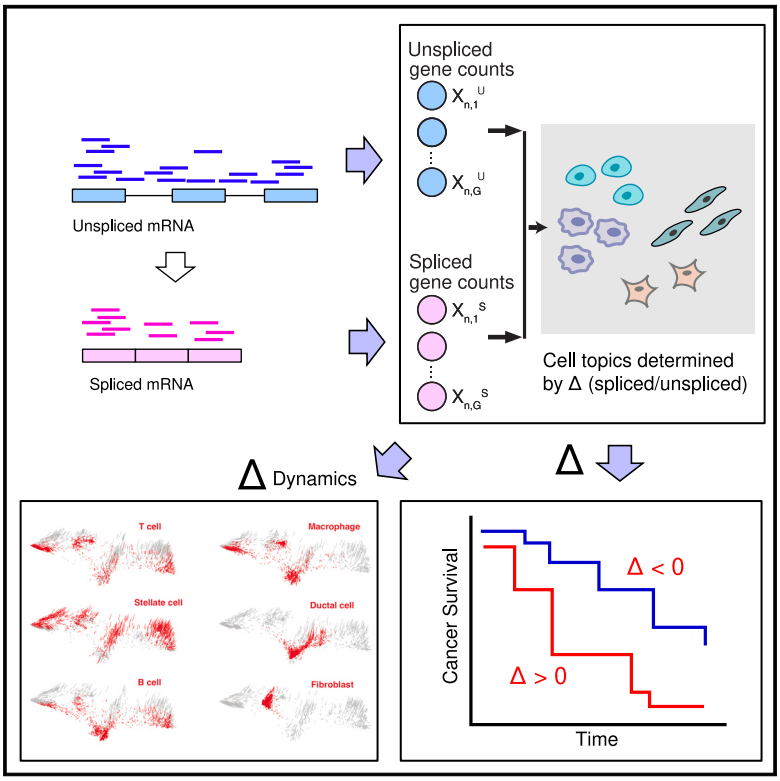
Cell Genomics

Article

Unraveling dynamically encoded latent transcriptomic patterns in pancreatic cancer cells by topic modeling

Graphical abstract



Authors

Yichen Zhang,
Mohammadali (Sam) Khalilitousi,
Yongjin P. Park

Correspondence

ypp@stat.ubc.ca

In brief

Zhang et al. present a new probabilistic model of cell topics that capture RNA-splicing dynamic patterns in single-cell sequencing data. Assigning 32 dynamic topics to 227,000 cells, they discovered four unique gene programs significantly correlated with cancer survival time.

Highlights

- A statistical framework to investigate short-term transcriptional dynamics
- A reusable machine-learning software to estimate interpretable cellular topics
- Gene programs pertinent to disease but independent of static cell states

Zhang et al., 2023, Cell Genomics 3, 100388
September 13, 2023
https://doi.org/10.1016/j.xgen.2023.100388

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Article

Unraveling dynamically encoded latent transcriptomic patterns in pancreatic cancer cells by topic modeling

Yichen Zhang,[1] Mohammadali (Sam) Khalilitousi,[2] and Yongjin P. Park[1,3,4,5,*]
[1]Department of Statistics, The University of British Columbia, Vancouver, BC, Canada
[2]Department of Biomedical Engineering, The University of British Columbia, Vancouver, BC, Canada
[3]Department of Pathology and Laboratory Medicine, The University of British Columbia, Vancouver, BC, Canada
[4]Department of Molecular Oncology, BC Cancer Research, Part of Provincial Health Care Authority, Vancouver, BC, Canada
[5]Lead contact
*Correspondence: ypp@stat.ubc.ca
https://doi.org/10.1016/j.xgen.2023.100388

SUMMARY

Building a comprehensive topic model has become an important research tool in single-cell genomics. With a topic model, we can decompose and ascertain distinctive cell topics shared across multiple cells, and the gene programs implicated by each topic can later serve as a predictive model in translational studies. Here, we present a Bayesian topic model that can uncover short-term RNA velocity patterns from a plethora of spliced and unspliced single-cell RNA-sequencing (RNA-seq) counts. We showed that modeling both types of RNA counts can improve robustness in statistical estimation and can reveal new aspects of dynamic changes that can be missed in static analysis. We showcase that our modeling framework can be used to identify statistically significant dynamic gene programs in pancreatic cancer data. Our results discovered that seven dynamic gene programs (topics) are highly correlated with cancer prognosis and generally enrich immune cell types and pathways.

INTRODUCTION

Single-cell RNA-sequencing (RNA-seq) technology has been successfully applied to profile regulatory genomic changes in studying many human disease mechanisms. Our capability to measure single-cell-level mRNA molecules has dramatically changed our research paradigm in genomics and translational medicine. A typical single-cell study implicitly assumes observed transcript levels as a static value, considering that every cell is fixed at a particular state. Recently, researchers have developed a complementary method to measure gene expression dynamics (the speed of splicing) by measuring the divergence of the spliced counts from the unspliced in single-cell RNA-seq (scRNA-seq) profiles.[1] More precisely, having the two types of mRNA counts, we can solve for ordinary differential equations of transcriptional dynamics and estimate the splicing and mRNA decay rate parameters. Several methods have extended the original method pioneered by La Manno and co-workers. Notably, the scVelo method generalizes to recover gene-level ordinary differential equation (ODE) models, allowing each gene to take independent timescales.[2]

Why is it difficult to estimate full-scale dynamics in datasets with limited snapshots?

However, probabilistic inference of full-scale dynamics often poses a substantial challenge, and the inferred rate parameters may greatly vary depending on the normalization and embedding methods.[3] Although a newly developed machine-learning method based on a mixture of ODE models improved the robustness and accuracy in single-cell data profiled in developmental processes,[4] existing velocity analysis methods rely on a critical assumption unmet by most single-cell datasets at a study design level. Most single-cell datasets, especially those collected from patient-derived cancer samples, only span several snapshots of full developmental, evolutionary, or disease progression processes. In human case-control studies, cells may not have reached steady states in the disease progression process and are likely to fail to provide enough information for most genes and pathways. Such discontinuity and sparsity in data collection somewhat force statistical inference algorithms to rely on an unrealistic steady-state assumption and on interpolated data points with high uncertainty.[3,5]

Why do we need a topic model for transcription dynamics?

Nevertheless, gene expression dynamics implicated by the transcript-level difference between the spliced and unspliced counts provide a valuable perspective in single-cell data analysis, making single-cell analysis more valuable beyond conventional static analysis. To overcome the limitations posed by incomplete temporal trajectories and poor quality of single-cell sequencing assays in scRNA velocity analysis, we propose a new modeling

framework, DeltaTopic, short for dynamically encoded latent transcriptomic pattern analysis by topic modeling. DeltaTopic combines two ideas: (1) latent topic analysis that will guide unsupervised machine learning for discovering new dynamic cell states and (2) application of first-order approximation to learn robust relationships between the spliced and unspliced counts instead of estimating a full trajectory of ODE models. For a latent topic model, we view each cell as a document and each gene as a word to make model parameters directly interpretable while keeping the Bayesian model's capability to impute missing information. The simplified dynamic model also permits an intuitive interpretation of spliced-unspliced differences as multiplicative "delta" parameters in the model.

We developed and applied our DeltaTopic approach to single-cell datasets on pancreatic ductal adenocarcinoma (PDAC), one of the most challenging cancer types with a poor prognosis. In the latent space, our model identified cancer-survival-specific topics marked by a unique set of gene expression dynamics. We also found that DeltaTopic further dissected subtopics clumped together in traditional clustering methods, implicating novel gene modules and cell states that are dynamically controlled along with the cancer progressions. With synthetic datasets, we demonstrate the effectiveness of DeltaTopic and Bayesian latent topic analysis with sparse association matrix (BALSAM) in cell label prediction and gene activity identification. Both methods significantly outperformed conventional principal-component analysis (PCA), with DeltaTopic showing particular strength in recovering both static and dynamic gene activities.

RESULTS

Single-cell transcriptomic dynamics in PDAC studies

We preprocessed single-cell expression datasets available in two large-scale, multi-individual studies.[6,7] We extracted both spliced and unspliced count vectors from the original short-read sequencing files for each cell using Kallisto[8] and UMI-BUS[9] tools. We consolidated all the samples/batches into one file set using our customized utility functions rcpp_mmutil_merge_file_sets, available in the mmutilR library (https://causalpathlab.github.io/mmutilR). Applying cell-level quality control steps, which filtered out cells with too few counts and with high mitochondrial gene expression activities, we retained 227,331 cells (91.13%), discarding 22,126 cells (8.87%) out of 249,457 cells by applying these quality control steps. The quantification algorithm results in two types of gene expression vectors for each cell—one for the spliced and the other for the unspliced counts. We measured 22,836 features (the spliced and unspliced genes) on a total of 227,331 cells, and only 329,824,833 elements were non-zero (6.35%).

Overall, we have two types of high-dimensional sparse matrices as input data on total G = 11,418 genes across N = 227,311 cells: (1) $X^{(U)}_{N\times G}$ for the unspliced counts and (2) $X^{(S)}_{N\times G}$ for the spliced transcript counts. Our goal is to identify latent factors/topics and the corresponding topic-specific gene expression frequency parameters—one for the unspliced and the other for the spliced components. No additional preprocessing steps, such as gene selection, batch adjustment, selection of principal components, or data transformation, were necessary for topic modeling since the underlying multinomial likelihood is less affected by potential effects of unwanted stochasticity than other probability models.[10,11]

Overview of our approach

A Bayesian approach to identify sparse cell topics in scRNA-seq data (BALSAM)

We developed a Bayesian topic modeling approach, extending the embedded topic model framework[12] with elementwise spike-and-slab prior probability.[13] Our BALSAM approach views cells as an admixture of gene topics to summarize static transcriptome patterns from raw gene expression count data (Figure 1A). BALSAM relies on variational autoencoders (VAEs) to learn cell topics and infer the cell-topic relationship. The encoder transforms the expression space into a latent topic space through a stack of non-linear layers (NN1), outputting a vector of relative topic proportions for each cell. The decoder generates a dictionary matrix β from a sparse-inducing prior called spike-slab to ensure that only a small number of genes are selected for each topic. The resulting dictionary matrix β is passed to a generalized linear model (GLM) along with topic proportion θ to estimate normalized gene frequency λ. Using λ as the parameter, we compute the likelihood of the expected gene count from a multinomial distribution. We provide detailed descriptions of the BALSAM, sparse priors, and variational inference algorithms in the STAR Methods.

DeltaTopic

In designing our DeltaTopic approach, we were inspired by common patterns that we repeatedly observed in gene-level data (Figure 1B).

- Sparsity due to technology and intrinsic difference: in early Drop-seq technology, dropout events may assign a substantial fraction of gene expression counts to a zero value, obfuscating our delineation between the undetectable and unexpressed genes. A small number of mRNA molecules per cell often result in gene expression profiles, including a large fraction of zero values and a high-dimensional vector with only a small number of non-zero elements, and also substantially deviate from conventional deep sequencing results. In the context of transcription dynamics, we need to handle zero values on both sides—the spliced and unspliced—further necessitating a model that can handle sparsity patterns without overfitting. For instance, the spliced counts of the *LDLRAD4* gene are essentially zero in many cells, even while the unspliced counts are positive (Figure 1B, left). Similarly, an excess of zero values in the unspliced were observed in *B2M* (Figure 1B, middle); for the *TC2N* case, both sides contain many zero values (Figure 1B, right). Although we believe that a substantial fraction of zero values reflect the underlying dynamics, it is impossible to completely rule out all the possibility of statistical biases due to technical factors.
- Sampling bias in a temporal axis: in contrast to what an analytical ODE solution predicts, we tend to have only a limited portion of the whole phase diagram of RNA velocity.[1] Parametric inference for an ODE model would remain

CellPress
OPEN ACCESS

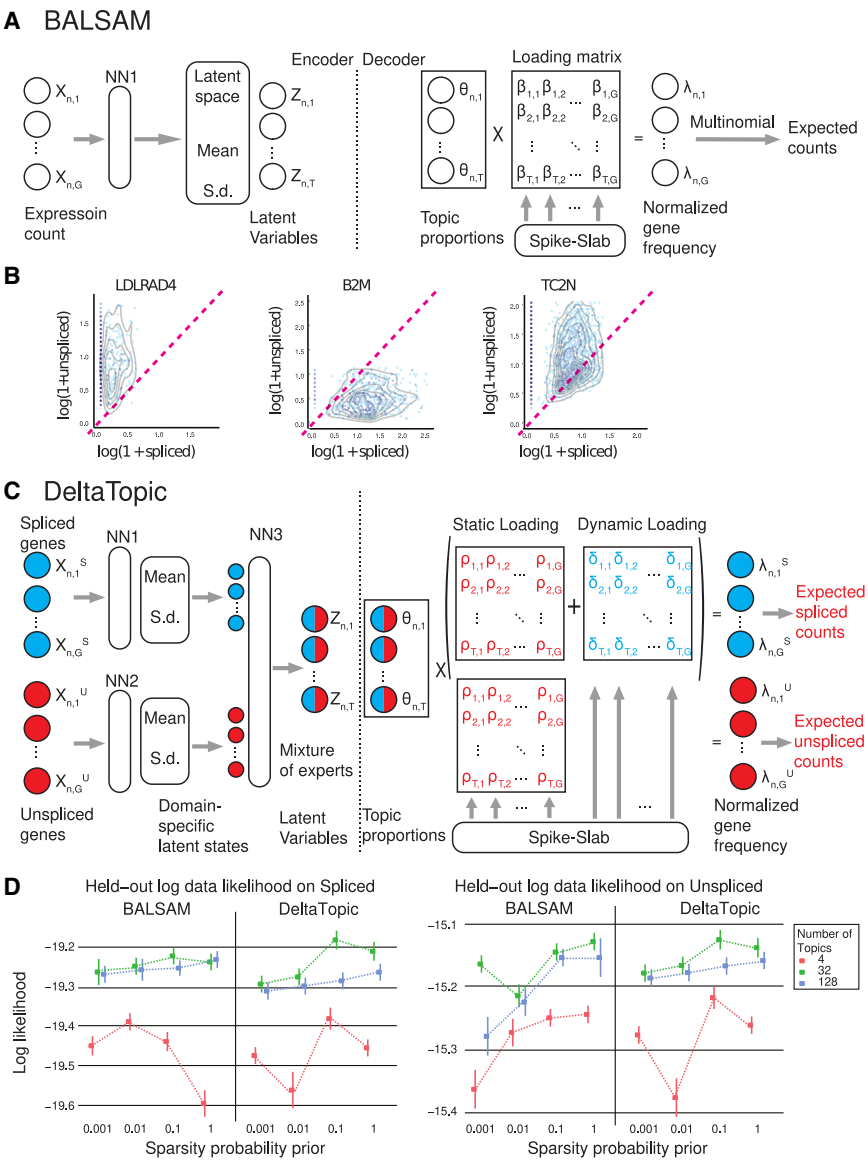


Figure 1. Modeling single-cell transcription dynamics with sparse probabilistic topic models

(A) BALSAM: given a raw gene expression count matrix, BALSAM learns cell topics to represent cell types or cell states using neural networks. The encoder transforms the expression space into a latent topic space through a stack of non-linear layers (NN1). The decoder (data-generative components) models single-cell data vectors as a probabilistic topic model (Dirichlet-multinomial). The Dirichlet parameters are modeled as a generalized linear model (with log link functions) as a linear combination of cell-topic-specific sparse factors ρ weighted by topic proportions θ.

(B) Here, representative examples of gene expression dynamics in the PDAC data are shown as a scatterplot of the spliced and unspliced counts. The x axis: the spliced gene count (log1p transformed); the y axis: the unspliced gene counts (log1p transformed). The red dashed line indicates where the spliced and unspliced genes are of the same amount (not a steady state).

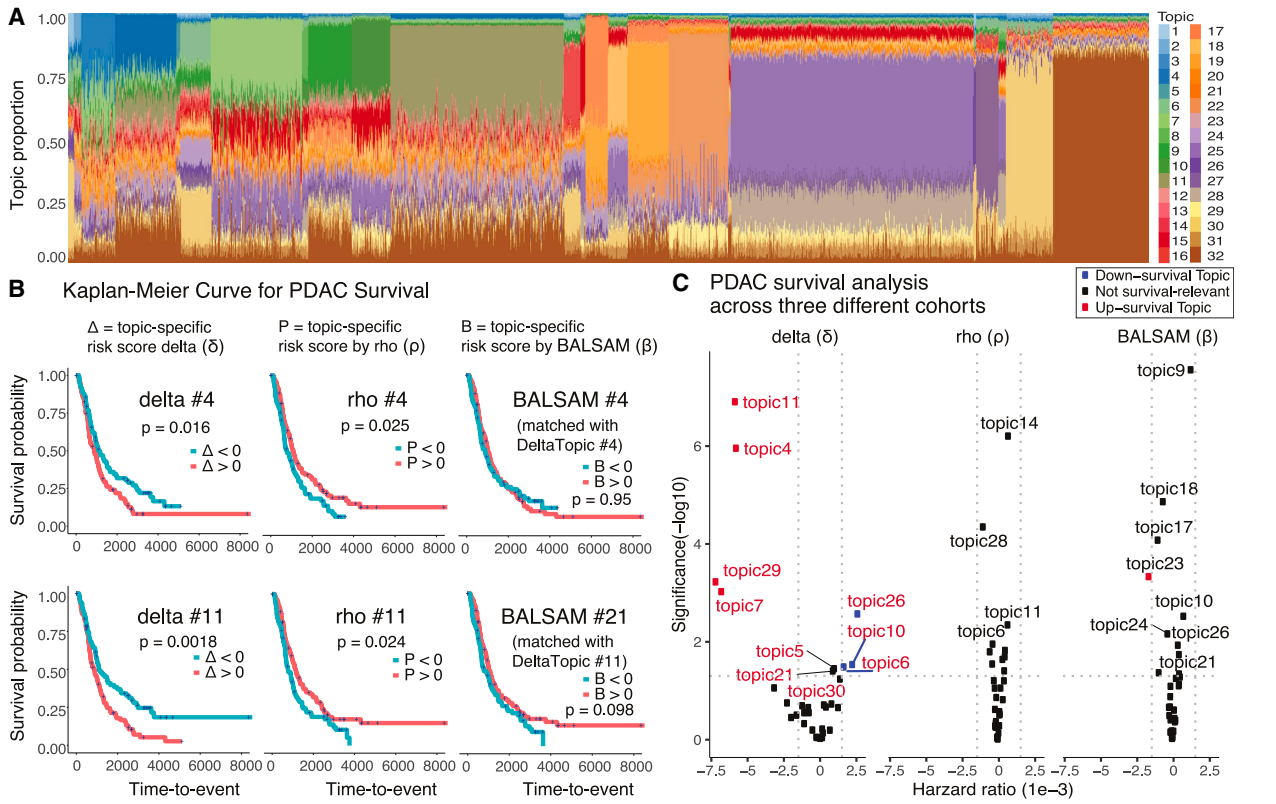
(C) DeltaTopic: given the spliced and unspliced gene expression count matrices, DeltaTopic's encoder layers embed a pair of the spliced and unspliced count vectors to latent space (NN1, NN2) and combine the information to form a shared latent space through a fusion layer (NN3). The decoder generates sparse gene factors—one for the static and the other for the dynamic ones—and constructs two gene-by-topic matrices, each corresponding to the spliced and the unspliced counts. The static topic matrix ρ sets a background level for the spliced and unspliced gene expressions. As for the spliced expression counts, the dynamic topic loading matrix is added to the static loading matrix to account for the divergence between the spliced and unspliced counts.

(D) Model evaluation on held-out data likelihood (spliced and unspliced). The y axis: the average held-out data likelihood and 95% confidence interval; the x axis: sparsity probability prior.

unidentifiable, suggesting multiple similarly likely solutions without a strong steady-state assumption. Again, as can be seen in the exemplary genes (Figure 1B), genes were differentially regulated in the same dataset, potentially suggesting the existence of multiple modes in transcription dynamics.

At least in our cancer study, it was difficult to identify characteristic phase diagrams[1] in all genes; hence, no cells in the given data have reached a steady state. Therefore, we focused on modeling short-term directional information implicated by RNA velocity in an intuitive model, regressing the spliced on the unspliced count data gene by gene and topic by topic. The DeltaTopic model extends the BALSAM model to ascertain common cellular topic space and topic-specific relationships between the unspliced and spliced data (Figure 1C). Since a GLM framework provides a flexible way to capture relationships across many data modalities, our approach can be easily extended to other similar tasks.

Given a pair of spliced and unspliced gene expression count matrices as input, DeltaTopic transforms each into a latent space through two independent BALSAM encoders (NN1 and NN2). The latent variables with encoded information from the two encoders were combined by taking their average[14,15] to obtain a shared topic vector as a mixture of experts from the spliced and unspliced latent space (Figure 1C, fusion layer). The decoder then generates two dictionary matrices from sparsity-inducing spike-slab priors to decompose spliced and unspliced transcription patterns into static and dynamic topics. The static topic dictionary matrix ρ sets a background level for both spliced and unspliced genes. The dynamic topic



(legend on next page)

dictionary matrix δ for spliced genes determines the directionality—activated vs. inhibited—compared with the background dictionary matrix ρ at a topic level. The decoder model generates the normalized frequency for the spliced and unspliced counts as a linear combination of multiple dynamic/static topics weighted by cell-level topic proportions. We implemented the model in PyTorch and performed posterior inference using an NVIDIA RTX 3080 GPU.

Training deep-learning models and generalization performance

To evaluate each model's generalization performance, we tested the likelihood of data that were held out during training for BALSAM and DeltaTopic. We randomly split the cells by a 9:1 ratio to form training and test sets. For BALSAM, we trained two independent models, one with the spliced gene count as input and the other with the unspliced gene count. We validated the held-out log data likelihood in their corresponding domain (e.g., spliced or unspliced). For DeltaTopic, we trained a unified model with the pairs of spliced and unspliced counts as input and validated the held-out log likelihood in both domains. To study the effects of the latent dimension and the sparsity probability prior on the model generalizability, we trained BALSAM and DeltaTopic with different numbers of latent topics (4, 32, 128) and set sparsity probability priors to $\pi = 0.001, 0.01, 0.1,$ and 1. Each training was repeated five times with different random seeds.

We found that DeltaTopic with 32 topics and sparsity probability prior $\pi = 0.1$ is overall the best-performing model compared with the other choices of latent dimension and sparsity level (Figure 1D), while for BALSAM, the best-performing model is the one with 32 topics and sparsity probability prior $\pi = 0.01$ in the spliced domain and $\pi = 1$ in the unspliced domain. With this choice of hyperparameters, DeltaTopic yields better hold-out log likelihood scores than BALSAM in both types—the spliced and unspliced data—suggesting that modeling the relationships between two related data types can improve robustness. For both models, finding the right model complexity (the number of topics) was necessary. Although a more fine-grained hyperparameter search, empowered by a Bayesian optimization method, is desired, our results suggest that underfitting with four topics and overfitting with 128 topics tend to yield inferior generalization performance. A wider error bar is observed in the underfitting model (with four topics), suggesting that the model is more sensitive to the choice of initialization. In general, Bayesian sparsity priors avoid overfitting issues, as we observed that a model with either 32 or 128 topics performs similarly regarding the hold-out log likelihood scores in both data types. Indeed, using a high number of topics and with stringent sparsity hyperparameters enforced, our Bayesian framework can automatically determine the right balance between bias and variance, shutting off unnecessary topics and leading to a slightly improved generalization performance.

DeltaTopic provides novel insights into pancreatic cancer etiology

Having an optimal configuration of hyperparameters established, we trained our topic models on the full dataset and resolved Δ (dynamic) and ρ (static) topics with the topic proportions for each cell θ (Figure 2A). We then asked whether the genes/features discovered by DeltaTopic are novel and different from the genes found by the topic models that were trained on either the spliced or unspliced counts alone. In order to answer the question, we linked the delta topics to the counterparts found by BALSAM. Using the cell-level topic proportion estimates, namely θ_i, per each model, almost all the topics found in the DeltaTopic were connected to the ones found in BALSAM (Figure S1). For instance, delta topic 4 and BALSAM topic 4, delta topic 10 and BALSAM topic 8, and delta topic 11 and BALSAM topic 21 are highly overlapping in their membership. However, topics with less than 2% of the total population cells were not included for further investigation for brevity. DeltaTopic marginally improves the resolution in cell clustering, providing a finer view of transcriptome dynamics. Some topics found by BALSAM can be portioned into two or more delta topics. For instance, BALSAM 32 can be dissected into two delta topics, 3 and 7; likewise, BALSAM 27 can be divided into three delta topics: 6, 14, and 30 (Figure S1).

To better understand how the directional information found by the parameter works differently from the static ones, such as ρ_t (the static component of DTM) and β_t (the parameters of BALSAM), we investigate therapeutic impacts of the three different gene sets. We estimated the topic-specific risk scores for pancreatic cancer samples available in the ICGC Data Portal and correlated the estimated risk scores with the survival outcomes of each individual. Bulk gene expression datasets for three independent pancreatic cancer studies are publicly available with donor-level information in the ICGC Data Portal (https://dcc.icgc.org/releases/current/Projects), including pancreatic cancer samples in the Canadian cohort (N = 234 donors on unique 53,800 transcripts), the Australian cohort (N = 91 donors on unique 42,346 transcripts), and the US cohort (N = 142 donors on unique 20,009 genes). We computed donor-level scores by taking average gene expression values weighted by topic-specific gene frequency vectors (vectors in the dictionary matrices): $\Delta_{it} = \sum_{g=1}^{G} Y_{ig} E[\delta_{tg}]$. Similarly, we can estimate the other two types of topic-specific scores by $P_{it} = \sum_{g=1}^{G} Y_{ig} E[\rho_{tg}]$ and $B_{it} = \sum_{g=1}^{G} Y_{ig} E[\beta_{tg}]$.

Figure 2. DeltaTopic approach identifies disease-relevant cell topics, implicating putative causative regulatory programs

(A) DeltaTopic model estimates topic proportions across 227,331 cells in the PDAC data.

(B) Kaplan-Meier survival curves for 234 donors in ICGC data differentially correlated with the positive and negative topic-specific risk scores implicated by DeltaTopic gene factors. The p values are computed by log-rank test comparing positive and negative risk groups in survival probability.

(C) A volcano plot summarizes the hazard ratios and p values testing the associations between topic-specific risk scores (derived from different topic models) and observed survival times across donors in three different cancer cohorts (ICGC-PDAC-US, ICGC-PDAC-CA, and ICGC-PDAC-AU; see the text). Each point represents an aggregated hazard ratio measure and a p value in the meta-analysis. The x axis: the hazard ratio estimate from the Cox proportional hazard model; the y axis: p values in negative log10 scale. Survival-relevant cell topics are colored red and blue for up- and down-regulation with respect to the PDAC survival time. The two vertical dashed lines correspond hazard ratio cutoff at $\pm 1.5 \times 10^{-3}$. The horizontal line marks the p value cutoff at 0.05.

We correlated these topic- and model-specific estimated risk scores with survival outcomes (time to event) using a regularized Cox proportional hazard regression model (Figure 2B). To account for cohort-specific bias in the survival data, we conducted meta-analysis, aggregating the hazard ratio statistics independently estimated in each cohort by an inverse variance-weighted (IVW) average approach. The IVW approach prioritizes topics found significant consistently across three cohorts while penalizing statistically significant topics only in one or two cohorts, denoting ψ_t^i be the hazard ratio estimate for topic t and cohort i and $\mathrm{se}(\widehat{\psi}^i)$ to be the standard error. For each topic t, we can obtain a summary hazard ratio estimate ψ_t^{IVW} by aggregating cohort-specific hazard ratio ψ_t^i: $\psi_t^{\mathrm{IVW}} = \sum_i W_t^i \widehat{\psi}_t^i / \sqrt{\sum_i W_t^i}$, where $W_t^i = \mathrm{se}(\widehat{\psi}_t^i)^{-2}$, and we computed the p value by $\mathrm{p}_t^{\mathrm{IVW}} = 2\Phi(-\left|\varphi_t^{\mathrm{IVW}}\right|)$.

We plotted each topic's hazard ratio estimates and p values. Topics with p values smaller than 0.05 and absolute hazard ratio estimates greater than 1.5×10^{-3} are interpreted as "survival relevant," colored in red and blue for up-regulated and down-regulated topics, respectively (Figure 2B). 7 survival-relevant topics were identified using δ_t, whereas none and only one topic were identified using ρ_t and β_t, respectively. Among seven survival-relevant delta topics, topics 11, 4, 29, and 7 were up-regulated survival topics correlated inversely with hazard ratio estimates. Topics 26, 6, and 19 were down-regulated survival topics.

Not all topics are directly comparable across different models. For some of them, we were able to compare those topics across different models if they are paired (Figure S1). Of those matched, we found many cases where the scores derived from DeltaTopic are significantly associated with the survival outcome, whereas the scores derived from the unimodal topic model are not strongly associated (Figures 2B and 2C). For instance, topic 4 in DeltaTopic and topic 4 in BALSAM predict the disease prognosis quite differently. Two patient groups stratified by DeltaTopic (Δ) 4 follow significantly different disease prognoses (p = 0.016). Similarly, the patient group with high topic-specific scores for delta topic 11 (greater than the median) tends to show shorter survival times than the other group with the low scores (p = 0.0018).

For the seven delta topics significantly associated with the survival outcome, topics 11, 4, 29, 26, 10, 7, and 6 (Figure 1C), we further investigated their top genes/features that define the characteristics of their topics, meaning strong $\delta_{tg} \vee$ values. Interestingly, these anchor genes show highly topic-specific activities (Figure 2D, left). However, the same set of genes showed weaker topic-specific patterns (Figure 2D, middle) and in the BALSAM model (Figure 2D, right). Our results suggest that our DeltaTopic approach can distinguish genes that are differentially regulated in transcriptomic dynamics from those differentially expressed at the static level and that they can play a pivotal role in cancer progression and metastasis.

Interestingly, the top genes in survival-relevant delta topics were related to PDAC and other cancer-related processes. *B2M*, among the top genes in the down-survival topic (delta 26), was associated with poor prognosis in PDAC.[16] *DPYD*, among the top genes in another down-survival topic (delta 10), was found to be overexpressed in the PDAC sample with a poor prognosis in the immunohistochemical analysis.[17] Furthermore, *NEDD9*, selected for delta 4 and 11, is a prognostic maker in pancreatic cancer progression.[18] *ERBB4* in topic 29 is known to accelerate PDAC development and progression.[19] These top genes were also related to other cancer cell processes. *BCL11B*, in up-survival delta topic 11, is a well-established tumor-suppressor gene in lymphoma and leukemia.[20] Another oncogene *TC2N* was also found in the same delta topic, 11, suggesting its role in helping tumor cells survive by suppressing the p53 signaling pathway.[21] *BTN3A* in topic 7 can be regulated by several signals induced by cancer cell or its microenvironment.[22]

DeltaTopic model identifies disease-relevant cell types and pathways

We next evaluated the extent to which the cell topics inferred by DeltaTopic reflect static and dynamic transcriptome patterns. In general, we found that the rho topic has high correspondence to cell-type differentiation (Figure 3A), while the delta topic loadings are more relevant to gene activities (Figures 3B1–3B3). Our results suggest that the ρ topics are suitable to capture cell-type-specific marker genes and recapitulated known immune cells, such as B cells, T cells, and macrophages, endocrine and endothelial cells, and multiple subtypes of ductal cells. Topics 15, 18, 25, and 28 correspond to one subtype of the ductal cells, and the other topics 14, 9, 27, 30, 26, and 6 are specifically matched with the second ductal cells found in the original study.[6] Topics 4 and 11 are likely associated with immune activities, as they mostly constitute immune cells' activities.

Gene set enrichment analysis of the Molecular Signatures Database[23–25] using the fgsea package[26] provided another line of evidence. We observed the top genes in topics 4 and 11 significantly overlap with the gene set "CD4 TCELL VS BCELL DN," which comprises genes differentially regulated between T and B cells. Unlike the static dictionary matrix ρ, the dynamic one, δ, is more likely to capture the gene expression changes more pertinent to immune mechanisms. For instance, topic 4 enriches two immune gene sets—"UNTREATED VS. IFNA STIM CD8 TCELL 90MIN UP" and "SIG BCR SIGNALING PATHWAY." Both were characterized to control tumor growth mechanisms.[27] Other topics 6 and 26 also bear many cancer-regulatory genes, constituting "ESTROGEN RESPONSE EARLY" and "ESTROGEN RESPONSE LATE" pathways, also well aligned with the previous study,[28] reporting high expression of estrogen receptor beta genes in pancreatic adenocarcinoma samples resected from a group with poor prognosis.

Vector fields reconstructed from DeltaTopic dictionary visualize distinctive disease trajectories

We projected the estimated transcriptome dynamics (velocities) onto a 2D space constructed by its eigenvectors (Figure 4A). Each headed arrow represents the estimated velocities for a cell, with the starting point for unspliced genes and the ending point for its spliced counterpart. We simply carried out the projection as follows: (1) performed singular value decomposition (SVD) on ρ dictionary matrix (with rank 2) $\rho = UDV^T$ to get its

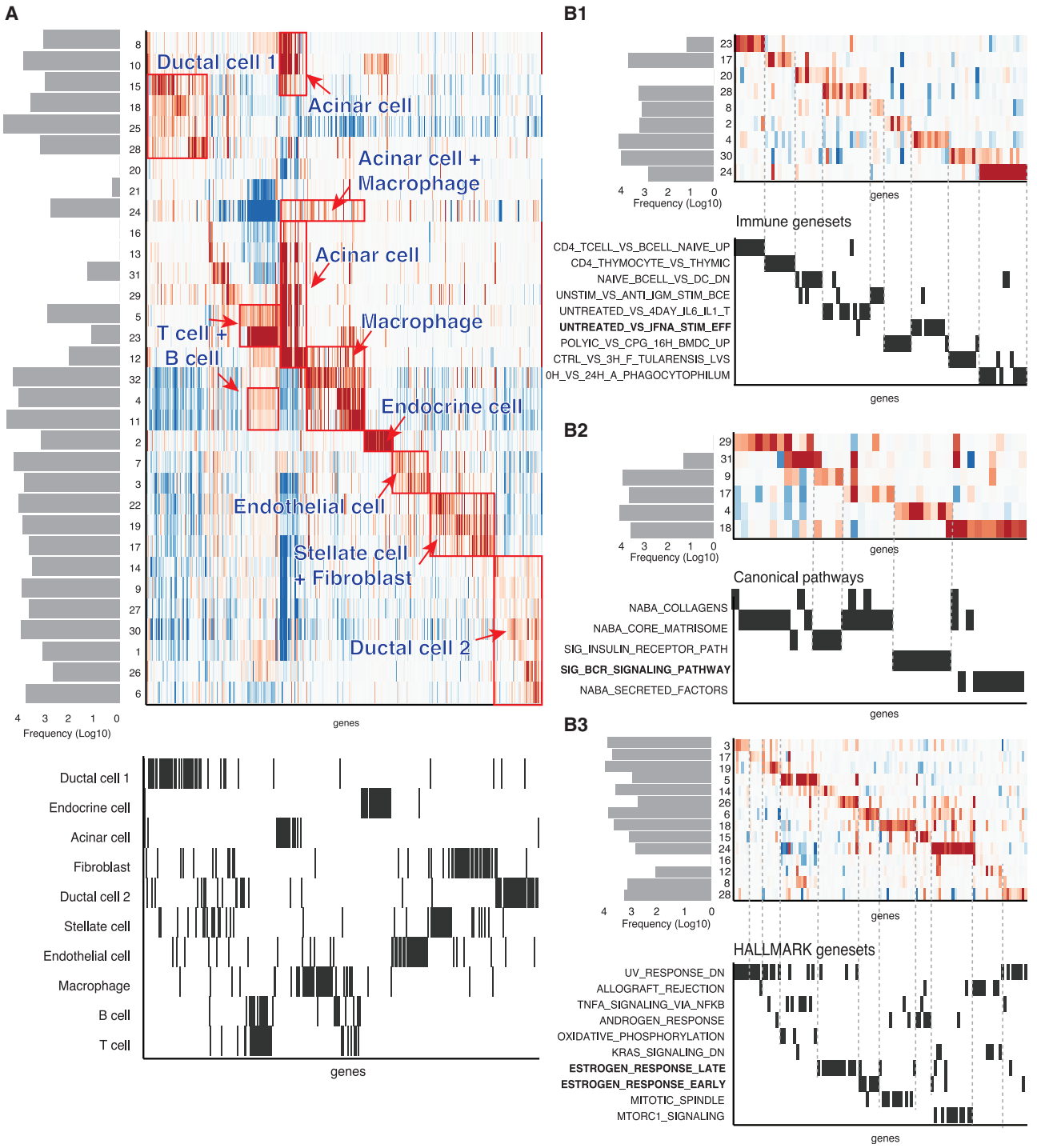


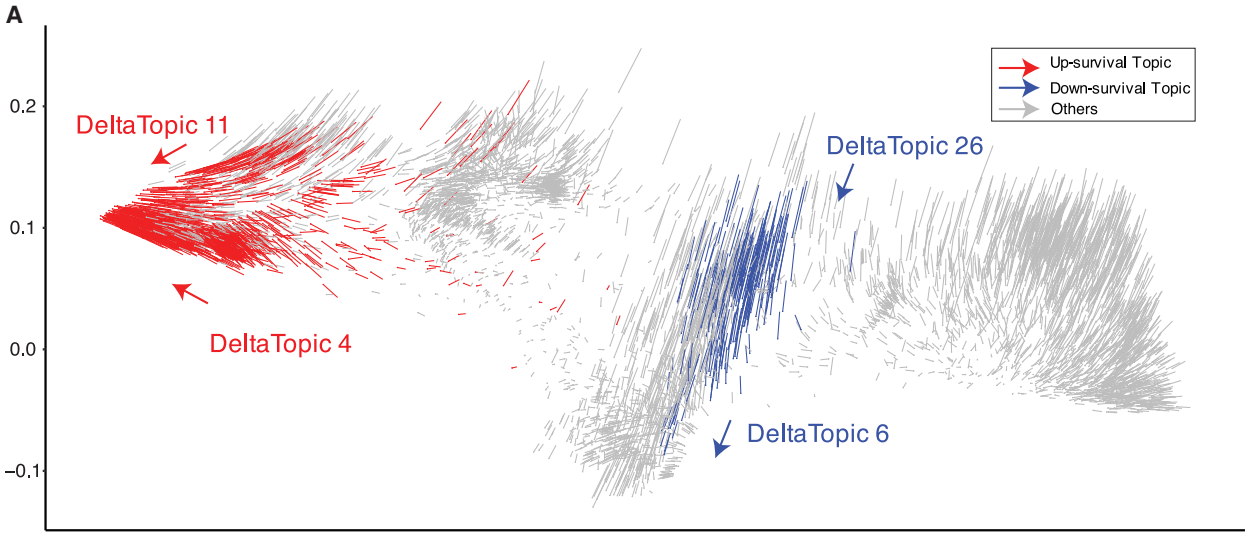
Figure 3. DeltaTopic approach uncover both static and dynamic transcriptome patterns

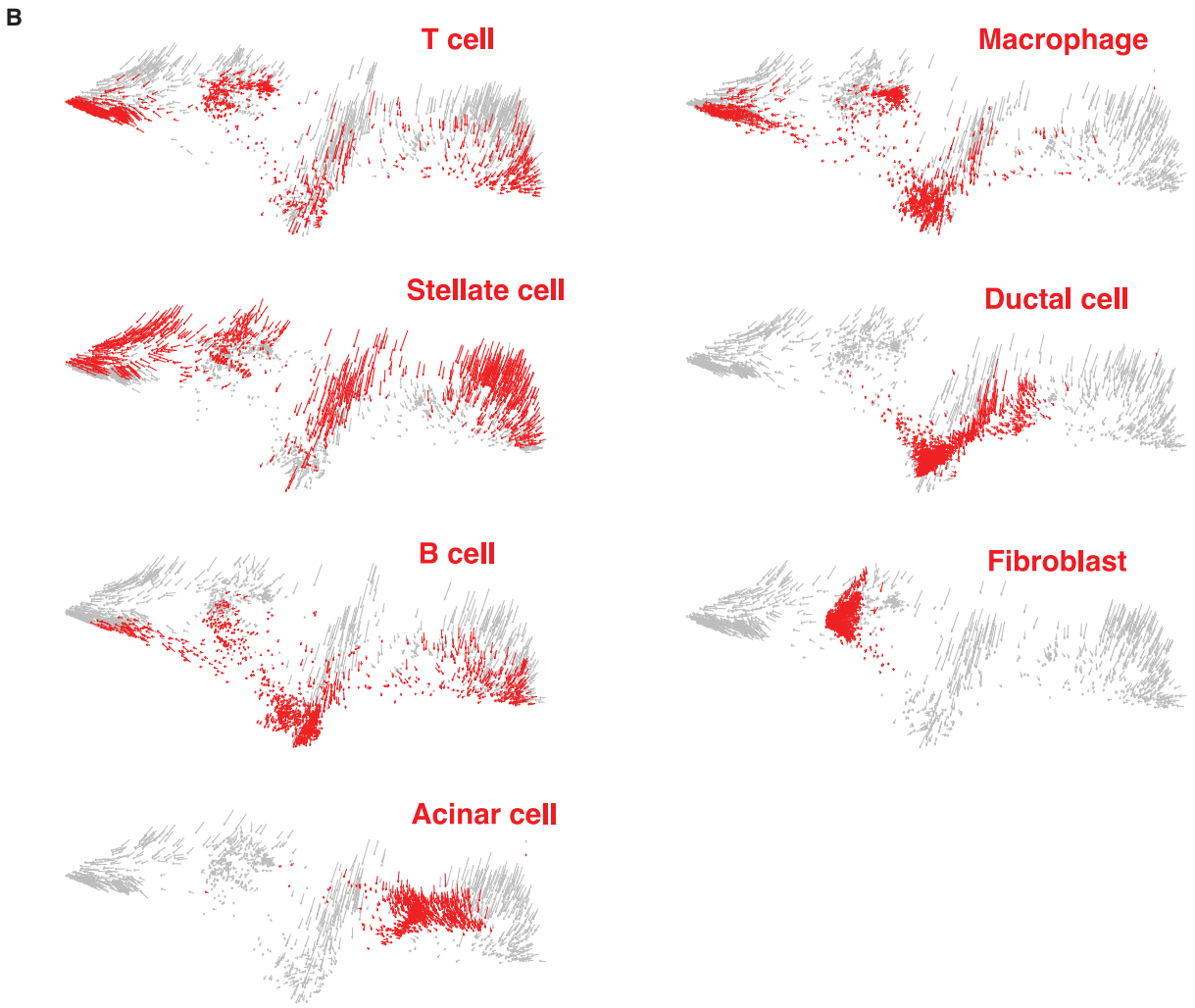
(A) The top heatmap: the static topic-by-gene parameters ρ; the bars on the left scale proportional to the size of each topic (log10 scale). As a comparison, the bottom heatmap indicates marker genes for the ten cell types assigned by the original PDAC study.[6].

(B) Gene set enrichment analysis of dynamic loading matrix. B1, ImmuneSig gene sets; B2, KEGG gene sets; B3, Hallmark gene sets. All three gene sets are from the MsigDB database.[23,24] For brevity, only significant gene sets and their corresponding cell topics are displayed.

2D representation V, (2) projected $\rho+\delta$ onto the same eigenspace for spliced genes by $V^{\text{spliced}} = (\rho+\delta)UD^{-1}$, and (3) connected V to V^{spliced} for each cell to visualize velocities.

Disease topics generally constitute two distinct flow patterns, consistent with our previous survival analysis results. Cell topics 4 and 11, which play the up-regulation role in PDAC survival,





(legend on next page)

converge to the left of the figure. On the other hand, the two down-regulating cell topics, 6 and 26, mostly flow in a downward direction (Figure 4A). Three immune cell types (T cell, B cell, and macrophage) align well with this disease-specific direction (Figure 4B). However, acinar cells and fibroblasts are relatively stagnant, not showing any salient flows in our visualization (Figure 4B).

Simulation studies confirm the topic-model-based approach can estimate dynamic and static cell topics

In order to gain confidence in our model's ability to identify dynamic and cell-type-specific gene programs, we conducted an extensive set of simulation-based benchmark studies. Not having a *de facto* simulation scheme suitable for generating matrices of both spliced and unspliced counts over all the genes, we simply generated synthetic datasets using the multinomial-Dirichlet hierarchical model as described previously. We possess limited knowledge of true transcriptomic dynamics in pancreatic cancer progression. Designing a realistic simulation method is also an active research area of single-cell genomics.

Treating the $\widehat{\rho}$ (static) and $\widehat{\delta}$ (dynamic) parameters as the ground truth for static and dynamic topic-specific gene programs, respectively, and assuming an equal distribution of cells across all topics, we randomly assigned each cell i to a topic t according to the Dirichlet distribution, namely topic proportion $\theta_i \sim \text{Dirichlet}(\alpha_i)$ with $\alpha_i =$, with $\sum_t \alpha_t = 1$. For a simpler interpretation, we randomly designated a major topic $t \in [T]$ for each cell so that the selected topic can most contribute to the transcriptomic variation for the cell, namely $\alpha_t = 0.9$, and the rest of variation can be explained by the other topics, namely $\alpha_{t'} = (1 - 0.9)/(T - 1)$ for all the other $t' \neq t$. We created distinct cell clusters for each topic. For simplicity, we simulated 1,000 cells per topic of a total of 32,000 cells and normalized the sequencing depth to the observed total number of genes, namely G. We repeated simulation experiments ten times with different random seeds.

For each simulation, we generated two single-cell expression matrices—the spliced and unspliced counts—and estimated our topic models and ran other relevant methods to compare performance in two prediction tasks: (1) how closely can we recover cell groups in the latent topic space, and (2) how robustly can we recover the dynamic and static gene programs in the model parameter matrices?

First we investigated the cell clustering problem, running the following methods.

- DeltaTopic (this work): DeltaTopic model trained on the spliced and unspliced data.
- BALSAM (this work): Bayesian sparse topic model trained on the spliced data only.
- PCA: PCA[29–31] on the spliced data only.
- PCA-concat: PCA run on the spliced and unspliced data concatenated with each other.
- NMF: non-negative matrix factorization[32] (vanilla version) on the splice data.
- LIGER: a variant of NMF[33,34] that can learn a unified latent space from two or more input datasets and identifies shared and dataset-specific latent factors.

For the results of topic models, we directly computed normalized mutual information scores with the true cell-type labels used in each simulation. For PCA and NMF, we had to resolve cell clusters by applying the Louvain algorithm[35] on the cell-cell network data based on similarity scores in the latent space. Since LIGER has can combine data across different data modalities, we performed the joint analysis implemented in the recent version of the pyLiger library: https://github.com/welch-lab/pyliger/blob/master/integrating_multi_scRNA_data.ipynb For all the methods, we computed normalized mutual information (NMI) scores by normalized_mutual_info_score, implemented in the scikit-learn library (v.1.2.2), which measured similarity between the predicted and true group membership. We would have the values 1 for an exact clustering result and 0 for completely random/independent clustering patterns between the predicted and true labels.

Since dynamic gene programs best determine some cell topics/clusters, our DeltaTopic method clearly outperformed other static methods, as well as the joint analysis conducted by the LIGER method (Figure 5A). Interestingly, among the unimodality methods trained on the spliced data only, BALSAM clearly outperformed other methods, such as PCA and NMF, demonstrating that the sparse Bayesian prior model generally improves clustering performance. Notably, DeltaTopic outperformed LIGER and PCA-concat, yielding substantially higher NMI scores, confirming that our strategy to build a GLM between the spliced and unspliced counts is more effective than the unpaired concatenation approaches. We also noted that LIGER performed additional pre- and post-processing quantized normalization steps to retain a subset of genes (8,248), which might have affected clustering results. At least in this benchmark analysis, if the data were generated from a sparse topic model, where only a small fraction of features define cell topics, we believe that probabilistic modeling, including Bayesian sparsity in the model, will be found to be beneficial not only in model interpretation but also in generalization performance.

For the second prediction problem, we assessed the quality of top-scoring genes/features. For the top K genes ranked by each method, we measured the precision by score as a fraction of the top K genes interacting with the top genes according to the true parameters, namely precision at K recall. We varied the K values from 10 to 1,000 and measured each precision value at each top

Figure 4. Velocities derived from the DeltaTopic identify distinct cell trajectories for disease development and cell-type differentiation

(A) Each segment corresponds to each cell uniformly sampled in each topic at a 0.5% rate to avoid visual clutter. The length of each segment (colored red, blue, and gray) scales proportionally to the estimated velocity projected onto two principal-component axes. We highlight cells constituting several disease-relevant topics identified by the previous survival analysis. We colored the cells red and blue according to their membership in the up-regulated and down-regulated topics, respectively.

(B) The same velocity plot colored by different cell types.

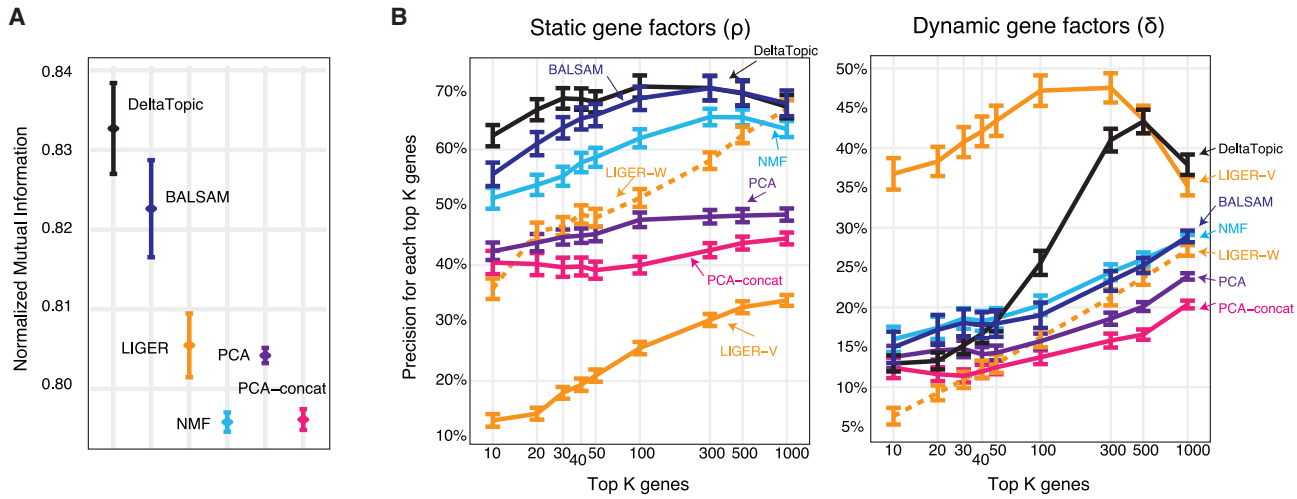


Figure 5. Benchmark results confirm that DeltaTopic and BALSAM can accurately predict cell-type labels and recapitulate true dynamic and static gene programs

(A) Normalized mutual information (NMI) scores between the predicted labels and true labels. The mean NMI scores and 95% confidence intervals are plotted for each method.

(B) Mean precision scores for static and dynamic gene activity identification. The mean and 95% confidence intervals are plotted for each method.

K cutoff (Figure 5B). In our simulation scheme, we specified that the static gene factors ρ_{gt} are manifested in both the spliced and unspliced expressions, whereas the dynamic factors δ_{gt} only exert actions in the spliced counts. For the static genes (Figure 5B, left panel), BALSAM and DeltaTopic generally achieve high precision values, followed by the NMF, PCA, and PCA-concat methods. For the dynamic gene prediction (Figure 5B, right panel), leveraging both types of counts, DeltaTopic clearly outperformed the other methods that were trained only on the spliced expression data.

However, it is interesting to note that the LIGER method led to more accurate prediction results for the dynamic genes—specifically those up-regulated in the spliced data—than DeltaTopic's results. A new version of the LIGER method, termed UNINMF, can differentiate common and data-type-specific components in a joint non-negative factorization setting.[34] We confirmed that the gene factors specific to the spliced count data (LIGER-V) are accurately captured by LIGER, especially for a few top genes ($K < 500$). For more than the top 500 genes, we found that DeltaTopic and LIGER perform equally well. Still, our simulation results show that DeltaTopic better recapitulated gene topics shared between two data modalities than LIGER (DeltaTopic vs. LIGER-W on the left of Figure 5B) without requiring additional post-processing steps. Moreover, DeltaTopic substantially outperformed LIGER in the cell-clustering predictions (Figure 5A).

Computing time and memory usage

The training of all models was accomplished utilizing an NVIDIA GeForce RTX 3080 GPU with a single-core configuration. Due to the implementation of spike and slab priors, both models are less prone to overfitting. We observed that 2,000 epochs were adequate for all models to reach convergence. For each simulated data, encompassing 32,000 cells, DeltaTopic (with 32 topics) finished in 6 h and 3 min, while BALSAM (32 topics) completed in 5 h and 12 min. The RAM usage for both models stood at 16.5 GB. For the full PDAC data, which comprises 227,311 cells and 11,418 genes, DeltaTopic took 16 h and 11 min, and BALSAM took 9 h and 43 min.

DISCUSSION

In this study, we propose a novel Bayesian topic model built on a deep-learning-based framework, firstly incorporating well-established sparsity prior distribution to model parameters (BALSAM) and secondly incorporating short-term dynamics implicated by the difference between the spliced and unspliced reads in scRNA-seq data (DeltaTopic). Considering technical limitations posed by the limited range of transcriptomic profiling, we believe that our Bayesian approach can provide a practical and statistical approach to estimating transcriptomic dynamics without requiring unattainable steady-state observations. In our case studies, we have demonstrated that DeltaTopic models built on two types of data modalities (the spliced and unspliced counts) achieved better generalization performance in terms of hold-out data reconstruction performance. In a broader sense, our approach can be considered a special case of a mixture of GLMs[36] that augments GLMs into traditional mixture components to express conditional probabilities between two different data types.

In benchmark analysis, we demonstrated that our GLM-based data integration framework is effective and often produces high-quality cell-clustering and gene prediction results. We can identify and differentiate dynamic and static gene programs in the DeltaTopic model parameters. However, a comparison with other joint analysis methods, especially a comparison with the recent version of LIGER,[34] also suggested that there is room for improvement in modeling. The next version of DeltaTopic can include additional factors on the side of the unspliced data

so that the model can better accommodate down-regulated gene programs than the current model.

By incorporating epigenomic and proteomic measurements, our approach can be straightforwardly extended to capture full information flows in regulatory genomics problems. Perhaps a harder challenge still lies in linking regulatory elements to target genes and selecting isoforms to target proteins and protein complexes. Nonetheless, we argue that employing multiple layers of GLMs to link data modalities can provide a principled way to address more complex multiomics data integration problems. As long as multiple stages of first-order approximation can constitute full-phase diagrams of a system of differential equations, we expect that such piecewise, layer-by-layer approaches will be used effectively in future research.

Interestingly, in our pancreatic cancer analysis, looking at short-term dynamics improved our understanding of disease progression. Our DeltaTopic model can pick up cancer-progression-specific latent factors, and the significance of our findings was validated in larger cohorts. Since we know relevant cell types where these disease-specific factors are selectively activated/repressed, we expect further *in vitro* or *ex vivo* validation experiments will further elucidate novel aspects of disease mechanisms. Although many existing latent variable models focus on clustering and subtype identification, our model especially highlighted that short-term dynamics (merely directional information) can provide an important clue in translation studies. Tracking full trajectories of disease progression may be intractable in current technology, but predicting immediate next steps at each cell-type/-state level appears possible. Conversely, dynamics analysis can complement a current way of investigating cellular heterogeneity, suggesting that there are other amiss axes in disease mechanisms besides cell-type composition changes.

As suggested by the previous embedded topic modeling approach,[37] prior knowledge can play an important role in dealing with stochastic, noisy datasets. In our case, eventually, we will need to understand transcription factors that drive topic-specific dynamics and disease mechanisms. Again, incorporating histone and DNA accessibility data will greatly benefit multimodal single-cell analysis.

Limitations of the study

Our topic model-based approach to single-cell data analysis carries several technical limitations, one of which is that we used a machine-learning library (torch) for most of our computation, which often requires specialized hardware resources, such as GPUs with sufficiently large memory, in order to achieve optimal performance. Here, we did not find a compelling reason for incorporating batch-specific or domain-specific technical bias terms in our modeling. However, batch effects are prevalent in single-cell genomics analysis; hence, one may need to consider a causal inference approach to delineate batch-specific effects from other biologically relevant signals, such as cell types and disease effects, in practice.

We also emphasize that DeltaTopic was specifically designed to capture immediate short-term dynamics between the spliced and unspliced counts for the same genes. Two types of extensions will greatly benefit: firstly, a model incorporating measurements from multiple time points will lead to statistically more robust inference results, provided that single-cell data were measured along a dense temporal axis. Moreover, we can generalize one-to-one correspondence between the spliced and unspliced variables because several regulatory regions are involved in a target gene's transcription, post-transcription, and translation processes.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Single-cell RNA-seq data preparation
  - Notations
  - Model descriptions
  - A bayesian extension for topic modeling in single-cell RNA-seq analysis
  - Amortized variational inference
  - Sampling from the latent variable model $q(\theta \vee \varphi)$
  - Global spike-and-slab parameters $q(\beta)$
  - Dynamically-encoded latent transcriptomic analysis by topic modeling (deltaTopic)
  - Kaplan-Meier survival analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100388.

ACKNOWLEDGMENTS

This work was supported by the BC Cancer Foundation and NSERC Discovery Grant (Y.P.P.). We are deeply grateful for support from Adrian Wan and the BC Cancer IT team.

AUTHOR CONTRIBUTIONS

Conceptualization, Y.P.P.; methodology, Y.Z., M.(S.)K., and Y.P.P.; investigation, M.(S.)K. and Y.Z.; writing – original draft, Y.Z. and Y.P.P.; writing – review & editing, Y.P.P.; funding acquisition, Y.P.P.; resources, Y.P.P.; supervision, Y.P.P.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 11, 2023
Revised: May 27, 2023
Accepted: July 31, 2023
Published: August 23, 2023

REFERENCES

1. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498.

2. Bergen, V., Lange, M., Peidli, S., Wolf, F.A., and Theis, F.J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. Nat. Biotechnol. *38*, 1408–1414.
3. Gorin, G., Fang, M., Chari, T., and Pachter, L. (2022). RNA velocity unraveled. PLoS Comput. Biol. *18*, e1010492.
4. Gu, Y., Blaauw, D.T., and Welch, J. (2022). Variational mixtures of ODEs for inferring cellular gene expression dynamics. In Proceedings of the 39th international conference on machine learning Proceedings of machine learning research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds. (PMLR), pp. 7887–7901.
5. Bergen, V., Soldatov, R.A., Kharchenko, P.V., and Theis, F.J. (2021). RNA velocity—current challenges and future perspectives. Mol. Syst. Biol. *17*, e10282.
6. Peng, J., Sun, B.-F., Chen, C.-Y., Zhou, J.-Y., Chen, Y.-S., Chen, H., Liu, L., Huang, D., Jiang, J., Cui, G.-S., et al. (2019). Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res. *29*, 725–738.
7. Chen, K., Wang, Q., Li, M., Guo, H., Liu, W., Wang, F., Tian, X., and Yang, Y. (2021). Single-cell RNA-seq reveals dynamic change in tumor microenvironment during pancreatic ductal adenocarcinoma malignant progression. EBioMedicine *66*, 103315.
8. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.
9. Melsted, P., Ntranos, V., and Pachter, L. (2019). The barcode, UMI, set format and BUStools. Bioinformatics *35*, 4472–4473.
10. Carbonetto, P., Sarkar, A., Wang, Z., and Stephens, M. (2021). Non-negative matrix factorization algorithms greatly improve topic model fits. Preprint at arXiv. https://doi.org/10.48550/arXiv.2105.13440.
11. Dey, K.K., Hsiao, C.J., and Stephens, M. (2017). Visualizing the structure of RNA-seq expression data using grade of membership models. PLoS Genet. *13*, e1006599.
12. Dieng, A.B., Ruiz, F.J.R., and Blei, D.M. (2020). Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics *8*, 439–453.
13. Carbonetto, P., and Stephens, M. (2012). Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Anal. *7*, 73–108.
14. Minoura, K., Abe, K., Nam, H., Nishikawa, H., and Shimamura, T. (2021). A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. Cell Rep. Methods *1*, 100071.
15. Kopf, A., Fortuin, V., Somnath, V.R., and Claassen, M. (2019). Mixture-of-Experts variational autoencoder for clustering and generating from similarity-based representations. Preprint at arXiv. https://doi.org/10.48550/arXiv.1910.07763.
16. Luchini, C., Mafficini, A., Chatterjee, D., Piredda, M.L., Sciammarella, C., Navale, P., Malleo, G., Mattiolo, P., Marchegiani, G., Pea, A., et al. (2022). Histo-molecular characterization of pancreatic cancer with microsatellite instability: Intra-tumor heterogeneity, B2M inactivation, and the importance of metastatic sites. Virchows Arch. *480*, 1261–1268.
17. Kato, H., Naiki-Ito, A., Suzuki, S., Inaguma, S., Komura, M., Nakao, K., Naiki, T., Kachi, K., Kato, A., Matsuo, Y., and Takahashi, S. (2021). DPYD, down-regulated by the potentially chemopreventive agent luteolin, interacts with STAT3 in pancreatic cancer. Carcinogenesis *42*, 940–950.
18. Radulović, P., and Krušlin, B. (2018). Immunohistochemical expression of NEDD9, e-cadherin and γ-catenin and their prognostic significance in pancreatic ductal adenocarcinoma (PDAC). Bosn. J. Basic Med. Sci. *18*, 246–251.
19. Hedegger, K., Algül, H., Lesina, M., Blutke, A., Schmid, R.M., Schneider, M.R., and Dahlhoff, M. (2020). Unraveling ERBB network dynamics upon betacellulin signaling in pancreatic ductal adenocarcinoma in mice. Mol. Oncol. *14*, 1653–1669.
20. Kominami, R. (2012). Role of the transcription factor bcl11b in development and lymphomagenesis. Proc. Jpn. Acad. Ser. B Phys. Biol. Sci. *88*, 72–87.
21. Hao, X.-L., Han, F., Zhang, N., Chen, H.-Q., Jiang, X., Yin, L., Liu, W.-B., Wang, D.-D., Chen, J.-P., Cui, Z.-H., et al. (2019). TC2N, a novel oncogene, accelerates tumor progression by suppressing p53 signaling pathway in lung cancer. Cell Death Differ. *26*, 1235–1250.
22. Blazquez, J.-L., Benyamine, A., Pasero, C., and Olive, D. (2018). New insights into the regulation of γδT cells by BTN3A and other BTN/BTNL in tumor immunity. Front. Immunol. *9*, 1601.
23. Dolgalev, I. (2022). Msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format.
24. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740.
25. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The molecular signatures database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425.
26. Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M.N., and Sergushichev, A. (2021). Fast Gene Set Enrichment Analysis (Cold Spring Harbor Laboratory), 060012.
27. Michaud, D., Mirlekar, B., Steward, C., Bishop, G., and Pylayeva-Gupta, Y. (2021). B cell receptor signaling and protein kinase D2 support regulatory B cell function in pancreatic cancer. Front. Immunol. *12*, 745873.
28. Seeliger, H., Pozios, I., Assmann, G., Zhao, Y., Müller, M.H., Knösel, T., Kreis, M.E., and Bruns, C.J. (2018). Expression of estrogen receptor beta correlates with adverse prognosis in resected pancreatic adenocarcinoma. BMC Cancer *18*, 1049.
29. Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. London, Edinburgh Dublin Phil. Mag. J. Sci. *2*, 559–572.
30. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. *24*, 498–520.
31. Jolliffe, I.T. (1986). Principal component analysis and factor analysis. In Principal component analysis, I.T. Jolliffe, ed. (Springer New York), pp. 115–128.
32. Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature *401*, 788–791.
33. Welch, J.D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E.Z. (2019). Single-Cell multi-omic integration compares and contrasts features of brain cell identity. Cell *177*, 1873–1887.e17.
34. Kriebel, A.R., and Welch, J.D. (2022). UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. Nat. Commun. *13*, 780.
35. Waltman, L., and van Eck, N.J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. Eur. Phys. J. B *86*, 471.
36. Hannah, L.A., Blei, D.M., and Powell, W.B. (2011). Dirichlet Process Mixtures of Generalized Linear Models. J. Mach. Learn. Res. *12*, 1923–1953.
37. Zhao, Y., Cai, H., Zhang, Z., Tang, J., and Li, Y. (2021). Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. Nat. Commun. *12*, 5261.
38. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (New York: Springer-Verlag). https://ggplot2.tidyverse.org.
39. Kingma, D.P., and Welling, M. (2013). Auto-Encoding variational bayes. Preprint at arXiv. https://doi.org/10.48550/arXiv.1312.6114.
40. Blei, D.M., Kucukelbir, A., and McAuliffe, J.D. (2017). Variational inference: A review for statisticians. J. Am. Stat. Assoc. *112*, 859–877.
41. Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning representations by back-propagating errors. Nature *323*, 533–536.
42. Kingma, D., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1412.6980.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Single-cell Pancreatic Ductal Adenocarcinoma data	Peng et al. (2019),[6] Chen et al. (2021)[7]	https://ngdc.cncb.ac.cn/gsa/browse/CRA001160
Software and algorithms		
Scanpy	Scanpy Development Team	https://scanpy.readthedocs.io/en/stable/
PyTorch	PyTorch Development Team	https://pytorch.org/
Numpy	Python package repository (PIP)	https://numpy.org/
Scipy	Python package repository (PIP)	https://scipy.org/
Pandas	Python package repository (PIP)	https://pandas.pydata.org/
Ggplot2	Wickham, 2016[38]	https://ggplot2.tidyverse.org/
Survminer	R package	https://cran.r-project.org/web/packages/survminer
Data.table	R package	https://cran.r-project.org/web/packages/data.table
Tidyverse	R package	https://www.tidyverse.org/
DeltaTopic and BALSAM	This work	https://deltatopic.readthedocs.io/en/latest/ https://doi.org/10.5281/zenodo.8173028

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Yongjin Park (ypp@stat.ubc.ca or yongjin.park@ubc.ca).

Materials availability

The study did not generate new unique reagents.

Data and code availability

We share our standalone Python package, called DeltaTopic https://causalpathlab.github.io/DeltaTopic/ with documentation pages, https://deltatopic.readthedocs.io/en/latest/. The development version is accessible on the project's GitHub page at https://github.com/causalpathlab/DeltaTopic. We also share all the codes used to generate the results in this article here (https://doi.org/10.5281/zenodo.8173028) https://zenodo.org/record/8173028.

METHOD DETAILS

Single-cell RNA-seq data preparation

We obtained the original FASTQ files for pancreatic ductal adenocarcinoma (PDAC) from the public repository (https://ngdc.cncb.ac.cn/gsa/browse/CRA001160) provided by two PDAC studies.[6,7] The spliced and unspliced count matrices were quantified by a scalable approach, namely kb-python (https://www.kallistobus.tools/), which coordinates the inputs and outputs of Kallisto[8] and UMI-BUS[9] tools:

```
$ kb count -i index.idx -g t2g.txt -x 10xv2 -o ${output} \ -c1 spliced_t2c.txt -c2 unspliced_t2c.txt \ –workflow lamanno –filter bustools \ ${fastq1} ${fastq2}
```

Notations

We measured G genes on a total of N cells. We denote the spliced one by X_{ng} for a gene g in a cell n, capturing only the reads mapped on exonic regions, gene data used in most single-cell expression analyses. We use $X_{ng}^{(U)}$ to denote an unspliced expression activity for a gene g and a cell n, concerning reads mapped on, or involving intronic regions. For brevity, we will use a row vector $x_n = (X_{n1}, \ldots, X_{nG})$ for expression data (spliced count) on each cell n. Similarly, we will use $x_n^{(U)}$ to denote an "unspliced" expression count, $x_n^{(U)} = (X_{n1}^{(U)}, \ldots, X_{nG}^{(U)})$.

Following the mapping protocol proposed by La Manno and coworkers,[1] we can separately quantify $X_{ng}^{(U)}$, the count of unspliced transcripts of a gene g in a cell n, and $X_{ng}^{(S)}$, the count of the spliced of the gene g in the same cell n.

Model descriptions

Review of topic modeling

Embedded Topic Model[12] built on a Variational Auto-Encoder (VAE) framework[39] provides a scalable approach for discovering latent topics from a large corpus of documents. We applied a similar topic model approach to single-cell data, treating each cell as a document, 20k genes as a total set of vocabularies, and short reads mapped on the genes as words. We model gene expression counts generated by independent multinomial probabilities (a bag of words assumption) since multinomial likelihood better preserves scale-invariant properties across different batches than other deep learning methods based on Poisson, Negative Binomial, and Gaussian distributions. Letting X_{ng} be a gene expression count of a gene g in a cell n, the data likelihood of a total expression count matrix can be defined by the corresponding multinomial probabilities r_{ng}:

$$\prod_{n}^{N} p(x_n \vee r_n) \propto \prod_{n}^{N} \prod_{g=1}^{G} r_{ng}^{X_{ng}}$$

where we have $\sum_g r_{ng} = 1$ for all $n \in [N]$ and $r_{ng} > 0$ for all cells n and genes $g \in [G]$.

The multinomial probability parameter, γ_{tg}, can be expressed as a linear combination of topic-specific vocabulary (gene) matrices weighted by topic portions in a document, θ_{nt}: $r_{ng} = \sum_{t=1}^{T} \theta_{nt}\gamma_{tg}$. The latent topic proportion vectors are assumed to follow Logistic Normal distribution *a priori*, which can enable reparameterized variational inference by taking stochastic gradient steps. We generate θ_n in two steps: $z_n \sim N(0, I)$ and $\theta_{nt} = \exp(z_{nt}) / \sum_{t'} \exp(z_{nt'})$. Since we restrict both θ and γ in a T- and G-dimensional simplex, namely, $\sum_t \theta_{nt} = 1$ and $\sum_g \gamma_{tg} = 1$, we can confirm that the resulting r_n vectors also result in valid probabilities across vocabulary (genes) within cells, i.e., $0 \leq \sum_g r_{ng} \leq 1$.

A bayesian extension for topic modeling in single-cell RNA-seq analysis

We extended the embedded topic model (ETM) in two ways:

(1) We introduced a Bayesian hierarchical prior on the model parameters, $r_n \sim \mathrm{Dir}(\lambda_n)$ while formulating the Dirichlet parameters as a generalized linear model,

$$\log \lambda_{ng} = \sum_{t=1}^{T} \theta_{nt}\beta_{tg} + b_g;$$

(2) We sought to improve the interpretability of the model parameters by introducing Bayesian sparsity on the linear models, β_{tg}, assuming a majority of the β_{tg} values are statistically zero with some prior probability π,

$$\beta_{tg} \sim \pi N(0, \tau) + (1 - \pi)\delta_0(\beta_{tg}).$$

The other effects not captured by spike-and-slab β parameters are simply represented by a gene-specific bias parameter b_g invariantly present across all the topics. We did not enforce any specific prior distribution on the bias parameters.

Exploiting the conjugate relationship between the multinomial and Dirichlet distributions, we can analytically integrate out the composite variable r and derive the following data likelihood:

$$p(x_n \vee \theta_n, \{\beta_{tg}\}, b_g) = \int dr_n p(x_n \vee r_n) p(r_n \vee \theta_n, \{\beta_{tg}\}, \{b_g\}) \propto \frac{\prod_g \Gamma(X_{ng}+\lambda_{ng})}{\Gamma\left(\sum_g X_{ng}+\lambda_{ng}\right)} \frac{\Gamma\left(\sum_g X_{ng}\right)}{\prod_g \Gamma(X_{ng})},$$

where $\lambda_{ng} = \exp(\sum_{t=1}^{T} \theta_{nt}\beta_{tg} + b_g)$, and $\Gamma(\cdot)$ is the Euler's gamma function.

Amortized variational inference

As the dimensionality increases, the exact inference of posterior probability of the latent cell-specific topics $p(\theta_n \vee x_n)$ quickly turns into a computationally-intractable inference problem. Stochastic variational inference confers a scalable approach to finding approximating distributions, which often leads to surprisingly accurate posterior inference results.[40] A reparametrization technique popularized by the VAE framework[39] cast an intractable inference problem of a latent variable model into an optimization problem in a deep belief network model, which can be solved by taking back-propagation steps with respect to the model parameters.[41] Here, we use two types of variational distributions, namely one for the local, latent variables, $q(\theta_n)$ and the other for global topic-specific gene activity parameters $q(\beta_{tg})$, and minimized the Kullbeck-Leibler (KL) divergence between these approximates and the actual data likelihood probability models.

Equivalently we can maximize the following evidence lower bound (ELBO) for total data likelihood (denoted by L):

$$\log\prod_n \int_{\theta,\beta} p(x_n,\theta_n,\beta)\frac{q(\theta_n \vee \varphi, x_n)q(\beta \vee \xi)}{q(\theta_n \vee \varphi, x_n)q(\beta \vee \xi)} \geq E_q\left[\sum_n \log p(x_n \vee \theta_n,\beta)\right] + E_q\left[\log\frac{p(\theta_n)p(\beta \vee \pi,\tau)}{q(\theta_n \vee \varphi)q(\beta \vee \xi)}\right] \triangleq L,$$

where we used φ and ξ to denote all the parameters of the latent state/parameter distributions. Amortized variational inference algorithm finds approximate posterior distributions by optimizing the ELBO objective, taking stochastic gradient steps with respect to the variational parameters, namely φ and *xi*. We adaptively scheduled the learning rates and step sizes by Adam optimizer.[42] For gradient calculation, we used PyTorch library.

Sampling from the latent variable model $q(\theta \vee \varphi)$

Since the exact evaluation the first expectation term is generally intractable, we approximate it by summing over the data log likelihood using the sampled instances of $\theta^{(s)} \sim q(\theta \vee \varphi, x_n)$ and $\beta^{(s)}$ for each minibatch sample $s \in [S]$:

$$E_q\left[\sum_n \log p(x_n \vee \theta_n,\beta)\right] \approx \frac{1}{S}\sum_s \log p\left(x_s \vee \theta^{(s)},\beta^{(s)}\right).$$

We parameterized the mean μ and variance σ functions for the latent variable inference in a deep encoder model taking inputs of the original high-dimensional data x. Using the reparamterized trick of Logistic Normal distribution, we sample the posterior sample of $\theta^{(s)}$ as follows:

- Sample ,I$\varepsilon_s \sim N(0)$
- Reparamterize εsz$_s \leftarrow \mu(x_s) + \sigma(x_s)\circ$
- Transform $\theta_t^{(s)} \leftarrow \exp(Z_{st}) / \sum_k \exp(Z_{sk})$.

Using the corresponding variational parameters $\varphi \equiv (\mu,\sigma)$, assuming $z_s \sim N(0,I)$ *a priori* for all s, we can derive KL divergence between the prior and variational distributions of latent states:

$$E_q\left[\log\frac{q(Z_{st} \vee \varphi)}{p(Z_{st})}\right] = D_{KL}(q \parallel p) = \sum_t \frac{1}{2}\left[\mu_{st}^2 + \sigma_{st}^2 - \log\sigma_{st}^2 - 1\right].$$

Global spike-and-slab parameters $q(\beta)$

We analytically derived the second term (the negative KL loss) involving the global parameters β by using fully-factored spike-and-slab distributions[13] as variational distributions for β_{tg} parameters. When β_{tg} is active/on, more precisely, a latent indicator variable $h_{tg} = 1$ with probability α_{tg}, we parameterize it by a Gaussian distribution:

$$q(\beta_{tg} \vee h_{tg} = 1) = N\left(\mu_{tg}^{\beta}, \nu_{tg}^{\beta}\right)$$

with probability $\alpha_{tg} \triangleq p(h_{tg} = 1)$; otherwise, we simply set β_{tg} to zero:

$$q(\beta_{tg} \vee h_{tg} = 0) = \delta_0(\beta_{tg})$$

with probability $1 - \alpha_{tg}$.

Given the variational parameters, $\xi \equiv (\alpha, \mu, \nu)$, we can characterize the mean, $E_q[\beta \vee \xi] = \alpha\mu$ and variance, $V_q[\beta \vee \xi] = \alpha\nu + \alpha(1-\alpha)\mu^2$.

Letting $h_{tg} = 1$ with probability π and $\beta \vee h = 1 \sim N(0,\tau)$ *a priori*, we get the KL loss for the global parameters:

$$D_{\mathrm{KL}}(q(\beta) \vee p(\beta)) = \left[\alpha\log\frac{\alpha}{\pi} + (1-\alpha)\log\frac{1-\alpha}{1-\pi}\right] - \frac{\alpha}{2}\left[1 + \log\frac{\nu}{\tau} - \frac{1}{\tau}\left(\mu^2 + \nu\right)\right].$$

Dynamically-encoded latent transcriptomic analysis by topic modeling (deltaTopic)

DeltaTopic is a hierarchical Bayesian model designed to capture transcription dynamics in topic space manifested in spliced and unspliced single-cell count matrices. Built on the Bayesian extension previously discussed, the goal of DeltaTopic is to characterize topic-specific relationships between spliced (S) and unspliced (U) gene expressions as a generalized linear model, delineating the static/shared and dynamic/directional topic-specific gene components.

We used the same Multinomial-Dirichlet hierarchical model that the unspliced and spliced vectors are parameterized by the rates of the unspliced and the unspliced, respectively:

$$L = \prod_n^N p\left(x_n^{(U)} \vee \lambda_n^{(U)}\right) p\left(x_n^{(S)} \vee \lambda_n^{(S)}\right)$$

where

$$p\left(x_n^{(U)} \vee \lambda_n^{(U)}\right) \propto \frac{\prod_g \Gamma\left(X_{ng}^{(U)} + \lambda_{ng}^{(U)}\right)}{\Gamma\left(\sum_g X_{ng}^{(U)} + \lambda_{ng}^{(U)}\right)} \frac{\Gamma\left(\sum_g \lambda_{ng}^{(U)}\right)}{\prod_g \Gamma\left(\lambda_{ng}^{(U)}\right)}$$

and

$$p\left(x_n^{(S)} \vee \lambda_n^{(S)}\right) \propto \frac{\prod_g \Gamma\left(X_{ng}^{(S)} + \lambda_{ng}^{(S)}\right)}{\Gamma\left(\sum_g X_{ng}^{(S)} + \lambda_{ng}^{(S)}\right)} \frac{\Gamma\left(\sum_g \lambda_{ng}^{(S)}\right)}{\prod_g \Gamma\left(\lambda_{ng}^{(S)}\right)}.$$

We can reasonably assume that each latent cell topic proportion vector θ_n is the property of a cell n. In the shared topic space, the unspliced and spliced counts are differently expressed by different topic-specific rate parameters. We represent shared transcription rates (before splicing) by ρ_{tg} of a gene g for a topic t; we explicitly capture splicing rates by δ_{tg} of a gene g for a topic t along with a gene-specific baseline activity b_g. Putting them altogether, we have the log-rates for the unspliced:

$$\log \lambda_{ng}^{(U)} = \sum_t \theta_{nt} \rho_{tg} + b_g,$$

and the log-rates for the spliced:

$$\log \lambda_{ng}^{(S)} = \sum_t \theta_{nt} \left(\rho_{tg} + \delta_{tg}\right) + b_g.$$

We pose spike-and-slab priors on the ρ and δ parameters[13]:

$$\rho_{tg}, \delta_{tg} \sim \pi N(0, \tau) + (1 - \pi)\delta_0\left(\rho_{tg}\right).$$

The latent vector θ_n is sampled from Logistic-Normal distribution. We used two independent encoder networks and combined stochastic latent vectors as a mixture of experts, equally weighting.[14,15] Each encoder network generates the mean μ and standard deviation σ.

- Sample $\varepsilon_n^{(S)}, \varepsilon_n^{(U)} \sim N(0, \mathrm{I})$
- Reparameterize for the unspliced, $z_n^{(U)} \leftarrow \mu^{(U)}(x_n^{(U)}) + \sigma^{(U)}(x_n^{(U)}) \circ \varepsilon_n^{(U)}$
- Reparameterize for the spliced, $z_n^{(S)} \leftarrow \mu^{(S)}(x_n^{(S)}) + \sigma^{(S)}(x_n^{(S)}) \circ \varepsilon_n^{(S)}$
- Combine and transform: $\theta_{nt} \leftarrow \exp(Z_{nt}^{(U)} + Z_{nt}^{(S)}) / \sum_k \exp(Z_{nk}^{(U)} + Z_{nk}^{(S)})$.

With the above sampling scheme, we optimized the following ELBO and estimated posterior distributions of the latent states and model parameters:

$$L^{\mathrm{Delta}} \approx \sum_{b=1}^{B} \log p\left(x_b^{(U)}, x_b^{(S)} \vee \theta_b, \rho, \delta\right) - K$$

where b denotes an index for a mini batch sample with the batch size B and K the KL loss.

$$K = \sum_b^B E_q\left[\log \frac{q\left(\theta_b \vee \varphi^{(S)}, \varphi^{(U)}\right)}{p(\theta_b)}\right] + \frac{B}{N} E_q\left[\log \frac{q\left(\rho \vee \xi^{(\rho)}\right)}{p(\rho \vee \pi, \tau)}\right] + \frac{B}{N} E_q\left[\log \frac{q\left(\delta \vee \xi^{(\delta)}\right)}{p(\delta \vee \pi, \tau)}\right].$$

Kaplan-Meier survival analysis

We obtained topic-specific gene loading scores, such as δ, *rho*, and β parameters, after fully training topic models. We then estimated individual-level scores by multiplying them to individual-level gene expression profiles available in larger ICGC cohorts. Followed by standardization of these individual-level scores, we can stratify these individuals into the positive ($\Delta_{it} > 0$) and negative activity ($\Delta_{it} < 0$) sets for each topic t. Using the same procedure, we can also partition individuals into positively and negatively correlated groups based on the other two types of scores (P_{it} and B_{it}). We estimated Kapler-Miere(KM) survival curve for each topic and tested the two-group difference in survival probabilities by log rank test.