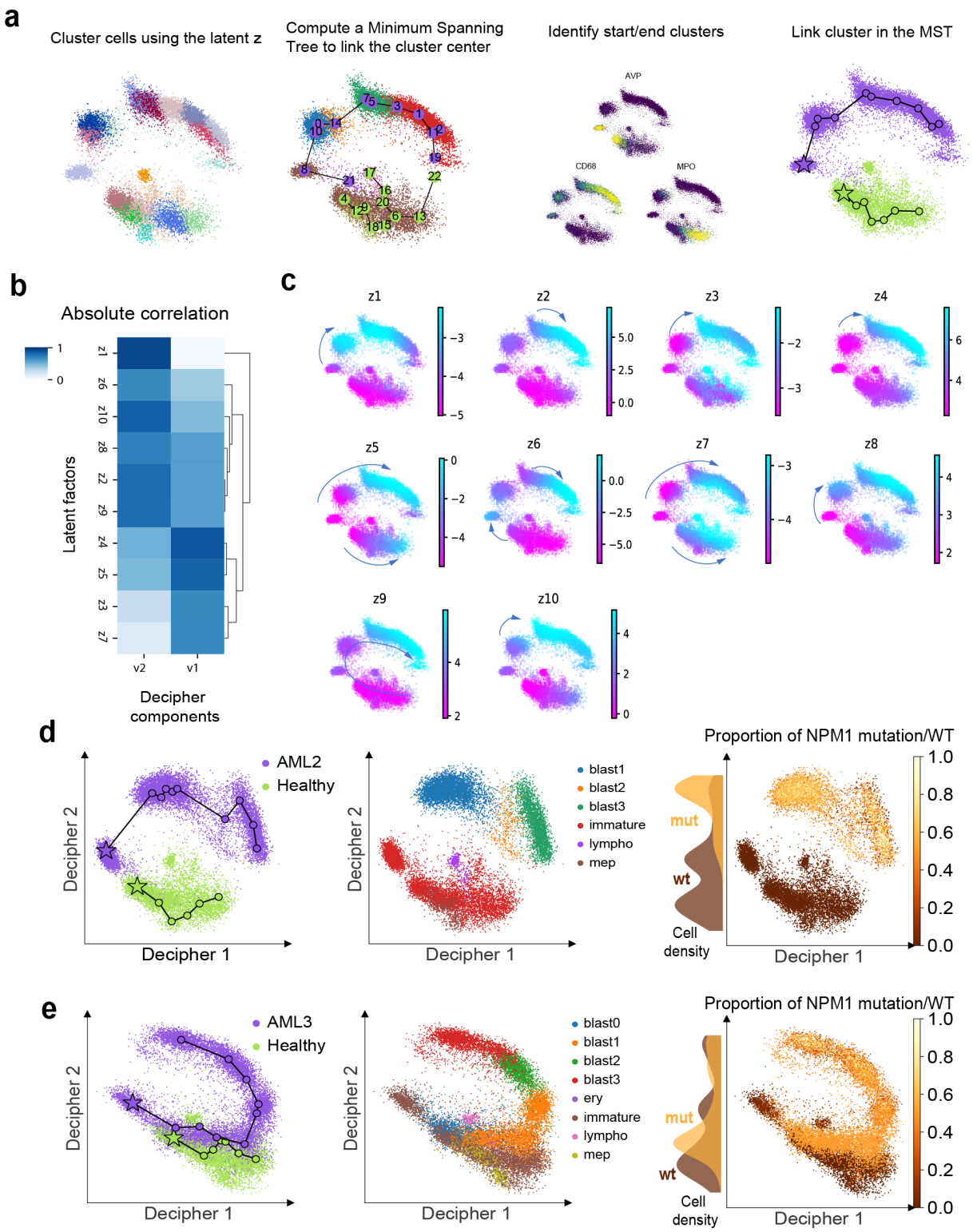


Figure 2. Decipher reconstructs derailed myeloid developmental trajectories in AML. (a) Decipher space embedding of 37,395 sorted cells from patient AML1 and a healthy donor (Methods), colored by sample of origin, cell type, proportion of *NPM1* mutation (Methods), and expression of select marker genes. Decipher space preserves the biological difference between the two samples while meaningfully organizing the cells along a maturation axis (Decipher 1) and a leukemia initiation axis (Decipher 2). Stars denote strating point of trajectories and dots show cluster centroids used for trajectory inference (Methods). Results for two other AML patients are shown in **Extended Data Fig. 4d,e**. **(b)** Comparison of scVI[35] (left) and Decipher (right) latent spaces for pairs of the first 4 latent dimensions (the other pairs are similar in resolving cell types). The scVI latent space collapses biological differences while Decipher preserves them. **(c)** Decipher space colored by latent factors z1, z10 and z2, each capturing a different state transition. **(d,e)** Benchmarking of Decipher against other dimensionality reduction and harmonization methods, using two metrics (d) that measure the interpretability of each method's latent space and faithfulness to the underlying biology (e; Methods).



Extended Data Figure 4. Details of trajectory inference, latent factors, and application to other patients. (a) Illustration of steps involved in trajectory inference, based on the inferred representation from Decipher (Methods). **(b)** Absolute value of correlation between the Decipher components and the latent factors. **(c)** Decipher space colored by all latent factors for AML1, a subset of which is also shown in **Fig. 2c**. Arrows indicate cell state transitions corresponding to an increase in the latent factor. **(d,e)** Decipher space visualization of cells from healthy donor and AML2 (c) or AML3 (d). Cells colored by condition (left), cell type (middle) and proportion of *NPM1* mutation (defined as number of neighbors with mutation/all neighbors for each cell, using 30 neighbors) (right). Lines, defined trajectories for normal and healthy samples (left).

Decipher delineates the altered landscape of transcriptional regulation in AML

To gain a deeper understanding of derailment from normal hematopoiesis, which exhibits heterogeneous trajectories across patients, we sought to compare individual gene expression patterns and underlying regulation, between normal and perturbed trajectories. This requires an accurate reconstruction of individual gene trends such that the relative expression, timing and temporal order of events is preserved, and comparable between trajectories - a challenging task for standard integration methods. Decipher's decoders can directly transform trajectories from the joint representation space into the gene expression space, obtaining the mean and standard deviation of gene expression (**Fig. 3a**; Methods), providing gene expression corresponding to any point on the trajectories, including sparsely sampled states. Decipher reconstructed gene trends faithfully match those of each individual dataset (healthy and AML) and the relative expression across datasets (**Fig. 3b,c** and **Extended Data Fig 5a,b**). For example, Decipher recapitulates the clear decline of CD34[70,71] and rise of MPO in hematopoiesis (**Fig, 3b,c**), matching the known behavior of these key markers. A similar gene reconstruction with the widely used batch-correction VAEs[35,72], does not faithfully recover gene trends (Methods). For instance, the batch-correction in scVI alters the relative expression, by design, for important markers such as *CD34* and *PROM1* (**Fig. 3b, Extended Data Fig. 5a**), which can lead to misleading conclusions on differentially expressed genes.

Another important feature of Decipher is its ability to infer a joint pseudotime (Decipher time; **Fig. 3a**) shared across trajectories, avoiding the need for additional trajectory alignment (Methods). Similar patterns for key marker genes for maturity: *AVP* (marker of stemness), *PABPC1* (protein synthesis in HSC differentiation), and *LYZ* (myeloid differentiation) in aligned segments of the trajectories confirm the inferred pseudotime is comparable between the two datasets (**Fig. 3d**), and highlights the multi-task capabilities of Decipher.

To shed light on the regulatory mechanisms underlying disease derailment, we focused on transcription factors (TFs) and used Decipher time to estimate the timing of maximum expression for individual TFs in each trajectory. Computing the distribution of peak times across all TFs, we find distinct locations along the normal hematopoietic trajectory where TFs are up-regulated in concert (**Fig. 3e**), revealing coordinated mechanisms underlying cell state transitions. In AML, however, we observe a fundamentally different landscape of TF timing, and a global loss of coordination among TFs in all patients (**Fig. 3f** and **Extended Data Fig. 5c)**. In particular, the second wave of TFs coordinated expression in the healthy hematopoiesis is lost in the AML trajectory precisely with the occurrence of NPM1 mutation (blast0).

We next examined reconstructed expression of individual TFs exhibiting altered peak timing and confirmed known increased expression of homeobox genes (*HOXA9*[73], *HOXB2*) and their cofactors *MEIS1*[71] and *PBX3*[71], and down-regulation of *GATA1*[74], at the precise location of NPM1 mutation acquisition (**Fig. 3g**). Moreover, we found altered TFs and other genes with various expression trends (e.g. *LCOR, AHR*; **Fig. 3g** and **Extended Data Figure 5d,e**). Since these complex patterns cannot be sufficiently summarized by a single peak time, we sought to develop a more systematic approach to quantify altered expression between two trajectories (Methods).

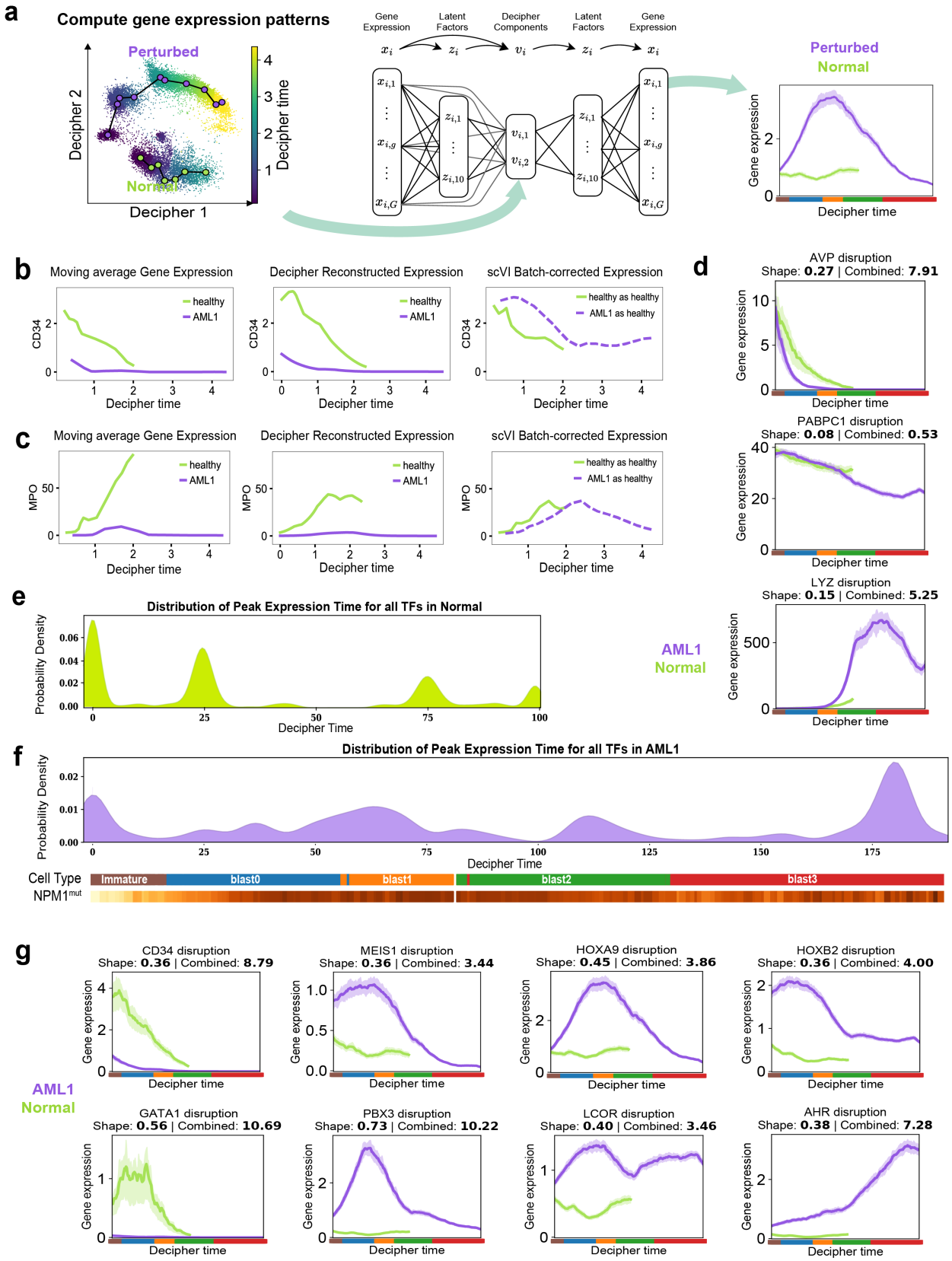
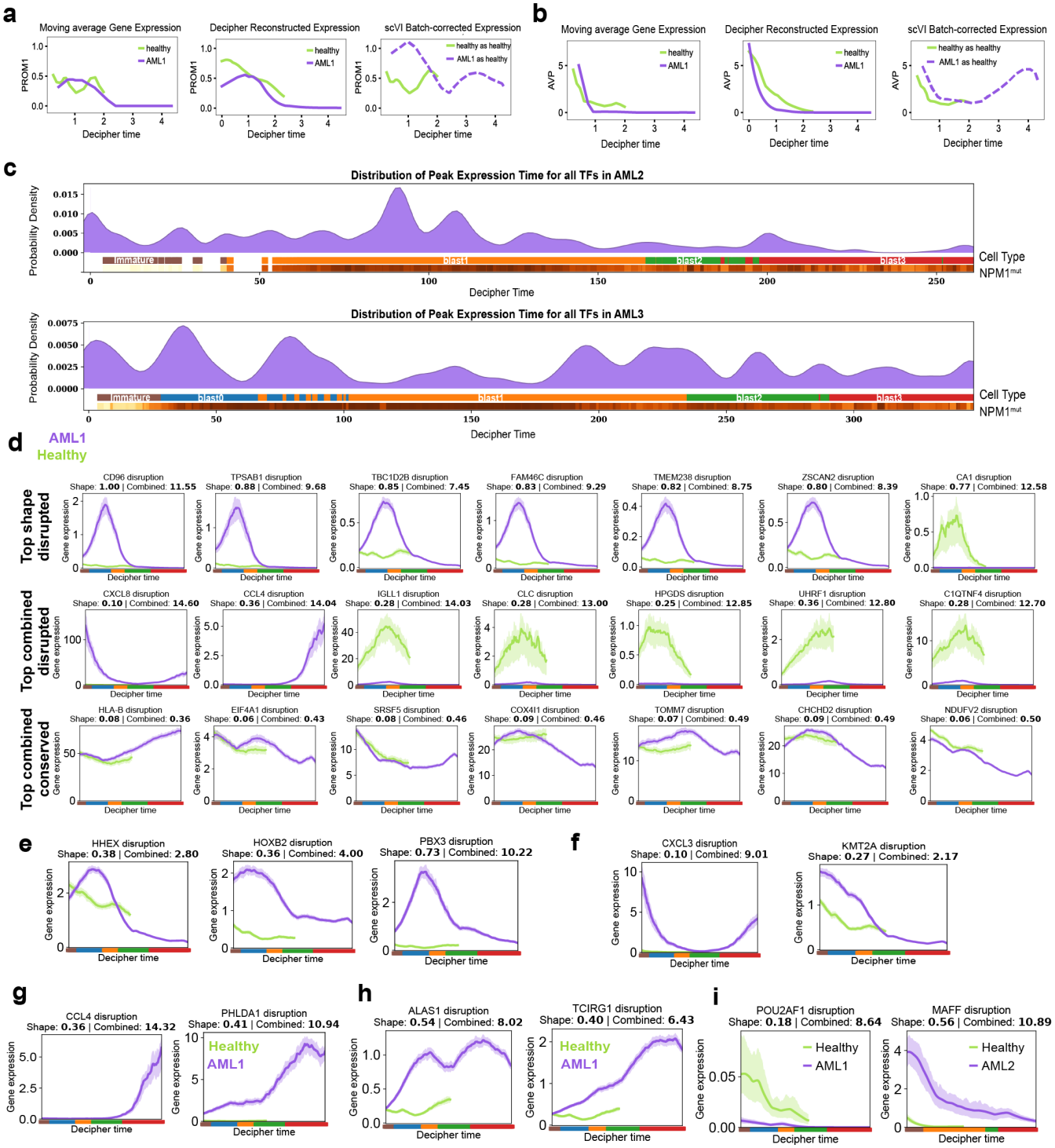


Figure 3. Reconstructing gene expression patterns and characterizing the regulatory landscape in AML compared to healthy HSPCs. **(a)** The Decipher generative model reconstructs gene expression along each trajectory, directly from the 2D Decipher representation. **(b,c)** Comparison of expression of major markers *CD34* (b) and *MPO* (c) along the Decipher time axis, reconstructed using different approaches: moving average of the raw gene expression (left), the Decipher expression

reconstruction (middle), and the scVI batch-corrected expression (right; Methods). **(d)** Reconstructed expression of example stemness and differentiation markers with Decipher support aligned pseudotimes between the two trajectories.The shaded bands represent the 25-75 interquartile range of the Decipher's model uncertainty (Methods) **(e-f)** Timing of peak expression of TFs, on the Decipher time axis, in Normal (e) and AML1 (f) samples. Figures (e,f) are aligned according to Decipher time (x-axis). The density plots show the distribution of local maxima across all TFs in each sample. Density plots for other patients are shown in **Extended Data Fig. 5c**. **(g)** Reconstructed gene expression of *HOXA9* and *MEIS1* (known deregulated TFs) and other disrupted TFs exhibiting various forms of dynamics. Solid lines show inferred mean and shaded areas reflect +/-1 SD.



Extended Data Figure 5. **Details of gene pattern analysis. (a-b)** Comparison of expression of *PROM1* (a) and *AVP* (b; stemness marker) along the Decipher time axis, reconstructed by different methods: the raw gene expression, the Decipher expression reconstruction, and the scVI batch-corrected expression (using a linear or a nonlinear decoder; see **Methods**). For

scVI, dotted lines show the reconstructed expression with batch correction, e.g., "AML1 as healthy" denotes the reconstruction gene expression from the AML1 cells, changing the batch ID to healthy in the decoder (see **Methods**). The relative expression of some markers are preserved with scVI without batch correction but altered when correcting for batch. **(c)** Timing of peak expression of TFs, on the Decipher time axis, in AML2 (top) and AML3 (bottom) samples. The density plots show the distribution of local maxima across all TFs in each sample. **(d)** Reconstructed gene patterns for top disrupted/conserved genes identified using the shape and combined disruption metrics. **(e-h)** Reconstructed gene patterns and disruption scores for homeobox genes and cofactors (e), example genes up-regulated in immature leukemic cells (f), TNFα pathway (g), oxidative phosphorylation (h) in AML1. **(i)** Example genes disrupted mainly in early immature state and thus unidentified in bulk analysis.

***NPM1*-mutated AML is associated with deregulation of cell differentiation and interferon response**

We define *disrupted genes* as genes exhibiting different expression patterns between the perturbed and normal trajectories, and genes with similar patterns as *conserved genes*. To quantify these differences, we devised a probabilistic framework with the key assumption that the expression pattern of each gene can be approximated as a weighted combination of a few representative patterns with distinct temporal dynamics e.g. ascending, descending, or peaking in intermediate states (**Fig. 4a**). We further assume that the representative patterns are shared between normal and perturbed trajectories, however, a gene can have a different set of weights and scale parameters in each trajectory. The representative patterns mathematically defined as *basis functions* (Methods) are simultaneously inferred with the decomposition parameters, and capture dominant transcriptional programs (e.g., pathways or co-regulated genes) and their dynamics along trajectories. The decomposition weights (β) indicate which programs are associated with each gene and the scale parameter (s) indicates the average expression of a gene across the entire trajectory. Specifically, we model the patterns using Gaussian Processes adapted from[75] and approximate them using neural networks (Methods).

Using the basis decomposition, we can distinguish changes in temporal dynamics of genes from change in the overall scale of expression. In particular, the *shape disruption* measures the distance between decomposition weights independent of the scale, while *combined disruption* considers both weights and scale (Methods). We computed disruption scores for all genes and identified conserved genes as well as those disrupted in both shape and combined measures (**Extended Data Fig. 5d)**. For instance, *HHEX* and *HOXB2* (other homebox members) exhibit a significant shape disruption score (**Extended Data Fig. 5e**). On the other hand, *CXCL3* and *CXCL8* which play key roles in inflammation as well as *KMT2A*, which interacts with Menin - a target of clinical trials for NPM1-mutant AML[76], are up-regulated in immature cells compared to normal, and have significant differences in scale and hence elevated combined disruption score (**Extended Data Fig. 5d,f**).

We note that many of the genes with the highest shape disruption (e.g. *HHEX*, *HOXB2*) rise in AML early (around initiation of *NPM1* mutation) and subsequently descend to similar expression levels in advanced blasts, highlighting the importance of enriching our data with early leukemic progenitors to detect such initial events. Others peak with full transformation to blasts, allowing us to identify genes associated with the final transition steps to disease; these include genes involved in the TNFα pathway (*CCL4, PHLDA1;* **Extended Data Fig. 5g**), and oxidative phosphorylation (*ALAS1, TCIRG1*) (**Extended Data Fig. 5h**). Thus the disruption scores can identify and associate specific genes with cell state transitions during AML progression.

The ability to pinpoint the order and timing of gene derailment offers the opportunity to understand regulatory events underlying AML. We focused on TFs with the most disrupted expression, contrary to coordinated waves of transcription found in normal hematopoiesis (**Fig 4b**), in AML we found a cascade of disrupted TFs (**Fig. 4c** and **Extended Data Fig. 6a,b**). At the top of this cascade we found *GATA1* and *KLF1*, key regulators of myelopoiesis (**Fig 4c**). Subsequently, peaking around the acquisition of NPM1

mutation, we found the homeobox (HOX) TFs, including *HOXA9* and their DNA-binding cofactor *MEIS1*, which are known to be altered specifically in NPM1-mutated AML[71] (**Fig. 4b, Fig. 3g** and **Extended Data Fig. 6a,b).** This suggests that TET2 mutation disrupts regulatory function of key TFs of hematopoiesis and these disruptions may propagate to additional TFs in the gene regulatory network. Importantly, we observe homeobox TFs among the most disrupted TFs in all 3 patients, suggesting that they may play a key role in the onset and progression of AML. This is in line with murine studies demonstrating the regulatory feedback loop from HoxA9 to the entire Hox-A locus which in turn perturb cell cycle and proliferation[77,71]. AML3 however presents additional disrupted regulators such as *MYC* and its interacting factor *MNT*, known to be overexpressed in AML[78] (**Extended Data Fig. 6b**).

To evaluate our disruption scores in the context of a larger cohort we performed differential gene expression analysis between *NPM1*^{mut} AML and normal samples using a publicly available cohort of 125 *NPM1*^{mut} AML samples and 16 normal HSC-enriched subpopulations (see **Data Availability; Methods**). Examining the top 20 disrupted TFs in AML1 (combined disruption score), we found that 18 out of 20 are differentially expressed (absolute fold-change>4; $p < 2.37e-03$) in the larger cohort. We also find 12 out of top 20 disrupted TFs in AML2 and 7 out of top 20 TFs in AML3 to be differentially expressed in bulk data (absolute fold-change>4; $p<1.43e-03$). Interestingly, TFs that are not detected in the larger cohort (e.g. *POU2AF1, MAFF*; **Extended Data Fig. 5i**), exhibit altered expression in early immature cells whose signal would be diluted in bulk data dominated by blasts. It is noteworthy however, that the disrupted TFs overlapping with the bulk differentially expressed are not the highest enriched genes, but rather among the top 26% of genes ranked according to absolute fold-change. The most differentially expressed genes reflect enrichment in terminal blasts, whereas our profiling of early immature cells and computational modeling of their dynamic gene expression deciphers regulators of leukemic initiation.

To further study altered TF activity in AML, we performed an integrated analysis of gene expression and chromatin accessibility patterns via single-cell Assay for Transposase-Accessible Chromatin (scATAC-seq; Methods). Using scATAC data, we explored possible regulatory relationships between disrupted TFs expressed in early immature cells and those expressed later in AML progression, that could explain the observed cascades of TF activity (**Fig. 4c; Extended Data Fig. 6**). Specifically, we identified top 50 disrupted TFs for each patient, and split them into those expressed before and after NPM1 mutation, labeled as '*early TF*s' and '*late TFs*' respectively. We then investigated putative regulatory relationships between the two groups of TFs by computing a motif enrichment score for each early TF detected within accessible regions in the vicinity of late TFs as targets. Then using bootstrapping analysis, we compared the motif scores to a null distribution generated from same-size random sets of target TFs, that are expressed in at least 1% of cells (Methods). In all three patients, we found significant motif enrichment in HOX TFs, including HOXB4 ($p < 0.0458$) and HOXB6 ($p < 0.032$) and FOSL1 ($p < 0.0057$), suggesting their putative role in regulating TFs driving AML transformation (**Fig. 4d**). In two out of three patients, we found significant motif enrichments in GATA2, MYC, NR4A2, MAFF, ATF3, NFKB2, RELB, KLF2, ZNF274 ($p < 0.05$; Methods), all of which show altered expression in immature cells (**Extended Data Fig. 7**).

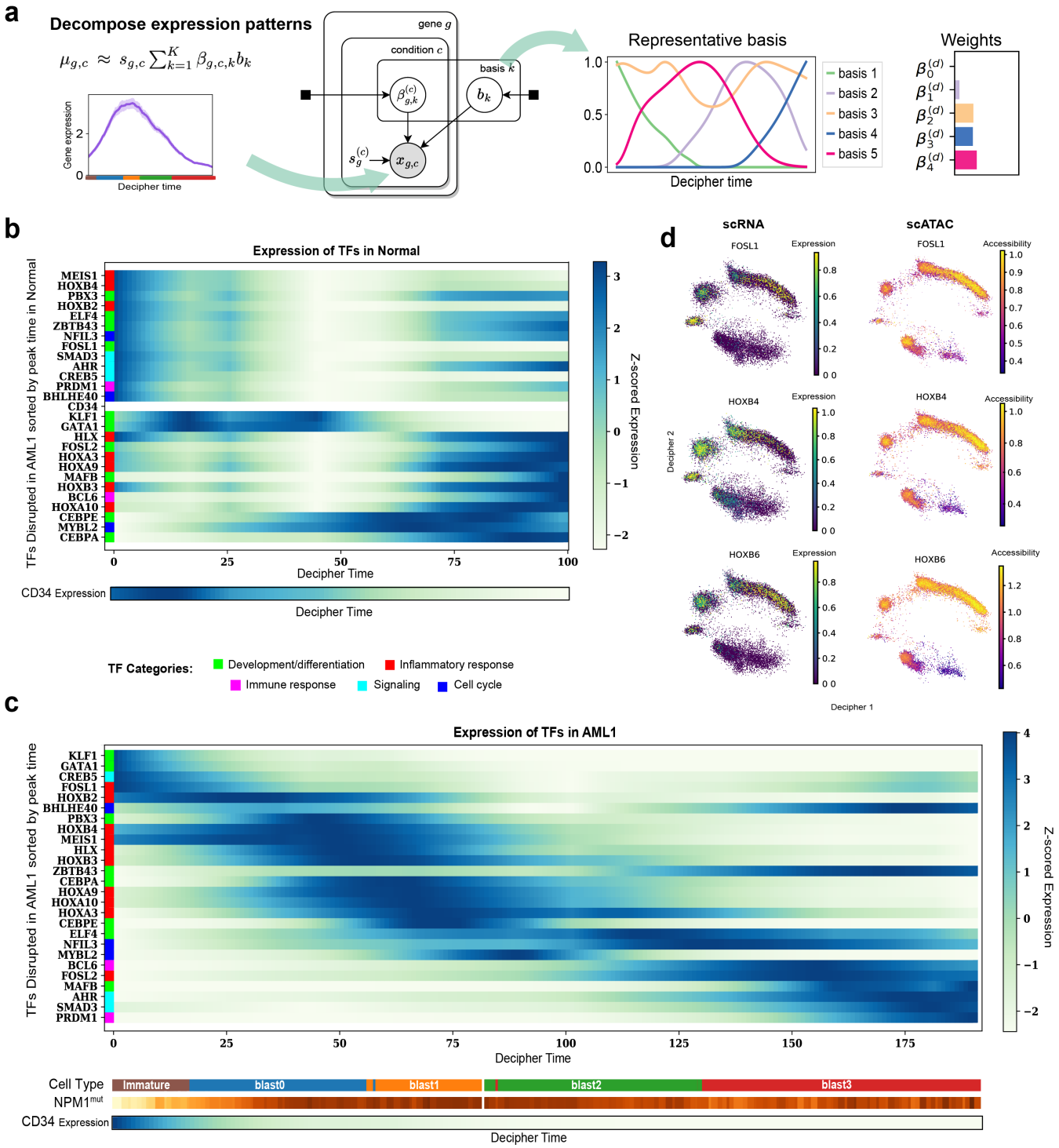
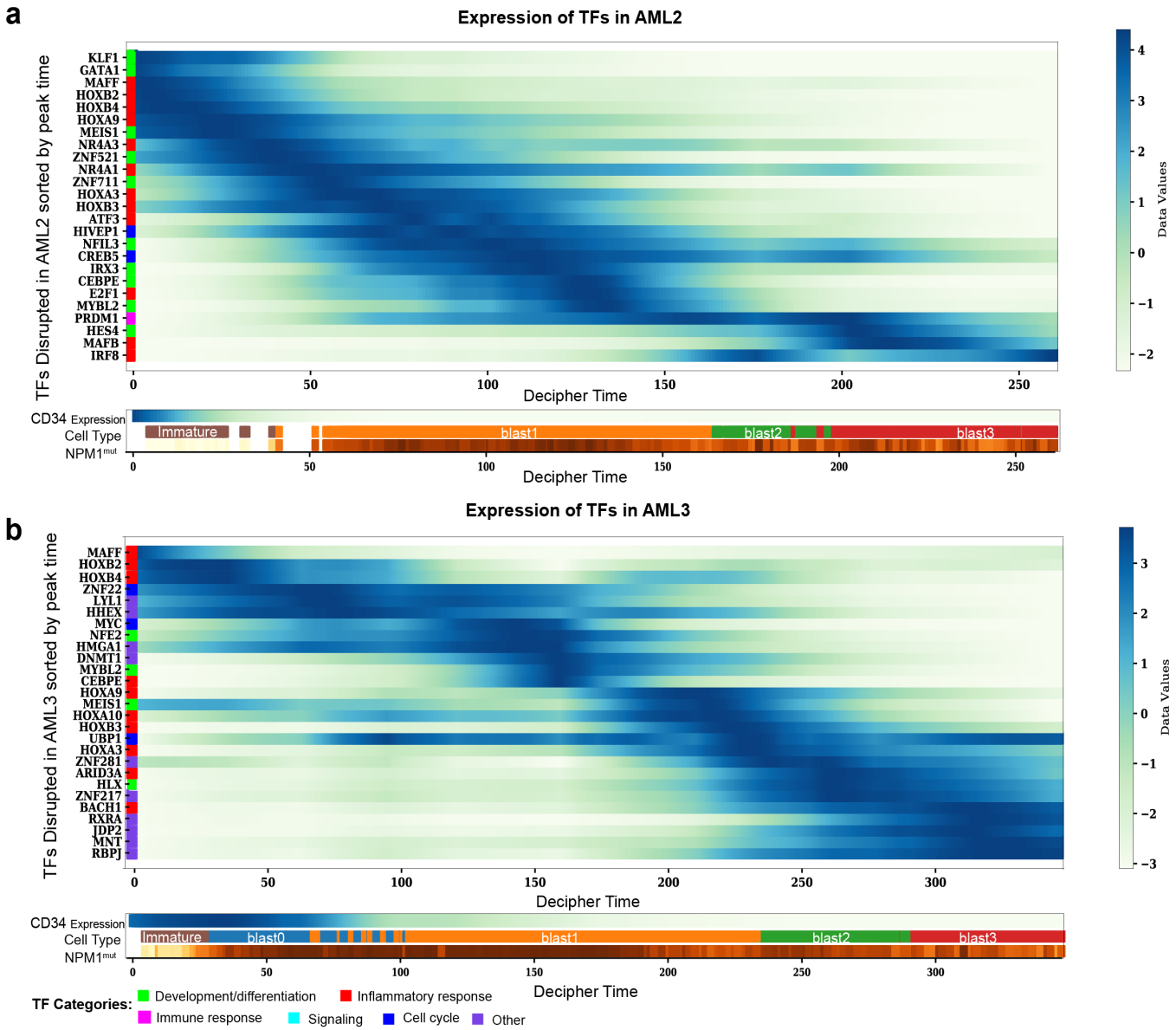


Figure 4: Decipher components can be used to unlock transcription factor (TF) dynamics. (a) Overview of Decipher's probabilistic basis decomposition and disruption quantification method. From left to right: the input is the neural basis decomposition learns the dominant representative patterns and decomposes each gene expression pattern onto them; the coefficients on the basis for each gene are compared between the normal branch and the AML branch to compute the disruption score (Methods). **(b-c)** Timing of TF expression on Decipher time axis, in AML1 (top) and Normal (bottom) samples. Heatmaps show log transformed and z-scored expression for top 20 TFs with highest combined disruption scores in AML1, sorted by timing of maximum expression in AML1. *CD34*[70,71] , a key marker of HPSCs is included for calibration. Colorbars correspond to cell type, and proportion of NPM1 mutated cells among the 30 nearest neighbors of each cell in AML1; both are smoothed over 50 nearest neighbors in the Decipher space. **(d)** scRNA and scATAC-seq of disrupted TFs which were found to have enriched

motifs in the vicinity of other disrupted TFs expressed after NPM1 mutation, in all patients (Methods). Scatterplots show expression (left) and accessibility (right) in the Decipher space of AML1 (Methods).

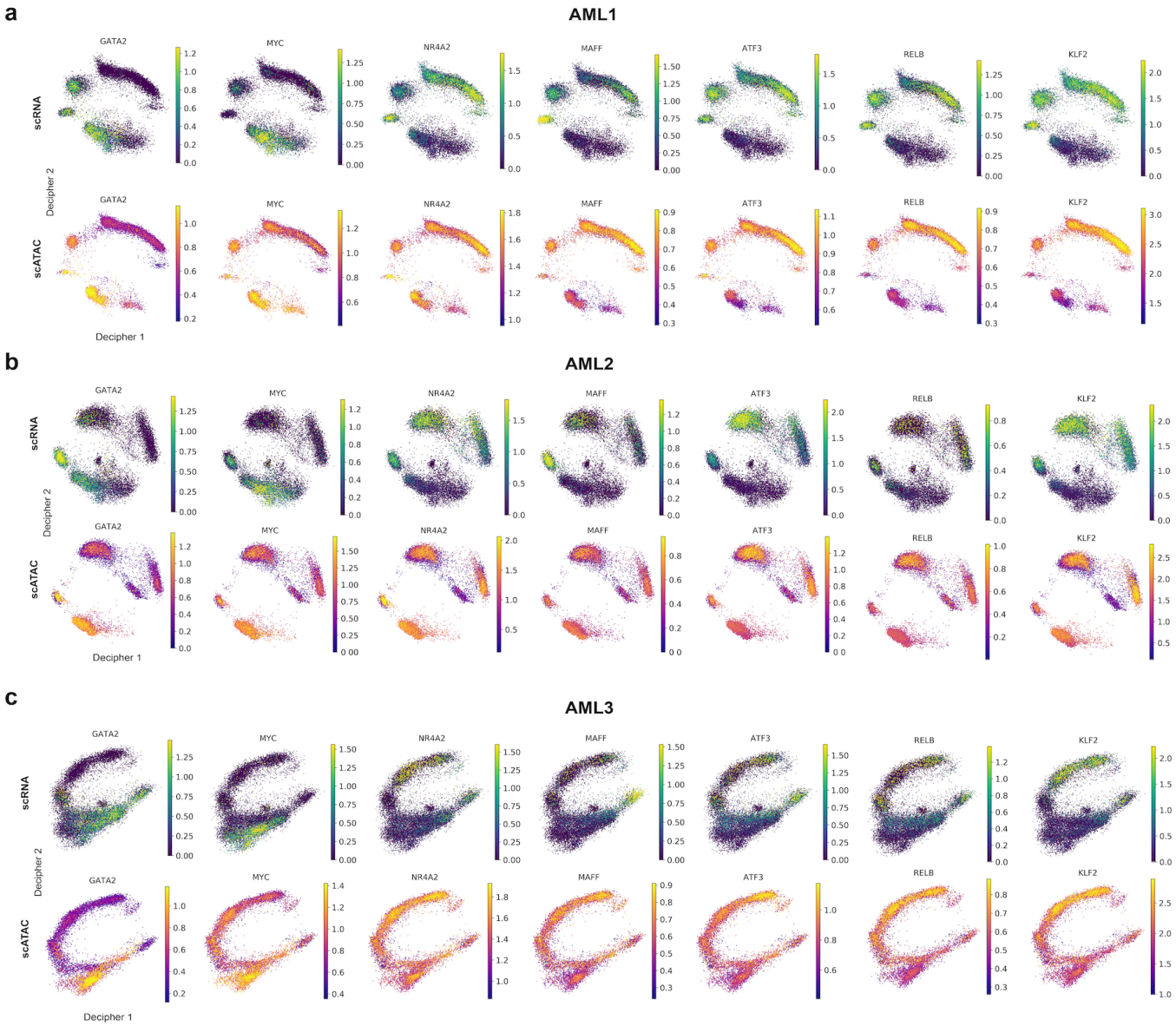


Extended Data Figure 6. **Altered TF dynamics in other patients (a-b)** Timing of TF expression on Decipher time axis, in AML2 (a) and AML3 (b) samples. Heatmaps show log transformed and z-scored expression for top 20 TFs with highest combined disruption scores in each sample, sorted by timing of peak expression. Colorbars correspond to cell type, and proportion of NPM1 mutated cells among the 30 nearest neighbors of each cell; both are smoothed over 50 nearest neighbors in the Decipher space.

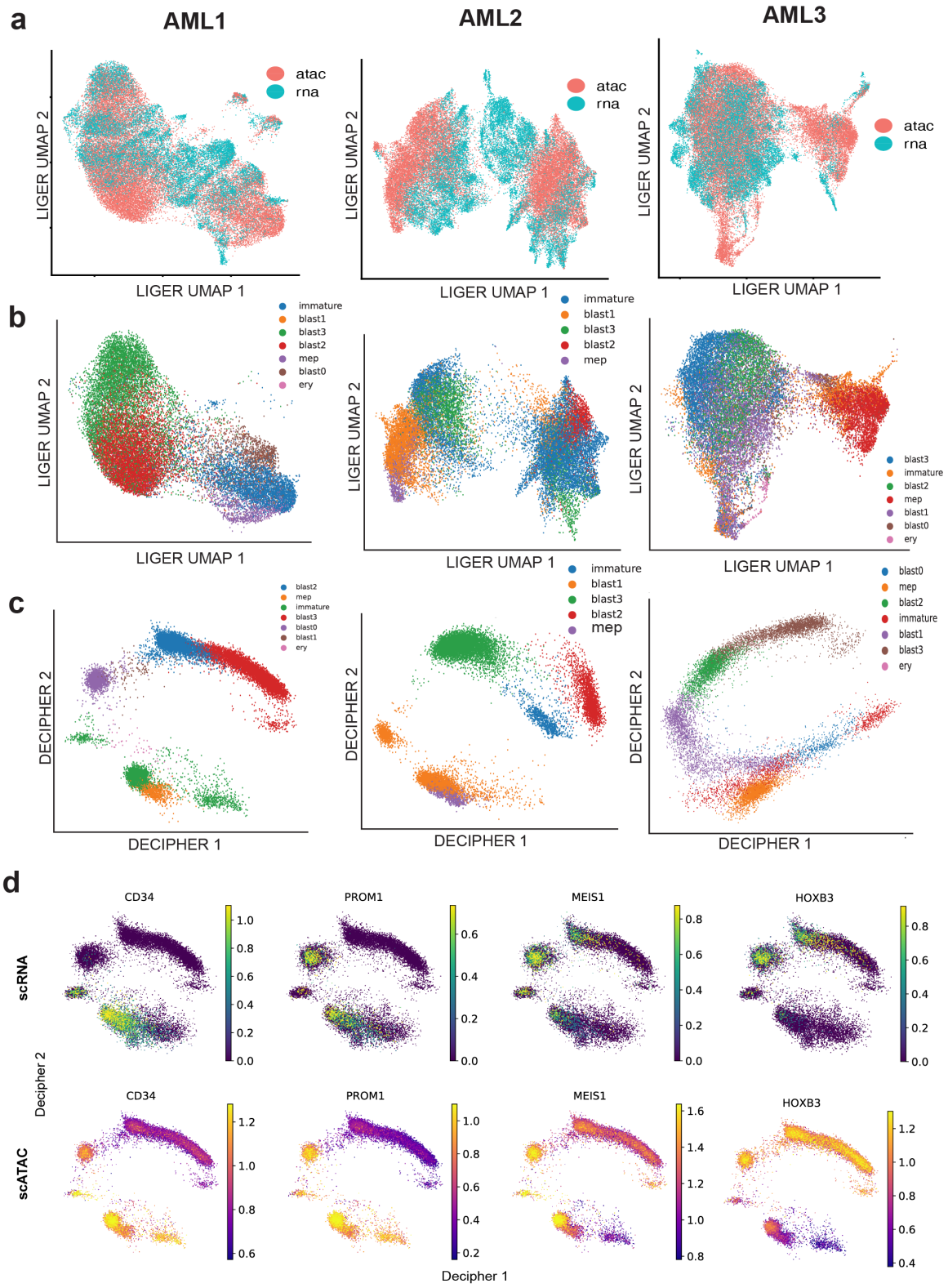
We then examined the dynamic accessibility of the TFs themselves during AML progression, through integration of scRNA- and scATAC-seq data using LIGER[34] (**Extended Data Fig. 8a,b, Methods**). Through this process, we are able to project scATAC data into the 2D Decipher space inferred from scRNA (**Extended Data Fig. 8c**), finding that the patterns of chromatin accessibility mirror gene expression dynamics for regulators with enriched motifs (**Fig. 4d; Extended Data Fig. 7**).

Using this framework, we can also identify gene programs co-expressed with aberrant TFs. Interestingly, the up-regulation of HOX TFs upon NPM1 mutation coincides with an increase in Interferon Type 1 signature genes[79,59] including *LY6E*, *FAM46C*, *ADAR*, and *TMEM238*, as well as TGFβ signaling

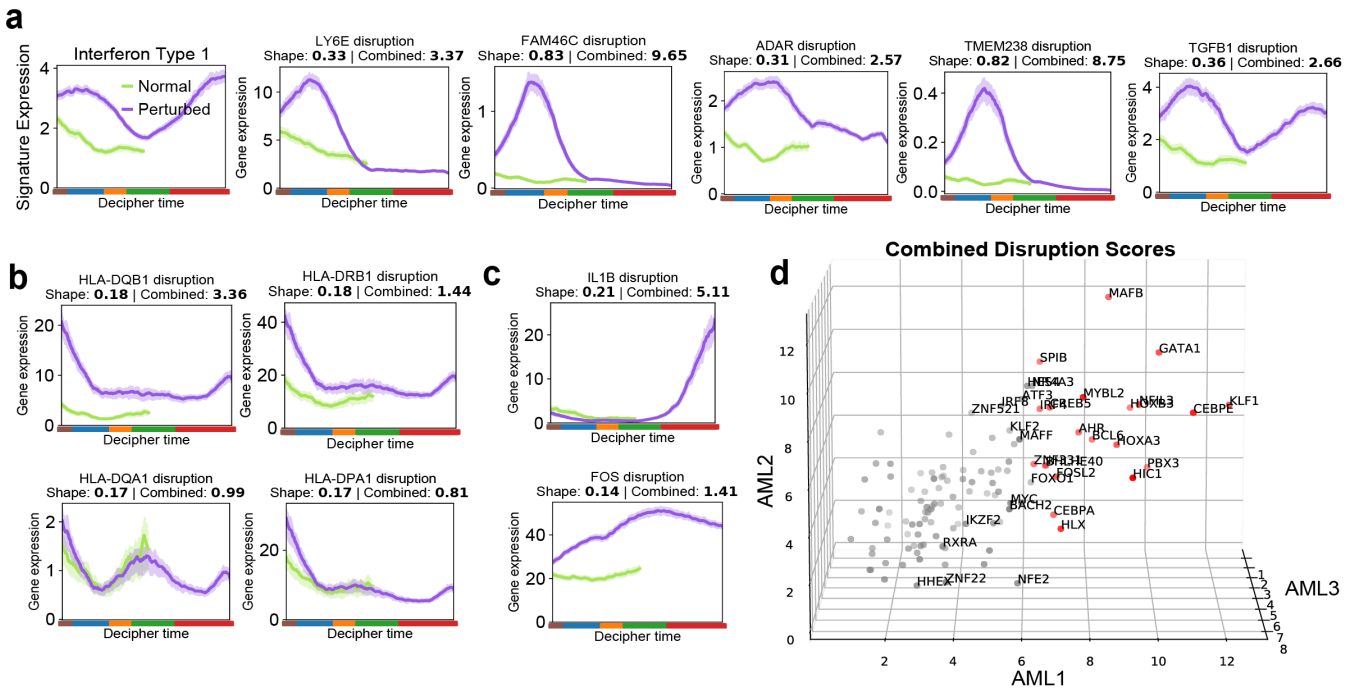
which is known to up-regulated epithelial-mesenchymal transition-related genes in NPM1 mutant AML[80] (**Extended Data Fig. 9a).** This is in contrast to components of Type II Interferon (IFNg) response, including MHC-II genes that are most up-regulated in early immature cells (**Extended Data Fig. 9b**), suggesting their link to TET2 mutation. Moreover, we observe genes in the proinflammatory TNFα pathway (**Extended Data Fig. 5g)**, the inflammatory cytokine IL-1, and AP-1 component FOS (**Extended Data Fig. 9c**) up-regulated towards the end of the trajectory in transformation to blasts. Interestingly, motif enrichment analysis using scATAC data further showed enrichment of motifs for known interferon-induced TFs such as STAT1 (MA0137.1 ,p<0.0357), and IRF1 (MA0050.1 ,p<0.0057; MA0050.2 ,p<0.0216) among the top disrupted TFs in AML1 and AML3. Our approach thus resolves the timing of TF activity with respect to significant events such as genetic mutations and activated signaling pathways, guiding further studies of regulatory relationships.



Extended Data Figure 7. **TFs with enriched motifs in all patients.** scRNA and scATAC-seq of TFs found to have enriched motifs in the vicinity of late TFs (expressed after NPM1 mutation), in two out of three AML patients (Methods). Scatterplots show expression (top) and accessibility (bottom) in the Decipher space for AML1 **(a)**, AML2 **(b)**, and AML3 **(c)**.



Extended Data Figure 8. **Integration of scRNA and scATAC data (a)** UMAP embedding of cells from scATAC and scRNA-seq in the LIGER integrated space in three AML patients. **(b)** UMAP projection of cells colored by cell types in ATAC-seq in the LIGER Integrated space in three AML patients. **(c)** Projection of cells in scATAC data in 2D decipher space colored by cell type. **(d)** scRNA and scATAC-seq of key marker genes in AML1. Scatterplots show expression (top) and accessibility (bottom) in the Decipher space (Methods).



Extended Data Figure 9. **Notable genes co-expressed with aberrant TFs and comparison of TFs across patients. (a)** Reconstructed expression patterns for IFN type 1 signature (left) and individual genes (right), found to be co-expressed with HOX TFs, under both normal (green) and AML (purple) conditions. **(b-c)** Reconstructed expression of genes involved in IFN type 2 (b) and TNFα pathway. **(d)** The combined disruption score (Methods) of transcription factors in each of three AML patients. Red points indicate TFs that are among the top 20 disrupted TFs in at least one patient.

Decipher provides a comparative analysis of early-occurring epigenetic mutations in AML

We sought to utilize Decipher to compare altered mechanisms found in the TET2^{mut} cohort to an independent cohort of AML patients with mutations in DNMT3A. While both TET2 and DNMT3A are epigenetic regulators, they have opposing roles in methylation, with DNMT3A adding and TET2 removing methyl groups to DNA[81]. Dysregulation of epigenetic factors is known to play a key role in AML initiation and maintenance. In fact, mutations in TET2 and DNMT3A, are also found in preleukemic lesions before the onset of AML[82]. Interestingly, epigenetic mutations involving DNMT3A, TET2, and ASXL1 are observed in healthy elderly patients, reflecting clonal hematopoiesis (CH). Individuals with CH are at increased risk of developing AML, which represents one of the several lines of evidence for the stepwise nature of AML development in which normal HSPCs sometimes acquire mutations in epigenetic modifiers prior to the acquisition of a transformative driver event such as NPM1 mutations[47]. Comparison of derailment in the context of these epigenetic mutations can thus elucidate how these pre-disease mutations with seemingly opposing roles lead to similar vulnerabilities and whether they harness shared mechanisms of leukemogenesis.

We thus performed single-cell profiling of both unsorted and CD34+ sorted cells from BM specimens of these DNMT3A^{mut} patients denoted as AML 13-17, three of whom also have NPM1 mutation (**Supplementary Table 1;** Methods). We used our original cohort to annotate the maturation stages of clusters identified in these patients (**Extended Data Fig. 10a; Methods)**. Applying Decipher to these patients, we confirmed the ability to align AML maturation trajectories (decline in *CD34)* with normal HSPC differentiation (decrease in *AVP* and increase in *MPO*) along the Decipher 1 axis and Decipher 2 representing disease progression (**Fig. 5a-d; Extended Data Fig. 10b**). In the two DNMT3A^{mut} and NPM1^{mut} patients with successful cell sorting (AML14, AML15), we also found *PROM1* marking blast0

states (**Fig. 5d; Extended Data Fig. 10b,c**) and overlap of disrupted TFs (8 out of top 20) including regulators of myeloid lineage commitment and differentiation (*CEBPE, HOXB3, AHR, KLF2, MYBL2)*, inflammatory response (*JUND*), homeobox co-factor (*MEIS1*), and ZBTB20. Interestingly, we found the reduction in oxidative phosphorylation as an enriched pathway along Decipher 2 in both DNMT3A^{mut} and TET2^{mut} patients (**Supplementary Table 4**).

To perform a more comprehensive comparison of the epigenetic mutations, we identified disrupted TFs according to both shape and combined disruption which revealed partial overlap in many combined disrupted genes between the two cohorts (**Fig. 5e**). For instance, we find deregulation of *CEBPE* and *HOXB3* in both cohorts, while interferon regulator *IRF8* shows higher combined disruption in TET2^{mut} and inflammatory response regulator *JUND* shows higher shape disruption in DNMT3A^{mut} [83,84] (**Fig. 5e**). In sum, Decipher provides a framework for unbiased characterization of patient-specific derailed trajectories in disease compared to normal, as well as comparative analysis of aberrant mechanisms between individual patients and genetic backgrounds.

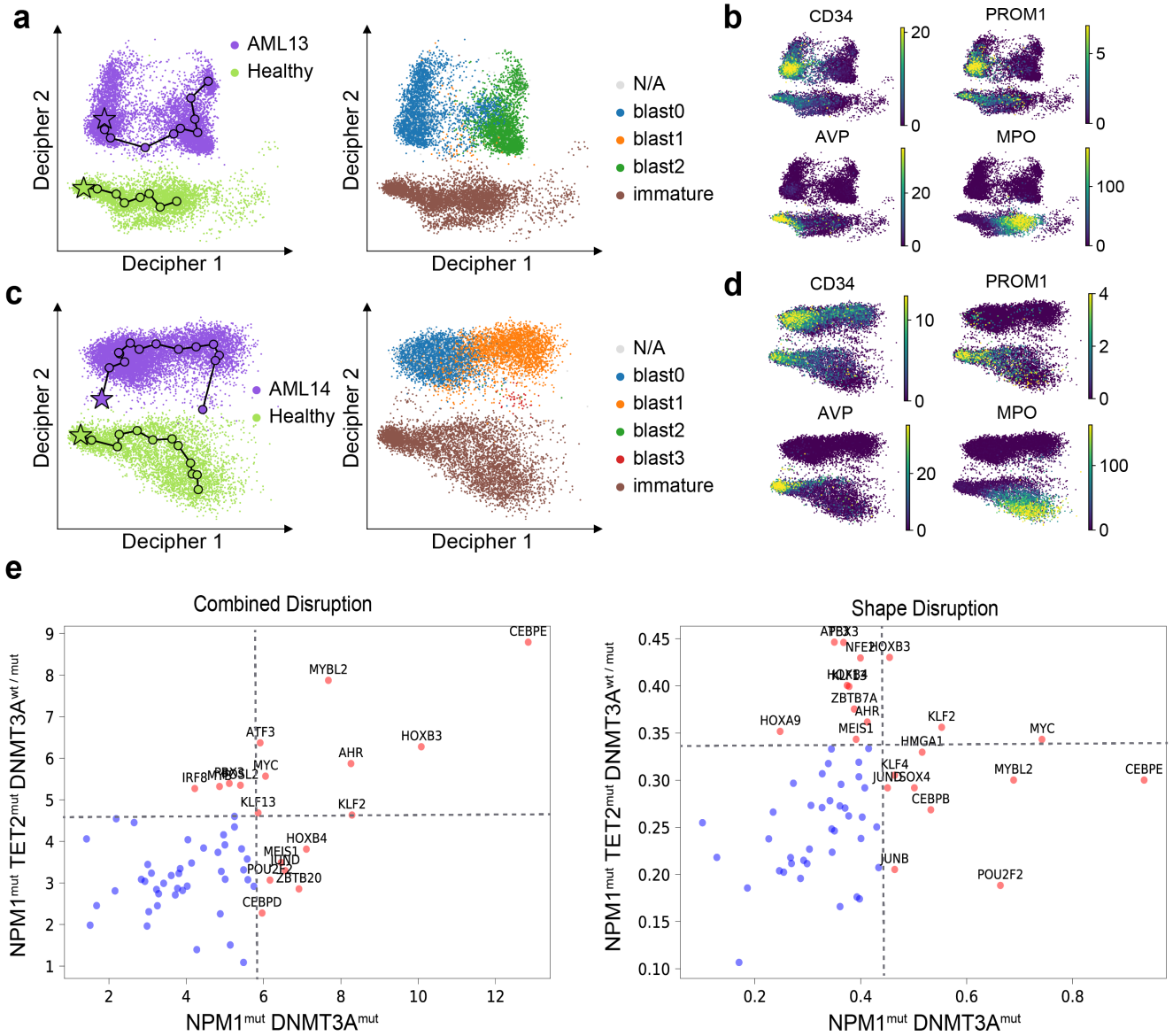
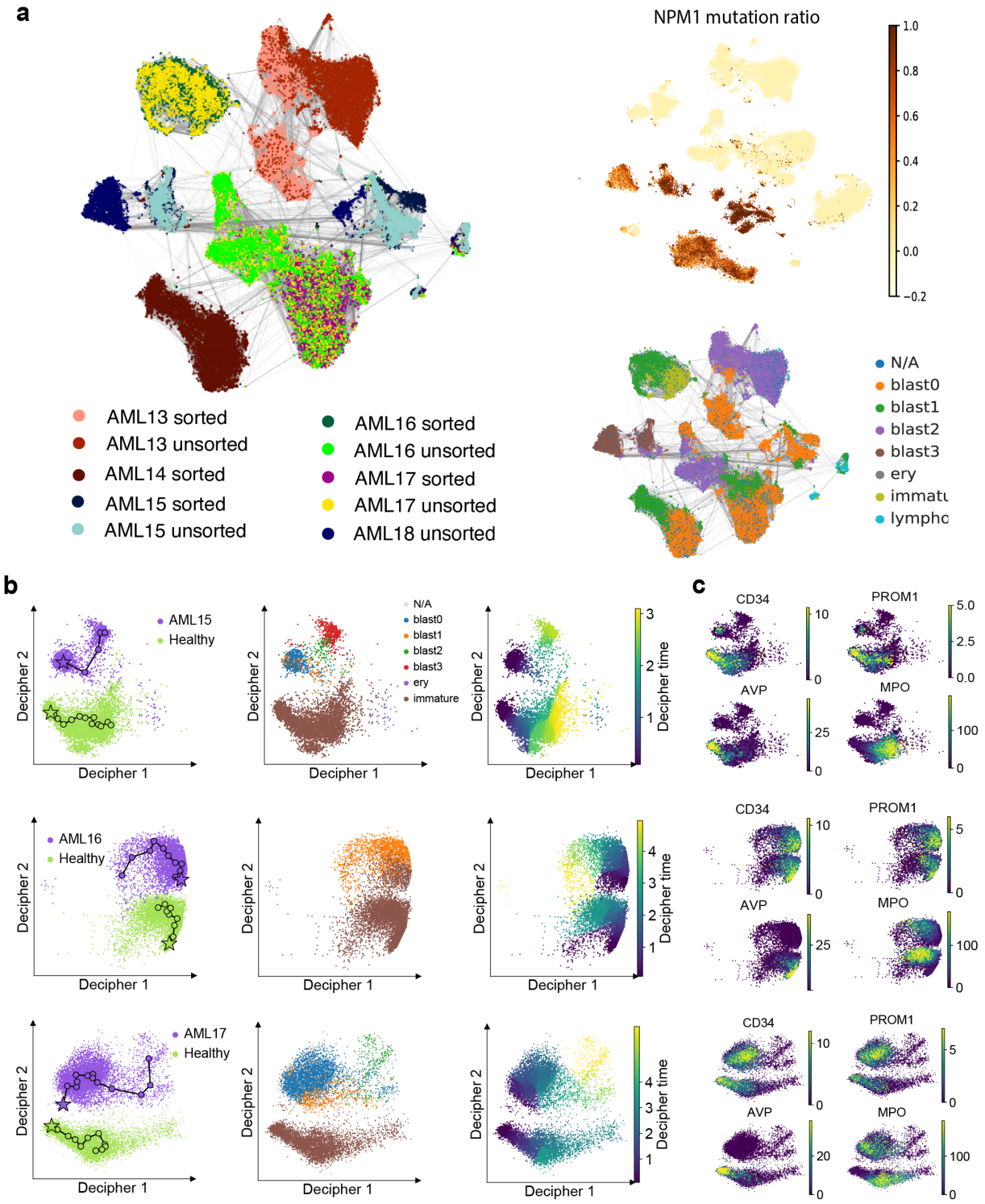


Figure 5. Comparative analysis of AML cohorts. (a-d) Decipher space embedding of sorted cells from two DNMT3A^{mut} AML patients (Supplementary Table 1) colored by sample of origin and cell type (a,c) and and expression of select marker genes (b,d). The Decipher spaces also identify and align trajectories of myeloid development in AML, even in the case of missing cell

state (blast1,3 in AML13; blast2,3 in AML14). Other patients are shown in **Extended Data Fig. 10b,c**. **(e)** Comparison of genes disrupted in patients bearing the NPM1 and DNMT3A mutations against patients with the NPM and TET2 mutations. The left scatterplot uses the combined disruption score, while the right uses the shape disruption score. Red points represent TFs in the 80th percentile of disruption in at least one category. Dotted lines show the 80th percentile threshold.



Extended Data Figure 10. **Annotation of DNMT3A^{mut} AML cohort and application of Decipher to other patients** (a) joint 2D tSNE map of all patients from the DNMT3A mutated cohort. Left indicates patient origin; Top right indicates the ratio of NPM1 mutated to NPM1 wild-type cells among the 30 nearest neighbors of a cell; bottom right indicates location of cell type annotations transferred from the TET2 cohort. (b) Decipher space embedding of sorted cells from 3 DNMT3A^{mut} AML patients

(Supplementary Table 1) colored by sample of origin and cell type (b) and expression of select marker genes (c). The Decipher spaces identify and align trajectories of myeloid development in AML, even in the case of missing cell states. Other patients are shown in **Fig. 5a-d**.

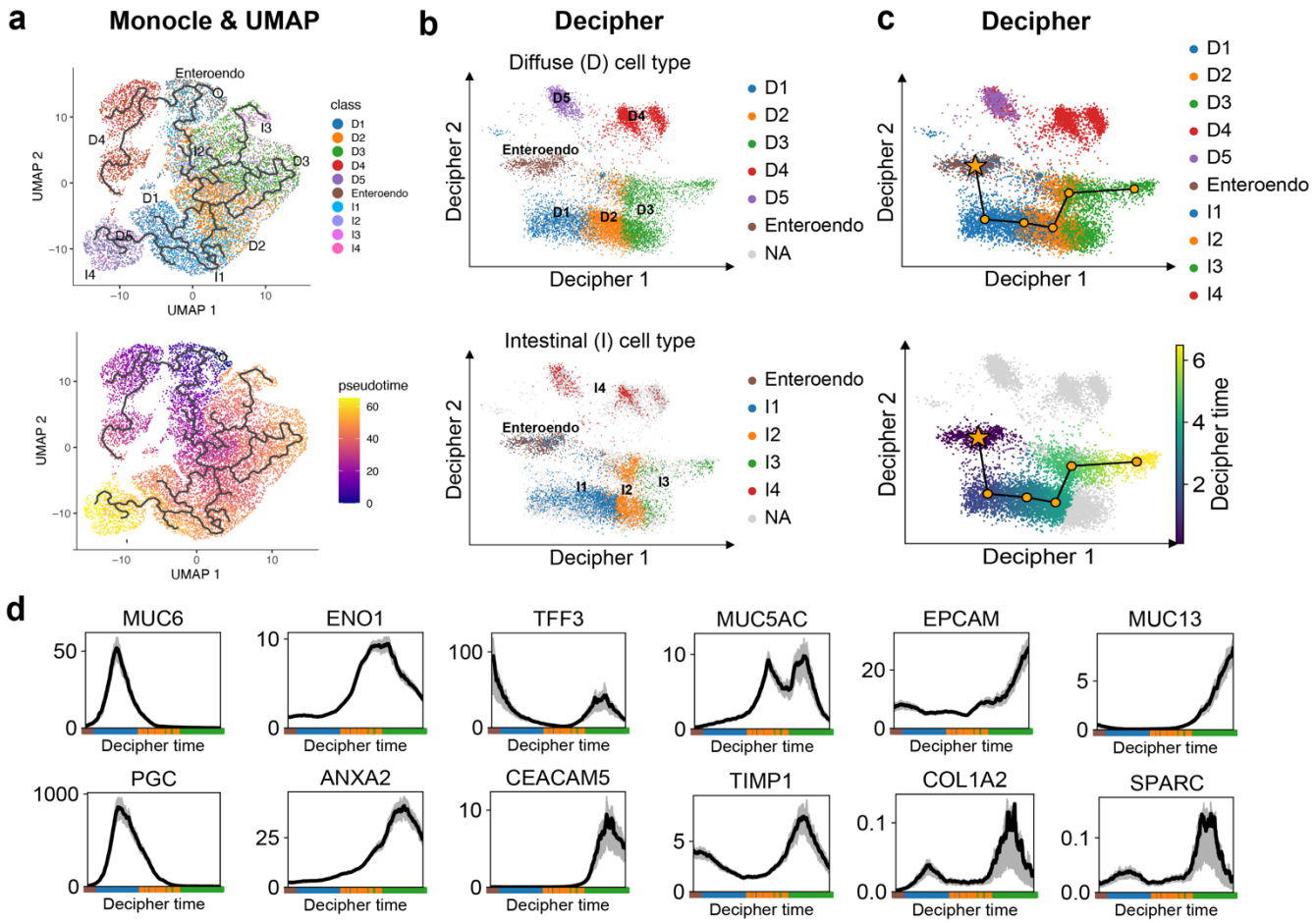


Figure 6. Application of Decipher to gastric cancer onset. (a) 2D UMAP projection and trajectory inference with Monocle applied to scRNA-seq data from 24 gastric cancer (GC) and precancerous lesions[85]. Single-cell data from both intestinal (IGC) and diffuse (DGC) types were merged to infer a joint tree using Monocle3[87] (in gray line). The order of disease progression states (I1-I3 in IGC and D1-D3 in DGC) is not captured by the tree. **(b-c)** Decipher can reveal the order of cancer stages, while still harmonizing the different types of cancer without requiring a batch correction method or a dimensionality reduction method for visualization. Cells in Decipher space are colored by cancer progression stages in DGC and IGC (b) and inferred Decipher trajectory and Decipher time on merged data is shown in (c). **(e)** Decipher's reconstructed gene patterns for known markers of progression states along the shared trajectory.

Joint analysis of disease onset in two gastric cancer types

Decipher can be applied to study early and step-wise transformations with any perturbation or disease beyond leukemia. To illustrate this, we applied Decipher to scRNA-seq data from primary gastric cancer tumors and paired adjacent non-malignant tissue from 24 patients[85] (**Fig. 6a**). This data involves specimens from two types of gastric cancer: intestinal and diffuse types (IGC and DGC, respectively). Application of Monocle2[86], as another trajectory inference method, to each GC type separately, was previously shown to capture a part of the disease progression trajectories, specifically cell state state transitions annotated as I1 to I3 in IGC and D1 to D3 in DGC. To infer early transitions shared between IGC and DGC, we first applied Monocle3[87] to the merged data from precancerous and cancerous lesions

with setting the initial state to non-malignant (enteroendocrine) cells. The inferred aligned trajectories do not fully preserve the correct order of progression, and rather suggest Enterendocrine transitioning to I2/D2 states which then branch into I3/D3, and I1/D1, with the latter further transforming to I4. The late diffuse-like state D4 is also misplaced as a separate branch off of enteroendocrine cells (**Fig. 6a** and [85]).

We utilized Decipher to achieve a more interpretable joint representation of tumor plasticity in the two GC types. The Decipher 1 axis correctly aligned cells along a continuous progression shared between the two types, with non-malignant enteroendocrine and I1 and D1 cells at the beginning of the trajectory transitioning to I2 and D2 intermediate states and further to I3 and D3 malignant states (**Fig. 6b)**. The inferred Decipher trajectory (**Fig. 6c**) reveals patterns of genes known to be related to the malignant process and up-regulated at the correct cell state: e.g., *MUC6* and *PGC* in normal gland mucous cells, *ENO1* marking intermediate state I2 and *MUC13* and *CEACAM5* marking malignant states (**Fig. 6d)**. The alignment of the two GC types is also informative about the relative timing of disease progression: we predict the transformation of premalignant cells to malignant cells in IGC occurs later than that in DGC, with I1 extending a longer range than D1 along Decipher 1 and Decipher time. The Decipher 2 axis, on the other hand, separates the diffuse-like malignant state (D4) in DGC, illustrating the drastic derailment from D3 upon up-regulation of hallmark DGC markers such as *COL1A2* in D4. Finally, fibroblasts and endothelial cell states (I4, D5), which are not a part of the progression of cancer, are preserved as distinct states, highlighting the ability to represent both continuums and distinct states with Decipher.

DISCUSSION

Decipher is a deep generative model designed to learn and visualize joint representations of normal and perturbed data. Unlike common single-cell data analysis approaches that typically employ two distinct sequential dimensionality reduction steps (a latent factorization and a 2D-projection), Decipher utilizes a hierarchical structure to merge the two steps and therefore finely control the degree of dimensionality reduction. As a result, it not only provides a ready to use 2D-visualization of the data, but Decipher can also capture and encode more intricate information while remaining interpretable and discovering the interdependencies between the underlying latent factors. The ability to capture complex relationships between latent factors is particularly important when analyzing data from different conditions, such as normal and disease states, as it can identify confounding mechanisms in the disparate conditions.

Decipher is also scalable to any large number of cells, due to its VAE model formulation which allows for stochastic variational inference[88]. Other approaches have deployed VAEs[35,41] or Gaussian Process Latent Variable Models (GP-LVMs)[89,90] for nonlinear dimensionality reduction, however, they are not capable of simultaneous 2D visualization, and GP-LVMs are not scalable to large datasets. In addition to providing a direct visualization of the cell states that is less distorted than other methods, Decipher can utilize the joint representation to infer trajectories of cell state transition and identifies genes with disrupted expression patterns using a novel basis decomposition technique.

We anticipate that Decipher will be a valuable discovery tool for understanding derailed development upon perturbation and initiation of disease. For instance, it can provide a framework for tracing back abnormal cell states to the cell of origin according to the cancer stem cell model[4] and identify markers for guiding further rigorous validation of functional properties of self-renewal and tumor initiation. Indeed, studying the early stages of tumor initiation is largely under-studied especially in primary AML, and findings from animal models do not entirely translate to humans[91].

While AML presents significant genomic and transcriptomic heterogeneity across patients, suggesting multiple vulnerable states and origins of derailment from normal hematopoiesis[16,92], we show

Decipher's ability to achieve personalized characterization of divergence from normal myeloid differentiation confirmed by genotyping, while the global geometry of trajectories is distorted by other integration methods. We characterized leukemic initiation marked with *PROM1* expression. *PROM1* is expressed in normal HSCs[93] and in LSCs[94]. However *PROM1* expression is variable between AML subgroups, and *NPM1*-mutated AML blasts are notoriously negative for *PROM1* and *CD34*[95]. This collection is the first documentation, both genotypic and phenotypic, of a rare subset of *PROM1*+ cells in NPM1 mutated samples that likely define a pre-leukemic cell population. The more refined latent factors further revealed that mutations in NPM1 is followed by up-regulation of genes involved in inflammation, and IFN response.

Theoretical models of cancer initiation have proposed that sequential mutations influence the temporal dynamics of cells[96]. Modeling temporal dynamics of gene expression using Decipher, revealed a cascade of disrupted transcription factors, specifically loss of coordinated expression of those driving myeloid lineage commitment with TET2 mutation, followed by abnormal up-regulation of HOX TFs upon NPM1 mutation. High expression of HOX has been previously linked to mutant NPM1 and its aberrant cytoplasmic localization, leading to maintained leukemic state in AML[97].

Recent studies have demonstrated loss of *Tet2* in mice induces the expansion of aberrant inflammatory monocytic populations in murine models, through establishing a pro-inflammatory microenvironment[59,58]. Similarly, we find up-regulation of IFN type 2 (specifically MHC-II) genes in early TET2-mutated cells in primary samples. NPM1 has been previously reported to regulate IFNg-inducible genes in HeLa cells[98], however, the link is not established in AML. Our pseudotime-resolved characterization of transcriptional dynamics shows genes involved in IFN type 1 response and TGFb to be highly expressed specifically in transition to aberrant NPM1-mutated progenitor cells, coinciding with expression of HOX TFs. In further transformation to blasts, we observe later up-regulation of TNFα, IL-1, and FOS. The diverse patterns of chemokine and cytokines along the leukemic transformation trajectory also point to possible dysregulated interactions among them[62]. Inflammatory cytokines such as IL1 can indeed regulate hematopoietic stem cells and promote disease progression in models[99].

In addition to the role of inflammation in shaping the AML microenvironment[58], cell-intrinsic inflammation has been reported to be induced with NPM1 perturbation in mice leading to MDS-like phenotypes[100], and driving progression from MDS to AML[101]. These observations motivate future studies on inflammatory response as a consequence of NPM1 perturbation, compared to epigenetic remodeling in clonal hematopoiesis, and disentangling the role of pre-existing epigenetic mutations in inducing an inflammatory environment crucial for disease transformation.

We also provide support for disrupted transcriptional programs using single-cell epigenetic data. Additionally, we demonstrate the broader applicability of Decipher to other systems, such as in joint mapping of transitions from premalignant to malignant cell states in two subtypes of gastric cancer. Extending application of Decipher to other primary cancer samples as well as animal models can guide therapeutic strategies for modulating the TME and cell-intrinsic effects by attenuating the inflammatory response, and in turn, inhibiting cancer progression or increasing sensitivity to treatments.

DATA AVAILABILITY

The data discussed in this manuscript will be deposited in the National Center for Biotechnology Information's Gene Expression Omnibus (GEO) upon publication. The bulk RNA-seq data for AML patients is publicly accessible at GEO with accession IDs GSE106272, GSE49642, GSE52656, GSE62190,

GSE66917, and GSE67039. Bulk RNA-seq for normal HSC-enriched subpopulations is accessible at GEO ID GSE48846.

CODE AVAILABILITY

Decipher is available at https://github.com/azizilab/decipher and has been deposited at DOI: 10.5281/zenodo.10079999. The specific code that produced the results and figures of this manuscript is available at https://github.com/azizilab/decipher_reproducibility and deposited at DOI: 10.5281/zenodo.10079971.

ACKNOWLEDGEMENTS

We thank Guy Sauvageau from the Leucegene group and Josée Hébert from the Banque de Cellules Leucémiques du Québec (BCLQ) for providing clinical samples. The BCLQ is supported by grants from the Cancer Research Network of the Fonds de recherche du Québec–Santé (FRQS). We are thankful to Tal Nawy, Benjamin Izar, José McFaline-Figueroa, Manu Setty, Cassandra Burdziak, and Sopho Kevlishvili for helpful feedback and discussions. A.N. acknowledges support from the Eric & Wendy Schmidt Center Ph.D. Fellowship and the Africk Family Fund. J.L.F. acknowledges support from the Columbia University Van C. Mow fellowship. V.P.L. received support from a Vanier Canada Graduate Scholarship and holds a Clinical Research Scholarship from the FRQS. R.L.L. was supported by NCI award R35CA197594. D.P. is an HHMI investigator and was supported by NCI grant U54 CA209975 Studies supported by MSK core facilities were supported in part by MSKCC Support Grant/Core Grant P30CA008748, the Alan and Sandra Gerry Metastasis and Tumor Ecosystems Center and the Marie-Josée and Henry R. Kravis Center for Molecular Oncology. E.A. was supported by National Institute of Health (NIH) NCI grant R00CA230195 and Columbia University Data Science Institute and Irving Institute for Cancer Dynamics Seed Funding.

AUTHOR CONTRIBUTIONS

A.N., J.L.F., D.P., E.A. conceived the study and wrote the manuscript. A.N., J.L.F., D.B., D.P., E.A. designed and developed Decipher. V.P.L., D.P., E.A. designed AML experiments and prepared samples. V.K., I.M., R.L.B., S.E., R.L.L. performed and assisted with single-cell genomics data acquisition experiments. A.N., J.L.F., V.P.L., J.C., J.W., L.S., A.E.C, D.P., E.A. analyzed and interpreted data.

CONFLICT OF INTEREST

D.P. is on the scientific advisory board of Insitro. R.L.L. is on the supervisory board of Qiagen and on the board of directors of Ajax Therapeutics, for which he receives compensation and equity support. He is or has recently been a scientific advisor to Imago, Mission Bio, Syndax. Zentalis, Ajax, Bakx, Auron, Prelude, C4 Therapeutics and Isoplexis for which he receives equity support. He has research support from Ajax and AbbVie, consulted for Janssen, and received honoraria from Astra Zeneca and Kura for invited lectures.

REFERENCES

1. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
2. Azizi, E. *et al.* Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* **174**, 1293–1308.e36 (2018).
3. Peng, J. *et al.* Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res.* **29**, 725–738 (2019).
4. Lawson, D. A., Kessenbrock, K., Davis, R. T., Pervolarakis, N. & Werb, Z. Tumour heterogeneity and metastasis at single-cell resolution. *Nature Cell Biology* vol. 20 1349–1360 Preprint at https://doi.org/10.1038/s41556-018-0236-7 (2018).
5. Zilbauer, M. *et al.* A Roadmap for the Human Gut Cell Atlas. *Nat. Rev. Gastroenterol. Hepatol.* **20**, 597–614 (2023).
6. Wilk, A. J. *et al.* A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat. Med.* **26**, 1070–1076 (2020).
7. Melms, J. C. *et al.* A molecular single-cell lung atlas of lethal COVID-19. *Nature* **595**, 114–119 (2021).
8. Lange, M. *et al.* CellRank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022).
9. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
10. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
11. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184–197 (2015).
12. Eppert, K. *et al.* Stem cell gene expression programs influence clinical outcome in human leukemia. *Nat. Med.* **17**, 1086–1093 (2011).
13. Tang, D. G. Understanding cancer stem cell heterogeneity and plasticity. *Cell Res.* **22**, 457–472 (2012).
14. Easwaran, H., Tsai, H.-C. & Baylin, S. B. Cancer epigenetics: tumor heterogeneity, plasticity of stem-like states, and drug resistance. *Mol. Cell* **54**, 716–727 (2014).
15. Kreso, A. & Dick, J. E. Evolution of the Cancer Stem Cell Model. *Cell Stem Cell* **14**, 275–291 (2014).
16. van Galen, P. *et al.* Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell* **176**, 1265–1281.e24 (2019).
17. Abelson, S. *et al.* Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400–404 (2018).
18. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).
19. Miles, L. A. *et al.* Single-cell mutation analysis of clonal evolution in myeloid malignancies. *Nature* **587**, 477–482 (2020).
20. Cancer Genome Atlas Research Network *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).
21. Bowman, R. L., Busque, L. & Levine, R. L. Clonal Hematopoiesis and Evolution to Hematopoietic Malignancies. *Cell Stem Cell* **22**, 157–170 (2018).
22. Zhang, S. *et al.* Single-cell transcriptomics identifies divergent developmental lineage trajectories during human pituitary development. *Nat. Commun.* **11**, 5275 (2020).
23. Duffy, M. F. *et al.* Divergent patterns of healthy aging across human brain regions at single-cell resolution reveal links to neurodegenerative disease. *bioRxiv* (2023) doi:10.1101/2023.07.31.551097.

24. Daniel, B. *et al.* Divergent clonal differentiation trajectories of T cell exhaustion. *Nat. Immunol.* **23**, 1614–1627 (2022).
25. Bachireddy, P. *et al.* Mapping the evolution of T cell states during response and resistance to adoptive cellular therapy. *Cell Rep.* **37**, 109992 (2021).
26. Cakir, B. *et al.* Comparison of visualization tools for single-cell RNAseq data. *NAR Genom Bioinform* **2**, lqaa052 (2020).
27. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4314.
28. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **10**, 5416 (2019).
29. Amir, E.-A. D. *et al.* viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31**, 545–552 (2013).
30. Chari, T., Banerjee, J. & Pachter, L. The Specious Art of Single-Cell Genomics. Preprint at https://doi.org/10.1101/2021.08.25.457696.
31. Argelaguet, R., Cuomo, A. S. E., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-021-00895-7.
32. Stuart, T. *et al.* Comprehensive integration of single cell data. Preprint at https://doi.org/10.1101/460147.
33. Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).
34. Liu, J. *et al.* Jointly defining cell types from multiple single-cell datasets using LIGER. *Nat. Protoc.* **15**, 3632–3662 (2020).
35. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
36. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
37. Persad, S. *et al.* SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01716-9.
38. Feng, C. *et al.* Dimension Reduction and Clustering Models for Single-Cell RNA Sequencing Data: A Comparative Study. *International Journal of Molecular Sciences* vol. 21 2181 Preprint at https://doi.org/10.3390/ijms21062181 (2020).
39. Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2022).
40. Svensson, V., Gayoso, A., Yosef, N. & Pachter, L. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics* **36**, 3418–3421 (2020).
41. Gayoso, A. *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods* **18**, 272–282 (2021).
42. Ranganath, R., Tang, L., Charlin, L. & Blei, D. M. Deep Exponential Families. (2014) doi:10.48550/arXiv.1411.2581.
43. Chari, T. & Pachter, L. The specious art of single-cell genomics. *PLoS Comput. Biol.* **19**, e1011288 (2023).
44. Novershtern, N. *et al.* Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144**, 296–309 (2011).
45. Falini, B., Brunetti, L., Sportoletti, P. & Martelli, M. P. NPM1-mutated acute myeloid leukemia: from bench to bedside. *Blood* **136**, 1707–1721 (2020).
46. Ivey, A. *et al.* Assessment of Minimal Residual Disease in Standard-Risk AML. *N. Engl. J. Med.* **374**,

422–433 (2016).
47. Busque, L. *et al.* Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. *Nat. Genet.* **44**, 1179–1181 (2012).
48. Uckelmann, H. J. *et al.* Therapeutic targeting of preleukemia cells in a mouse model of NPM1 mutant acute myeloid leukemia. *Science* **367**, 586–590 (2020).
49. Shlush, L. I. *et al.* Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506**, 328–333 (2014).
50. Khoury, J. D. *et al.* The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms. *Leukemia* **36**, 1703–1719 (2022).
51. Bennett, J. M. *et al.* Proposals for the classification of chronic (mature) B and T lymphoid leukaemias. French-American-British (FAB) Cooperative Group. *J. Clin. Pathol.* **42**, 567–584 (1989).
52. Rodriguez-Meira, A. *et al.* Single-cell multi-omics identifies chronic inflammation as a driver of TP53-mutant leukemic evolution. *Nat. Genet.* **55**, 1531–1541 (2023).
53. Petti, A. A. *et al.* A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. *Nat. Commun.* **10**, 3660 (2019).
54. Heo, S.-K. *et al.* CD45CD34CD38CD133 cells have the potential as leukemic stem cells in acute myeloid leukemia. *BMC Cancer* **20**, 285 (2020).
55. Kühn, M. W. M. *et al.* Targeting Chromatin Regulators Inhibits Leukemogenic Gene Expression in NPM1 Mutant Leukemia. *Cancer Discov.* **6**, 1166–1181 (2016).
56. Zhang, Q. *et al.* Tet2 is required to resolve inflammation by recruiting Hdac2 to specifically repress IL-6. *Nature* **525**, 389–393 (2015).
57. Jaiswal, S. *et al.* Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N. Engl. J. Med.* **377**, 111–121 (2017).
58. Lasry, A. *et al.* An inflammatory state remodels the immune microenvironment and improves risk stratification in acute myeloid leukemia. *Nat Cancer* **4**, 27–42 (2023).
59. Yeaton, A. *et al.* The Impact of Inflammation-Induced Tumor Plasticity during Myeloid Transformation. *Cancer Discov.* **12**, 2392–2413 (2022).
60. Velten, L. *et al.* Identification of leukemic and pre-leukemic stem cells by clonal tracking from single-cell transcriptomics. *Nat. Commun.* **12**, 1366 (2021).
61. Forristal, C. E. *et al.* Hypoxia inducible factor (HIF)-2α accelerates disease progression in mouse models of leukemia and lymphoma but is not a poor prognosis factor in human AML. *Leukemia* **29**, 2075–2085 (2015).
62. Binder, S., Luciano, M. & Horejs-Hoeck, J. The cytokine network in acute myeloid leukemia (AML): A focus on pro- and anti-inflammatory mediators. *Cytokine Growth Factor Rev.* **43**, 8–15 (2018).
63. Raffel, S. *et al.* BCAT1 restricts αKG levels in AML stem cells leading to IDHmut-like DNA hypermethylation. *Nature* **551**, 384–388 (2017).
64. Corradi, G. *et al.* Release of IFNγ by Acute Myeloid Leukemia Cells Remodels Bone Marrow Immune Microenvironment by Inducing Regulatory T Cells. *Clin. Cancer Res.* **28**, 3141–3155 (2022).
65. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One* **9**, e98679 (2014).
66. Van Der Maaten Laurens, G. H. & Hinton, G. Van Der Maaten. *Visualizing data using t-sne. JMLR*.
67. Kuchroo, M. *et al.* Multiscale PHATE identifies multimodal signatures of COVID-19. *Nat. Biotechnol.* **40**, 681–691 (2022).
68. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic

data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
69. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
70. Gao, Y., Zhou, J.-F., Mao, J.-Y., Jiang, L. & Li, X.-P. Identification of the Thyrotropin-Releasing Hormone (TRH) as a Novel Biomarker in the Prognosis for Acute Myeloid Leukemia. *Biomolecules* **12**, (2022).
71. Dovey, O. M. *et al.* Molecular synergy underlies the co-occurrence patterns and phenotype of -mutant acute myeloid leukemia. *Blood* **130**, 1911–1922 (2017).
72. Gayoso, A. *et al.* A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.* **40**, 163–166 (2022).
73. Collins, C. T. & Hess, J. L. Role of HOXA9 in leukemia: dysregulation, cofactors and essential targets. *Oncogene* **35**, 1090–1098 (2016).
74. Qian, Z., Fernald, A. A., Godley, L. A., Larson, R. A. & Le Beau, M. M. Expression profiling of CD34+ hematopoietic stem/ progenitor cells reveals distinct subtypes of therapy-related acute myeloid leukemia. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 14925–14930 (2002).
75. Nazaret, Fan, Pe'er & Azizi. Probabilistic basis decomposition for characterizing temporal dynamics of gene expression. in *ICML Workshop on Computational Biology* (2022).
76. Issa, G. C. *et al.* The menin inhibitor revumenib in KMT2A-rearranged or NPM1-mutant leukaemia. *Nature* **615**, 920–924 (2023).
77. Zhong, X. *et al.* HoxA9 transforms murine myeloid cells by a feedback loop driving expression of key oncogenes and cell cycle control genes. *Blood Adv* **2**, 3137–3148 (2018).
78. Pan, X.-N. *et al.* Inhibition of c-Myc overcomes cytotoxic drug resistance in acute myeloid leukemia cells by promoting differentiation. *PLoS One* **9**, e105381 (2014).
79. Platanias, L. C. Mechanisms of type-I- and type-II-interferon-mediated signalling. *Nat. Rev. Immunol.* **5**, 375–386 (2005).
80. Yang, L. *et al.* Up-regulation of EMT-related gene VCAN by NPM1 mutant-driven TGF-β/cPML signalling promotes leukemia cell invasion. *J. Cancer* **10**, 6570–6583 (2019).
81. Zhang, X. *et al.* DNMT3A and TET2 compete and cooperate to repress lineage-specific transcription factors in hematopoietic stem cells. *Nat. Genet.* **48**, 1014–1023 (2016).
82. Eriksson, A., Lennartsson, A. & Lehmann, S. Epigenetic aberrations in acute myeloid leukemia: Early key events during leukemogenesis. *Exp. Hematol.* **43**, 609–624 (2015).
83. Castillo, D. *et al.* Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level. *PLoS One* **14**, e0212127 (2019).
84. Shih, A. H. *et al.* Mutational cooperativity linked to combinatorial epigenetic gain of function in acute myeloid leukemia. *Cancer Cell* **27**, 502–515 (2015).
85. Kim, J. *et al.* Single-cell analysis of gastric pre-cancerous and cancer lesions reveals cell lineage diversity and intratumoral heterogeneity. *NPJ Precis Oncol* **6**, 9 (2022).
86. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
87. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
88. Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. Stochastic variational inference. *J. Mach. Learn. Res.*
89. Lalchand, V., Ravuri, A. & Lawrence, N. D. Generalised GPLVM with Stochastic Variational Inference. in *International Conference on Artificial Intelligence and Statistics* 7841–7864 (PMLR, 2022).

90. Lawrence, N. Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models. *J. Mach. Learn. Res.* **6**, 1783–1816 (2005).
91. Almosailleakh, M. & Schwaller, J. Murine Models of Acute Myeloid Leukaemia. *Int. J. Mol. Sci.* **20**, (2019).
92. Zhai, Y. *et al.* Longitudinal single-cell transcriptomics reveals distinct patterns of recurrence in acute myeloid leukemia. *Mol. Cancer* **21**, 166 (2022).
93. Yin, A. H. *et al.* AC133, a novel marker for human hematopoietic stem and progenitor cells. *Blood* **90**, 5002–5012 (1997).
94. Herrmann, H. *et al.* Delineation of target expression profiles in CD34+/CD38- and CD34+/CD38+ stem and progenitor cells in AML and CML. *Blood Adv* **4**, 5118–5132 (2020).
95. Falini, B. *et al.* Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *N. Engl. J. Med.* **352**, 254–266 (2005).
96. Teimouri, H. & Kolomeisky, A. B. Temporal order of mutations influences cancer initiation dynamics. *Phys. Biol.* **18**, (2021).
97. Brunetti, L. *et al.* Mutant NPM1 Maintains the Leukemic State through HOX Expression. *Cancer Cell* **34**, 499–512.e9 (2018).
98. Abe, M., Lin, J., Nagata, K. & Okuwaki, M. Selective regulation of type II interferon-inducible genes by NPM1/nucleophosmin. *FEBS Lett.* **592**, 244–255 (2018).
99. Carey, A. *et al.* Identification of Interleukin-1 by Functional Screening as a Key Mediator of Cellular Expansion and Disease Progression in Acute Myeloid Leukemia. *Cell Rep.* **18**, 3204–3218 (2017).
100. Morganti, C. *et al.* NPM1 ablation induces HSC aging and inflammation to develop myelodysplastic syndrome exacerbated by p53 loss. *EMBO Rep.* **23**, e54262 (2022).
101. Barreyro, L., Chlon, T. M. & Starczynowski, D. T. Chronic immune response dysregulation in MDS pathogenesis. *Blood* **132**, 1553–1560 (2018).

METHODS

The Decipher framework

Given a dataset of single-cell gene expression $(x_{i,g})$ of N cells and G genes, Decipher models the expression of genes in cells by learning multiple hidden representations of each cell, at increasingly finer detail: the Decipher components v gives a high-level two-dimensional representations, and the ten-dimensional Decipher latent factors z are more refined representations of cells. To achieve this, Decipher successively combines multiple neural networks in a probabilistic framework and successively generates hidden representations with increasing dimensions (**Fig. 1b**).

Decipher is a generative model that extends traditional Variational AutoEncoder models (such as scVI and others [22, 23]) by adding an additional neural network, encoding a higher level two-dimensional latent space on top of their latent space. Importantly, this extra layer allows Decipher to model more complex cell state distributions with potentially dependent latent space factors; it provides an interpretable decomposition of the latent factors and a ready-to-use visualization of the data. Moreover, the top two-dimensional Decipher components enable visualization of the space directly from the model without additional dimensionality reduction. To fit Decipher to data from multiple samples, cells are simply concatenated.

Fitting the probabilistic model defined by Decipher is done with amortized variational inference [5]. In other words, the Decipher model and the Decipher inference together form a special kind of Variational Auto-Encoder [19], with two nested encoders and two decoders, allowing it to encode and decode from any level of representation to any other level of representation (**Fig. 1b**), which is a novelty of our approach.

The Python code implementing our method is available at https://github.com/azizilab/decipher. For each method presented below, we reference its corresponding Python function. Our Python code follows the architecture of the popular `scanpy` package [44], with computation functions in a `.tl` submodule and the plotting functions in a `.pl` submodule. In the code snippets of the methods below, we assume that we have imported the `decipher` package as follows `import decipher as dc` and that the data of interest is in an AnnData object called `adata`. For instance, training Decipher is done with `dc.tl.decipher_train(adata)` and plotting the Decipher space colored by cell type is performed with `dc.pl.decipher(adata, color="cell_type")`.

The generative model

Decipher first models each cell i as a two-dimensional standard normal variable v_i, termed *Decipher components*, representing the two largest axes of cell heterogeneity, such as cell type or cancer progression stages. This representation (v_i) directly enables two-dimensional visualization of the data. A learnable neural network f – a *decoder* – transforms each v_i into a distribution over medium-dimensional vectors, a latent space representing cell state.

Decipher samples a cell state z_i conditionally on v_i from the distribution induced by $f(v_i)$; the z_i contains richer information about cell i than v_i. The z_is are medium-dimensional. They are substantially lower dimensional than the number of genes but higher than v_i and can thus capture more information regarding cell variability (we set the dimension to 10 in our experiments). We refer to the medium-dimensional variables z_i as *Decipher latent factors*. These are comparable to the latent variables of other VAE-based or matrix-factorization-based methods [21, 23, 29].

Then, a second neural network h – a second decoder – with a softmax output layer, transforms the state z_i into the gene expression values $(\mu_{i,g})_{1 \leq g \leq G} = h(z_i)$ expressed in cell i. In practice, we measure a scaled version of this expression, specifically scaled by the (observed) library size of the cell, l_i, and the observed gene counts x_{ig} are sampled from a negative Binomial distribution with mean $\mu_{i,g}$ and dispersion θ_g specific to each gene g. This generative process is represented in **Fig. 1b** and is described mathematically as follows:

θ_g		$\triangleright$	gene dispersion	$\in \mathbb{R}_{>0}$
$h,\ f$		$\triangleright$	decoder neural networks	
v_i	$\sim \mathcal{N}(0, I_2)$	$\triangleright$	Decipher components	$\in \mathbb{R}^2$
z_i	$\sim \mathcal{N}(f_{\texttt{mean}}(v_i), f_{\texttt{var}}(v_i))$	$\triangleright$	Decipher latent factors	$\in \mathbb{R}^L \approx \mathbb{R}^{10}$
$\mu_{i,g}$	$= h(z_i)_g$	$\triangleright$	normalized gene expression	$\mu_i \in \mathbb{R}^G$
$x_{i,g}$	$\sim NB(\ell_i \cdot \mu_{i,g}, \theta_g)$	$\triangleright$	observed gene expression	$x_i \in \mathbb{R}^G$

The mapping functions f and h are neural networks. In practice, Decipher uses a single linear layer for h as suggested by Svensson *et al.* [36] to improve interpretability, i.e. insight into relationships between variables. Meanwhile, f has two linear layers interleaved with ReLU activations. The last layer of f produces a vector in $\mathbb{R}^{2L}$ that is split to form two outputs: $f_{\texttt{mean}} \in \mathbb{R}^L$ and $f_{\texttt{var}} \in \mathbb{R}^L$. The choice of negative binomial distribution can be replaced by other distributions if the user believes it is more appropriate for the data at hand.

High-level summary. The Decipher components v_i represent the high-level organization of the cells and form the Decipher space. This space provides a ready-to-use 2D representation of the data without requiring further projection methods such as UMAP or t-SNE. It offers direct visual access inside the probabilistic model. Then, through the neural network f, each v_i induces a cell state z_i, a more detailed representation of cell i. The space of the z_i corresponds to the latent space of traditional variational-autoencoders.

Inference for Decipher's deep probabilistic model

Given the observed gene expression data $\mathcal{D} = (x_{i,g})_{1 \le g \le G}^{1 \le i \le N}$ and the parameters $(\theta_g)_{1 \le g \le G}$ f,h, the Decipher's probabilistic model defines a posterior $p\,(v, z \mid \mathcal{D})$ over the latent variables $v = (v_i)_{1 \le i \le N}, z = (z_i)_{1 \le i \le N}$. We use variational inference [5, 15, 41] to approximate this exact posterior with a variational approximation $q\,(v, z)$.

The structure of the variational family. We use amortization over the local variables v_i, z_i in function of the observations x_i, such that the variational family becomes: $q(v, z) = \prod_{i=1}^N q(v_i, z_i|x_i)$ which always factorizes as

$$q(v, z) = \prod_{i=1}^{N} q(v_i|z_i, x_i) q(z_i|x_i).$$

The amortized distributions are set to diagonal Gaussian distributions with parameters (mean and variance) given by neural networks. The first neural network transforms x_i to the mean and the variance of the distribution $q(z_i|x_i)$, and the second neural network transforms (z_i, x_i) to the mean and the variance of the distribution $q(v_i|z_i, x_i)$. We denote them as: $d^{\to z}_{\texttt{mean}}(x), d^{\to z}_{\texttt{var}}(x), d^{\to v}_{\texttt{mean}}(x, z)$, and $d^{\to v}_{\texttt{var}}(x, z)$, such that:

$$q(z_i|x_i) = \mathcal{N}(d^{\to z}_{\texttt{mean}}(x_i), d^{\to z}_{\texttt{var}}(x_i)), \quad \text{and} \quad q(v_i|z_i, x_i) = \mathcal{N}(d^{\to v}_{\texttt{mean}}(x_i, z_i), d^{\to v}_{\texttt{var}}(x_i, z_i))$$

The variational inference objective. Variational inference seeks to minimize the KL divergence between the variational posterior q and the exact posterior $p(\ \cdot\ \mid \mathcal{D})$. It is equivalent to maximizing a lower bound of the evidence, called the ELBO [5]:

$$ELBO(q) = \sum_{i=1}^{N} \mathbb{E}_{q(v_i, z_i)} \left[\sum_g \log p(x_{i,g}|z_i, \theta_g) + \log \frac{p(z_i|v_i)}{q(z_i|x_i)} + \beta \log \frac{p(v_i)}{q(v_i|z_i, x_i)} \right],$$

where β is a scalar controlling the importance of the prior $p(v_i)$, between 0 (no prior) and 1 (standard ELBO) [12].

Because we chose the variational posteriors q to be Gaussian distributions, we can reparametrize the expectations to sample unbiased low-variance estimates of the ELBO. To obtain a sample for (z_i, v_i) from $q(v_i, z_i|x_i) = q(v_i|z_i, x_i)q(z_i|x_i)$, we first sample a reparametrized z_i from $q(z_i|x_i)$ and then sample a reparametrized v_i from $q(v_i|z_i, x_i)$ [19].

The gradients are then computed using automatic differentiation. To scale up to large datasets of cells, we further subsample the outer sum using a minibatch size of 64 observations to perform stochastic variational inference [13]. We use the Adam optimization algorithm [18] to execute the gradient updates. The code is implemented in Python using Pyro [3].

Because we have little prior on the distribution of the Decipher components v (remember that a limitation of other methods is that the prior enforces unrealistic independence between latent variables), we set β to a low value $1e-1$ in our experiments.

The Decipher model can be fitted using the function `dc.tl.decipher_train(adata)`.

Generating the Decipher space v and the latent space z. Once the inference is performed, the variational posteriors $q(v, z|x)$ are fitted to the data. The "decoders" $d^{\to z}_{\texttt{mean}}(x), d^{\to z}_{\texttt{var}}(x), d^{\to v}_{\texttt{mean}}(x, z)$, and $d^{\to v}_{\texttt{var}}(x, z)$ give the posterior expected values of v_i and z_i given each cell x_i. For each cell x_i, we compute (as in any auto-encoder architecture): $\hat{z}_i = d^{\to z}_{\texttt{mean}}(x_i)$ and $\hat{v}_i = d^{\to v}_{\texttt{mean}}(x_i, z_i)$.

The Decipher space v and latent z are automatically computed when calling `dc.tl.decipher_train`. They are stored in `adata.obs["decipher_v"]` and `adata.obs["decipher_z"]`.

Rotating and aligning the Decipher space. Decipher does not use sample or batch IDs in learning latent variables. However, in a post-processing step, the sample IDs can be optionally used to align Decipher components to represent the most shared and most distinct information between the samples (e.g. perturbed and normal conditions), thus facilitating downstream analysis. This is accomplished by rotating or flipping the v components. Like most auto-encoder models (e.g., scVI [23]), the axes of the latent spaces v and z can be rotated or flipped without changing the likelihood of the data. To automatically rotate and flip the Decipher components given the user preferences, the user can specify if some cell labels should be aligned with a given component. For example, in our analysis, we choose to align the cells' sample labels (Healthy and AML) along Decipher 2 and the cells' cell-type labels (ordered from blast0 to blast3) along Decipher 1.

Given cell labels and their target alignment axis (e.g., ordered cell types along Decipher 1, ordered cell sample IDs along Decipher 2), we try 100 rotations (from 0 to 2π) and all possible axis flips (2 for v_1, 2 for v_2), and pick the setting that maximizes the correlations between the cell labels and their target Decipher axis.

This is accomplished by calling the `dc.tl.decipher_rotate_space` function.

Trajectory inference

The Decipher components (v_i) organize cells along visual trajectories. For instance, in the joint AML-healthy data, there are two trajectories: one for healthy maturation and one for AML progression (**Fig. 2a**). The trajectories could hypothetically be traced manually by the user. Still, we propose a simple automated determination of the trajectories using, as input, the marker genes for the beginning and the end of the trajectories.

Given the cell representations (v_i) in the Decipher space and (z_i) in the latent space , we first cluster cells using the Leiden algorithm [37] on the latent representations (z_i) – we use the representation (z_i) to cluster the cells because they contain more detailed information about the cells than the (v_i) (**Extended Data Fig. 4a**, left). We then compute a minimum spanning tree between the clusters' centroids using the distances in the Decipher space – we use the distance in Decipher space because the high-level geometry of the data is better captured by the (v_i) (**Extended Data Fig. 4a**, middle left). Finally, we use the provided marker genes to identify the trajectories' beginning and end, from which we compute the shortest path in the minimum spanning tree (**Extended Data Fig. 4a**, middle right). We use linear interpolation to form parameterized trajectories $\gamma : t \mapsto v(t)$ in the Decipher space (**Extended Data Fig. 4a**, right). The time t that parametrizes the trajectories is called the *Decipher time*, and we compute one trajectory per sample in our analysis (γ_{AML} and γ_{healthy}). If the analysis requires it, more trajectories or less could be computed.

The procedure is described in Algorithm 1 and is visually represented in Extended Data Fig. 4a. The trajectories are computed using the function `dc.tl.trajectories`.

Algorithm 1 Trajectory inference with Decipher

Input: (v_i), (z_i), start and end marker genes for each trajectory j: $(g_j^{(s)}, g_j^{(e)})$.
Output: Trajectories $(\gamma_j : v \mapsto \gamma_j(v))_j$
  Partition the cells in clusters (C_m) using the Leiden algorithm applied on the (z_i).
  Compute the centroid of the clusters in the Decipher space: $\mu_m = \frac{1}{|C_m|} \sum_{k \in C_m} v_k$
  Compute a minimum spanning tree between the clusters' centroids (distances computed in Decipher space)
  for each trajectory j **do**
    Identify starting and ending cluster expressing the markers:
    $s_j \leftarrow \arg\max_m \frac{1}{|C_m|} \sum_{k \in C_m} x_{k,g_j^{(s)}}$
    $e_j \leftarrow \arg\max_m \frac{1}{|C_m|} \sum_{k \in C_m} x_{k,g_j^{(e)}}$
    Find shortest path from C_{s_j} to C_{e_j} in the tree.
    Create a parametrized curve $v \mapsto \gamma_j(v)$ linearly interpolating the obtained path.
  end for

Trajectory alignment using Decipher time. While trajectory alignment approaches e.g. with dynamic time warping [7, 16], can alter the relative lengths of trajectories by locally compressing or stretching them and potentially missing rare cell states, our jointly inferred Decipher time obtains a common time axis for both trajectories. Decipher assigns a pseudo-time value to each location along the trajectories, called the Decipher time. This is determined by calculating the curvilinear coordinate t along each trajectory within the Decipher space. To assign a Decipher time to any cell from our data, we project the cells onto the trajectories. That is, we compute for each cell i in each sample j (healthy or AML), the closest point $\gamma_j(t^*)$ on the dataset-specific trajectory (γ_{healthy} or γ_{AML}), and assign the time of this trajectory point t^* to cell i (**Fig. 1c**). The Decipher time is computed using the function `dc.tl.decipher_time`.

Reconstruction of gene expression by Decipher. Given a trajectory γ of Decipher components, we can obtain the gene expression along this trajectory using the decoder neural networks in the Decipher probabilistic model. Indeed, we recall that the encoders and the decoders in the Decipher model can convert any gene expression into Decipher components and vice versa. Mathematically, given some Decipher components on the trajectory $v := \gamma(t)$, we use the decoders to compute the associated expected latent factors $z := f_{\texttt{mean}}(v)$, followed by the expected normalized gene expressions $\mu := h(z)$, that can further be scaled up to the desired library size by multiplying it by ℓ. Because Decipher is a probabilistic model, it is also possible to obtain probabilistic gene expression samples instead of a single estimate. This is particularly useful to obtain some model uncertainty around the reconstructed gene expression (**Fig. 3a**). To achieve this, we sample multiple latent factors given the Decipher components as $z_m \sim \mathcal{N}(f_{\texttt{mean}}(v), f_{\texttt{var}}(v))$ (instead of just $f_{\texttt{mean}}(v)$) and we compute their expected gene expression $\mu_m = h(z_m)$.

The Decipher gene expression patterns are computed using the function `dc.tl.gene_patterns`.

Reconstruction of gene expression by scVI. Like Decipher, scVI [9] has a latent space and a decoder that can reconstruct gene expression from a latent representation z. Instead of having a low-dimensional space v that jointly represents the healthy and disease sample, scVI is given the label (healthy or disease) as a batch variable. Once the model is learned, scVI produces batch-corrected gene expression data for each sample by manually changing the batch label from one sample to the other, as described in `https://docs.scvi-tools.org/en/stable/api/reference/scvi.model.SCVI.html#scvi.model.SCVI.get_normalized_expression` (accessed on July 10, 2023).

Basis decomposition

To quantify the difference in patterns between two trajectories, we need a metric that accounts for the temporal order of cells – two genes may have the same mean expression value but opposite patterns, e.g., ascending versus descending along trajectories. Existing methods that encode temporal dependencies are limited in modeling assumptions and scalability. For instance, tradeSeq [39] performs trajectory-based differential expression; however, approximating gene patterns with splines may not be appropriate for complex transcriptional programs, e.g.,

with cascades of mutations leading to cancer. Additionally, relying on built-in denoising limits compatibility with preprocessed data (e.g., from VAEs). Methods such as DPGP [25] utilize Gaussian Processes model distributions of all functions over time; however, they are computationally expensive. This motivates a metric that not only accounts for the order of cell states and uncertainty but is generalizable and scalable to the size of standard single-cell datasets. For this, we develop a method that decomposes every gene pattern on a basis of dominant patterns learned with neural networks. Comparing different patterns then becomes a comparison of their basis weights.

The data. The input data for the basis decomposition method is the collection of gene patterns observed over a time or pseudotime axis in different conditions, such as normal, disease, or perturbed (**Fig. 4a**). Similar to the trajectory $\phi : t \mapsto \phi(t) =: v$ that maps pseudo-time to Decipher components, we define a gene pattern $\mu_{g,c}$ for a gene g under condition c to be the function $t \mapsto \mu_{g,c}(t)$ that represents the expected expression of gene g at time t in condition c. With Decipher, these can be the gene expression patterns reconstructed from the inferred trajectories. In a more general setting, these patterns can be obtained with time-series bulk RNA-seq or trajectory inference methods [34, 38], applied in advance to single-cell RNA-seq data, or other dynamic features, such as chromatin accessibility (measured with ATAC-seq) or protein expression (CITE-seq).

The set of all genes is G, and the set of conditions is C. For simplicity of the exposition, we restrict the conditions to $C = \{\text{healthy}, \text{disease}\}$. The input data is the collection of functions $\mathcal{D} = (\mu_{g,c})_{g \in G, c \in C}$. The data has $|G| \cdot |C|$ observations, each of which is a function. In the proposed probabilistic model, each pattern $t \mapsto \mu_{g,c}(t)$ is considered a single observation.

The model. In light of commonly used generative models [4], the proposed model is a linear factor model operating in function space. For gene g and condition c, the model associates the data point $\mu_{g,c}$ with a latent scalar $s_{g,c}$ – the *gene scale* – and a latent vector $\beta_{g,c}$ of K dimensions – the *gene shape*. Each dimension k corresponds to a latent basis pattern $t \mapsto b_k(t)$. The $\beta_{g,c,k}$ are coefficients for the function $\mu_{g,c}$ in this basis. The coefficient $s_{g,c}$ is the intrinsic scale of gene c in condition g, which will scale up or down the pattern computed from the bases, which, in contrast, are constrained to be between 0 and 1. The observations $\mu_{g,c}$ and the latent basis b_k are functions. The scale $s_{g,c}$ and the weights $\beta_{g,c,k}$ combine the latent bases b_k to generate the observed function $\mu_{g,c}$. That is, informally, $\mu_{g,c} \approx s_{g,c} \sum_{k=1}^{K} \beta_{g,c,k} b_k$. Each basis function b_k forms a representative pattern shared by multiple genes.

The weights $\beta_{g,c,k}$. To ensure the interpretability of the weights, the model mimics methods like mixture or topic models [6] and draws positive weights that sum to 1. With this, a non-zero weight $\beta_{g,c,k}$ signifies that gene g in condition c exhibits the representative pattern k. Specifically, the weights vectors $\beta_{g,c}$ are drawn independently from a Dirichlet distribution with concentration parameter $[\eta, \eta, ..., \eta]$, denoted as $\text{Dir}(\eta)$, and with density

$$p(\beta|\eta) = \frac{1}{B(\eta)} \prod_{k=1}^{K} \beta_k^{\eta-1},$$

where $B(\eta) = \Gamma(\eta)^K / \Gamma(K\eta)$ and Γ is the Gamma function. In terms of negative log-likelihood, this prior induces a regularization of the coefficients that will lead to sparse β_k when $\eta < 1$ and β closer to $\frac{1}{K}$ when $\eta > 1$. We choose $\eta < 1$ to associate the basis with dominant patterns and thus obtain a better interpretability of the basis.

The function basis b_k. To sample the basis functions b_k, we represent them as neural networks and sample each b_k by drawing its neural network parameters. Concretely, each basis function is modeled by a one-dimensional neural network with two hidden layers of 32 units each, followed by the `tanh` activation. The neural network b_k is of the form

$$b_k : \mathbb{R} \to \mathbb{R}^{32} \to \mathbb{R}^{32} \to \mathbb{R}$$

and its parameters are denoted ϕ_k. Each ϕ_k is sampled from a centered diagonal normal distribution, and the variance of each of its coordinates is set to the inverse of the input dimension of the linear layer in which it appears. The study of infinite neural networks in Neal [26] demonstrates that with wide hidden units (here $32 \gg 1$) and such a prior on the parameters, the induced prior in function space is close to a Gaussian process. Gaussian processes are used in other methods [25] but are hardly scalable. Using neural networks for efficient computations solves the problem. The induced prior in function space is denoted by φ. Finally, to design more interpretability for the gene scale $s_{g,c}$, we normalized the sampled basis by their maximum value so that the maximum value reached by a basis is 1.

The gene scales $s_{g,c}$. The gene scales $s_{g,c}$ are learned as variational parameters (no variational distributions), as the gene scales can greatly vary between genes.

The observations $\mu_{g,c}$. Finally, the gene pattern $\mu_{g,c}$ is generated from a distribution parameterized by $\sum_k \beta_{g,c,k} b_k$ and the scale $s_{g,c}$. More specifically, $\mu_{g,c}$ is sampled from a Gaussian process[1] with mean $s_{g,c} \cdot \sum_k \beta_{g,c,k} b_k$ and with a white Gaussian noise kernel $(x, x') \mapsto \sigma^2 \delta_{x,x'}$ of variance σ^2.

The generative process is represented graphically in **Fig. 4a** and proceeds as follows:

1. For each factor dimension $k \in [\![1, K]\!]$, draw a basis function b_k from the function prior: $b_k \sim \varphi(b_k)$ (that is draw weights ϕ_k according to the prior detailed above)
2. For each gene g, and condition c do:
   (a) For each factor dimension k, draw the basis weight $\beta_{g,c,k} \sim \mathcal{E}(\eta)$
   (b) Draw the observed function $\mu_{g,c}$ from $\mu_{g,c} \sim \mathcal{GP}\left(s_{g,c} \cdot \sum_k \beta_{g,c,k} b_k, (x, x') \mapsto \sigma^2 \delta_{x,x'}\right)$.

For simplicity of notations, the ϕ_k are grouped in parameter ϕ, and the $\beta_{g,c,k}$ into parameter β.

The inference. We learn an approximate posterior on the model variables $q(b_k, \beta_{k,g,c}, s_{g,c})$ using variational inference implemented in the Python probabilistic modeling library Pyro [3], with automatic guides.

The basis decomposition is computed using the function `dc.tl.basis_decomposition`.

The disruption scores. From the inferred model parameters, we design multiple disruption scores that inform us of different types of disruptions for the same gene across two conditions $c_1 =$ healthy and $c_2 =$ disease.

- The scale disruption highlights the difference in gene scale between the two conditions, e.g., a gene that is up-regulated in one of the conditions. It is defined as $|\log s_{g,c_1} - \log s_{g,c_2}|$.
- The shape disruption highlights the difference in gene shape between the two conditions, e.g., a gene that activated later in one of the conditions and earlier in another. It is defined as $||\beta_{g,c_1} - \beta_{g,c_2}||$.
- The combined disruption is a combination of both disruption scores to capture a general high-level disruption score, including both shape and scale. It is defined as $||\log(s_{g,c_1}\beta_{g,c_1}) - \log(s_{g,c_2}\beta_{g,c_2})||$.

The disruption scores are computed using the function `dc.tl.disruption_scores`.

Data generation

Simulated data

To evaluate Decipher's ability to identify cell state evolution trajectories within its latent space, we simulate data with ground-truth trajectories, fit the Decipher model on this data, and evaluate the quality of the trajectory reconstruction.

General simulation. To generate simulated data, we employ a semi-synthetic approach. First, we fit Decipher to the healthy-AML1 dataset to obtain the neural networks h and f. Then, we utilize the trajectory inference method to derive the normal cell maturation trajectory $t \mapsto v^{(1)}(t)$ and the AML progression trajectory $t \mapsto v^{(2)}(t)$ within the Decipher space v. We use these trajectories to sample ground-truth times and ground-truth cell states. We simulate data along each trajectory as follows:

For each trajectory $c \in 1, 2,$:

[1] Because the Gaussian Process is used here only to define the distribution of the observations, and not to sample an unobserved latent variable, there is no computational difficulty in using it.

- Uniformly sample ground-truth Decipher times t_i between the trajectories' beginning and end.
- Use the trajectories $v^{(1)}$ and $v^{(2)}$ to associate a cell state to each t_i.
- Sample Decipher components v_i along the trajectory from a normal distribution as $v_i \sim \mathcal{N}(v^{(c)}(t_i), 0.1)$.
- Utilize the generative model of Decipher, with the learned neural networks h and f, to sample gene expression for each gene from the Decipher components v_i. This involves sampling $z_i \sim \mathcal{N}(f_{\texttt{mean}}(v_i), f_{\texttt{var}}(v_i))$, calculating $\mu_{i,g} = h(z_i)$, and generating $x_{i,g} \sim NB(\ell_i \cdot \mu_{i,g}, \theta_g)$. The gene-specific dispersion parameter (θ_g) are drawn from a gamma distribution $\Gamma(\text{shape} = 2, \text{scale} = 0.5)$, and the library sizes ℓ_i are set to 10,000.

We then have gene expression (x_{ig}) and ground-truth time t_i for each synthetic cell i. We evaluate Decipher's success by computing the Spearman correlation between the inferred Decipher time and the ground-truth time and by measuring Decipher's ability to learn a meaningful space of v where the trajectory inference method accurately recovers the true trajectories.

Simulation with rare cell states. To examine the robustness of Decipher in the context of rare cell states, we perform additional simulations where 90% of cells from two intermediary states were removed. That is, we simulate the trajectory as described and then randomly remove 90% of cells from the cell states *blast0* and *blast2* (**Extended Data Fig. 3d**).

AML data collection

The TET2^{mut} AML cohort consists of 12 cryopreserved (DMSO) BM AML samples from the *Banque de cellules leucémiques du Québec* (BCLQ) biobank, with 10 patient specimens collected at the time of diagnosis, and two specimens from the same patient, at diagnosis and relapse (Supplementary Table 1). The DNMT3A^{mut} AML cohort consists of 5 cryopreserved (DMSO) BM AML patient samples. FAB information for samples was provided by BCLQ (Supplementary Table 1). Karyotyping, as well as mutation (variant) calling, was performed via bulk RNA-sequencing as part of the Leucegene project with data deposited on GEO with accession IDs GSE106272, GSE49642, GSE52656, GSE62190, GSE66917, and GSE67039.

All samples, as well as sorted cells, were profiled using 10X Genomics Chromium Single-cell 3' for scRNA-seq. For scATAC-seq, cells were subjected to 10X Genomics Chromium Single Cell ATAC Reagent Kits User Guide (v1.1 Chemistry). The resulting nuclei suspension was subjected to a transposition reaction for 60 min at 37°C and then encapsulated in microfluidic droplets using a 10X Chromium instrument following the manufacturer's instructions with a targeted nuclei recovery of approximately 5,000. Barcoded DNA material was cleaned and prepared for sequencing according to the Chromium Single Cell ATAC Reagent Kits User Guide (10X Genomics; CG000168 RevA). Purified libraries were assessed using a Bioanalyzer High-Sensitivity DNA Analysis kit (Agilent) and sequenced on an Illumina HiSeq 2500 (High Output) and NovaSeq platform at approximately 100 million reads per sample (around 5,000 nuclei) at MSKCC's Integrated Genomics Operation Core.

Flow cytometry activated cell sorting (FACS). For immature cell enrichment, FACS-purified CD34+ or PROM+ cells were subjected to single-cell RNA sequencing. Cryopreserved mononuclear cells were thawed into $10ml$ of prewarmed FACS buffer (phosphate-buffered saline (PBS) + 2% fetal bovine serum). Cells were pelleted at $300 \times G$ for 5 minutes and washed again with FACS buffer. Cells were then resuspended in FACS buffer containing Human TruStain FcX™ (Fc Receptor Blocking Solution; Biolegend #422301) for 15 minutes at 4°C. Antibodies against CD34 (Clone 561; FITC Biolegend 343603) and CD133 (clone 7; PE Biolegend 372803) were subsequently added, and cells were stained for an additional 15 minutes at 4°C. Cells were then washed twice with $3ml$ of FACS buffer and resuspended in FACS buffer with DAPI. Cell sorting was performed on a Sony SH800.

Data pre-processing and analysis

Data pre-processing Quantification of counts was done with SEQC [2], and 10X Genomics Cellranger [45]. Counts outputs were loaded into AnnData format using `scanpy` 1.7.2 [44]. Cells with low library size were filtered out, with the filtering threshold being selected by the knee-point of a histogram of the log10 of the total counts per cell. We obtained a median of 10,504 cells per sample and a median of 5165 molecules per cell after filtering. Data were then normalized by median library size using `sc.pp.normalize_per_cell`. Doublet detection was performed using DoubletDetection [10], with 25 iterations. scATAC FASTQ files for each sample were preprocessed to a cell-by-peak count matrix through the CellRanger ATAC pipeline [32] with modifications as described in [1].

Annotation of AML TET2 cohort All cells from unsorted AML samples were considered **(Fig. 1e,f)**. PhenoGraph [20] clustering was run using 100 principal components and 15 nearest neighbors. Annotation of clusters with low counts or high mitochondrial reads was performed by visual analysis of boxplots for the log10 of counts per cluster, as well as the fraction of reads belonging to mitochondrial genes compared to all genes. We further annotated lymphoid and erythroid clusters using `scanpy`'s dotplot function to visualize key gene markers. We were able to identify these clusters by analyzing the fraction of cells in each cluster expressing key marker genes as well as the mean expression. Cells forming distinct low-count clusters and additional clusters with high mitochondrial fraction, using Phenograph clustering on a per-sample basis, were additionally removed, resulting in a global cohort of 104,116 cells.

To annotate the maturation stages of leukemic blasts, we computed correlations (`scipy.stats.pearsonr`) between the mean expression of each cluster and bulk gene expression data from sorted HSPCs [27]. The correlation calculation was limited to the 5277 most varying genes, 3475 of which overlapped with bulk data. Non-significant values ($p > 0.0005$) were removed (**Fig. 1f**). To control for cluster size, Shannon Diversity (**Fig. 1f**) was computed for the distribution of patient IDs in subsamples of N = 1000 (approximating the median cluster size) cells from each cluster and averaged across 20 iterations. Paired diagnosis-relapse samples (AML9, AML10) (**Extended Data Fig. 1a**) annotations were considered together as they are phenotypically very similar (**Fig. 1e**). Clusters are ordered within cell-type by decreasing diversity (**Fig. 1f**).

Mutation identification and metrics

We implemented a mutation calling protocol in order to identify mutations in NPM1 and DNMT3A. We first sorted and indexed the patient bam files using `samtools` [8], then reduced the file to the region containing the gene of interest. The NPM1 gene was analyzed from chromosome position 5:171387116-171411137, and the DNMT3a gene was analyzed from chromosome position 2:25230961-25344590. Files were then merged and indexed, then converted to FASTA format. The indexed bam file was then loaded into the Integrative Genomics Viewer [30], and the alignments were visually analyzed for the presence of the mutation. If a mutation is present, a range of 5-20 base pairs are selected for subsequent single-cell analysis (**Supplementary Tables 2,3**). For single-cell annotation of mutations, we used the previously generated FASTA file and the mutated sequence identified for each patient to search for the presence of the mutated sequence in individual cells.

Because the mutation state may be heterozygous (a cell may have both mutant and wild-type labels), most of our subsequent analysis utilizes our defined *mutation proportion* for each cell. Since our detection of mutation is dependent on expression, which is affected by dropouts in scRNA-seq, we compute an average mutation proportion in the neighborhood of each cell. To compute the mutation proportion, we find the 30 nearest neighbors of each cell on the truncated SVD decomposition (100 components) of the normalized data. The mutation proportion for each cell is then $m/(m + w + 1e^{-10})$, where $m =$ the number of cells bearing the mutated copy of the gene and $w =$ the number of cells bearing the wild-type copy of the gene among the 30 neighbors. The heterozygous NPM1 mutation is detected in 5-39 % of cells in each of the unsorted samples.

Verification of immature cell enrichment in sorted samples

The primary purpose of the cell sorting was to enrich the populations of CD34+ and PROM1+ immature cells in the data (**Extended Data Fig. 1c**). To verify that this enrichment was achieved, we first performed a visual analysis of the UMAP computed on the subset of cells in AML1 originating from the unsorted collection process first, and compared it with the UMAP of cells once the sorted cells were included with the unsorted cells. We verified that enrichment of PROM1 and CD34, along with cells with low NPM1 mutation proportion, was achieved in the UMAP (**Extended Data Fig. 1c**). We also quantified the expression of CD34 and PROM1 in each of four

categories: immature and non-immature cells in the unsorted cells only and immature and non-immature cells in both sorted and unsorted cells. All visualization was performed using `scanpy` [44].

Analysis of key gene markers with mutation onset

To understand the dynamics of expression in key gene markers in leukemic cell states, we performed a preliminary study on the full sorted dataset of each patient in the TET2 cohort (**Extended Data Fig. 2**). For each patient, we computed a Force Directed Layout using `scanpy` [44] and visualized the distributions of cell types, key marker genes, and mutation of NPM1 (**Extended Data Fig. 2a,b**). We also produced violin plots to show the distributions of expression for each marker gene as well as the proportion of NPM1 mutation within each of the cell types (**Extended Data Fig. 2c**). Plots were performed using `scanpy` and Seaborn [42]. Other notable markers enriched in NPM1 WT subset are CD34, MLLT3, and HMGA2, which confirm HSC-like phenotypes, whereas MPO is higher in expression in the NPM1 mutated subset, along with HOXA9, PBX3, MEIS1, NKX2-3.

Cell type mapping onto the DNMT3A cohort

To extend the annotations from the TET2 cohort to patients in the DNMT3a cohort, we combined the data for all patients in the TET2 cohort, normalizing by median library size and log transforming across the entire cohort. Cells were then grouped based on their prior cell type annotations, and 700 differentially expressed genes were identified per cell type using the T-test version of `scanpy`'s [44] *rank_genes_groups()* function. Mitochondrial and Ribosomal genes were excluded from the gene sets. Cells in the DNMT3A cohort were also combined across patients, normalized by median library size, and log-transformed. Cells were then split into clusters using PhenoGraph [20], computed using 100 principal components and k=5. We then computed the cluster centroids of the PhenoGraph clusters in the DNMT3a cohort and the cell types of the TET2 cohort by taking the mean across cells in the cluster, limiting to the differentially expressed genes. Pearson correlation (using `scipy` 1.7.0 [40]) was computed between the centroids of the two cohorts, and each DNMT3A cluster was labeled with the cell type of the TET2 cell type cluster to which it had the greatest correlation coefficient (**Extended Data Fig. 10a**).

Benchmarking. We evaluated the performance of Decipher on simulated data. To further benchmark the performances of Decipher on real data, we define two metrics based on our prior knowledge of AML and compare Decipher to a large spectrum of commonly used methods.

Since we do not have ground-truth trajectory values for the real data, we build metrics on prior knowledge of AML progression, AML marker genes, and our independently curated cell state annotations: immature, blast0, blast1, blast2, and blast3. Among the healthy immature cells, we further use the markers CD34 and MPO to distinguish early cells (CD34+), late cells (MPO+), and intermediary cells. Our metrics are based on the distances between annotated cell states.

- **Ordering score**: We expect the cell states in a latent space to be spatially ordered along the known cell maturation trajectories. For instance, blast1 should be between blast0 and blast2. Given the orders $o_1 =$ [immature, blast0, blast1, blast2, blast3] and $o_2 =$ [early, intermediary, late], we want the total distances between consecutive cell states to be smaller than the distances between non-consecutive cell states. The triangular inequality guarantees that the ratio of these two quantities (the second over the first one) is maximized when the clusters are perfectly aligned in the right order.

$$\text{order}_j = \frac{\sum_{|i_1-i_2|>1} \text{distance}(o_j[i_1], o_j[i_2])}{\sum_{|i_1-i_2|=1} \text{distance}(o_j[i_1], o_j[i_2])}.$$

- **Divergence score**: We expect the AML trajectory to diverge from the healthy trajectory. That is, the immature cells of the AML sample are close to the early immature cells of the healthy sample. But then, the blast3 cells of the AML sample are far from the late immature cells of the healthy sample.

$$\text{divergence}_j = \sum_{\substack{c_1 \in o_1 \\ c_1 \neq \text{immature}}} \sum_{\substack{c_2 \in o_2 \\ c_2 \neq \text{early}}} \text{distance}(c_1, c_2) - 2 * \text{distance}(\text{immature(AML)}, \text{immature early (healthy)}).$$

  This metric is higher when non-immature AML cells and non-early healthy cells are far from each other and when the immature AML cells and early healthy cells are close to each other.

These metrics attempt to capture our high-level prior knowledge of AML. They summarize the latent space of each method in two numbers: the ordering score and the divergence score. For further details of each method, one can also directly analyze the visualization of the latent space of each method (**Extended Data Fig. 3a,b**).

Below are the benchmarked methods, the implementation we used, the hyperparameters, and which latent space we used to compute the metrics:

- **PCA.**
  - We run PCA with 50 components (default) using `scanpy`.
  - `sc.tl.pca(adata)`
  - The latent space is the space of 50 PCA components (comparable to our decipher z space).
  - `latent = adata.obsm["X_pca"]`
- **TSNE.**.
  - We run TSNE on the 50-dimensional PCA space using `scanpy` using a knn-graph with $k = 10$.
  - `sc.pp.neighbors(adata, n_neighbors=10); sc.tl.tsne(adata)`
  - The latent space is the 2d TSNE space (comparable to our decipher v space).
  - `latent = adata.obsm["X_tsne"]`
- **UMAP.**
  - We run UMAP on the 50-dimensional PCA space using `scanpy` using a knn-graph with $k = 10$.
  - `sc.pp.neighbors(adata, n_neighbors=10); sc.tl.umap(adata)`
  - The latent space is the 2d UMAP space (comparable to our decipher v space).
  - `latent = adata.obsm["X_umap"]`
- **Force Atlas.**
  - We run Force Atlas in `scanpy` using a knn-graph with $k = 10$.
  - `sc.pp.neighbors(adata, n_neighbors=10); sc.tl.draw_graph(adata)`
  - The latent space is the 2d force-directed layout space (comparable to our decipher v space).
  - `latent = adata.obsm["X_draw_graph_fa"]`
- **scVI with batch correction.**
  - We run scVI with two layers, a latent space of dimension 10, and batch correction on the *origin* label (AML vs Healthy), using `scvi-tools`.
  - `scvi.data.setup_anndata(adata, batch_key="origin"); vae = scvi.model.SCVI(adata, n_layers=2, n_latent=10); vae.train()`
  - The latent space is the 10-dimensional latent space (comparable to our decipher z space).
  - `latent = vae.get_latent_representation()`
- **scVI without batch correction.**
  - We run `scVI` with two layers, a latent space of dimension 10, and without batch correction, using `scvi-tools`.
  - `scvi.data.setup_anndata(adata); vae = scvi.model.SCVI(adata, n_layers=2, n_latent=10); vae.train()`
  - The latent space is the 10-dimensional latent space (comparable to our decipher z space).
  - `latent = vae.get_latent_representation()`
- **Phate.**
  - We run Phate using the `phate` Python package.
  - `phate_ = phate.PHATE()`

  - The latent space is the 2-dimensional latent space (comparable to our decipher v space).
  - `latent = phate_.fit_transform(adata)`

- **Harmony.**
  - We run Harmony using `scanpy` on the PCA with 50 components.
  - `sc.tl.pca(adata); sce.pp.harmony_integrate(adata, 'origin')`
  - The latent space is the 50-dimensional PCA-corrected latent space (comparable to our decipher z space).
  - `latent = adata.obs["X_pca_harmony]`

- **Seurat.**
  - We run Seurat using `Seurat` R package.
  - `adata <- FindVariableFeatures(data); adata.list <- SplitObject(adata, split.by = "origin"); features <- SelectIntegrationFeatures(object.list = adata.list); adata.anchors <- FindIntegrationAnchors(object.list = adata.list, , anchor.features = features);adata.combined <- IntegrateData(anchorset = adata.anchors); adata.combined <- ScaleData(adata.combined, verbose = FALSE)`
  - The latent space is the PCA-corrected latent space (comparable to our decipher z space).

Application of Decipher to AML patient data

Before applying Decipher to the AML patient data, we first performed a gene filtering step to include the most important genes representative of all cell types. For each patient, we performed PhenoGraph [20] clustering using 40 principal components and 30 neighbors. Then, using `scanpy`'s [44] *rank_genes_groups()* function, we performed a T-test to identify the top 400 most differentially expressed genes for each cluster. This list of genes was pooled with a list of known marker genes to produce the final set of genes on which the model was run. We also removed erythrocytes and lymphocytes from the data, as they were not relevant for the analysis of AML derailment.

We obtained:

- 3130 genes for the joint dataset AML1 and normal,
- 3264 genes for the joint dataset AML2 and normal,
- 3258 genes for the joint dataset AML3 and normal,
- 2863 genes for the joint dataset AML13 and normal,
- 2532 genes for the joint dataset AML14 and normal,
- 3291 genes for the joint dataset AML15 and normal,
- 2944 genes for the joint dataset AML16 and normal,
- 2664 genes for the joint dataset AML17 and normal.

Aside from these filtering steps, we emphasize that no other pre-processing or normalization was performed, as the model is always run on raw counts data. For our results **(Figure 2)**, we use latent factors z of dimension 10, Decipher components v of dimension 2, a neural network decoder from v to z with one hidden layer of dimension 64, a linear decoder from z to x, a neural network decoder from x to z with one hidden layer of dimension 128 and another neural network decoder from (z, x) to v with one hidden layer of dimension 128. BatchNorm was applied after each hidden layer in the neural networks, followed by a ReLU activation. We set $\beta = 0.1$ and a batch size of 64. The code to reproduce the results in this manuscript is available at https://github.com/azizilab/decipher_reproducibility.

Interpretation of Decipher components, latent dimensions, and basis

To identify pathways associated with the latent components of Decipher, we computed the covariance of each gene with each of the Decipher components, the latent dimensions, and the results from basis decomposition (**Supplementary Table 5**). Precisely, we computed for each gene g the covariance over cells between x_g and each variable v_1, v_2 and z_j for $j \in [10]$. We then ran Gene Set Enrichment Analysis (GSEA) [35], with genes preranked by covariance with each latent component. Next, to interpret the learned basis functions, we ranked genes by their weights in each basis to identify pathways most associated with each basis. For all use cases, GSEA was run on the pre-ranked setting against the Hallmarks Database, with 1000 permutations and no collapse (**Figure 2d, Supplementary Table 4**). To select genes for visualization in **Extended Data Fig. 5e**, we selected a top pathway for each component/basis function with known biological importance and found the top disrupted genes belonging to that pathway.

We highlight the usage of Decipher in reconstructing gene patterns over a temporal dimension. These analyses necessitated the translation of cell-level metadata to the temporal dimension. In order to analyze observations such as cell type, mutation proportions, etc., along the temporal dimension, we applied the projection method outlined in the trajectory inference section to obtain a cell-level Decipher time. This transformation allowed for observations to be directly studied along the temporal axis. For discrete observations such as cell type, we first performed nearest-neighbor smoothing using Scikit-learn 0.24.0 [28], with 50 neighbors and a radius of 0.2. A smoothed label was obtained for each cell by taking the mode of the labels among its neighbors. We then visualized the observations along a temporal axis by producing a scatterplot of cell observations, where the x-axis is the computed pseudotime of each cell and the color corresponds to the smoothed label (**Figure 3e,f**).

A key feature of Decipher is its ability to produce Decipher components that can be rotated to align with axes of disease maturation and development. We extended our analyses of the NPM1 mutational status of cells to examine its correlation with Decipher component 2 in AML1, 2, and 3. Specifically, we specified a cutoff threshold of 0.4 for the mutation proportion and classified cells with proportions greater than that as belonging to the mutated class and cells with proportions less than as being wild-type. We then binned cells by their pseudotime projections (with a bin size of 0.5 and a sliding window of 0.05). We then counted the number of cells classified into mutated and wild-type by our threshold in each bin and smoothed the resulting counts by time curves using a 1d Gaussian kernel (using `scipy` [40]) with a standard deviation of 2. We visualized the results as distributions along the Decipher component 2 axis to emphasize the shift in NPM1 mutational status (**Fig. 2a, Extended Data Fig. 4d,e**).

Distribution of TF peak expression over time

To study the patterns of TF expression over time, we directly utilized the gene patterns produced by Decipher that showed the expression values of each gene over the learned Decipher time axis. For each TF present in the data, expression patterns were first smoothed using a 1d Gaussian filter (using `scipy` [40]) with the standard deviation of the Gaussian kernel set to 3. This smoothing is performed only for detecting peak expression and is not applied to expression plots. Local maxima of expression were then identified by searching for points at which the first derivative of the curve switches from positive to negative. We furthermore filter points by using the midrange of expression (defined as the mean of the minimum and the maximum expression) as a threshold: local maxima whose expression values were less than this value. We additionally included the starting point as a maxima if the expression value was greater than the threshold and the first derivative was negative or if the maximum value of the expression was at the start; we included the ending point as a maxima if the maximum value of expression was at the end. For visualization, we plotted the kernel density estimation (with a bin width of 0.05) of all TF maxima along the Decipher time axis to show the points in time where overall TF activity is concentrated (**Fig. 3e,f; Extended Data Fig. 5c**). Kernel density estimation was performed using `seaborn` [42], and plotting was done using `matplotlib` [14].

Analysis of temporal TF co-regulation

In conjunction with the visualization of the timing of TF activity, we also sought to determine if families of similar TFs demonstrated coordinated activation times and if these temporal dynamics could be utilized to derive insight into regulatory wiring. We focused our analyses on the top 20 most disrupted TFs by the combined disruption metric, as well as the known disrupted TFs from the literature. The Decipher gene pattern for each TF was visualized as rows in a heatmap, with the horizontal axis representing the pseudotime axis and the color representing the z-scored expression value. The rows were sorted based on the time at which the maximum peak occurred, and the TFs were labeled with colors based on their biological function (**Fig. 4b,c; Extended Data Fig. 6a,b**).

Estimation of uncertainty in expression patterns

Because Decipher is a probabilistic model, it learns the uncertainty about the gene expression induced by a cell representation v. Given a location v in the Decipher space, the distribution of the expected gene expression $\mu_g(v)$ of gene g in a cell with representation v is given by,

$$\mu_g|v \sim h(z)_g|v$$

where $z|v \sim \mathcal{N}(f_{\text{mean}}(v), f_{\text{var}}(v))$.

To compute this uncertainty for each v, we sample 100 values for z from $z|v \sim \mathcal{N}(f_{\text{mean}}(v), f_{\text{var}}(v))$ and compute $\mu_g = h(z)_g$ for each of them. In **Fig. 3d,g**, the shaded bands represent the interquartile range (25-75) of the 100 samples.

Analysis of bulk AML data

We applied DeSeq2 [24] to obtain metrics characterizing the AML data at the cohort level. Using the resulting L2FC and p-values, we were able to confirm whether or not expected genes, as well as our newly identified disrupted genes, were also differentially detected in the bulk data. For both sets of genes, we selected genes whose absolute L2FC was greater than 1 and reported the minimum p-value.

Analysis of single-cell ATAC-seq

We utilized ArchR [11] for pre-processing of scATAC-seq data. In order to analyze accessibility trajectories from immature cell states to mature blast cells, we combined each patient AML sample with scATAC-seq data collected from a healthy bone marrow sample. ArchR was run using default parameters and the HG38 genome for peak alignment. UMAP visualization was run using ArchR's implementation with iterative LSI in order to provide a preliminary visualization of the distribution of cells in the ATAC space. The peak counts and gene scores were then exported and stored in an AnnData object for downstream use in `scanpy`. For our next steps, we extracted the merged gene scores across samples from the ArchR object.

Integration of scATAC and scRNA. To integrate the scATAC data with the scRNA data, we applied LIGER [43], using the gene scores from ArchR and the counts from the scRNA data as the two inputs. We applied a log2(x+1) transformation to the gene scores (where x=gene scores) to improve the signal. Running with standard parameters yielded an integrated space that could be visualized using UMAP (**Extended Data Fig. 8**) to confirm the overlapping of modalities and improved separation of cell types. From this step, we exported the matrix of normalized cell factors, H, that represents the LIGER shared factors across modalities that can be subsequently analyzed to compare cells from the scATAC data to cells in the scRNA data.

Projection of Accessibility scores in Decipher space. The first step of this analysis was to project cells in scATAC-seq to the Decipher space by finding the most similar cells in the RNA-seq to each cell in the ATAC-seq, with the goal of obtaining sharing metadata and Decipher values from the RNA-seq to the ATAC-seq. To achieve this, we applied a nearest neighbors algorithm using Scikit-learn [28] to accomplish the projection as follows: we first trained a neighbors classifier on the H matrix obtained from LIGER for RNA alone, H_{RNA}, (with the number of neighbors set to 3). Then, for each cell in scATAC-seq, we applied the classifier to the row corresponding row H_{ATAC}. The classifier subsequently returned the 3 nearest cells in RNA for each cell in ATAC-seq. We then compute the projections by taking the mean Decipher v component value for each cell, as well as the mean projected value onto the trajectory (Decipher time). The annotation for a given cell was simply taken to be the mode cell type of the neighbors.

Because this projection step may be a noisy process, we also implemented a filtering step wherein we computed the maximum pairwise Euclidean distance using `scipy` [40] of the Decipher space coordinates in each set of 3 nearest neighbors. The cells with distances greater than a selected threshold were removed. This step was particularly important for those Decipher embeddings that take on a curved or horseshoe shape, where a cell from ATAC-seq may be assigned to cells originating from two opposing regions. Finally, to obtain a better signal for visualizations, we again implemented a nearest-neighbors smoothing. We first built a nearest neighbors graph (k=50) on the ATAC peaks matrix. Then, the accessibility values for a given cell are obtained by taking the mean gene score for each gene across all neighbors.

Comparison of scRNA-seq and scATAC-seq in the Decipher space

For key transcription factors and marker genes, we sought to compare the patterns in their expression and their accessibility. For the expression plots, we combined the original unfiltered scRNA data for each patient's AML sample as well as the healthy bone marrow sample. We normalized these samples together by median library size and applied a natural log transformation. It was necessary to return to the unfiltered data for this task since the highly variable gene filter was applied before Decipher removed some genes of interest. We then subsetted the cells to match those that were run through Decipher previously and plotted the expression of the selected genes (from the unfiltered set) in the previously obtained Decipher space. `scanpy` plotting was utilized, with the colorbar range set to the .02-.09 percentiles of the data. For the scATAC-seq data, we created an AnnData object from the smoothed gene scores as described above and plotted using the projected Decipher coordinates. `scanpy` plotting was utilized, with the colorbar range set to .02-.98 percentiles in AML1; .05-99.5 percentiles of the data in AML2; and .05-98.5 in AML3. (**Fig. 4d, Extended Data Fig. 7, Extended Data Fig. 8d**).

Motif Analysis in scATAC-seq. To obtain the set of TFs for motif analysis, we examined the time of maximum expression in each of the top 50 disrupted TFs (according to the combined disruption metric). If the time of maximum expression was before the estimated time of NPM1 mutation (e.g., at pseudotime 50 for AML1), we designated the TF to be "early peaking." We then investigated whether the early peaking TFs could be putative regulators of other disrupted TFs, explaining the propagation of dysregulated mechanisms leading to the altered transcriptional landscape (**Fig. 4d, Extended Data Fig. 6a,b**). Potential targets were thus defined as the set of remaining TFs (peaking after NPM1 mutation). Motif searching was performed using motifmatchR [33], with the JASPAR2020 [31] database of motifs and the Hg38 genome.

For each patient, we identified clusters obtained from the ArchR Iterative LSI analysis that corresponded to either blast 0 (AML1,3) or blast 1 (AML2). The ArchR reproducible peak set for that cluster, containing pseudo-bulked peak data, was used as the input to motifmatchR. To transform the result into a matrix of motifs by targets, we grouped motif scores by the target's nearest gene and calculated the mean score for all scores belonging to the same gene.

To evaluate the significance of motif enrichment, we defined a background (null) set of TFs as follows: for each patient scRNA-seq data, we filtered out genes expressed in fewer than 1 percent of cells. This threshold was selected based on the proportion of PROM1-expressing cells in the patient samples, e.g., 125 out of 13834 cells in AML3, since we expected PROM1 to be a good example of a gene expressed in the rare immature cells of interest.

We then limited the null set to TFs that were not among the target set. P-values were computed by bootstrapping: we sampled the null set 10,000 times for a subset of TFs that were the same size as the target set. We then computed the mean motif score for all TFs in the null set to create the null distribution. The p-value was obtained by counting the number of samples in the null set whose score was greater than the mean motif score of our target set of TFs. We reported the p-values for the TFs that were significantly enriched in either all three patients (in which case the maximum p-value was reported) (**Fig. 4d**) or significantly enriched in two out of three patients (in which case the maximum among the two significant p-values was reported) (**Extended Data Fig. 7**).

In all three patients, we found significant motif enrichment in HOX TFs, including HOXB4 (motif JASPAR ID MA1499.1, $p < 0.0458$) and HOXB6 (MA1500.1, $p < 0.032$) and FOSL1 (MA0477.1, $p < 0.0057$; MA0477.2 0.0086), suggesting their putative role in regulating TFs driving AML transformation (**Fig. 4d**).

In two out of three patients, we found significant motif enrichments in GATA2 (MA0036.1, $p < 0.0115$; MA0036.2, $p < 0.0423$; MA0036.3, $p < 0.0309$), MYC (MA0147.3, $p < 0.045$), NR4A2 (MA0160.1, $p < 0.0272$), MAFF (MA0495.1, $p < 0.0365$), ATF3 (MA0605.2, $p < 0.0316$), NFKB2 (MA0778.1, $p < 0.0292$), RELB (MA1117.1, $p < 0.0109$), KLF2 (MA1515.1, $p < 0.0468$), ZNF274 (MA1592.1, $p < 0.009$) (**Extended Data Fig. 7**).

Application to gastric cancer evolution. We applied Decipher on the gastric cancer data from Kim *et al.* [17]. We pooled the data from the 24 patients in the study, each with pre-malignant cells and cancerous cells. 9 of these patients have *intestinal* cancer and the other *15* patients have *diffuse-like* cancer. The resulting data has 12,612 cells and 8,705 genes. We ran Decipher with its default hyperparameters: z of dimension 10, v of dimension 2, and for 30 epochs **(Fig. 6b-d)**.

In sum, Decipher on the gastric data was applied on

- 12614 cells,
- 8705 genes,

- from 24 patients,
- with 2 major types of cells (pre-malignant, cancerous)
- with 2 types of cancer (diffuse-like or intestinal),
- with different stages of cancer.

Application of Monocle. H5 files of gastric data were converted to a Seurat object with the `convertFormat` function from `sceasy` package. The Seurat object was converted to a cds object to run monocle with `as.cell_data_set()` function from `SeuratWrappers` package.

Gene names were assigned to the rowData of cds object, and `estimate_size_factor()` were performed to calculate the size factor. Then, the standard monocle analysis pipeline can be applied. 10 dimensions were used when running the `preprocess_cds()` function. Dimension reduction was performed with `reduce_dimension(cds, umap.min_dist = 1, umap.n_neighbors = 10L)`. Trajectory was learned with `learn_graph()` function. Lastly, the inferred pseudotime was plotted with `plot_cells()` function **(Fig. 6a)**.

References

1. Alonso-Curbelo, D., Ho, Y.-J., Burdziak, C., Maag, J. L., Morris IV, J. P., Chandwani, R., Chen, H.-A., Tsanov, K. M., Barriga, F. M., Luan, W., *et al.* A gene–environment-induced epigenetic program initiates tumorigenesis. *Nature* (2021).
2. Azizi, E., Carr, A. J., Plitas, G., Cornish, A. E., Konopacki, C., Prabhakaran, S., Nainys, J., Wu, K., Kiseliovas, V., Setty, M., *et al.* Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* (2018).
3. Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P. & Goodman, N. D. Pyro: deep universal probabilistic programming. *The Journal of Machine Learning Research* (2019).
4. Blei, D. M. Build, compute, critique, repeat: data analysis with latent variable models. *Annual Review of Statistics and Its Application* (2014).
5. Blei, D., Kucukelbir, A. & McAuliffe, J. D. Variational inference: a review for statisticians. *Journal of the American Statistical Association* (2017).
6. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research* (2003).
7. Cacchiarelli, D., Qiu, X., Srivatsan, S., Manfredi, A., Ziller, M., Overbey, E., Grimaldi, A., Grimsby, J., Pokharel, P., Livak, K. J., *et al.* Aligning single-cell developmental and reprogramming trajectories identifies molecular determinants of myogenic reprogramming outcome. *Cell Systems* (2018).
8. Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M. & Li, H. Twelve years of SAMtools and BCFtools. *GigaScience.* giab008 (2021).
9. Gayoso, A., Lopez, R., Xing, G., Boyeau, P., Valiollah Pour Amiri, V., Hong, J., Wu, K., Jayasuriya, M., Mehlman, E., Langevin, M., Liu, Y., Samaran, J., Misrachi, G., Nazaret, A., Clivio, O., *et al.* A python library for probabilistic analysis of single-cell omics data. *Nature biotechnology* (2022).
10. Gayoso, A., Shor, J., Carr, A. J., Sharma, R. & Pe'er, D. Doubletdetection (version v2. 4). *Zenodo, DOI* (2018).
11. Granja, J. M., Corces, M. R., Pierce, S. E., Bagdatli, S. T., Choudhry, H., Chang, H. Y. & Greenleaf, W. J. Archr is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature genetics* (2021).
12. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S. & Lerchner, A. *Beta-vae: learning basic visual concepts with a constrained variational framework* in *International conference on learning representations* (2016).
13. Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research* (2013).
14. Hunter, J. D. Matplotlib: a 2d graphics environment. *Computing in Science & Engineering* (2007).
15. Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. An introduction to variational methods for graphical models. *Machine Learning* (1999).
16. Kanton, S., Boyle, M. J., He, Z., Santel, M., Weigert, A., Sanchís-Calleja, F., Guijarro, P., Sidow, L., Fleck, J. S., Han, D., *et al.* Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* (2019).
17. Kim, J., Park, C., Kim, K. H., Kim, E. H., Kim, H., Woo, J. K., Seong, J. K., Nam, K. T., Lee, Y. C. & Cho, S. Y. Single-cell analysis of gastric pre-cancerous and cancer lesions reveals cell lineage diversity and intratumoral heterogeneity. *NPJ Precision Oncology* (2022).
18. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
19. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
20. Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., ad D. Amir, E., Tadmor, M. D., Litvin, O., Fienberg, H. G., Jager, A., Zunder, E. R., Finck, R., Gedman, A. L., Radtke, I., Downing, J. R., Pe'er, D. & Nolan, G. P. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* (2015).

21. Levitin, H. M., Yuan, J., Cheng, Y. L., Ruiz, F. J., Bush, E. C., Bruce, J. N., Canoll, P., Iavarone, A., Lasorella, A., Blei, D. M., *et al.* De novo gene signature identification from single-cell rna-seq with hierarchical poisson factorization. *Molecular systems biology* (2019).

22. Lopez, R., Nazaret, A., Langevin, M., Samaran, J., Regier, J., Jordan, M. I. & Yosef, N. A joint model of unpaired data from scrna-seq and spatial transcriptomics for imputing missing gene expression measurements. *arXiv preprint arXiv:1905.02269* (2019).

23. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods* (2018).

24. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology* (2014).

25. McDowell, I. C., Manandhar, D., Vockley, C. M., Schmid, A. K., Reddy, T. E. & Engelhardt, B. E. Clustering gene expression time series data using an infinite gaussian process mixture model. *PLoS computational biology* (2018).

26. Neal, R. M. in *Bayesian Learning for Neural Networks* (1996).

27. Novershtern, N., Subramanian, A., Lawton, L. N., Mak, R. H., Haining, W. N., McConkey, M. E., Habib, N., Yosef, N., Chang, C. Y., Shay, T., *et al.* Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* (2011).

28. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* (2011).

29. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications* (2018).

30. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. Integrative genomics viewer. *Nature biotechnology* (2011).

31. Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. & Lenhard, B. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research* (2004).

32. Satpathy, A. T., Granja, J. M., Yost, K. E., Qi, Y., Meschi, F., McDermott, G. P., Olsen, B. N., Mumbach, M. R., Pierce, S. E., Corces, M. R., *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral t cell exhaustion. *Nature biotechnology* (2019).

33. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. Chromvar: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature methods* (2017).

34. Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., Choi, K., Bendall, S., Friedman, N. & Pe'er, D. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature biotechnology* (2016).

35. Shi, J. & Walker, M. G. Gene set enrichment analysis (gsea) for interpreting gene expression profiles. *Current Bioinformatics* (2007).

36. Svensson, V., Gayoso, A., Yosef, N. & Pachter, L. Interpretable factor models of single-cell rna-seq via variational autoencoders. *Bioinformatics* (2020).

37. Traag, V. A., Waltman, L. & Van Eck, N. J. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports* (2019).

38. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S. & Rinn, J. L. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* (2014).

39. Van den Berge, K., Roux de Bézieux, H., Street, K., Saelens, W., Cannoodt, R., Saeys, Y., Dudoit, S. & Clement, L. Trajectory-based differential expression analysis for single-cell sequencing data. *Nature communications* (2020).

40. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., *et al.* Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods* (2020).

41. Wainwright, M. J. & Jordan, M. I. *Graphical Models, Exponential Families, and Variational Inference* (2008).

42. Waskom, M. L. Seaborn: statistical data visualization. *Journal of Open Source Software* (2021).

43. Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C. & Macosko, E. Z. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* (2019).

44. Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology* (2018).

45. Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature communications* (2017).