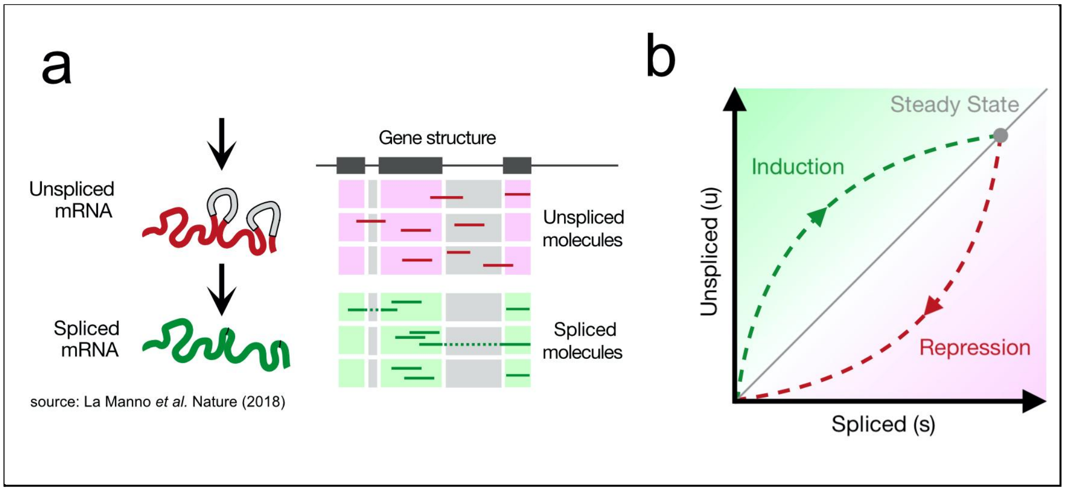
This is the draft of the paper. It includes any sources that were not transferred to the current version. Paragraphs have been significantly changed since this version.

Background

Having a five-year survival rate of less than 5%, pancreatic cancer is the third leading cause of cancer mortality in the United States (1). Symptoms of this rapidly fatal cancer which include jaundice, weight loss, and pain in the abdomen often present themselves at later stages, making treatment difficult. This lack of early diagnosis, limited treatment response, and a deficient understanding of the molecular mechanisms make pancreatic cancer a challenging disease to treat (2). Further RNA-seq data analysis and exploration of new bioinformatics lenses to view genetic and pathological features are crucial to elucidate the nature of this "silent killer" and improve its prognosis.

The traditional steady-state methods employed to analyze single-cell RNA-seq data explain some but not most of the genetic variation or heritability in disease. The use of a dynamical model can provide additional directed dynamic information by leveraging unspliced pre-mRNA counts. Specifically, such a dynamical model also referred to as RNA velocity allows for the analysis of the change in the rate of gene expression in genes at a given time point given the spliced and unspliced mRNA counts. This enables cellular dynamics analysis without needing time-series single-cell RNA data.



source: La Manno *et al.* Nature (2018)

Figure 1a shows the underlying model that is the basis for RNA velocity analysis. Essentially, the lifecycle of a messenger RNA can be roughly explained by the transcription, splicing and degradation phases, where two main forms of mRNA, spliced and unspliced, can be found. At different points in time, different ratios of unspliced and spliced RNA could be present in a cell. The former molecule is characterized by the presence of introns, shown in the gene structure diagram of figure 1a. Simply put, if there is more unspliced mRNA than spliced mRNA for a given gene, we can assume that this gene is about to be upregulated. We can therefore also assume that the gene is about to be downregulated if there is less unspliced mRNA than spliced mRNA.

These phenomena are shown on the phase plot in figure 1b, where a gene is in induction if the unspliced count is greater than the steady state slope. The opposite is also true, with repression of a gene. Using this model, and by obtaining this phase diagram for all genes, it would be possible to guess or gather information on the future state of the cell.

This model of mRNA lifecycle can be explained mathematically as shown in figure 3. Two main ordinary differential equations can be used to describe the mRNA lifecycle. These two equations include the transcription rate alpha which can depend on the on or off transcription state in the cell, the splicing rate beta, which is the rate of transformation of unspliced to spliced mRNA (assumed constant), and a degradation rate constant gamma which represents the breakdown of the spliced mRNA by the cell. The first equation essentially states that the rate of change of the unspliced mRNA is increased by the transcription rate alpha and is decreased by the splicing rate beta which is in accordance with the model shown in figure 1. Based on the same model, the second equation shows that rate of change of spliced RNA is increased by splicing rate beta and decreased by the degradation rate gamma.

By making use of both spliced and unspliced mRNA count data (also known as mature and pre-mRNA counts respectively) in a dynamical model, additional directed dynamical data can be recovered from common scRNA-seq protocols. A higher relative abundance of unspliced mRNA for a specific gene, detectable by the presence of introns, can indicate that a gene is undergoing upregulation. Contrarily, a lower relative abundance of unspliced mRNA can indicate downregulation (3).

Topic modeling is a generative probabilistic method for identifying abstract topics within a dataset. A topic model can be used to discover groups of similar documents within a text corpora by taking into account the words present in each document. Similarly, it can also allow the modeling of groups of cells that share similar genotypic characteristics by making use of their RNA count information. Using the ratios of the abundance of unspliced and spliced mRNA counts as features in fitting topic models, we aim to identify cell type dynamics in pancreatic ductal adenocarcinoma, being the most prevalent type of pancreatic cancer.

Results

Gene-level Topic Modeling Analysis

Topic modeling on a feature matrix composed of the spliced and unspliced gene count ratios was used to uncover 10 topics. The below structure plots display the different topic proportions for each cell, each represented by a vertical bar (Figure X). We share the constructed feature matrices in the Data Availability section.

a

b

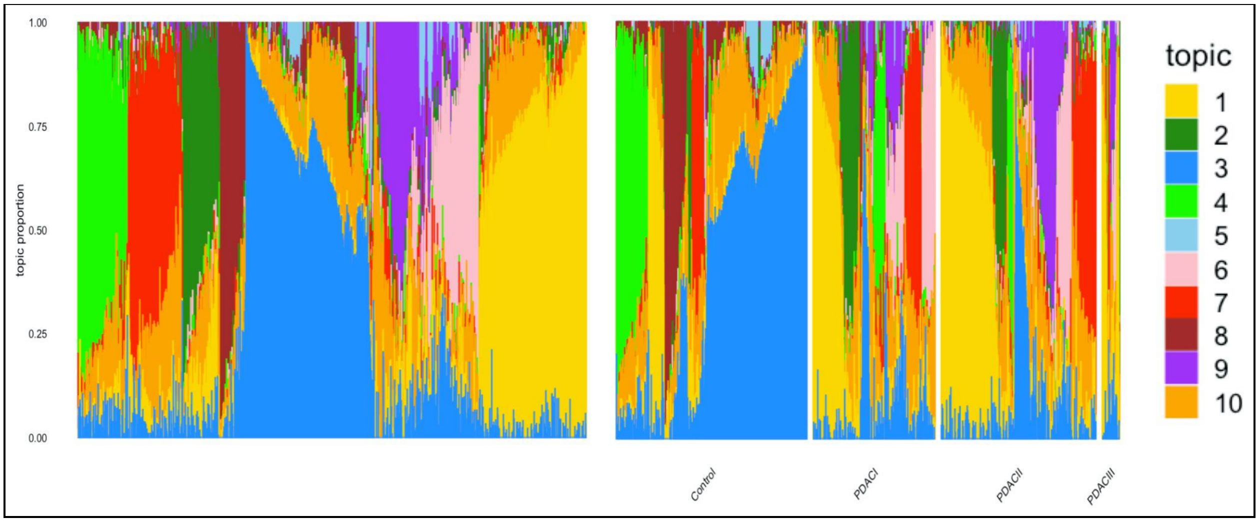


Figure X

Gene Level Topics Structure Plot

(a) Structure plot (or stacked bar chart) obtained using a combination of selected PDAC associated and high variance genes (See Materials and Methods for details). Each of the 10 topics are represented by a different colour bar. As it can be seen, the total height of all bars for each cell is the same, representing the mixture proportions summing up to 1. The observed patterns are a result of the arrangement of cells with similar mixture proportions close to each other. (b) The structure plot in plot A rearranged in order to group cells based on their disease labels (See Materials and Methods for details).

From the structure plots displayed in figure X, it is evident that gene ratios alone could allow us to segregate cells into distinct groups having similar dynamical expression landscapes. Upon further separating the cells based on their disease label (control, PDAC I, PDAC II, and PDAC III), we notice that some topics are almost exclusively associated with either control or disease cells (Fig. Xb). For example, it can be seen that the group of cells with gene ratios largely explained by topics 3 and 8 (darker blue and maroon) are mainly contained in the control group. This is as opposed to cells associated with topics 2, 6, and 9 (darker green, pink and purple) which are almost exclusively present in the PDAC labelled groups.

Analyzing the top genes (those with the highest relative gene ratio values per topic) can provide useful insight about the cell types associated with each topic as well as elucidate the roles that groups of similar genes perform in regards to PDAC. Figure X below shows the top genes associated with each topic.

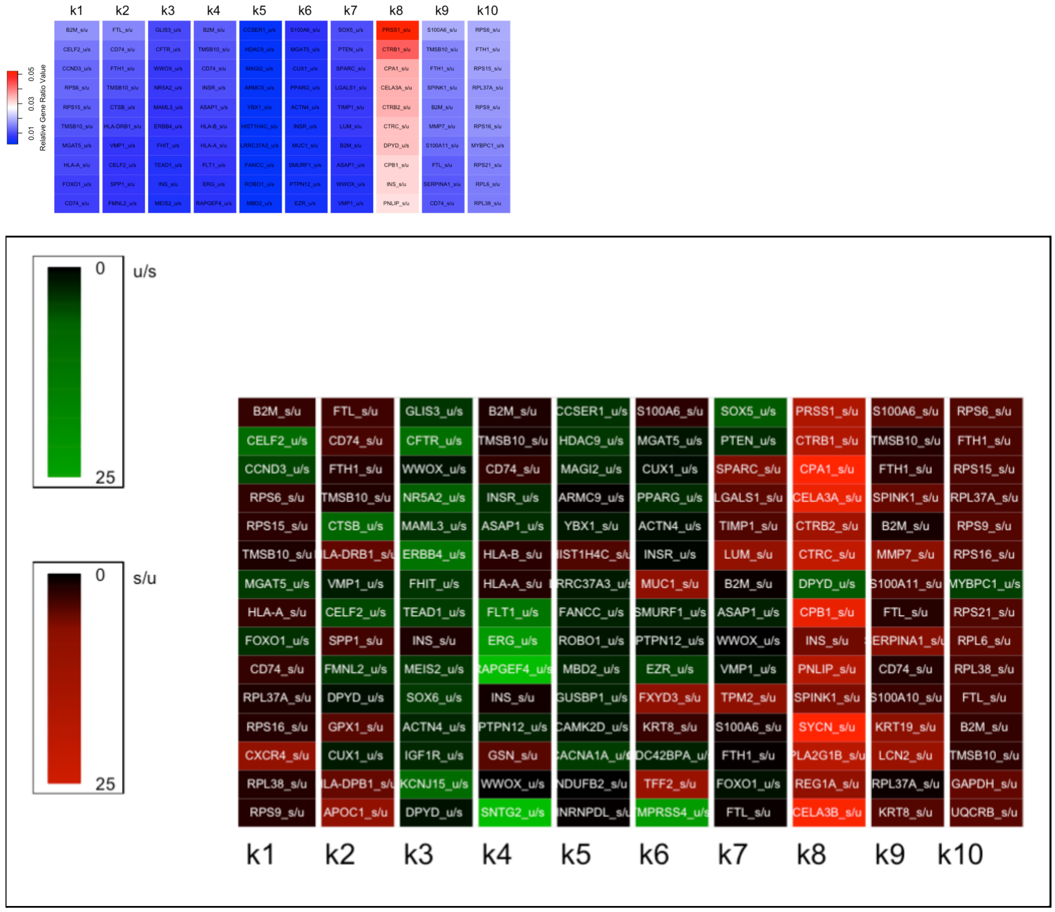


Figure X

Top 10 Genes per Topic

The top gene ratios are displayed for each topic, ranked by their relative gene ratio values. Gene names end with either "u/s" or "s/u" indicating potential gene activation or deactivation respectively. The colors of the cells indicate the relative gene ratio value of each gene, with values ranging from 0.01 to 0.05 in the top 10 genes displayed. Being mixture proportions, the relative gene ratios for all genes sum to 1 for each topic.

With the possible exception of topic 8, it can be seen that topics are not characterized by a single or small number of genes. In fact, most of the genes have small and similar relative gene ratio values of less than 0.05. We also note that each topic tends to include groups of similar genes, such as those being activated or those encoding ribosomal proteins.

It is worth noting that each topic contains rankings for all genes, therefore the same gene could have a high relative gene ratio in two or more topics, as is seen with the "B2M_s/u" gene ratio in topics 1 and 4. This is an advantage that topic modeling offers in contrast to hard clustering techniques. In other words, this allows the possibility that topics, which provide dynamical or expression information, have overlap in terms of genes. This takes into account the fact that genes that are activated in a specific gene expression process could also be activated in other processes, and in this case the same gene would be enriched in two or more topics.

Housekeeping Topics

With a quick literature search the characteristics of most topics can be found. Topics 1 and 10 are mostly associated with ribosomal protein genes (starting with “RP”) and known housekeeping genes (such as B2M and FTH1).[1] [2] The presence of topics 1 and 10 (yellow and orange) in both control and disease samples further confirms these associations (Fig. XB). However, while topic 10 seems to focus on deactivated genes marked by high spliced/unspliced (s/u) gene ratios, topic 1 seems to highly rank both activated and deactivated genes.

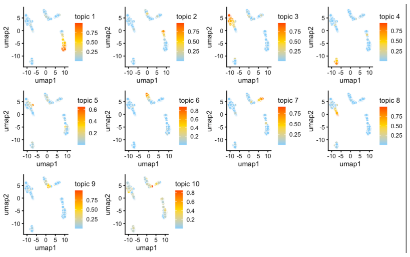
Control Topics

Topics 3 and 8, which are mostly seen in control samples, focus on activated and deactivated genes respectively. Tumour suppressor genes such as CFTR and WWOX are activated in control cells highly associated with topics 3.[3] [4] In topic 8, pancreatitis-associated genes such as PRSS1 and CTRC which can lead to oncogenic mutations are seen to be deactivated, explaining the association of control cells to topic 8.[5]

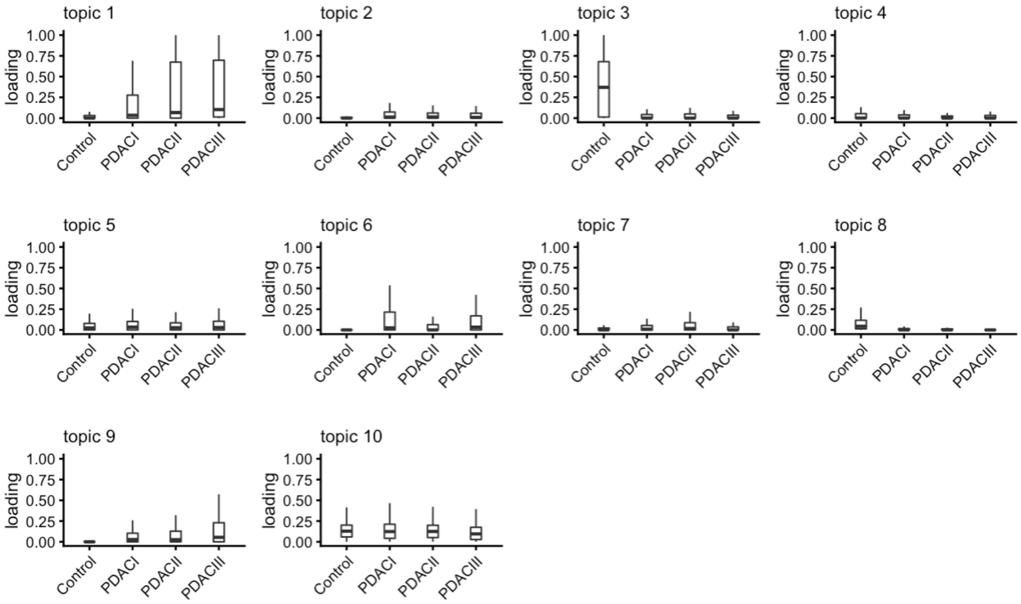
PDAC Topics

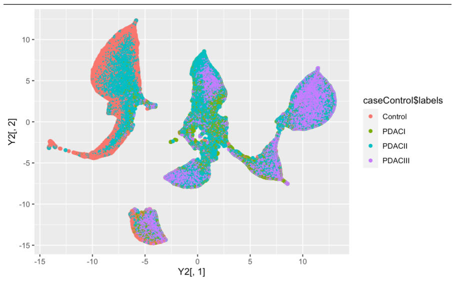
Topics 2, 6, and 9 are mostly seen in PDAC samples of various stages. Topic 6 is mostly associated with activated genes known to be highly expressed in pancreatic cancer tissue such as MGAT5 and CUX1.[6] [7] Interestingly, top genes in topics 2 and 9 consist of a mixture of housekeeping genes, deactivated tumor suppressor genes such as S100A11 and KRT19 , and activated genes involved in tumorigenic processes such as CTSB. [8] [9] [10]

It is worth highlighting that although many of the top genes in each topic were selected due to an already existing association with pancreatic cancer, a significant number of top genes consist of those showing high variability within the dataset. The handpicked pancreatic cancer genes can be seen to have driven the formation of different topics, where other high variability genes with similar characteristics were placed. One example of such a gene is the deactivated tumour suppressor TFF1 which was found as a top gene in topic 6 along other genes associated with disease samples. [11]



[1] https://www.genomics-online.com/resources/16/5049/housekeeping-genes/
[2] https://en.wikipedia.org/wiki/Housekeeping_gene
[3] https://pubmed.ncbi.nlm.nih.gov/24044959/
[4] https://pubmed.ncbi.nlm.nih.gov/26751771/#:~:text=Our%20results%20indicate%20that%20Cftr,in%20%3E60%25%20of%20mice.
[5] https://www.frontiersin.org/articles/10.3389/fphys.2014.00070/full
[6] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7082313/
[7] https://gut.bmj.com/content/59/8/1101
[8] https://clincancerres.aacrjournals.org/content/12/18/5417
[9] https://www.frontiersin.org/articles/10.3389/fonc.2019.01239/full
[10] https://www.ncbi.nlm.nih.gov/gene/3880
[11] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4170596/

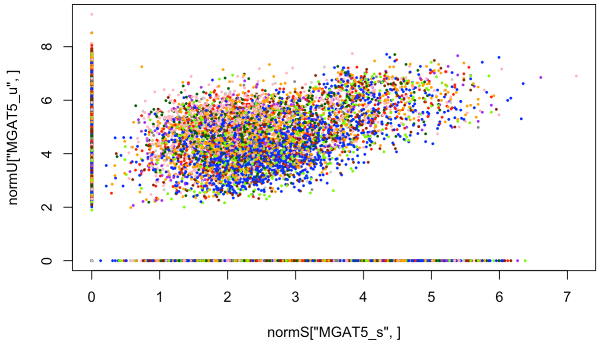


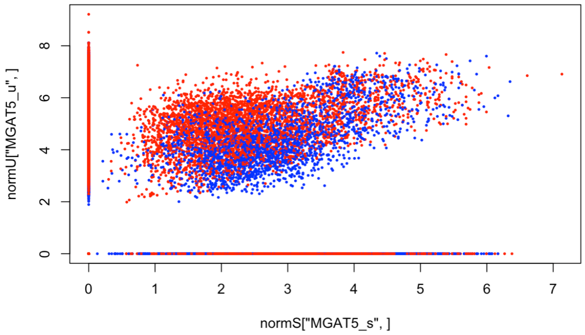


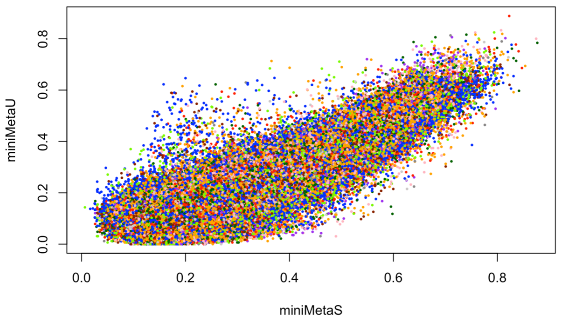
Summarize the number of activated and inactivated genes per topic here?

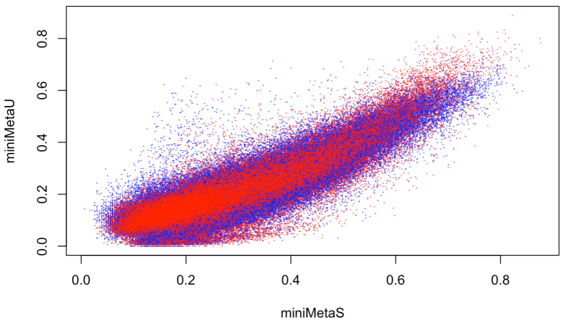
Phase diagrams

Topic driven clusters and case/control labelled cells





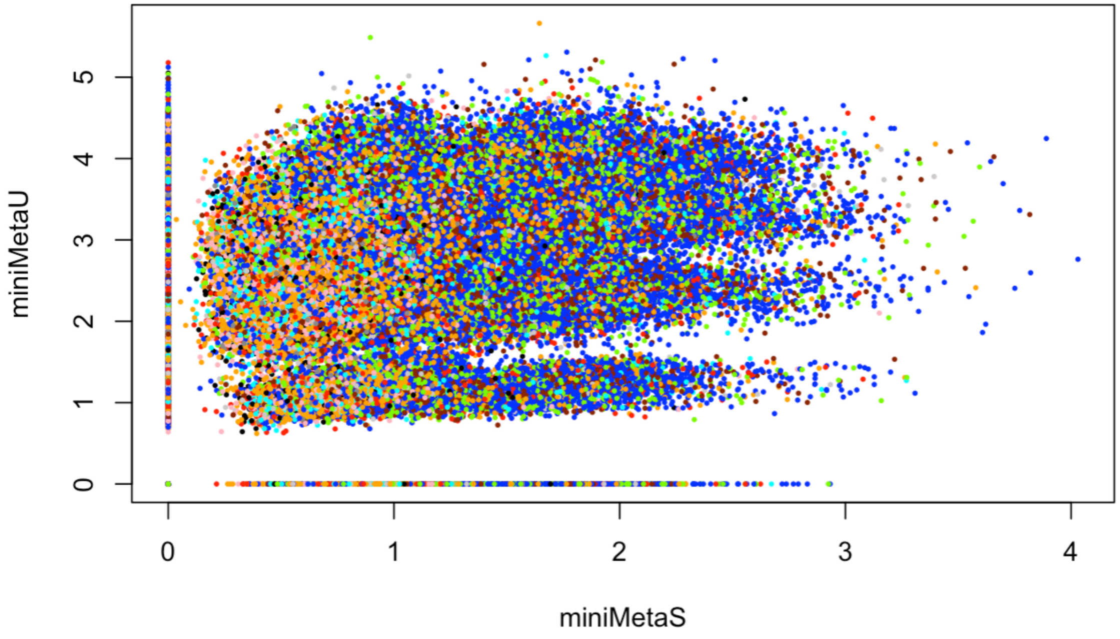




WE WANT DIRECTIONALITY. The clustering is great and LDA is great, somehow we wanna get to: this gene is really driving this topic then activation of this topic is leading to cascade of this other topic…

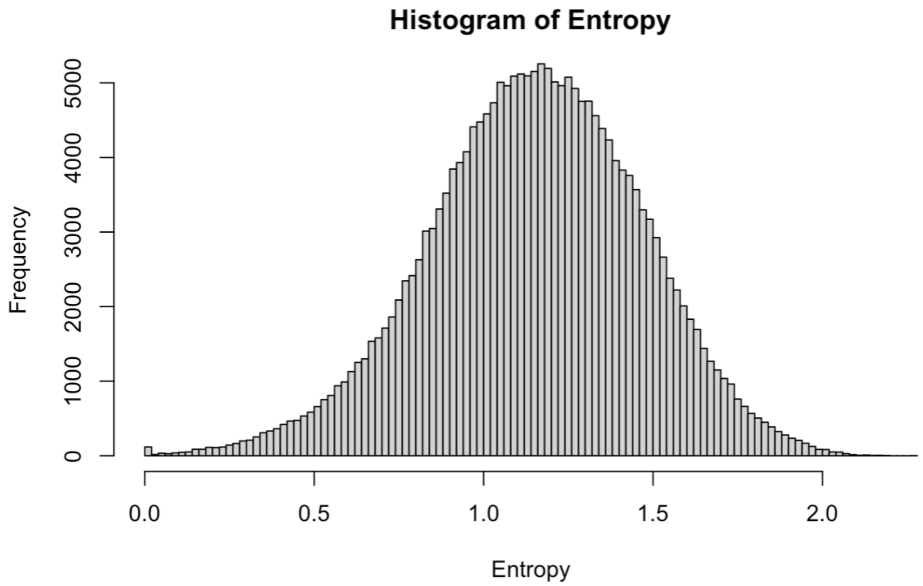
Selecting a few key genes could show us that directionality:

"CFTR", "WWOX", "CTRC", "MGAT5", "CUX1"



We identified 10 topics containing average XX active genes and XX repressive genes

We define active genes in each topic if they have a ratio of unspliced to spliced counts greater than one. Repressive genes are therefore defined by a ratio of spliced to unspliced counts greater than one. Since we take the log of these ratios (GIVE REASON Because values can be close to 0 and to avoid floating-point values), there would be negative log ratios, which are simply removed. For each cell, there is either a positive log ratio of unspliced over spliced counts indicating activation or a positive log ratio of spliced of unspliced counts indicating repression. These gene log-ratios marked by the suffixes "_u/s" or "_s/u" can be counted to find the number of activated or repressed genes in each topic.



Discussion

Another direction, many papers consider GWAS genes as done deal, however we could reinterpret them (and say look GWAS was kinda wrong) Highlight the genes that were not that significant in our topics

Pathway annotation is imperfect, primarily driven by static information or meta analysis, which disregards dynamic context of gene regulation, using our approach we can revise using pathways as units

- Heterogeneity in the topic proportions for each cell, also entropy

- Highlight evidence of new functionality and experimental or theoretical evidence of performance improvement
- State the limitations of the approach and suggest directions for future research.
- Limitations
- Related work - LDA applied to other data
- Conclusion

Materials and Methods

Data preprocessing

Human PDAC FASTQ files were retrieved from the Genome Sequence Archive (GSA) database (accession number CRA001160) (4). The 35 available samples obtained from the Department of General Surgery of Peking Union Medical College Hospital (PUMCH) included 11 control patients, 9 patients with PDAC I, 12 patients with PDAC II, and 3 patients with PDAC III (5).

RNA velocity (spliced and unspliced RNA) count matrices were generated from the FASTQ files using the kb-python python wrapper package for Kallisto-Bustools (6). 5 samples were not chosen for further analysis due to differing numbers of cells found in their spliced and unspliced RNA count matrices. The remaining samples were merged into a single matrix containing the spliced & unspliced (differentiated by the "_s" and "_u" suffixes respectively) genes (rows) and cells (columns). The ENSEMBL Gene IDs generated from kb-python were converted to gene symbols using the online Biotools database (7).

The combined count matrix was then imported in the mmutil (v0.3.0) R package, where quality control was performed (8).[12] High quality cells containing more than 300 non-zero spliced and unspliced genes expressed with a mitochondrial gene activity of less than 50% were selected. Out of the 238,135 cells available, 221,244 (92.91 %) were retained, discarding 16,891 cells (7.09 %). Out of the 58,367 detected genes, a subset of 1,434 known pancreatic cancer-associated (causal) genes from previous genome-wide association studies was selected (9).[13] This subset of genes was chosen to minimize the accidental identification of passenger genes while also limiting the runtime of the LDA algorithm. Considering the spliced and unspliced counts separately, doubling the number of features, a final MTX dataset containing 2,868 features, 238,133 cells and 42,513,902 non-zero elements (6.22%) was obtained. Similarly, another MTX dataset was obtained containing additional background genes with high variability, containing 7,248 features, 238,133 cells and 53,669,636 non-zero elements (3.11%).

Topic Modeling

[12] https://satijalab.org/seurat/

[13] https://maayanlab.cloud/Harmonizome/gene_set/pancreatic+cancer/DISEASES+Text-mining+Gene-Disease+Assocation+Evidence+Scores

Here we focus on the analysis of topic models as opposed to the conventional single-cell RNA-seq data analysis. We can create a model of the count data by directly applying latent Dirichlet Allocation (LDA) on the spliced and unspliced counts (or their ratio) arising from the RNA sequencing assay. Therefore, no conventional preprocessing step was required, including log-transformation or normalization of the UMI counts (10).[14]

The preprocessed feature matrix was further modified to obtain a gene-level ratios feature matrix. Each gene's ratio of the spliced and unspliced counts (and vice versa) was taken and gene names either included a "_s/u" or "_u/s" suffix. The ratios were then log-transformed to avoid floating-point values resulting from small ratios. The negative ratios were replaced with a value of 0 to obtain a non-negative matrix. In addition to the gene-level feature matrix, a canonical pathway-level feature matrix was created from the BioCarta, KEGG and Reactome gene sets containing a total of 2,082 pathways (11). Similar to the transformation made to the gene-level feature matrix, the ratio of the sums of the spliced and unspliced gene counts in each pathway were taken and pathway names either were given a "_s/u" or "_u/s" suffix.

We argue that these ratios can provide additional information to traditional single-cell RNA-seq analysis. The unspliced to spliced mRNA count ratio for a particular gene can serve as an estimation for the velocity or change in expression of that particular gene.

After performing preliminary experiments with different topic modelling packages, and testing different numbers of topics, the fastTopics (12) (v0.5-52) R package was chosen for further analysis of our feature matrix (https://github.com/stephenslab/fastTopics) (13).[15] This package was chosen due to its ability to exploit data sparsity to deliver efficient and scalable topic model optimization algorithms. Additionally, fastTopics integrates many useful tools to identify key features in topics and to perform generalized differential expression analysis allowing partial membership to groups.

The optimal number of topics providing the largest number of easily identifiable groups of cells was determined to be 10 topics. Gene-level and pathway-level models for datasets containing 10 topics were fitted using 144 RcppParallel threads on a compute cluster. The total elapsed time to obtain the gene-level topic model using 2,868 features (s/u and u/s versions of 1,434 genes) was 50955.01s (14h). For the pathway-level model using 4,164 features (s/u and u/s versions of 2,082 pathways), the elapsed time was 120819.8s (33h).

Data Availability

Analysis scripts and data: (https://github.com/causalpathlab/pdac_velocity_topic)
The mmutil (Matrix Market Utility) C++ program: (https://github.com/causalpathlab/mmutil)
For further questions on analysis scripts, please contact Sam Khalilitousi (samkt@student.ubc.ca)

References

1. *Bengtsson A, Andersson R, Ansari D. The actual 5-year survivors of pancreatic ductal*

[14] https://stephenslab.github.io/fastTopics/articles/single_cell_rnaseq_basic.html

[15] https://github.com/stephenslab/fastTopics

adenocarcinoma based on real-world data. Sci Rep. 2020 Oct 2;10(1):16425.

2. *Sarantis P, Koustas E, Papadimitropoulou A, Papavassiliou AG, Karamouzis MV. Pancreatic ductal adenocarcinoma: Treatment hurdles, tumor microenvironment and immunotherapy. World J Gastrointest Oncol. 2020 Feb 15;12(2):173–81.*

3. *La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. Nature. 2018 Aug;560(7719):494–8.*

4. *Wang Y, Song F, Zhu J, Zhang S, Yang Y, Chen T, et al. GSA: Genome Sequence Archive<sup/>. Genomics Proteomics Bioinformatics. 2017 Feb;15(1):14–8.*

5. *Peng J, Sun B-F, Chen C-Y, Zhou J-Y, Chen Y-S, Chen H, et al. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res. 2019 Sep;29(9):725–38.*

6. *Melsted P, Ntranos V, Pachter L. The barcode, UMI, set format and BUStools [Internet]. Vol. 35, Bioinformatics. 2019. p. 4472–3. Available from: http://dx.doi.org/10.1093/bioinformatics/btz279*

7. *ENSEMBL ID to Gene Symbol Converter - Genomics Biotools [Internet]. [cited 2021 Nov 11]. Available from: https://www.biotools.fr/human/ensembl_symbol_converter*

8. *mmutil [Internet]. Github; [cited 2021 Nov 16]. Available from: https://github.com/causalpathlab/mmutil*

9. *Gene Set - pancreatic cancer [Internet]. [cited 2021 Nov 10]. Available from: https://maayanlab.cloud/Harmonizome/gene_set/pancreatic+cancer/DISEASES+Text-mining+Gene-Disease+Assocation+Evidence+Scores*

10. *Analysis of single-cell RNA-seq data using a topic model, Part 1: basic concepts [Internet]. [cited 2021 Nov 11]. Available from: https://stephenslab.github.io/fastTopics/articles/single_cell_rnaseq_basic.html*

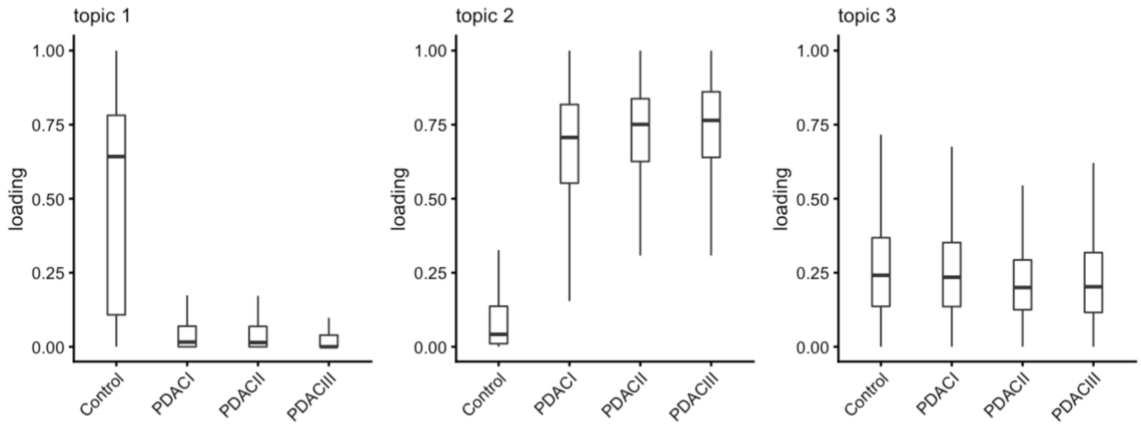
11. *Dolgalev I. MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format [R package msigdbr version 7.4.1]. 2021 May 5 [cited 2021 Nov 11]; Available from: https://cran.r-project.org/web/packages/msigdbr/index.html*

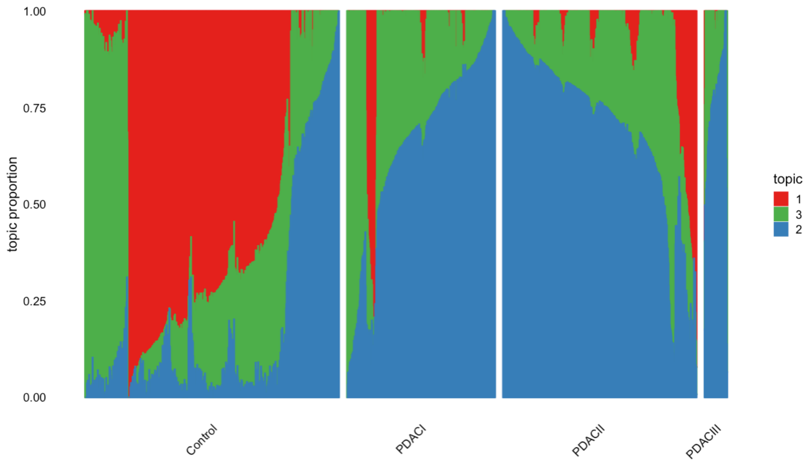
12. *Carbonetto P, Sarkar A, Wang Z, Stephens M. Non-negative matrix factorization algorithms greatly improve topic model fits [Internet]. arXiv [stat.ML]. 2021. Available from: http://arxiv.org/abs/2105.13440*

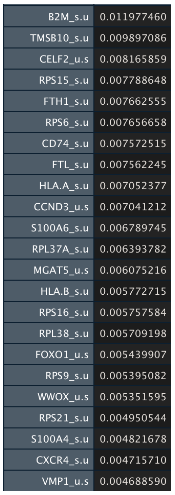
13. *fastTopics: Fast algorithms for fitting topic models and non-negative matrix factorizations to count data [Internet]. Github; [cited 2021 Nov 10]. Available from: https://github.com/stephenslab/fastTopics*

Extra things we could include

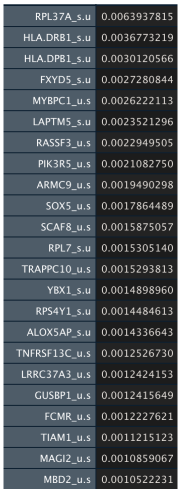
Merging similar topics together to obtain this 3 topic model.



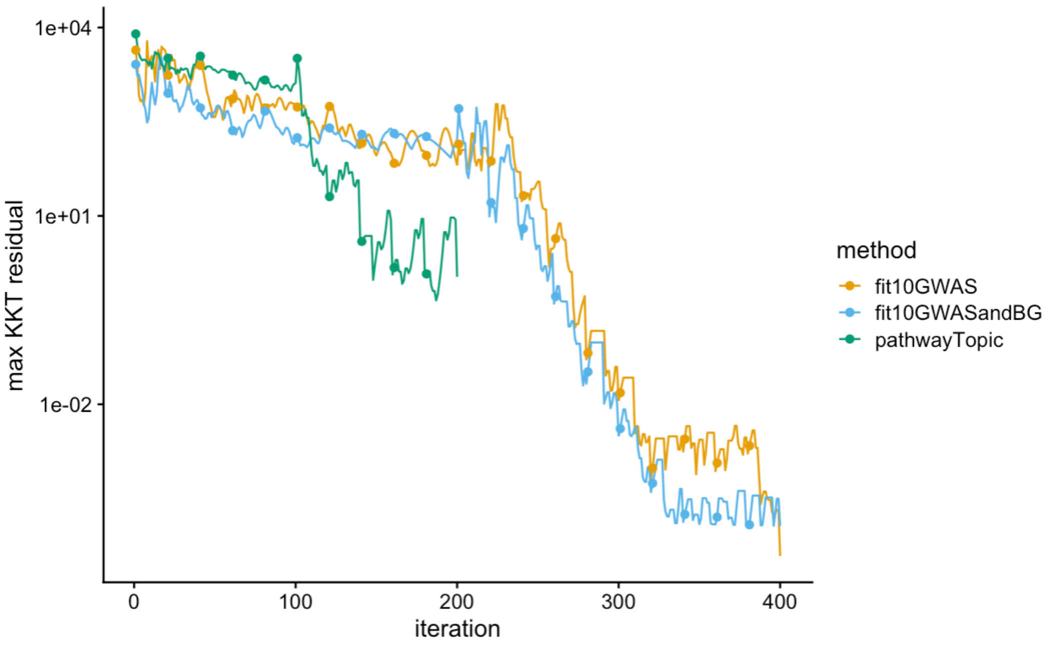




B2M_s.u	0.011977460
TMSB10_s.u	0.009897086
CELF2_u.s	0.008165859
RPS15_s.u	0.007788648
FTH1_s.u	0.007662555
RPS6_s.u	0.007656658
CD74_s.u	0.007572515
FTL_s.u	0.007562245
HLA.A_s.u	0.007052377
CCND3_u.s	0.007041212
S100A6_s.u	0.006789745
RPL37A_s.u	0.006393782
MGAT5_u.s	0.006075216
HLA.B_s.u	0.005772715
RPS16_s.u	0.005757584
RPL38_s.u	0.005709198
FOXO1_u.s	0.005439907
RPS9_s.u	0.005395082
WWOX_u.s	0.005351595
RPS21_s.u	0.004950544
S100A4_s.u	0.004821678
CXCR4_s.u	0.004715710
VMP1_u.s	0.004688590



RPL37A_s.u	0.0063937815
HLA.DRB1_s.u	0.0036773219
HLA.DPB1_s.u	0.0030120566
FXYD5_s.u	0.0027280844
MYBPC1_u.s	0.0026222113
LAPTM5_s.u	0.0023521296
RASSF3_u.s	0.0022949505
PIK3R5_u.s	0.0021082750
ARMC9_u.s	0.0019490298
SOX5_u.s	0.0017864489
SCAF8_u.s	0.0015875057
RPL7_s.u	0.0015305140
TRAPPC10_u.s	0.0015293813
YBX1_s.u	0.0014898960
RPS4Y1_s.u	0.0014484613
ALOX5AP_s.u	0.0014336643
TNFRSF13C_u.s	0.0012526730
LRRC37A3_u.s	0.0012424153
GUSBP1_u.s	0.0012415649
FCMR_u.s	0.0012227621
TIAM1_u.s	0.0011215123
MAGI2_u.s	0.0010859067
MBD2_u.s	0.0010522231


1e+04
1e+01
1e-02
max KKT residual
0
100
200
300
400
iteration
method
fit10GWAS
fit10GWASandBG
pathwayTopic