Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Non-negative Independent Factor Analysis for single cell RNA-seq

Weiguang Mao [1,3], Maziyar Baran Pouyan [2,5], Dennis Kostka [1,2,3,4] and Maria Chikina [1,3,*]

[1]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA
[2]Department of Developmental Biology, School of Medicine, University of Pittsburgh,Pittsburgh, PA, USA
[3]Joint Carnegie Mellon-University of Pittsburgh Ph.D. Program in Computational Biology, Pittsburgh, PA, USA
[4]Center for Evolutionary Biology and Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA
[5]Current address: Accenture AI Laboratory, Accenture, San Francisoc, CA, USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Single cell RNA sequencing (scRNA-seq) enables transcriptional profiling at the level of individual cells. With the emergence of high-throughput platforms datasets comprising tens of thousands or more cells have become routine, and the technology is having an impact across a wide range of biomedical subject areas. However, scRNA-seq data are high-dimensional and affected by noise, so that scalable and robust computational techniques are needed for meaningful analysis, visualization and interpretation. Specifically, a range of matrix factorization techniques have been employed to aid scRNA-seq data analysis. In this context we note that sources contributing to biological variability between cells can be discrete (or multi-modal, for instance cell-types), or continuous (e.g. pathway activity). However, no current matrix factorization approach is set up to jointly infer such mixed sources of variability.
Results: To address this shortcoming, we present a new probabilistic single-cell factor analysis model, **N**on-negative **I**ndependent **F**actor **A**nalysis (NIFA), that combines features of complementary approaches like Independent Component Analysis (ICA), Principal Component Analysis (PCA), and Non-negative Matrix Factorization (NMF). NIFA simultaneously models uni- and multi-modal latent factors and can so isolate discrete cell-type identity and continuous pathway-level variations into separate components. Similar to NMF, NIFA constrains factor loadings to be non-negative in order to increase biological interpretability. We apply our approach to a range of data sets where cell-type identity is known, and we show that NIFA-derived factors outperform results from ICA, PCA and NMF in terms of cell-type identification and biological interpretability. Studying an immunotherapy dataset in detail, we show that NIFA identifies biomedically meaningful sources of variation, derive an improved expression signature for regulatory T-cells, and identify a novel myeloid cell subtype associated with treatment response. Overall, NIFA is a general approach advancing scRNA-seq analysis capabilities and it allows researchers to better take advantage of their data. NIFA is available at https://github.com/wgmao/NIFA.
Contact: mchikina@pitt.edu

1 Introduction

Single-cell RNA sequencing (scRNA-seq) techniques have allowed researchers to query the complexity of transcription regulation at an unprecedented level of detail. scRNA-seq technologies have the power to reveal both distinct cell types and transcriptional heterogeneity within a defined cell population. However, as individual transcript measurements are noisy and often difficult to interpret in isolation, scRNA-seq analysis methods rely heavily on multivariate techniques.

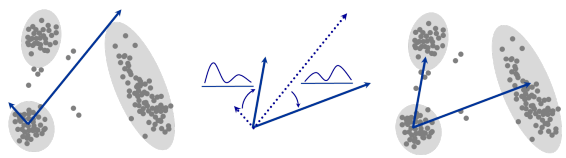

Fig. 1. Illustration of the NIFA. Left: We assume there are three hypothetical cell clusters and there are two latent components which point to arbitrary directions. Middle+Right: By imposing multi-modal prior, we force the latent factors to rotate and align with the directions that can best separate the cell-type identity.

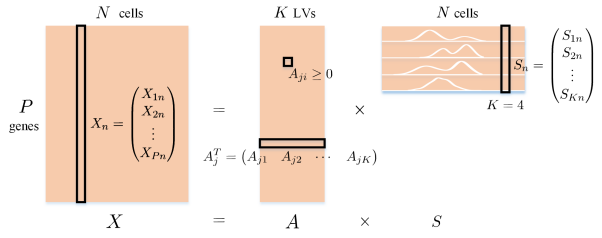

Fig. 2. A schematic representation of the NIFA model. The gene × cells matrix X is decomposed as a non-negative loading matrix A and a factor matrix S. We impose multi-modal priors on the rows of S, but the exact number of modes is automatically determined and thus can be one.

As the the number and size of single-cell datasets increases, it becomes important to develop methods that can quickly summarize the *biological* information embedded in a scRNA-seq dataset as a set of interpretable variables that can be used for downstream analysis. One kind of summary measure is the identity and number of cell types present in a datasets. In recent years there has been a proliferation of clustering methods designed to address this problem (Kiselev *et al.*, 2017; Butler *et al.*, 2018). Clustering approaches assume that the data is well described by a discrete set of cell types, but in many cases, questions about continuous biological variation, such as developmental trajectories or levels of pathway activation are also of interest.

Such continuous variables do not conform to the assumptions of clustering algorithms but can be effectively modeled as latent factors. For example, cell-cycle variation has been repeatedly discovered in single-cell data, both using sophisticated latent variable models (Buettner *et al.*, 2015) and simple Principle Component Analysis (PCA) (Kowalczyk *et al.*, 2015).

Of course, cell-type identity can also be thought of as a latent factor and this observation underlies the popularity of Independent Component Analysis (ICA) in single-cell pipelines. Unlike PCA which seeks directions that maximize variance, ICA finds maximally independent or maximally non-Gaussian directions (Hyvärinen and Oja, 2000).This property is well suited for the analysis of single-cell datasets as directions that maximally separate cell types are multi-modal and thus highly non-Gaussian. For this reason ICA is used as dimensionality reduction pre-processing step (Butler *et al.*, 2018). However, the ICA formulation is not a proper likelihood framework as it has no reconstruction error. A side effect of this is that it requires a loading orthogonalization step to prevent latent variables from collapsing. This rigid formulation restricts the interpretability of individual components a criticism is also valid for PCA/SVD. For the case of PCA/SVD, there are a number of alternative factor analysis methods that produce more interpretable components by relaxing orthogonality and introducing additional constraints, for example, NMF (Lee and Seung, 1999) and SPC (Witten *et al.*, 2009). It is natural to ask if analogous approaches can be applied to find interpretable multi-modal factors.

We propose **N**on-negative **I**ndependent **F**actor **A**nalysis (NIFA) that combines properties of ICA, PCA and NMF. As illustrated in Fig. 1, our approach simultaneously models uni- and multi-modal factors thus isolating discrete cell-type identity and continuous pathway-level variations into separate components. Furthermore, our model constrains the factor loading to be non-negative providing greater biological interpretability.

2 Methods overview

2.1 The statistical model

X represents a scRNA-seq matrix with dimension P-by-N, where P is the number of genes and N is the number of cells. Given X, we want to infer A which denotes loading matrix with dimension P-by-K and S which stands for sources or latent variables with dimension K-by-N. We denote the n_{th} column of X as $X_n = (X_{1n}, X_{2n}, \ldots, X_{Pn})^T$, the j_{th} row of A as $A_j^T = (A_{j1}, A_{j2}, \ldots, A_{jK})$ and the n_{th} column of S as $S_n = (S_{1n}, S_{2n}, \ldots, S_{Kn})^T$ (see Fig. 2). We assume the noise model to be Gaussian with a single precision parameter β.

$$X_n | A, S_n, \beta \sim N(0, \Sigma), \Sigma^{-1} = \text{diag}(\beta)_{P \times P} \tag{1}$$

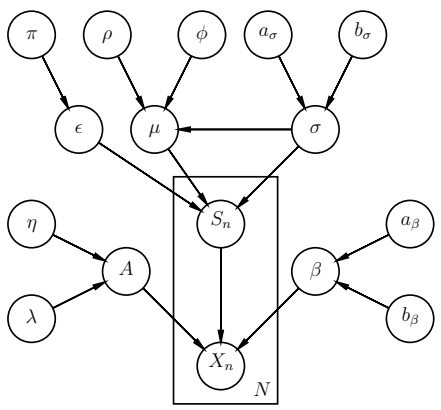

Fig. 3. Parameters of the NIFA model are summarized in a directed acyclic graph.

2.2 The prior distribution

Each latent variable is associated with M component distributions which we assume follows a Gaussian distribution with μ_{im} as the mean and σ_{im} as the inverse of the variance. ϵ_{imn} is a set of binary latent variables, and $\sum_{m=1}^{M} \epsilon_{imn} = 1$. If S_{in} is generated from component j, then $\epsilon_{imn} = 1$ if $m = j$ and $\epsilon_{imn} = 0$ if $m \neq j$.

$$P(S_n | \epsilon, \mu, \sigma) = \prod_{i=1}^{K} \prod_{m=1}^{M} N(S_{in} | \mu_{im}, \sigma_{im})^{\epsilon_{imn}} \tag{2}$$

The loading matrix A is modelled with a truncated normal prior where $a = 0$ and $b = \infty$ indicating each entry A_{ji} falls within the interval $[0, \infty)$. η and λ denotes the mean and the inverse of the variance. $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

$$\begin{aligned} &P(A_{ji} | \eta_{ji}, \lambda_i) \\ =&(2\pi)^{-\frac{1}{2}} (\lambda_i)^{\frac{1}{2}} \exp(-\frac{1}{2} \lambda_i (A_{ji} - \eta_{ji})^2) \cdot \frac{1(A_{ji} \geq 0)}{1 - \Phi(-\eta_{ji} \lambda_i^{\frac{1}{2}})} \end{aligned} \tag{3}$$

The dependency structure of the IFA model is summarized in Fig. 3. We assume the noise parameter β comes with a Gamma prior with parameter

a_β and b_β. The membership indicator ϵ_{imn} comes with a Bernoulli prior with mixing proportion π_{im}. The μ_{im} is assumed to follow a Gaussian distribution with parameters ρ_{im} and ϕ_{im} and the inverse of the variance σ comes with a Gamma distribution with parameters $a_{\sigma_{im}}$ and $b_{\sigma_{im}}$.

2.3 Parameter Inference

The joint likelihood $P(X, A, S, \epsilon, \mu, \sigma, \beta)$ is as follows (Eq. 4).

$$P(X, A, S, \epsilon, \mu, \sigma, \beta) = \prod_{n=1}^{N} P(X_n|A, S_n, \beta)P(S_n|\epsilon, \mu, \sigma)$$
$$\prod_{j=1}^{P}\prod_{i=1}^{K} P(A_{ji}|\eta_{ji}, \lambda_i) \prod_{i,m,n} P(\epsilon_{imn}) \prod_{i,m} P(\mu_{im})P(\sigma_{im})P(\beta) \tag{4}$$

In order to efficiently infer the parameters, we apply variational inference technique, more specifically, mean-field approximation (Blei *et al.*, 2017). By assuming each variational parameter is independent of each other, we formulate the joint posterior distribution $Q(S, A, \epsilon, \mu, \sigma, \beta)$ (see Eq. 5) for the model and minimize the KL-divergence between Eq.4 and Eq.5 to derive the expression $q(\cdot)$ for each variational parameter as an approximation of single posterior distribution.

$$Q(S, A, \epsilon, \mu, \sigma, \beta) = \prod_{n=1}^{N} q(S_n) \cdot \prod_{j,i} q(A_{ji}) \cdot \prod_{i,m,n} q(\epsilon_{imn})$$
$$\cdot \prod_{i,m} q(\mu_{im})q(\sigma_{im}) \cdot q(\beta) \tag{5}$$

The derivations of variational updates for our model are detailed in the Supplement.

2.4 Hyper-parameters

Our model has a number of hyper-parameters, however, most of them are Bayesian priors and have relatively little impact on the results. One of the main hyper-parameters of considerable relevance is the number of latent factors K. For the independent factor analysis model, the typical approach is to calculate the likelihood or ELBO (variation lower bound), comparing values directly or with BIC criterion (Krumsiek *et al.*, 2012) and selecting K corresponding to the optimal values. Since scRNA-seq data often has thousands of cells the computation for likelihood-based or ELBO-based tuning method is time-intensive and impractical. Instead we can rely on variance-based method SVD with BIC criterion (Allen *et al.*, 2014) to figure out a conservative estimate of the number of latent factors. Importantly we use this only as a reference value, we perform all our evaluations across a range of K parameters.

We have one more discrete hyper-parameter which is the number of Gaussian mixtures M for each latent factor. However, this parameter needs to be just the maximum number of components one can expect to find. Since our model fits the Gaussian mixtures by variational inference, it has the desirable property that the number of mixture components is determined automatically as some of the mixing coefficients go to 0. In experiment we set this hyper-parameter to be be 4 and find that for non-developmental datasets, where we expect to find discrete cell types, the final number of modes is usually either one or two. This conforms to the biological intuition that cell types differ from each other by a set of (not necessarily unique) "marker" genes. Such marker genes typically have a bimodal distribution corresponding to high and low expression. While the distributions may overlap due to technical noise we typically do not observe intermediate expression modes.

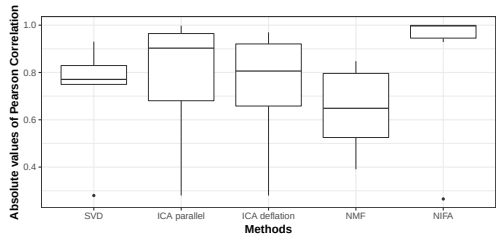

Fig. 4. Evaluation on a simulated dataset. Boxplot of the correlation between simulated S and those recovered by SVD, ICA (parallel), ICA (deflation), NMF and NIFA. We find that the best performance is achieved by NIFA compared with all the other common methods.

3 Results

3.1 Simulation Studies

We simulate data with $P = 2000$, $N = 500$, $K = 6$ (number of factors) and $M = 2$ (number of mixtures associated with each factor). We simulate the latent factor independently but the columns of the loading matrix A are correlated. Simulation details are described in section 5.3. For the decomposition, we set $K = 8$ (all methods) and $M = 3$ to see if NIFA can robustly recover the right latent factors given larger K and M, which is often the case in practice. As shown in Fig. 4, none of the methods can recover all latent factors since the loadings are highly correlated. But NIFA is able to accurately recover most of latent factors compared with alternative decomposition methods. NIFA also correctly recovers the number of mixture components as one of the mixing coefficients goes to 0 (not shown).

3.2 Datasets included

We test NIFA on several gold or silver standard scRNA-seq dataset (Gong *et al.*, 2018). The gold-standard dataset contains relatively homogeneous cell lines or the experimental conditions are well controlled. The silver-standard dataset defines cell types based on expert knowledge. The data is mostly downloaded through (https://github.com/hemberg-lab/scRNA.seq.datasets) or corresponding GEO repository. We also include simulated datasets generated by Splatter (Zappia *et al.*, 2017) using Kumar (Kumar *et al.*, 2014) and Zheng (Zheng *et al.*, 2017) as simulation input. The complete sets of datasets are described in Table 1.

3.3 Evaluation

There is a number of ways to evaluate factor analysis models. One natural evaluation is the reconstruction error (see Levitin *et al.* (2019) for example). However, for any decomposition there are infinitely many alternatives with exactly the same reconstruction error, yet these may differ greatly with respect to the individual factors and loadings. Instead of reconstruction error we focus on evaluating the biological utility in several different ways.

3.3.1 Cell-type identification

One of the desirable properties of an interpretable factor analysis is that there is one-to-one correspondence between the factors and a known data generating variable. In the case of scRNA-seq data the gold or silver standard of cell-type identity is one such variable. In the ideal case each cell-type corresponds to a unique factor in the model. In order to evaluate this property we compute maximum one-to-one correlations between factors and cell-type assignments (Fig. 5). We find that on average our NIFA model performs better than NMF and ICA at this cell-type detection

Abbreviation	Protocol	Evidence	Type	Tissue	Cells	Cell Types
Camp (Camp *et al.*, 2017)	SMARTer	Homogeneous cell line	Gold	Human: Liver	777	7
ImmunoTherapy (Sade-Feldman *et al.*, 2018)	Smart-Seq2	k-means clustering	Silver	Human: Metastatic melanoma	16,291	11
Klein (Klein *et al.*, 2015)	inDrop	Principal genes identified by PCA	Silver	Mouse: Embryonic Stem Cells	2,717	4
Kolodziejczyk (Kolodziejczyk *et al.*, 2015)	SMARTer	Homogeneous cell line	Gold	Mouse: Embryonic Stem Cells	704	3
Li (Li *et al.*, 2017)	SMARTer	Homogeneous cell line	Gold	Human: Colorectal Tumors	561	9
Liu (Liu *et al.*, 2019)	10x drop-seq	k-means clustering & specific markers	Silver	Mouse: Tumor immune cells	1,607	13
Nestorowa (Nestorowa *et al.*, 2016)	Smart-Seq2	hierachical clustering & specific markers	Silver	Mouse: Hematopoietic Stem Cells	1,656	9
Olsson (Olsson *et al.*, 2016)	SMARTer	Flow cytometry & cell sorting	Gold	Mouse: Hematopoietic Stem Cells	382	4
SimKumar4easy (Duò *et al.*, 2018)	NA	Simulated	Gold	NA	500	4
SimKumar4hard (Duò *et al.*, 2018)	NA	Simulated	Gold	NA	499	4
SimKumar8hard (Duò *et al.*, 2018)	NA	Simulated	Gold	NA	499	8
Zhengmix4eq (Duò *et al.*, 2018)	NA	Simulated	Gold	NA	3,555	4
Zhengmix4uneq (Duò *et al.*, 2018)	NA	Simulated	Gold	NA	6,414	4
Zhengmix8eq (Duò *et al.*, 2018)	NA	Simulated	Gold	NA	3,971	8

Table 1. The complete set of datasets evaluated in this study. Gold datasets are those where cell-types are determined due to the experimental design (for example by sorting cells). Silver datasets are those where cell-types were assigned from the data using biological prior knowledge.

task. Importantly, while there can be large differences between NMF and ICA the performance of NIFA (which combines features of both) always tracks with the best method.

3.3.2 Pathway enrichment

Of course, one important feature of factor analysis models is that the factors should be interpretable even in the absence of any ground truth knowledge. In such cases the factors are interpreted by inspecting the genes in their loading. The expectation is that for a factor that captures a unique biological variable (which could be binary cell-type or continuous pathway activation) the top loading genes are enriched for a few known functional modules. We evaluate this property by computing pathway enrichment for each factor as a hypergeometric test with the top 500 genes as foreground. This evaluation strategy allows us to evaluate the general biological validity of the model, independently of cell-type annotation. In this way the model can be credited for finding factors which capture pathway or cell-type signals even if these do not correspond to an annotated cell type. The pathway databases we use are "canonical pathways" from MsigDB (Liberzon *et al.*, 2011) and a comprehensive set of cell-type markers from xCell (Aran *et al.*, 2017). For canonical pathways we excluded pathways that had greater than 20% overlap with ribosomal or mitochondria genesets (defined as "KEGG_RIBOSOME" and "KEGG_OXIDATIVE_PHOSPHORYLATION" respectively). We found that these are consistently enriched but provide little biological insight as variation is these pathways is often technical.

We then quantify the overall biological enrichment of a single loading vector as the mean fold enrichment for pathways that are significant at FDR<0.05. Pathway enrichment metrics summarized across all factors are plotted in in Fig. 6 and Fig. 7 for canonical pathways and xCell respectively. We find that not surprisingly the performance of all methods is much better for real biological datasets than simulated ones (Zhengmix4eq, Zhengmix4uneq, Zhenmix8eq). We also find that among the biological datasets NIFA is a consistently top performer in both "canonical pathway" and xCell evaluations, though the effect is more dramatic for xCell.

3.4 In-depth evaluation of the Sade-Feldman *et al.* immunotherapy dataset

Antibodies that block immune checkpoint proteins, including CTLA4, PD-1, and PD-L1 are increasingly used to treat a variety cancers. While checkpoint inhibitor (CI) therapy can be remarkable effective not all patients respond (Larkin *et al.*, 2015). Determining the biological factors that facilitate or impede response to CIs remains an important and unresolved problem

In order to demonstrate how NIFA can be used to gain biological insight we performed several in-depth analyses of the Sade-Feldman *et al.* immunotherapy dataset. This dataset consists of 16,291 individual immune cells from 48 tumor samples of melanoma patients treated with checkpoint inhibitors. The dataset contains both pre-treatment and post-treatment samples and the patients are classified into responders and non-responders.

We applied our NIFA model to the entire single-cell dataset using K=25 which corresponds to the k with maximal correlation with known cell-type annotations (see Fig. 5). The distributions of the inferred factors and the corresponding inferred Gaussian mixture fits are plotted in Fig. 8. Each NIFA factor that has the best correspondence to human annotations is given the same name. NIFA finds both uni- and multi-modal factors and as expected the multi-modal factors are more likely to correspond to cell types.

In order to investigate which variables are associated with immunotherapy response the resulting factors were mean aggregated to a single value for each unique patient sample. We also summarized the human annotated cell-type indicators as their mean values, corresponding to fraction of cells in sample. The resulting summary statistics were tested for association with response using Wilcoxon ranksum and Benjamini-Hochberg FDR adjusted (separately for NIFA factors or human annotations). Pre and post-treatment samples were analyzed separately and the resulting variables that were significant in either the pre-treatment or post-treatment comparison at FDR<0.2 are plotted in Fig. 9A. Top loading genes corresponding to each significant NIFA factor are show in Table 3.4.

Each NIFA factor that has the best correspondence to human annotations is given the same name and the results are grouped with grey ovals in Fig. 9. We find that for these matched variables there is an overall good correspondence between the results of NIFA factors and human cell-type annotations. Specifically, both methods discover B-cells as the variable most positively predictive of response and a CD8 T-cell/exhaustion/cell-cycle signature (termed "Lymphocytes exhausted/cell-cycle" in the original study) as the most negatively predictive.

For some subtle pattern the results of NIFA and human annotations can diverge. Human annotations such as "Lymphocytes" and "Cytotoxicity" were not well reproduced by NIFA (correlation of 0.46 and 0.46 respectively) and the corresponding NIFA variables are not significant. On the other hand, NIFA found three different T-cell signatures (8, 19 and 20) which were all associated with the "memory T-cell" human annotation and all were significantly predictive of response. One of these signatures has TCF7 as a top loading gene and thus NIFA was able to independently discover one of the key findings of the original study – that the fraction of TCF7 positive T-cells is highly associated with response.

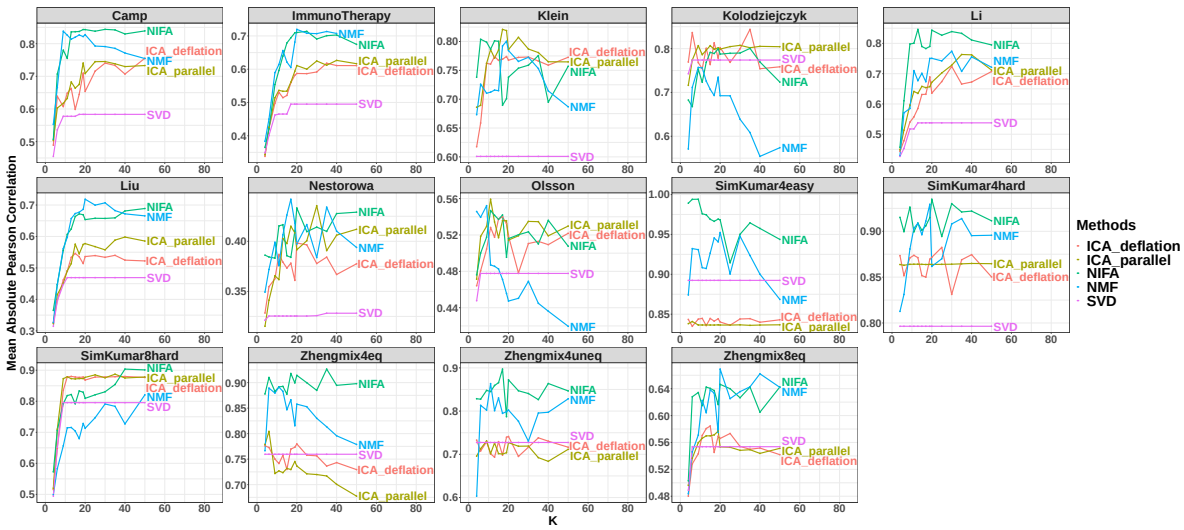

Fig. 5. Evaluation of one-to-one correspondence between factors and cell-types. Given a set of factors and a set of cell-type labels we evaluate the maximum correlation between each cell-type and a factor. For clarity, we plot the mean correlation value across all cell-types. In order, to account for the possibility that different models may need different number of factors (K) we report the results at varying K. We compare NIFA with ICA, NMF (KL-loss), and SVD as a baseline. We find that on average our NIFA model performs better than NMF and ICA and importantly while there can be large differences between NMF and ICA the performance of NIFA (which combines features of both) always tracks with the best method.

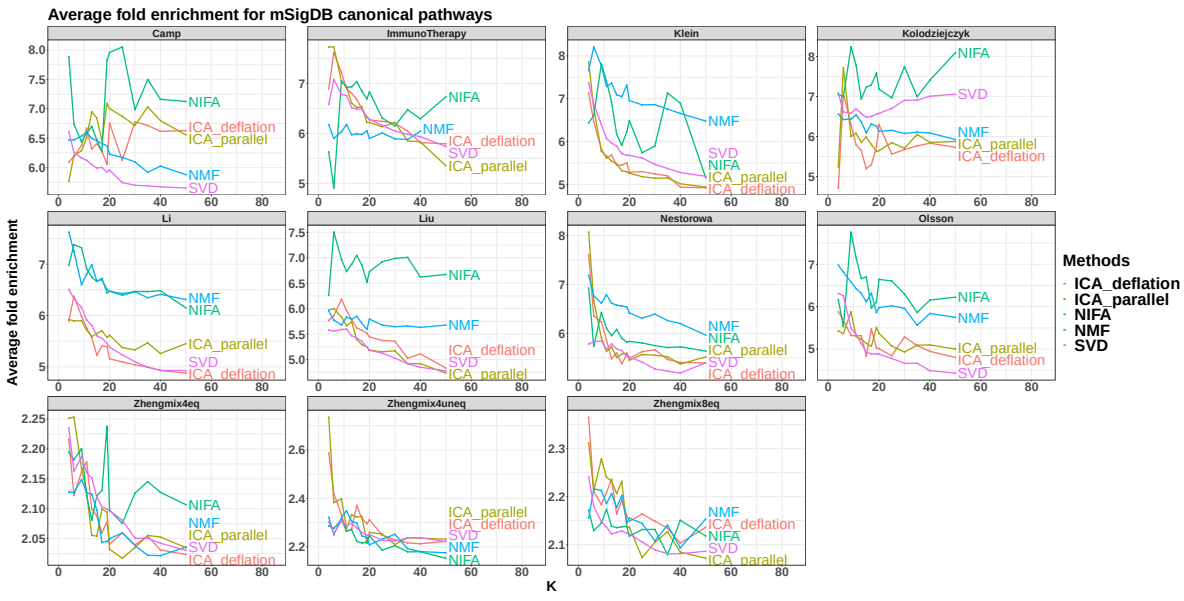

Fig. 6. Pathway enrichment of "canonical pathways" from mSigDB (Liberzon et al., 2011). Enrichment is quantified as average fold enrichment among all factor-pathways pairs where the pathway is significantly over-represented in the top 500 loading genes (hypergeometric test, FDR<0.05). The first two rows are biological datasets. The last row (Zhengmix4eq, Zhengmix4uneq and Zhenmix8eq) are simulated datasets. All SimKumar datasets are excluded from this evaluation as they were not supplied with real gene names.

Aside from generally reproducing the main findings of the original study NIFA was able to uncover additional patterns. For example, we find that presence of Tregs is negatively associated with response in the post-treatment samples. The corresponding human annotation is however not significant despite the fact that the two variables are highly correlated (Pearson correlation = 0.71). Human regulatory T-cells are difficult to identify from a transcriptional profiles. There are no genes that are *unique* to this cell-type. The canonical transcription factor (FOXP3) and surface marker (CD25/IL2RA) can also be transiently expressed by non Treg CD4 cells (Chen and Oppenheim, 2011); on the other hand, because of noise in scRNA-seq data the absence of these markers doesn't exclude Treg status. Upon closer inspection, we find that NIFA is more conservative in designating Tregs than the human annotation counterpart. Using the NIFA mixture components we can perform a hard cell-type assignment based on the probability of being in the high-expression component being >0.5. Using this cutoff, NIFA finds on 1,418 Tregs, in contrast to 1,740 of human annotated ones. We find that these discrepancy is highly non-random and

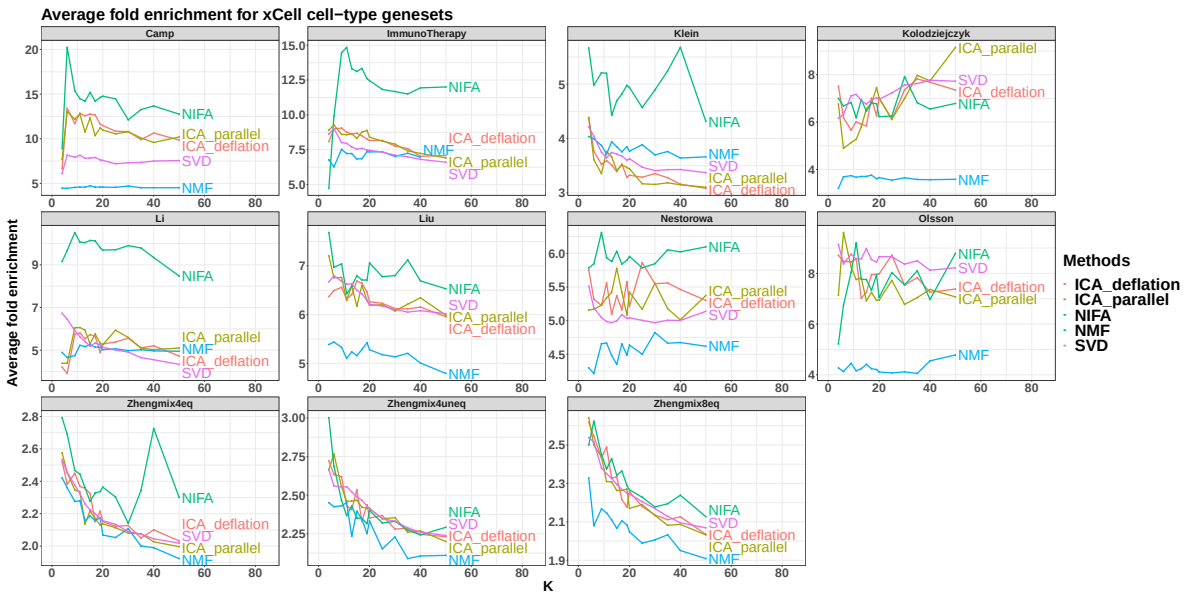

Fig. 7. Pathway enrichment for xCell cell-type signatures. This figure is generated identically to Fig. 6 but using the xCell genesets.

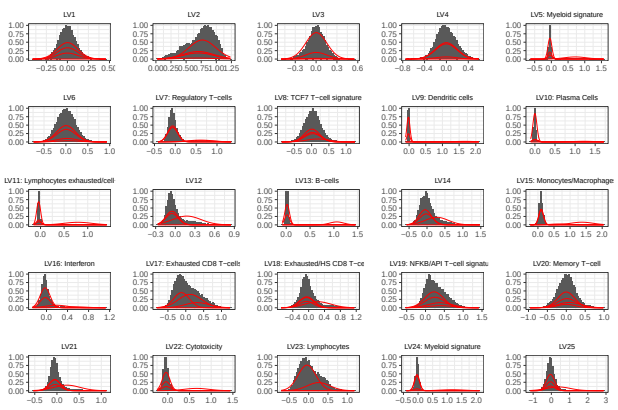

Fig. 8. Factor histograms and mixture model fits for the Sade-Feldman et al. immunotherapy dataset. Factors that best correspond to cell-types identified in the original study are labeled accordingly. NIFA finds both multi-modal and unimodal factors and as expected the multimodal factors are more likely to represent cell-types

that the NIFA Tregs are more likely to express both FOXP3 and IL2RA (Fig. 9C) indicating that the NIFA Treg signature is more specific.

Overall, within this dataset a large number of the human annotated cell-types and NIFA factors are associated with response but some general patterns emerge. Specifically, the presence of myeloid cell-types is negatively associated with response while presence of lymphocytes, exclusive of those with an exhaustion-like phenotype (for example B-cells, CD4 memory cells), is positively associated with response (see Fig. 9A). The general trend that a high myeloid to lymphocyte ratio is associated with worse outcome is observed across a variety of cancers (Thorsson *et al.*, 2018). Our NIFA based analysis however finds a myeloid signature (NIFA latent factor 24) that correspond to a subset of annotated "Monocytes/Macrophages" cells is *positively* associated with response, with an effect size that is similar to the lymphocyte populations (Fig. 9 A).

This myeloid subset is identified by high levels of metallothionein genes (MT1X, MT1F, MT1E and MT2A) and some metabolic genes (see Fig. 9B). Metallothioneins are a family of small proteins that play important roles in metal homeostasis and protection against heavy metal toxicity, DNA damage and oxidative stress (Si and Lang, 2018). Their induction in tumor-associated macrophages (TAMs) has been noted (Ge *et al.*, 2012) but to our knowledge this is the first report of an association with clinical outcome.

4 Discussion and Conclusion

We propose a factor analysis model designed specifically for single-cell data. The model combines features of PCA, ICA and NMF. Specifically, our model optimizes the PCA-like matrix reconstruction objective with mixture of Gaussians priors on the factors which encourages decomposition along multi-modal directions. We also adopt truncated Gaussian priors on the loadings thus imposing an NMF-like strict non-negativity constraint. Using a variational Bayes framework allows us to automatically fit hyper-parameters such as the number of Gaussian mixtures. We evaluate our model using both known cell identity and pathway information and demonstrate that NIFA generates biologically coherent factors that align well this prior knowledge.

One additional feature of our model is that the fully Bayesian framework is readily extensible. For example, it easily supports gene-specific priors for the loadings. This makes it possible to use known biological pathways as additional constraints. We plan on developing this extension in our future work.

5 Method Details

5.1 Preprocessing Pipeline

The data preprocessing pipeline is illustrated in Fig. 10. Some preprocessing steps were only applied to certain methods. For example, we employed SVD smoothing (that is reconstructing the input as a truncated SVD with rank=50) because it makes the NIFA constant variance Gaussian

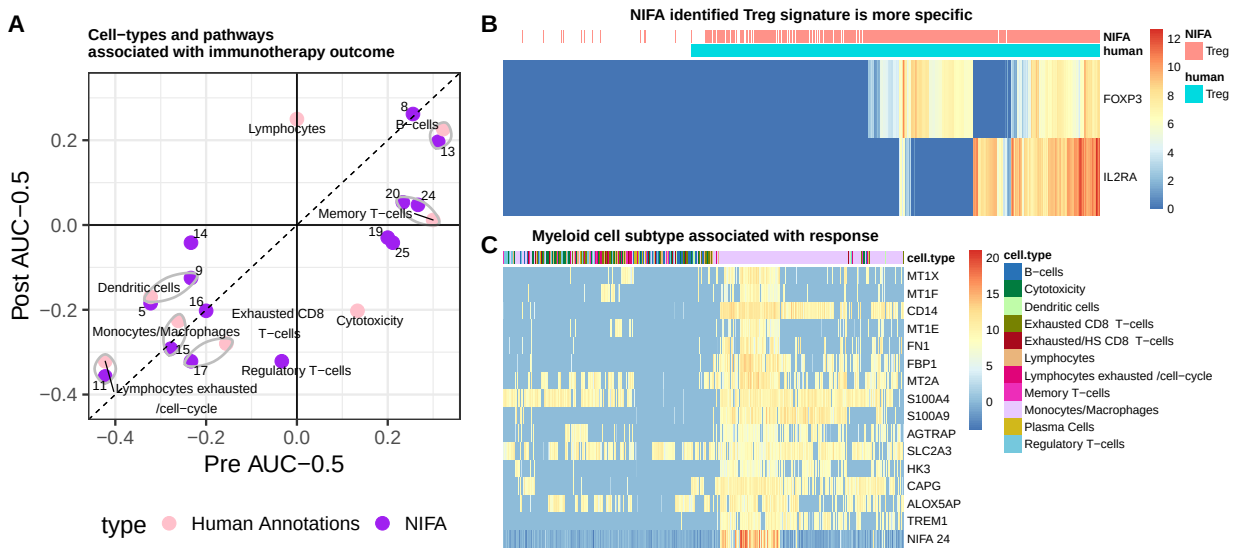

Fig. 9. NIFA analysis of signatures associated with immunotherapy response. (A) NIFA-derived signatures of human-annotated cell-types are mean aggregated per patient sample and the resulting summary statistics are tested for association with immunotherapy response. Variables that are significant at FDR<0.2 are shown with their respective normalized and centered ranksum statistics (ranksum/(number-of-positives × number-of-negatives)-0.5, equivalent to binary classification AUC-0.5). Pre-treatment effects are on the x-axis and post-treatment effects are on the y-axis. NIFA variables most closely matched to a human annotation are grouped with grey ellipses. (B) Differences in Treg (Regulatory T-cells) identification between NIFA and human annotations. Heatmap of canonical Treg marker genes (FOXP3 and IL2RA) across all cells annotated as Tregs by either method and 1000 randomly sampled other T cells. Overall, NIFA identifies fewer Treg cells and has a higher correlation with FOXP3 and IL2RA expression. While the NIFA Treg factor is significantly negatively associated with response in post-treatment samples, the corresponding human annotation is not (panel A). (C) A new myeloid signature positively associated with response. Heatmap of top loading genes along with the factor values for NIFA factor 24 across all cells identified as "Monocytes/Macrophages" and 1500 randomly sampled cells. NIFA identifies a subset of the Monocytes/Macrophages calls with unique gene expression. While general myeloid signatures (that is Monocytes/Macrophages and Dendritic cells) were negatively associate with response, the NIFA-24 signature has the opposite pattern (see panel A).

ID	name	genes
5	Myeloid signature	FCN1, LYZ, TIMP1, S100A9, S100A8, SERPINA1, VCAN, IL1B, IFI30, PLAUR
7	**Regulatory T-cells**	CTLA4, TNFRSF18, RGS1, TIGIT, CD4, BATF, PIM2, PRDM1, FOXP3, ARID5B
8	TCF7 T-cell signature	DGKA, DDX17, SMG1P1, DENND2D, ARHGEF1, DOCK8, NPIPB5, NLRC5, TCF7, N4BP2L2
9	**Dendritic cells**	GZMB, IGJ, PLAC8, NAPSB, ALOX5AP, GPR183, AC096579.7, IRF7, BCL11A, CLIC3
11	**Lymphocytes exhausted/cell-cycle**	STMN1, RRM2, TUBA1B, TYMS, KIAA0101, TUBB, HIST1H4C, HMGB2, NUSAP1, CDK1
13	**B-cells**	CD74, IGHM, MS4A1, CD79A, IGKC, IRF8, CD79B, CD37, BCL11A, CD52
14		HSPD1, FLNA, BIRC3, REL, HSPE1, COTL1, WARS, PSME2, HSPB1, SLC25A3
15	**Monocytes/Macrophages**	CD74, FTL, CTSB, B2M, FTH1, PSAP, S100A11, IFI30, VIM, ALDOA
16	Interferon	ISG15, IFI44L, MX1, IFI6, XAF1, STAT1, ISG20, IFITM1, TRIM22, IFI44
17	**Exhausted CD8 T-cells**	GZMA, RAC2, CLIC1, NKG7, CORO1A, IL32, ARPC1B, CNN2, LCK, PSMB9
19	NFKB/AP1 T-cell signature	VIM, NFKBIA, FOS, TNFAIP3, ANXA1, SLC2A3, CD52, B2M, JUNB, S100A4
20	**Memory T-cells**	EEF1B2, GAS5, TOMM7, LDHB, SELL, IL7R, EIF3E, COX7C, EIF3L, FAIM3
24	Myeloid signature	MT1X, MT1F, CD14, MT1E, FN1, FBP1, MT2A, S100A4, S100A9, AGTRAP
25		C1QBP, NME1, HSP90AB1, GTF3A, NHP2, PPA1, CCT7, CNBP, CCT2, SNHG1

Table 2. Top loading genes for each factor found to be associated with immunotherapy response. To facilitate biological interpretability ribosomal genes and genes that we not provided with HGNC symbols were not considered. Factors that best match the known human annotations were named accordingly and are in bold. Other factors could be clearly identified as coherent biological pathways based on the loadings. Factors 14 and 25 did not have a clear correspondence to any pathways or cell-type signatures and are left unnamed.

error assumption more valid. However, we found that empirically this had little effect on the results. For NMF we used KL divergence (or equivalently a Poisson error model) which is most appropriate for unsmoothed data. For NIFA we chose to z-score the input by row (gene). The z-scoring operation theoretically makes it easier for to pick up on small variance but highly deferentially expressed genes and it produced modest improvement in most (though not all) benchmark datasets. Row z-scoring was not applied to NMF as it produces negative numbers. It was also not applied to ICA as the ICA objective function references only the shape of the factor distribution and thus is invariant under row scaling.

5.2 NIFA initialization

NIFA updates are relatively expensive and thus a good initialization can significantly affect running time. We initialize NIFA with a simple matrix decomposition with non-negativity constraints on the loadings. Specifically, we initialize with a solution to a simpler optimization

problem.

$$\min_{A,S} ||X - AS||_F + \lambda_1 ||A||_F + \lambda_2 ||S||_F \qquad \text{subject to} A \geq 0 \quad (6)$$

This problem can be solved quickly by alternating least squares.

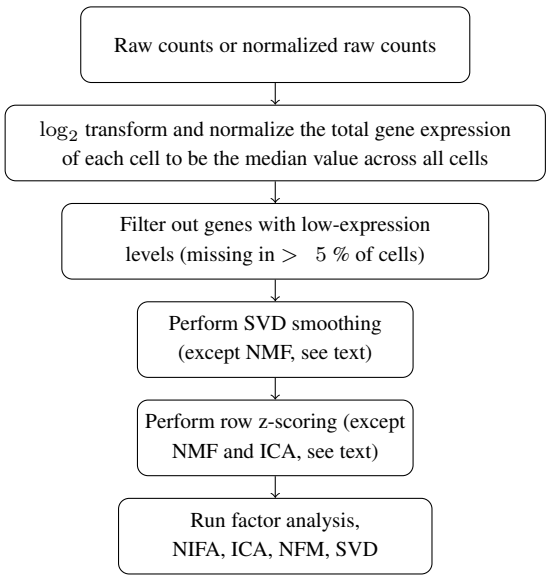

Fig. 10. Preprocessing workflow.

5.3 Simulation Details

The dimension of the simulated matrix X is set to be 2000-by-500 (gene-by-sample). There are 6 latent factors, each of which contains 2 Gaussian mixtures. The dimension of loading matrix A is 2000-by-6 with each column corresponding to a single loading and the matrix S is 6-by-500 with each row corresponding to a single latent factor. We draw the first loading vector from Gamma distribution $\Gamma(5, 1)$. Then the subsequent loadings are simulated by adding noise following Gaussian distribution $N(0, 2)$ to the first simulated loading. We take the absolute values of noise to make sure loadings are kept positive. In this way, we can control the collinearity to simulate correlated loadings. Each latent factor is generated hierarchically. First we draw mean and variance parameters for the first Gaussian mixture associated with each latent factor from Gaussian distribution $N(2, 10)$ and Gamma distribution $\Gamma(10, 1)$ correspondingly. Regarding each latent factor the rest of mixtures are generated by adding noise drawn from a uniform distribution $U(2, 4)$ to the mean of the first mixture. Then each entry in the latent factor is assigned to any of the mixtures with probability and the exact value is drawn based on the distribution of assigned mixture. Finally X is generated as the sum of AS and noise drawn from Gaussian distribution $N(0, 0.1)$. In order to generate non-negative NMF input we offset the matrix by a minimum constant $c = \min(C)$ which makes the result $X + c$ non-negative.

5.4 Evaluation

We compare NIFA with NMF (Gaujoux, 2018) and ICA (fastICA implementation, Marchini *et al.* (2019)). For NMF we used KL loss which we found dramatically outperformed the square loss alternative.

Cell-Type Correspondence We compute the absolute Pearson correlations between each cell type and decomposed factors and we annotate the factor with maximum absolute correlation with the corresponding cell type.

Pathway Enrichment We perform a hypergeometric test on each of the loading with the top 500 genes as foreground and the rest as background. For SVD and ICA, we use the absolute values of loadings to perform the test. The p values are adjusted with Benjamini-Hochberg procedure and we denote pathways with adjusted p-val < 0.05 as significant enriched pathways. The pathway enrichment is summarized as average fold enrichment across the significant loading-pathway associations.

Funding

Conflict of Interest: none declared.

References

Allen, G. I. *et al.* (2014). A generalized least-square matrix decomposition. *Journal of the American Statistical Association*, **109**(505), 145–159.

Aran, D. *et al.* (2017). xcell: digitally portraying the tissue cellular heterogeneity landscape. *Genome biology*, **18**(1), 220.

Blei, D. M. *et al.* (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, **112**(518), 859–877.

Buettner, F. *et al.* (2015). Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, **33**(2), 155.

Butler, A. *et al.* (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, **36**(5), 411.

Camp, J. G. *et al.* (2017). Multilineage communication regulates human liver bud development from pluripotency. *Nature*, **546**(7659), 533–538.

Chen, X. and Oppenheim, J. J. (2011). Resolving the identity myth: key markers of functional cd4+ foxp3+ regulatory t cells. *International immunopharmacology*, **11**(10), 1489–1496.

Duò, A. *et al.* (2018). A systematic performance evaluation of clustering methods for single-cell rna-seq data. *F1000Research*, **7**.

Gaujoux, R. (2018). An introduction to nmf package. *URL: https://cran. r-project. org/package= NMF. R package version 0.20*, **6**.

Ge, Y. *et al.* (2012). Induction of metallothionein expression during monocyte to melanoma-associated macrophage differentiation. *Frontiers in biology*, **7**(4), 359–367.

Gong, W. *et al.* (2018). Drimpute: imputing dropout events in single cell rna sequencing data. *BMC bioinformatics*, **19**(1), 220.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, **13**(4-5), 411–430.

Kiselev, V. Y. *et al.* (2017). Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, **14**(5), 483.

Klein, A. M. *et al.* (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**(5), 1187–1201.

Kolodziejczyk, A. A. *et al.* (2015). Single cell rna-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell stem cell*, **17**(4), 471–485.

Kowalczyk, M. S. *et al.* (2015). Single-cell rna-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome research*, **25**(12), 1860–1872.

Krumsiek, J. *et al.* (2012). Bayesian independent component analysis recovers pathway signatures from blood metabolomics data. *Journal of proteome research*, **11**(8), 4120–4131.

Kumar, R. M. *et al.* (2014). Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature*, **516**(7529), 56–61.

Larkin, J. *et al.* (2015). Combined nivolumab and ipilimumab or monotherapy in untreated melanoma. *New England journal of medicine*, **373**(1), 23–34.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**(6755), 788.

Levitin, H. M. *et al.* (2019). De novo gene signature identification from single-cell rna-seq with hierarchical poisson factorization. *Molecular systems biology*, **15**(2).

Li, H. *et al.* (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature genetics*, **49**(5), 708.

Liberzon, A. *et al.* (2011). Molecular signatures database (msigdb) 3.0. *Bioinformatics*, **27**(12), 1739–1740.

Liu, C. *et al.* (2019). Treg cells promote the srebp1-dependent metabolic fitness of tumor-promoting macrophages via repression of cd8+ t cell-derived interferon-γ. *Immunity*, **51**(2), 381–397.

Marchini, J. *et al.* (2019). Package 'fastica'.

Nestorowa, S. *et al.* (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood, The Journal of the American Society of Hematology*, **128**(8), e20–e31.

Olsson, A. *et al.* (2016). Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature*, **537**(7622), 698–702.

Sade-Feldman, M. *et al.* (2018). Defining t cell states associated with response to checkpoint immunotherapy in melanoma. *Cell*, **175**(4), 998–1013.

Si, M. and Lang, J. (2018). The roles of metallothioneins in carcinogenesis. *Journal of hematology & oncology*, **11**(1), 107.

Thorsson, V. *et al.* (2018). The immune landscape of cancer. *Immunity*, **48**(4), 812–830.

Witten, D. M. *et al.* (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**(3), 515–534.

Zappia, L. *et al.* (2017). Splatter: simulation of single-cell rna sequencing data. *Genome biology*, **18**(1), 174.

Zheng, G. X. *et al.* (2017). Massively parallel digital transcriptional profiling of single cells. *Nature communications*, **8**, 14049.