Bioinformatics, 33(2), 2017, 235–242
doi: 10.1093/bioinformatics/btw607
Advance Access Publication Date: 23 September 2016
Original Paper

Gene expression

Robust classification of single-cell transcriptome data by nonnegative matrix factorization

Chunxuan Shao[1,2,*] and Thomas Höfer[1,2,*]

[1]Division of Theoretical Systems Biology, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany and [2]Bioquant Center, University of Heidelberg, 69120 Heidelberg, Germany

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on March 16, 2016; revised on September 15, 2016; accepted on September 16, 2016

Abstract

Motivation: Single-cell transcriptome data provide unprecedented resolution to study heterogeneity in cell populations and present a challenge for unsupervised classification. Popular methods, like principal component analysis (PCA), often suffer from the high level of noise in the data.
Results: Here we adapt Nonnegative Matrix Factorization (NMF) to study the problem of identifying subpopulations in single-cell transcriptome data. In contrast to the conventional gene-centered view of NMF, identifying metagenes, we used NMF in a cell-centered direction, identifying cell subtypes ('metacells'). Using three different datasets (based on RT-qPCR and single cell RNA-seq data, respectively), we show that NMF outperforms PCA in identifying subpopulations in an accurate and robust way, without the need for prior feature selection; moreover, NMF successfully recovered the broad classes on a large dataset (thousands of single-cell transcriptomes), as identified by a computationally sophisticated method. NMF allows to identify feature genes in a direct, unbiased manner. We propose novel approaches for determining a biologically meaningful number of subpopulations based on minimizing the ambiguity of classification. In conclusion, our study shows that NMF is a robust, informative and simple method for the unsupervised learning of cell subtypes from single-cell gene expression data.
Availability and Implementation: https://github.com/ccshao/nimfa
Contacts: c.shao@Dkfz-Heidelberg.de or t.hoefer@Dkfz-Heidelberg.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The measurement of gene expression at the single-cell level has emerged as a new approach to study cellular heterogeneity and cell-fate decisions. Quantitative PCR (RT-qPCR) is a sensitive method for quantifying mRNA transcripts and has been used in several single-cell studies. For example, single-cell RT-qPCR coupled with microfluidic chips was employed to study the cell fate decisions from zygote to blastocyst in the mouse (Guo *et al.*, 2010), and the immune response of individual CD8+ T lymphocytes was characterized with selected surface markers and transcriptional regulators (Arsenio *et al.*, 2014). Single-cell RT-qPCR is readily available and can be applied to comparatively large numbers of cells; however, only a limited number of genes can be profiled, requiring a prior choice of informative genes (Moignard *et al.*, 2015). Single-cell transcriptomics by next generation sequencing (single-cell RNA-seq) has been developed in the past years and has already become a popular tool to study cell-to-cell heterogeneity. Applications include the classification of thousands of cells from mouse spleen (Jaitin et al., 2014) or brain (Zeisel *et al.*, 2015) into distinct cell types as well as the reconstruction of lineage hierarchies and the identification of novel lineage-specific genes in lung epithelium (Treutlein *et al.*, 2014). These data pose new computational problems for statistical learning.

The identification of subgroups (e.g. cell types or functional cell states) from single-cell transcriptomes is an unsupervised classification problem. Principal component analysis (PCA) has been a popular tool here. However, the level of technical noise (due to mRNA capture efficiency, amplification bias, sequencing efficiency) is usually very high in these data, certainly much larger than for bulk-population measurements due to low amount of starting materials (Bacher and Kendziorski, 2016; Brennecke *et al.*, 2013; Grün *et al.*, 2014). Additionally, there is true biological variability of gene expression from cell to cell. As a consequence of technical and biological noise, the first few PCA components typically explain only a small fraction of the variance (often only a few percent even in the first component). Therefore, cell subpopulations identified through PCA projection on the first few dimensions may not be readily distinguished and remain ambiguous. Tandem PCA analyses, where genes are selected from a first PCA and then a second PCA is performed with this smaller gene set, have produced interesting results (Treutlein *et al.*, 2014). Nevertheless, bias could be introduced through feature selection. Clique-based classification methods with shared nearest neighbor similarity showed improved performance in identifying cell types in single cell transcriptome data (Xu and Su, 2015). Dimension reduction of the data by diffusion maps, which stresses continuity of cell states along putative developmental pathways, or by independent component analysis (ICA) have also been used (Haghverdi *et al.*, 2015; Trapnell *et al.*, 2014). Sophisticated methods that involve iterative clustering have been proposed to study subtype classification and the detection of relationships between the subtypes (Macosko *et al.*, 2015; Tasic *et al.*, 2016; Zeisel *et al.*, 2015). These multiple approaches illustrate that the unsupervised learning of biologically relevant features from single-cell gene expression data presents a formidable challenge.

Arguably the most common task is to decompose the high-dimensional transcriptome data into (biologically) interpretable parts. Among the various methods for decomposing multivariate data, Nonnegative Matrix Factorization (NMF) is distinguished by specifically aiming at detecting interpretable individual parts (localized features), rather than holistic eigenstates (global features) as in PCA (Lee and Seung, 1999). While both NMF and PCA are matrix decomposition techniques, they are constrained differently, with non-negativity of feature superposition for NMF and orthogonality of principal components for PCA. NMF decomposes a nonnegative data matrix (V) into two nonnegative matrices: a basis matrix (W) and a coefficient matrix (H). It thus interprets the data as a superposition of distinct parts, which are either absent in particular setting (entry 0) or present to a specific degree (entry positive). NMF has been successfully applied to mRNA microarray data for classification and signature identification (Brunet *et al.*, 2004; Carmona-Saez *et al.*, 2006; Nik-Zainal *et al.*, 2012). Here the data, represented as a genes-by-samples matrix, with a much larger number of genes than samples, have been described in terms of a handful of metagenes (groups of genes that behave similarly). It was already noted that there is a possible dual viewpoint to find ‘metasamples’ from the transposed samples-by-genes matrix instead of metagenes (Brunet *et al.*, 2004). We hypothesize that this alternative use of NMF on a samples-by-genes matrix might serve to find natural groupings of cells based on single-cell transcriptome data. Here, we explore this metasample perspective of NMF applied to single-cell transcriptome data measured by RT-qPCR and RNA-seq, and propose novel heuristic methods to estimate the appropriate number of metasamples (which, as we will argue below, might be better characterized as metacells). We find that NMF robustly and accurately detects functional cell subgroups while simultaneously guiding the identification of biologically relevant features in the data.

2 Methods

2.1 Nonnegative matrix factorization

Given a nonnegative $m \times n$ matrix V and positive rank r, NMF finds two nonnegative matrices W and H to approximate V as $V \approx WH$. A widely used error function is the L^2 norm $E(W, H) = \|V - WH\|^2$ (squared Euclidean distance). Assuming that W and H are available, each column v_i of V can be approximated as $v_i \approx Wh_i$; thus W is referred to as the basis matrix and H as the coefficient matrix. In this study, V is a samples-by-genes matrix, where m is the number of cells and n is the number of genes (note that this formulation of the problem can simply be viewed as the ‘transpose’ of the widely studied problem of finding metagenes). Several algorithms have been developed to perform NMF, with emphasis on sparseness or/and speed. Initially, Lee and Seung proposed the multiplicative-update algorithm, which we refer to as standard NMF (STNMF) (Lee and Seung, 1999). Lin proposed the projected-gradient approach to solve NMF from the perspective of alternating nonnegative least squares (referred to as LSNMF) and showed its improved performance compared to STNMF (Lin, 2007); Kim and Park enforced sparseness for both basis and coefficient matrices, obtaining better clustering results (Kim and Park, 2007). An overview of NMF algorithms is provided in Berry *et al.* (2007).

2.2 Data and software

Guo *et al.* (2010) explored cell fate decisions from zygote to blastocyst via single-cell gene expression measured by RT-qPCR. We downloaded the corresponding expression data on early embryonic development and removed one cell with duplicated name. Two ambiguously located cells in the PCA projection plot were removed as well. Finally, we constructed the 156 by 48 data matrix, in which each row is a cell and each column represents a gene (embryonic development data).

Single-cell transcriptome analysis yield important insights in the development of distal lung epithelium, where previous unknown markers were discovered (Treutlein *et al.*, 2014). We downloaded the log2 transformed single-cell RNA-seq transcriptome data for mouse embryonic day 18.5 lung epithelial cells and constructed a cells by genes (80 by 8772) matrix, filtering out genes with low variance and expression (complete lung epithelium data). Moreover, we constructed a reduced set of lung epithelium data, in which the genes in columns are proposed signature genes kindly provided by Barbara Treutlein (personal communication).

We analyzed a larger single-cell RNA-seq transcriptome data of 3005 cells, which was used to study the specialized cell types in mouse cortex and hippocampus (Zeisel *et al.*, 2015) (mouse cortex-hippocampus data). The corresponding UMI (unique molecular identifier)-count data was normalized to the library size and logarithm transposed with base 2, i.e. log2(count/library_size * 10000 +1). In addition, we filtered out genes expressed in less than 10% of cells, and with variance smaller than 0.125.

To perform NMF, we used the python module nimfa (Žitnik and Zupan, 2012) in which several algorithms and seeding methods are implemented. We modified this module to calculate the cophenetic correlation coefficient of the basis matrix.

2.3 Metacell classification

We applied NMF to the pre-processed data described above. The basis matrix W returned by NMF had size $m \times r$, while m denotes the number of cells and r is rank. For ease of interpretation, we normalized the basis matrix to make each row sum to unity and obtained $\bar{W}$. The normalized value in each row can be thought of as the nonnegative contribution of this metacell to the single cell in question, where higher values suggest larger contribution. For classifying the single cells we used the largest metacell contribution in $\bar{W}$. We constructed the confusion matrix by intersecting the metacell classification obtained via NMF against an independent subpopulation classification using traditional markers (see Section 3.1 for details). Generally, the LSNMF algorithm with the NNDSVD seeding method (Boutsidis and Gallopoulos, 2008) was used for NMF, if not otherwise specified, with default value for other parameters except rank. The choice of an appropriate rank r (number of metacells or cell types) is discussed in detail below.

To obtain the layout of single cells in two-dimensional space and compare with PCA, we performed nonmetric multidimensional scaling (NMDS) in R (R Core Team, 2015) using 1 – Pearson's correlation coefficient calculated from the normalized basis matrix as distance measure.

2.4 Feature selection

We normalized the $r \times n$ coefficient matrix to make columns sum to unity and obtained $\bar{H}$, where n is the number of genes. We view the normalized value as the relative expression of a gene in a metacell. Genes were assigned as subgroup-specific markers by largest relative expression. For the lung epithelium data, we selected feature genes based on the cutoff of relative expression ≥ 0.6 in $\bar{H}$ and mean $\log_2$RPKM ≥ 0.15 in the data matrix.

2.5 Rank estimation

The LSNMF algorithm with random seeding was applied to the embryonic development data and complete lung epithelium data to calculate the cophenetic correlation coefficients of the basis matrix for ranks from two to ten. We found that running NMF for 500 times yielded a reproducible optimized result. We applied the LSNMF algorithm with NNDSVD seeding to calculate the proportion of non-classified cells, entropy and cell sparseness for each of the three data for the given ranks of the normalized basis matrices. For each rank, the non-classified cells were identified as having the largest contribution in metacells smaller than a predefined cutoff (0.5, 0.6, 0.7 or 0.8). We calculated the cell sparseness for each cell as

$$\sqrt{(\sum_{j=1}^{r}(\bar{W}_{ij})^2)}$$

In order to correct for the effect of the discrete rank we normalized the raw cell sparseness by projecting to the [0, 1] interval via

$$(sparseness - a)/(1 - a)$$

where $a = 1/\sqrt{r}$. In addition, the entropy of each cell i was calculated as

$$-\sum_{j=1}^{r}\bar{W}_{ij}\log_2\bar{W}_{ij}$$

in which r is the rank value, and $\bar{W}$ is the normalized basis matrix. In a completely non-informative classification each metacell would have the same contribution in $\bar{W}$, thus the entropy is $\log_2$r. We defined the normalized information gain of a given rank r as

$$1 - \frac{1}{r\log_2 r}\left(-\sum_{i}^{m}\sum_{j}^{r}\bar{W}_{ij}\log_2\bar{W}_{ij}\right)$$

2.6 Seeding methods and algorithms comparison

We compared two different initiation methods for W and H: random and nonnegative double singular value decomposition (NNDSVD). The random seeding method generates the initial W and H from a uniform distribution, while the NNDSVD method has no randomization in generating initiation matrixes. NNDSVD contains two SVD processes approximating the target matrix and positive parts of resulting SVD factors, respectively (Boutsidis and Gallopoulos, 2008). We ran LSNMF on the complete and reduced lung epithelium data with different seeding methods. For NMF with random seeding, we ran NMF for 500 times to obtain optimized results.

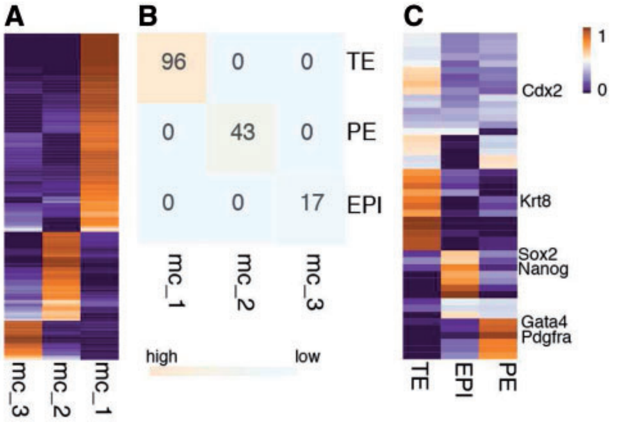
We compared the performance of different algorithms for NMF: BD (Schmidt *et al.*, 2009), LFNMF (Wang *et al.*, 2004), STNMF, LSNMF, ICM (Schmidt *et al.*, 2009), SNMF_L and SNMF_R (Kim and Park, 2007) on the complete and reduced lung epithelium data with NNDSVD seeding. The Python time module was used to estimate the CPU time, and each algorithm was run three times to eliminate possible influences of the running environments.

3 Results

3.1 NMF robustly identified subpopulations in RT-qPCR and RNA-seq data

Guo *et al.* (2010) studied gene expression in ~64-cell murine blastocysts, quantifying 48 genes (mainly transcription factors) by RT-qPCR. The cells were grouped into three subgroups based on the expression of known markers for trophectoderm (TE), epiblast (EPI), and primitive endoderm (PE). PCA analysis on the data recovered these subgroups but showed ambiguity in the classification of a small number of cells as EPI or PE. We applied NMF to the dataset with a rank value of three. The classification into metacells was unequivocal for each individual cell, and Figure 1A showed the normalized basis matrix $\bar{W}$ after hierarchical clustering. 96 of 156 cells were clustered as Metacell 1, 43 cells were clustered as Metacell 2, and 17 cells were clustered as Metacell 3. Most of the cells were faithfully assigned to metacells with contribution higher than 0.7, while a few cells (mainly EPI cells) had a lower contribution value, possibly indicating a more transient differentiation state. Importantly, the metacell classification recovered without error the grouping into the three biological subgroups based on the previously known markers, as seen by the confusion matrix (Fig. 1B). In addition, the coefficient matrix returned by NMF correctly identified the canonical markers in each group, e.g. Cdx2 and Krt8 highly expressed in TE, Gata4 and Pdgfra highly expressed in PE, while Sox2 and Nanog predominately expressed in EPI (Fig. 1C).

Next, we used the lung epithelium data to test the performance of NMF on single-cell RNA-seq data (Treutlein *et al.*, 2014). Tandem PCA performed by the authors separated three groups of cells that were further characterized by known markers. Two of the PCA groups appeared to contain only one cell type, Ciliated cells and Clara cells, respectively. The third group contained AT1 cells, AT2 cells and cells containing both AT1 and AT2 markers, which were classified as bipotential progenitor cells (BP cells). We firstly applied NMF to the reduced lung epithelium data and obtained a clear separation of cells with rank five based on the normalized basis matrix (Supplementary Fig. S1A). The comparison of NMF with the independent, marker based assignment of phenotypes showed a remarkable concordance (Supplementary Fig. S1B). These promising results encouraged us to apply NMF directly to the complete lung epithelium dataset containing more than 8000 genes (Fig. 2A).



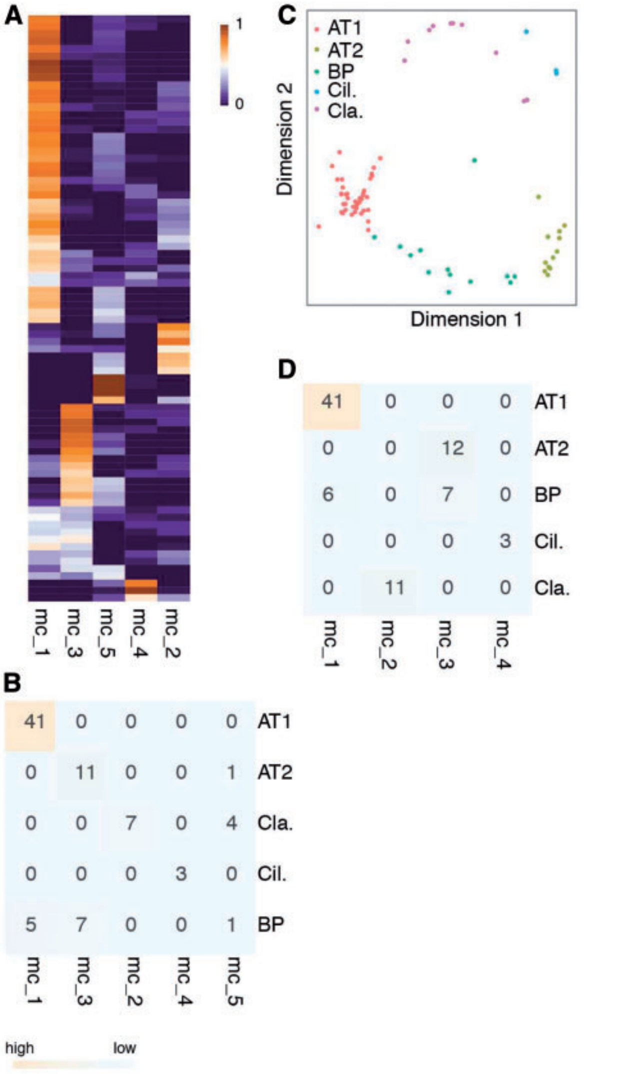
	mc_1	mc_2	mc_3
TE	96	0	0
PE	0	43	0
EPI	0	0	17

Fig. 1. NMF recovers subpopulations in single cell data measured by RT-qPCR. (**A**) The heatmap represents the normalized basis matrix returned by NMF with the mouse embryonic development data. Each row is a single cell, color indicates the contribution of the metacells calculated by NMF. Cells were grouped by hierarchical clustering. (**B**) Confusion matrix generated by comparing NMF results and known classification results in Guo *et al.* (2010). (**C**) Canonical markers in TE, PE, EPI. mc: metacell

Remarkably, NMF unequivocally recovered all cell types but Clara cells, which were split into Metacells 2 and 5 (Fig. 2B). NMF clearly distinguished AT1, AT2 and Ciliated cells (Metacells 1, 3 and 4, respectively). The fifth NMF metacell contained four cells assigned previously as Clara, one cell assigned as AT2 and one cell assigned as BP, while five and seven previously assigned BP, respectively, were classified as Metacell 1 (AT1) and Metacell 3 (AT2). This result very likely indicates the status of BP cells as progenitors of both AT1 and AT2 cells, as discussed by Treutlein *et al.* (2014). Thus, unlike PCA, which requires a feature selection step, NMF yielded a direct, unbiased classification of the major cell types in the sample based on the noisy data, with a clear indication of a transitory cell type.

We further mapped these cells onto a two-dimensional space by applying nonmetric multidimensional scaling (NMDS) on the normalized basis matrix. Interestingly, the neighbor relations of the 80 cells in this abstract space governed by gene expression reflected their relative spatial location in the lung epithelium (Fig. 2C). BP cells spanned the region between clearly separated AT1 and AT2 cells, while Clara and Ciliated cells were located at the opposite end. Due to the rather fixed cell positions in the epithelium, the spatial proximity could indeed reflect lineage relationships that would be mirrored in the gene expression pattern. Together with the assignment of BP, AT1 and AT2 cells (see above), the congruence between cell relationships via gene expression and spatial proximity indicate that the NMF analysis and the subsequent low-dimensional embedding of the cells might be informative on lineage relationships. Straightforward PCA analysis of the data missed these interesting features (Supplementary Fig. S2). Collectively, our findings show that NMF is an easily interpretable and robust method to dissect heterogeneity of single-cell RNA-seq data. For the representative datasets studied, NMF is superior to both straightforward PCA and tandem PCA.

However, prior information on the appropriate number of metacells might not generally be available. Therefore, we asked how the lung-epithelium cells would be clustered with a different number of metacells. Interestingly, choosing $r = 4$, the classification was largely identical or even improved, as Clara cells were clearly separated (Fig. 2D). Ciliated, AT1 and AT2 cells were unambiguously



	mc_1	mc_3	mc_2	mc_4	mc_5
AT1	41	0	0	0	0
AT2	0	11	0	0	1
Cla.	0	0	7	0	4
Cil.	0	0	0	3	0
BP	5	7	0	0	1

	mc_1	mc_2	mc_3	mc_4
AT1	41	0	0	0
AT2	0	0	12	0
BP	6	0	7	0
Cil.	0	0	0	3
Cla.	0	11	0	0

Fig. 2. NMF recovers subpopulations in single cell data measured by RNA-seq. (**A**) The heatmap represents the normalized basis matrix returned by NMF on the lung epithelium data. Each row is a single cell, while color indicates the contribution of Metacells. Cells were grouped by hierarchical clustering. (**B**) and (**D**) Confusion matrices generated by comparing NMF results with rank value five and four to the known classification results in Treutlein *et al.* (2014). Numbers in heatmaps are common cells in corresponding clusters. (**C**) The NMDS plot shows spatial location of lung epithelium cells basing on NMF results on complete lung epithelium data. Cells are colored according to Treutlein *et al.* (2014). mc, metacell; Cil., Ciliated cells; Cla., Clara cells

clustered into three different metacells as well. BP cells, the least definite category in the five-metacells classification, were assigned to AT1 or AT2 cells, again consistent with their presumed progenitor status for both cell types. Thus, the NMF classification with a reduced number of metacells preserved prominent biological features, demonstrating its robustness. We return to the problem of choosing the appropriate number of metacells without prior biological information in more detail in Section 3.3.

3.2 Feature selection

Since NMF, via the coefficient matrix, defines the expression of specific genes in metacells, it could be used to select informative feature genes for the different subgroups. Therefore, we investigated the biological function of genes with entries in the normalized coefficient matrix $\bar{H}$. We began by applying NMF to the reduced lung epithelium data. The following genes had high values in specific metacells: Rtkn2, Cav1, Pdpn and Hopx in Metacell 1 (AT1 cells); Lamp3, Cd36, Sftpb, and Sftpc in Metacell 2 or Metacell 5 (AT2 and BP cells); Krt15, Scgb3a2, Scgb1a1, Chad, Upk3a in Metacell 3 (Clara cells); and Ccdc39 in Metacell 4 (Ciliated cells) (Fig. 3A). Indeed, these were the signature genes identified by the tandem PCA (Treutlein *et al.*, 2014). Furthermore, NMF directly provided a summary of gene profiles in different groups; for example, Clara and Ciliated cells have highest and second highest expression of Scgb1a1 in the RNA-seq data. By contrast, quantitative expression values in subgroups cannot be easily extracted from PCA. In addition, we have observed that NMF successfully selected the marker genes in the RT-qPCR based data (Fig. 1C). Thus NMF results lend themselves to straightforward interpretation in terms of signature genes.

Next we asked whether NMF could be used for unbiased feature selection, that is, for identifying a relatively small number of genes that preserve the key information in the original data. To this end, we used the complete lung epithelium data and selected 164 metacell-specific genes based on their contributions in the normalized coefficient matrix and mean expression (see Methods section). Simply applying a PCA to these putative signature genes, we qualitatively recovered the grouping of the cells (Fig. 3B) that we previously found by applying NMF directly to the complete lung epithelium data (see also Fig. 2E). The signature genes returned by NMF significantly overlapped with the tandem PCA results (hypergeometric test P-value = 0). These findings suggest that NMF can be used for unbiased identification of marker genes.

3.3 Rank estimation

The rank value specifies the number of metacells (or subtypes) in a cell population and is a key parameter in the application of NMF. However, the correct number of subtypes is generally not known for unsupervised classification problems. Approaches based on cophenetic correlation or residual sum of squares have been proposed to estimate the proper rank value in NMF decomposition (Brunet *et al.*, 2004; Hutchins *et al.*, 2008). We found that these approaches successfully uncovered the correct (i.e. biologically plausible) rank for the PCR-based mouse embryonic data; however, they failed to find the appropriate rank value for noisy, high-dimensional RNA-seq data (Supplementary Figs. S3–4).

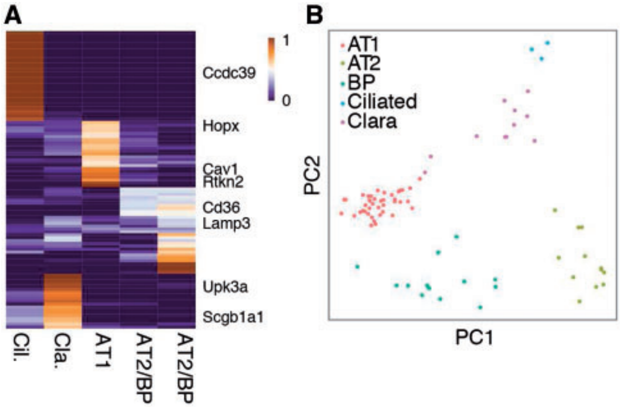


Fig. 3. Feature selection by NMF. (**A**) Normalized coefficient matrix recovers subpopulations marker genes based on the reduced lung epithelium data. Each row is a single gene, and selected marker genes for different groups are labeled on the right-hand side. Genes are grouped by hierarchical clustering. (**B**) PCA projection revealed five subpopulations based on feature genes selected from NMF. Cells were colored according to Treutlein *et al.* (2014). Cil., Ciliated cells; Cla., Clara cells

Therefore, we looked for alternative approaches to choose an appropriate number of metacells (rank) for the NMF analysis. We hypothesized that a biological meaningful classification should yield a mapping of the individual cells to metacells that is as unambiguous as possible. We defined the non-classified cells in given ranks via examining the largest contribution in metacells (see Methods section). Indeed, we observed that too high rank values increased the ambiguity in the classification via NMF, in the sense that individual cells would no longer have a clearly dominant metacell contribution (Supplementary Figs. S5–6). Ideally, an unambiguously assigned cell has contribution of 1 in one and only one metacell and 0 in all other metacells. In contrast, a completely non-informative classification will be the case where all metacells have equal contribution of $1/r$. Thus, we propose two approaches based on different measurements of uncertainty, cell sparseness and entropy, to estimate the proper rank. In the first approach, we calculated the cell sparseness from the normalized basis matrix for each single cell as the measurement of classification ambiguity. Sparseness was then adjusted for rank bias and compared among different ranks (see Methods section). A higher cell sparseness indicated a lower ambiguity in classification results, thus a significant drop of cell sparseness would suggest at a certain rank the cell population should not be further divided into subtypes. Cell sparseness showed a clear drop from rank three to four in the mouse embryonic development data (Fig. 4A), fully consistent with a biological subdivision into three distinct groups. Interestingly, this approach yielded a rank value of four when applied to the complete lung epithelium data (Fig. 4B), suggesting that a more unambiguous classification is achieved with rank four than the five subgroups chosen in the original publication (Treutlein *et al.*, 2014). Indeed, with four subgroups the bipotential progenitor cells were assigned to either AT1 or AT2, indicating that their gene-expression profile already contains signatures of differentiation bias (see Fig. 2C). This result implies that cell sparseness is a useful measure for rank choice even in noisy data.

In the second approach, we calculated the entropy of each cell in the normalized matrix $\bar{W}$ as a measurement of ambiguity. In the worst scenario of classification in which a cell has identical contribution by all metacells, the entropy is considered as non-informative background (see Methods section). The normalized information gain, which is the normalized difference of entropy between background and NMF classification, indicates how well NMF performs in unambiguously classifying cells, and we were interested in a clear drop of information gain as a function of the rank value. The information gain was almost at the maximum level in the mouse embryonic development data at rank three and clearly declined after that (Supplementary Fig. S7). In the more complicated lung epithelial

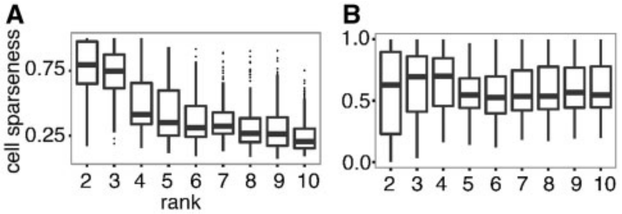


Fig. 4. Rank estimation. (**A**) and (**B**) Boxplots of normalized cell sparseness in the mouse embryonic development data and complete lung epithelium data. A clear drop of cell sparseness from rank three and four in these two data coincides with proposed subgroups in Guo *et al.* (2010) and Treutlein *et al.* (2014), respectively

system that contains five subpopulations defined by marker genes, we observed a clear drop from rank four (Supplementary Fig. S7). Thus both sparseness and entropy criteria for rank choice agree in above examples.

NMF classification with rank four in the lung epithelium data suggested an interesting biological question: are the fates of BP cells already determined to AT1 or AT2, as they are not selected as a separate cell type? To address this question, we examined the expression of marker genes in AT1-like BP cells and AT2-like BP cells. Indeed, we found Cxcl15, a ligand reported to be expressed in AT2/BP cells, was highly expressed in AT2-like BP cells; in addition, Fgfr2, an important receptor expressed in AT2/BP cells, showed higher expression in AT2-like BP cells than in AT1-like BP cells (Supplementary Fig. S8). These findings are consistent with a pre-existing differentiation bias in BP cells that is detected by NMF. It might be interesting to study BP cell heterogeneity in larger cell populations, together with lineage tracing methods.

3.4 NMF performance on a large RNA-Seq dataset

The numbers of cells sequenced in single-cell RNA-seq have recently increased from hundreds to thousands per sample, due to the improvement of experimental methods (Hashimshony *et al.*, 2016; Jaitin et al., 2014; Macosko *et al.*, 2015). To deal with the increased complexity and noise in the data, complex analysis methods have been developed, most of which iteratively apply clustering on subgroups. To test the utility of NMF, we applied it to a large dataset studying cell type complexity in the mouse cortex and hippocampus based on 3005 single cells (Zeisel *et al.*, 2015). To find the appropriate number of metacells for NMF, we performed NMF with rank values in the range of (2, 50) on the 3005 × 8839 data matrix (see Methods section). The sparseness and entropy plots suggested that seven metacells reasonably subdivided the single-cell data (Supplementary Fig. S9A–B). In this straightforward manner, NMF recovered all major cell groups that were described by the authors of the original publication using a more complicated biclustering method (BackSPIN) augmented by additional manual examination of the data (Fig. 5A). Metacells 2 to 7 were groups of oligodendrocytes, CA1 pyramidal neurons, microglia-endothelial-mural cells, astrocytes, interneurons and S1 pyramidal neurons, respectively. Of note, NMF outperformed BackSPIN in classifying CA1 pyramidal neurons: the latter method grouped the CA1Pyr2 and CA2Pyr2 cells together with S1 pyramidal neurons whereas NMF clearly distinguished S1 (Metacell 7) and CA1 (Metacell 3) neurons. However, some difficulty in classifying pyramidal neurons was also noticed with NMF, which grouped several cells that were identified by classical markers as CA1Pyr2 and S1PyrDL in Metacell 1, thus distinguishing them from the major S1 and CA1 pyramidal cells (Supplementary Fig. S9C). Interestingly, BackSPIN was unable to cluster 189 cells (labeled as category ‘none’ in Zeisel *et al.*, 2015), which NMF placed into Metacells 1, 5, 7 (Supplementary Fig. S9C). Thus with a very large single-cell RNA-Seq dataset, NMF proved to be a robust method for inferring major cell types without manual interference.

Zeisel *et al.* (2015) attempted to find a more fine-grained subdivision of the major cell types by applying their BackSPIN algorithm separately to each cell type, describing a total of 47 subgroups. We also applied NMF iteratively to the major metacells. To this end, we selected genes specifically expressed in each major metacell based on their value in the normalized coefficient matrix. Examining first Metacell 5, the sparseness dramatically declined from rank three to four, suggesting three was a meaningful choice for the number of subgroups (Supplementary Fig. S9D). The first subgroup coincided with astrocyte 1 described in the original paper while their astrocyte 2 category was split into two subtypes (Fig. 5B). A small number of cells previously assigned to two additional subtypes (ependymal cell, choroid cell) were not separated by NMF.

A similar picture was obtained for Metacells 6 and 4, where NMF obtained clear subdivisions that were fewer in number than the subtypes introduced in the original paper. However, there was a clear relation between our result and the previous analysis: NMF metacells usually combined several previously defined subgroups by Zeisel *et al.* (2015) rather than dividing them (Metacell 6_1 contained Int12-16, Metacell 6_2 contained Int5, 7, 8, Metacell 6_3 contained Int1-4, the last Metacell contained Int6 and Int10; Fig. 5C, Supplementary Fig. S9E). Thus our findings confirm the further subdivision of interneurons but also indicate that there may be fewer robust subgroups than previously introduced.

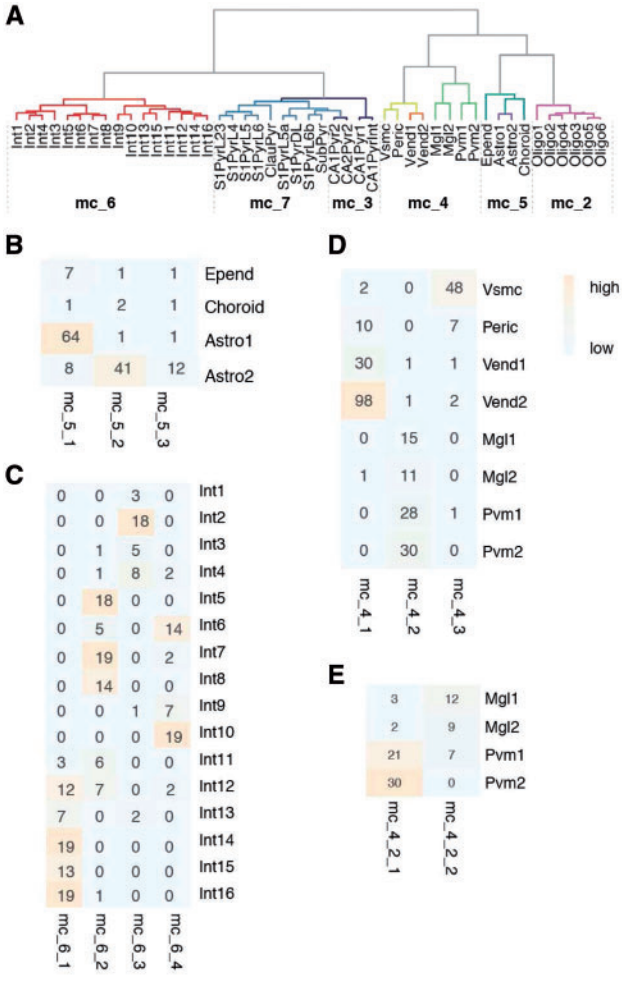


Fig. 5. NMF classification of mouse cortex-hippocampus data. (**A**) Comparison of BackSPIN with NMF results. The dendrogram shows the original classification (modified from Zeisel *et al.*, 2015) while the NMF results are indicated as metacells (mc_2-mc_7). Metacell 1 (not shown) collected many cells not identified by Zeisel *et al.* (2015). (**B**) NMF reveals the two subtypes of astrocytes in Metacell 5. (**C**) NMF classification applied to Metacell 4 shows three major cell subtypes. (**D**) NMF reveals four metacells for interneurons; (**E**) NMF classification on the sub-Metacell 2 separates microglia and perivascular macrophages. Abbreviations: mc, metacell; Mgl, micorglia; Pvm, perivascular macrophages; Vsmc, vascular smooth muscle cells; Peric, pericytes; Vend, vascular endothelial cells; Astro, astrocytes; Int, interneurons; CA1/SA1Pyr2, CA1/SA1 pyramidal neurons; Epend, ependymal cells

For Metacell 4, rank three was indicated by the sparseness plot (Supplementary Fig. S9F), and NMF found the subgroup of vascular endothelial cells (Vend1 and Vend2), microglia-perivascular macrophages (Mgl1, Mgl2, Pvm1, Pvm2), and vascular smooth muscle cells (Vsmc), while a subtype of mural cells (pericytes) was grouped together with smooth muscle cells or endothelial cells (Fig. 5D). We therefore applied NMF again to the 'mixed' sub-Metacell 4_2 and, interestingly, now found a separation between microglia and perivascular macrophages cells (Fig. 5E, Supplementary Fig. S9G). NMF did not further split microglia (Mgl) or perivascular macrophages (Pvm) in contrast to the original study. Moreover, we noticed that cell sparseness saturated for rank larger than two in some cell groups, e.g. vascular endothelial cells (Vend) and vascular smooth muscle cells (Vsmc), arguing against further dividing these subgroups (Supplementary Fig. S9H–I). NMF did not recover the previously suggested subtypes in Metacells 2, 7 (Supplementary Fig. S9K–M) with the rank values suggested by sparseness plots.

Taken together, NMF proved to be a robust, straightforward and fully automated method to divide a large collection of single-cell RNA-Seq data into the major cell types. Moreover, iterative application of NMF achieved further refinement of subtypes, which generally agreed with subdivisions detected by a more complex method (BackSPIN). However, we noticed that NMF is not able to identify the same fine classes (especially within neuronal subtypes) as BackSPIN. On one hand this shows the limitation of NMF in classifying cells with very similar transcriptome profile, one the other hand, it raise the interesting question of meaningful and robust definition of cell subtypes with biological functionality.

3.5 Comparison of seeding methods and NMF algorithms

The final section is devoted to practically important computational issues. The initiation of the *W* and *H* (which is referred to as seeding) is an important step in NMF which influences convergence of algorithms and running time. The simplest random seeding method generates *W* and *H* randomly from a uniform distribution; however, in this case NMF needs to be run multiple times to obtain an optimized result. In contrast, the nonnegative double singular value decomposition (NNDSVD) seeding method is a deterministic seeding approach without randomization. Briefly, NNDSVD contains two SVD processes approximating the target matrix and positive parts of resulting SVD factors, respectively. It was shown that NNDSVD leads to rapid convergence of many NMF algorithms (Boutsidis and Gallopoulos, 2008). We compared the performance of random and NNDSVD seeding methods in the context of single-cell transcriptome data. NMF results were comparable with these two seeding methods on the reduced lung epithelium data (Supplementary Fig. S10). However, the random seeding method led to poor classification on the complete lung epithelium data, causing AT1 cells to become mixed with other cell types (Supplementary Fig. S11). Our results indicate that the NNDSVD method is preferred for complex and large data because of both speed and accuracy.

As there are several algorithms available for NMF analysis, it is of interest to examine their performance in practical applications. We applied seven algorithms (BD, LFNMF, ICM, LSNMF, STNMF, SNMF_L and SNMF_R) to the reduced and complete lung epithelium dataset. These algorithms produced similar results in terms of confusion matrix, explained variance (evar), sparseness of basis matrix (sparse_*W*) and coefficient matrix (sparse_*H*) (Supplementary Figs. S12 and S13, Supplementary Tables S1–S3). BD and LSNMF were the fastest algorithms and LFNMF was impractically slow. We noticed that BD runs significantly faster than LSNMF on the larger cortex-hippocampus dataset, and provided comparable classification results (Supplementary Fig. S14). Taken together, the LSNMF algorithm together with the NNDSVD seeding method offer a robust and fast implementation of NMF.

4 Discussion

Single-cell transcriptome data pose a challenge to existing unsupervised classification approaches. PCA suffers from the large noise inherent in the data, which, to a large part, is due to incomplete sampling (and hence dropout events) as well noise generated by amplification (Brunet *et al.*, 2004) (Kharchenko *et al.*, 2014). In addition, the biological noise, which stems from stochastic gene expression and is not necessarily cell type-specific, contributes to the variance in single-cell data as well (Munsky *et al.*, 2012). Here we explored the performance of NMF for the purpose of subgroup identification in single-cell transcriptome data. NMF decomposes a target matrix into two matrices under the non-negative constraint. Notwithstanding the fact that this is an NP-hard and ill-posed problem in principle, studies on the practical implementation and applications of NMF abound (Cai *et al.*, 2011; Choi, 2008; Ding et al., 2006; 2010). Instead of adopting the conventional gene-centered view, in which thousands of genes are summarized by a handful of metagenes, we focused here on the sample-centered view by grouping individual cells to metacells. We found that NMF correctly classified cells to biologically meaningful metacells in RT-qPCR and RNA-seq data. Compared to PCA, NMF presents the classification results as a non-negative matrix that is straightforward to interpret. In addition, NMF proved to be very robust to the noise in the single-cell RNA-seq data. Remarkably, NMF performed well even without prior dimension reduction (e.g. by choosing signature genes through a first PCA), thus providing an easy-to-understand, unbiased method for biologically interpreting single-cell transcriptome data.

We found that the coefficient matrix generated by our NMF approach allows one to identify genes specifically expressed in subgroups. For example, we were able to recover the signature genes in mouse blastocysts and lung epithelium data. Thus NMF offers an unbiased approach to select feature genes that preserve key information about cell subtypes in the original data.

The rank value in the sample-centered view of NMF corresponds to the number of metacells in cell populations. In most of the applications the number of subpopulation is not known. We presented two approaches for finding a meaningful number of subpopulations – sparseness and entropy gain, which are both based on the idea of minimizing the ambiguity of cell classification. Compared to the previously suggested use of the cophenetic correlation or rss, our approaches were clearly superior in finding biological meaningful rank values.

Sophisticated methods have been proposed to analyze the single-cell RNA-seq data, most of which involving iterative clustering, feature selection and subsampling (Macosko *et al.*, 2015; Tasic *et al.*, 2016; Zeisel *et al.*, 2015). We compared the classification results with one cutting-edge method, BackSPIN, on a fairly large dataset. NMF successfully recovered the major cell types; in addition, iteratively applied NMF recovered biological meaningful subtypes in some cell groups, similar to BackSPIN results, through NMF generally found fewer subdivisions than BackSPIN, showing the limitation of NMF on the high similarity transcriptome data. NMF provided different cell subtypes for CA1/S1 pyramidal neurons and oligodendrocytes. This might due to the fact that BackSPIN always iteratively splits the target cells into two groups until a predefined stop criteria reached, which increases the sensitivity but also involves a risk of over-splitting. It is a general open question of how to choose the number of iterative steps in subdividing cell populations. Applying NMF, we noticed that the saturation of sparseness value might be used a signal for stopping iterative classification (Supplementary Fig. S9H–J).

We accessed the performance of NMF basing on different seeding methods and found the deterministic NNDSVD (Boutsidis and Gallopoulos, 2008) method provided clearer and more robust classification than the random seeding methods when applied to a large data matrix. Another advantage of NNDSVD seeding is time saving as it only needs to run one time, while NMF with random seeding need multiple runs to obtain an optimized result. We found most of the NMF algorithms provided similar results in term of important metrics, e.g. explained variance, sparseness of basis matrix and coefficient matrix. However, there are significant differences in running time among algorithms, among which LSNMF and BD are fastest in the tested data. In the tested data, LSNMF with NNDSVD provided the most reasonable classification results. In addition, as showed in the cortex-hippocampus data, it makes more sense to iteratively apply NMF with small rank values than working directly with a larger one. Although run time has not been a constraint here even with the largest dataset (3005 cells), we note that new approaches are being proposed to further reduce the NMF running time. For example, a recent GPU-based NMF implementation runs about 120 times faster than the traditional methods (Mejía-Roa *et al.*, 2015), and a parallel computing method is proposed to solve the coefficient matrix (Bauckhage, 2014). In conclusion, our study shows that NMF is a robust, informative, simple method for the unsupervised learning of cell subtypes from single-cell gene expression data.

Acknowledgements

We thank Dr. Barbara Treutlein for providing help on building the lung epithelium data matrix, as well as Dr. Michael Floßdorf and Lena Appel for helpful discussions.

Funding

This work has been supported by the German Ministry for Education and Research (BMBF) through e:Med (Sysmed-NB).

Conflict of Interest: none declared.

References

Arsenio,J. *et al.* (2014) Early specification of CD8+ T lymphocyte fates during adaptive immunity revealed by single-cell gene-expression analyses. *Nat. Immunol.*, **15**, 365–372.

Bacher,R. and Kendziorski,C. (2016) Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.*, **17**, 63.

Berry,M.W. *et al.* (2007) Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Stat. Data Anal.*, **52**, 155–173.

Boutsidis,C. and Gallopoulos,E. (2008) SVD based initialization: a head start for nonnegative matrix factorization. *Pattern Recognit.*, **41**, 1350–1362.

Brennecke,P. *et al.* (2013) Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods*, **10**, 1093–1095.

Brunet,J.P. *et al.* (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 4164–4169.

Bauckhage,C. (2014) A purely geometric approach to non-negative matrix factorization. *Proceedings of the 16th LWA Workshops: KDML, IR and FGWM*, 125–136.

Cai,D. *et al.* (2011) Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **33**, 1548–1560.

Carmona-Saez,P. *et al.* (2006) Biclustering of gene expression data by non-smooth non-negative matrix factorization. *BMC Bioinformatics*, **7**, 78.

Choi,S. (2008) Algorithms for orthogonal nonnegative matrix factorization. *IEEE*, pp. 1828–1832.

Ding,C. *et al.* (2006) Orthogonal Nonnegative Matrix T-factorizations for Clustering. In: *ACM, New York, NY, USA*, pp. 126–135.

Ding,C. *et al.* (2010) Convex and semi-nonnegative matrix factorizations. *IEEE Trans. Pattern Anal. Mach. Intell.*, **32**, 45–55.

Grün,D. *et al.* (2014) Validation of noise models for single-cell transcriptomics. *Nat. Methods*, **11**, 637–640.

Guo,G. *et al.* (2010) Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell*, **18**, 675–685.

Haghverdi,L. *et al.* (2015) Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, **31**, 2989–2998.

Hashimshony,T. *et al.* (2016) CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.*, **17**, 77.

Hutchins,L.N. *et al.* (2008) Position-dependent motif characterization using non-negative matrix factorization. *Bioinformatics*, **24**, 2684–2690.

Jaitin,D.A. *et al.* (2014) Massively parallel single-cell RNA-Seq for marker-free decomposition of tissues into cell types. *Science*, **343**, 776–779.

Kharchenko,P.V. *et al.* (2014) Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, **11**, 740–742.

Kim,H. and Park,H. (2007) Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, **23**, 1495–1502.

Lee,D.D. and Seung,H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.

Lin,C. (2007) Projected gradient methods for nonnegative matrix factorization. *Neural Comput.*, **19**, 2756–2779.

Macosko,E.Z. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.

Mejía-Roa,E. *et al.* (2015) NMF-mGPU: non-negative matrix factorization on multi-GPU systems. *BMC Bioinformatics*, **16**, 43.

Moignard,V. *et al.* (2015) Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.*, **33**, 269–276.

Munsky,B. *et al.* (2012) Using gene expression noise to understand gene regulation. *Science*, **336**, 183–187.

Nik-Zainal,S. *et al.* (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell*, **149**, 979–993.

R Core Team (2015) *R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing*. Vienna, Austria.

Schmidt,M.N. *et al.* (2009) *Bayesian Non-negative Matrix Factorization*. Springer, Proceedings of the 9th International Conference on Independent Component Analysis and Signal Separation, 540–547.

Tasic,B. *et al.* (2016) Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.*, **19**, 335–346.

Trapnell,C. *et al.* (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, **32**, 381–386.

Treutlein,B. *et al.* (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, **509**, 371–375.

Wang,Y. *et al.* (2004) Fisher non-negative matrix factorization for learning local features. In: *Proc Asian Conf on Comp Vision*, pp. 27–30.

Xu,C. and Su,Z. (2015) Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, **31**, 1974–1980.

Zeisel,A. *et al.* (2015) Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, **347**, 1138–1142.

Žitnik,M. and Zupan,B. (2012) Nimfa: A python library for nonnegative matrix factorization. *J. Mach. Learn. Res.*, **13**, 849–853.