SPECK: An Unsupervised Learning Approach for Cell Surface Receptor Abundance Estimation for Single Cell RNA-Sequencing Data

Azka Javaid[1] **and H. Robert Frost**[1]

[1]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, NH 03755, USA

The rapid development of single cell transcriptomics has revolutionized the study of complex tissues. Single cell RNA-sequencing (scRNA-seq) can profile tens-of-thousands of dissociated cells from a single tissue sample, enabling researchers to identify the cell types, phenotypes and interactions that control tissue structure and function. A key requirement for these applications is the accurate estimation of cell surface protein abundance. Although technologies to directly quantify surface proteins are available, this data is uncommon and limited to proteins with available antibodies. While supervised methods that are trained on both transcriptomics and proteomics data (e.g., Cellular Indexing of Transcriptomes and Epitopes by Sequencing or CITE-seq) can provide the best performance, this training data is also limited by available antibodies and may not exist for the tissue under investigation. In the absence of protein measurements, researchers must estimate receptor abundance from scRNA-seq data. We thereby developed a new unsupervised method for receptor abundance estimation using scRNA-seq data, Surface Protein abundance Estimation using CKmeans-based clustered thresholding (SPECK), and evaluated its performance against other unsupervised approaches on up to 215 human receptors and multiple tissue types. This analysis reveals that techniques based on a thresholded reduced rank reconstructed (RRR) output of scRNA-seq data are effective for receptor abundance estimation with the SPECK method providing the best overall performance.

single cell transcriptomics | receptor abundance estimation

Correspondence: ***azka.javaid.gr@dartmouth.edu***

Introduction

Transcriptome profiling has conventionally been performed using RNA sequencing (RNA-seq) of bulk tissue samples (1–3). While bulk RNA-seq facilitates genome-wide gene expression profiling, it measures the average gene expression for all cells within a tissue sample. To address the limitations of conventional RNA-seq, single cell RNA-sequencing (scRNA-seq) technologies (4, 5), such as the 10X Chromium System (6), have been developed that can quantify gene expression within thousands of cells from a single tissue sample. The resolution provided by single cell assays is critical for unraveling the biology of diseases associated with complex tissues, e.g., the tumor microenvironment, with important implications for advancing precision medicine, e.g., understanding the association between specific driver mutations and the composition and phenotype of tumor infiltrating immune cells. Key applications of scRNA-seq data include cataloging the cell types in a tissue (7, 8), reconstructing dynamic processes (9), and analyzing cell-cell signaling (10). Many of these tasks are based on the abundance of cell surface proteins, which follows the traditional use of immunohistochemical profiling for characterizing dissociated cells (e.g., Fluorescence Activated Cell Sorting (FACS) (11)) or intact tissue. Although techniques such as Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq) (12) can measure protein abundance for individual cells, in addition to quantifying single cell transcriptome, using barcoded antibodies, this type of data is uncommon, and targets a limited number of proteins that have available antibodies.

If receptor protein measurements are not available for a given sample, researchers can estimate receptor abundance from scRNA-seq data using supervised or unsupervised approaches. Supervised methods, such as the single cell Transcriptome to Protein prediction with deep neural network (cTPnet) tool (13), fit a predictive model on joint transcriptomics and proteomics CITE-seq training data and use this model to generate estimates for target scRNA-seq datasets. While supervised methods can provide excellent predictive performance, they are only feasible if sufficient training data exists and, may be susceptible to over-fitting and low transparency. For example, cTPnet only supports 24 immune-related receptors and uses a deep learning approach that makes model interpretation challenging. For cases where training data is not available, which includes many receptors that lack CITE-seq antibodies, an unsupervised approach that leverages RNA expression to determine protein levels must be employed. While the heterogeneity induced by processes such as transcriptional bursting may reduce the otherwise approximately 56%-84% reported variation accounted for by RNA variance measured using canonical quantification tools (14, 15), advances in scRNA-seq and CITE-seq technology may make it increasingly more feasible to use RNA expression to quantify protein abundance. Abundance estimation using RNA data is further legitimized in the context of relative abundance estimation. As experimentally observed, while post-transcriptional regulations have a more substantial influence on changes in absolute protein concentrations, these processes do not seem to considerably affect relative protein levels (16, 17).

One simple unsupervised strategy uses the expression of the associated RNA transcript as a proxy for protein abundance. While this approach can work in some situations, the sparsity of scRNA-seq data often leads to very poor quality estimates. Although this sparsity can be mitigated by a cluster-based approach, that approximates abundance using the average expression across all cells in a cluster, such a cluster-level analysis is sensitive to the number of computed clusters, and, importantly, ignores within cluster heterogeneity. A second promising family of unsupervised methods for generating cell-specific estimates are scRNA-seq data imputation

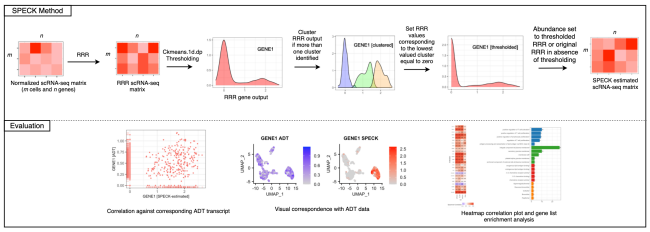

Fig. 1. SPECK Method: SPECK performs normalization, rank selection, RRR and thresholding on a $m \times n$ scRNA-seq counts matrix with m cells and n genes. The reconstructed and thresholded matrix is of size $m \times n$. Evaluation: SPECK-based receptor abundance estimates are visually assessed using feature plots and heatmaps and their correspondence to ADT data is quantified using the Spearman rank correlation. Gene list enrichment analysis can be used to identify predominant functional signatures of the SPECK-estimated profiles.

techniques. In this scenario, the imputed value of the receptor transcript is used to estimate protein abundance under the assumption that the imputation process reduces sparsity without inflating false positives. Although a large number of scRNA-seq imputation approaches exist (18), we have found that the class of reduced rank reconstruction (RRR) techniques, which assumes that the intrinsic dimensionality of scRNA-seq data is much lower than the empirical rank, provides superior performance relative to other methods. RRR-based imputation methods include Markov Affinity-based Graph Imputation of Cells (MAGIC) (19) and Adaptively thresholded Low-Rank Approximation (ALRA) (20). While MAGIC and ALRA significantly outperform the naïve approach of directly using the receptor transcript, they have a few limitations. ALRA, for example, applies the same quantile probability of 0.001 to threshold each gene, thereby overlooking considerable individual, gene-level expression variability. While MAGIC, like ALRA, is scalable, its performance is largely inferior to ALRA's as a single cell imputation strategy (20).

To address the limitations of existing unsupervised approaches for relative receptor protein abundance estimation, we developed a new technique named Surface Protein abundance Estimation using CKmeans-based clustered thresholding (SPECK). Similar to ALRA, the SPECK method utilizes a singular value decomposition (SVD)-based RRR but includes a novel approach for thresholding of the reconstructed gene expression matrix that improves receptor abundance estimation. A second important contribution of this paper is a comprehensive evaluation of unsupervised receptor abundance estimation performance across 215 human cell surface receptors and multiple tissue types.

Methods

SPECK method for scRNA-seq data. The SPECK method estimates abundance profiles for m cells and n genes in a $m \times n$ matrix of scRNA-seq counts using the procedure detailed below. See Figure 1 and Algorithm 1 for the implementation overview and the associated pseudocode, respectively.

Rank estimation and Reduced rank reconstruction (RRR). Given a $X_{m,n}$ scRNA-seq count matrix, consisting of the number of molecules detected for each gene in each cell, we first perform log-normalization using Seurat's normalization pipeline (21), which consists of dividing the feature counts for each cell by the total cell count, multiplying this value by a scale factor of $10{,}000$ and then natural-log transforming the product. Reduced rank reconstruction (RRR) is next performed on the resulting $X_{m,n}$ normalized matrix to create a low-rank representation of the original gene expression data. This RRR leverages the generally prevalent observation that since genes act in a concerted and interdependent manner, the original expression values can be sufficiently approximated by a low-rank matrix (22, 23). While principal component analysis (PCA), based on the eigendecomposition of the covariance matrix of a non-square gene expression matrix, is a reasonable option for low-rank decomposition, PCA is not feasible for large scRNA-seq data due its dependence on the computation of a covariance matrix (24). Given this limitation, alternative decomposition approaches based on techniques such as the SVD, which may even be more numerically stable, are preferable for high-throughput data (25, 26). While SVD is computationally less intensive than PCA, a full or even a truncated SVD, that extracts the first k components from the RRR data, is too costly to the order of $O(mn^2)$ for a large $m \times n$ matrix where $m \geq n$. As an alternative to these traditional SVD approaches, randomized SVD algorithms have recently been proposed to obtain a more efficient and parallelizable rank-k approximated decomposition with a $O(mnk)$ computational cost (27, 28).

We leveraged an implementation of a randomized SVD algorithm using the rsvd package (28, 29), which was observed to be accurate and computationally fast relative to other non-randomized techniques, to perform a rank-100 approximation. We then transformed the singular values from this SVD decomposition to eigenvalues for subsequent rank estimation. To compute these eigenvalues, we first squared the singular values and then divided by one minus the number of cells in the scRNA-seq matrix. We then computed the square root of these values to derive the eigenvalue standard deviation and performed a rate of change computation of each successive standard deviation estimate relative to the previous estimate. Rank-k was then selected to represent a numeric value, ranging from 1 to 100, for which the absolute value of the difference between consecutive eigenvalue standard deviation estimates was at least 0.01 for two or more estimate pairs. The estimated k columns from each of the three rank-100 rsvd-based decomposition matrices were subsequently used for reconstruction, thereby resulting in a $X_{m,n}$ RRR matrix.

Cluster-based thresholding. We next performed thresholding on the RRR values for each of the n genes from the $X_{m,n}$ RRR matrix. This step was inspired by the bimodality often exhibited by protein expression distributions, which functionally reflects the presence of two different subpopulations with varying individual response mechanisms to factors like stress or a particular drug treatment (30–32). While bimodality in protein expression at the population level has historically been linked to stochastic switching (33), defined by the transition of cells between multiple phenotypes as a result of environmental fluctuations, it is now more recently at-

tributed to cell-to-cell variability that affects the sustained oscillation frequencies in a heterogeneous cell population (30). We sought to leverage this idea of population-level bimodality exhibited by proteomics data in our abundance estimation technique by performing a thresholding step on the RRR values to ensure that the resulting estimates are representative of the corresponding ADT data.

For this purpose, we utilized the Ckmeans.1d.dp algorithm from the Ckmeans.1d.dp package (34) to perform thresholding on each RRR gene expression estimate with a specified value for the Ckmeans.1d.dp parameter k, which defines the minimum and maximum number of clusters to be examined. We selected the Ckmeans.1d.dp algorithm, a dynamic programming-based one-dimensional clustering approach which functions by minimizing the sum of squares of within-cluster distances from an element to its associated cluster mean, due to its optimality and reproducibility over more heuristic clustering methods such as k-means (35), which are not always optimal nor repeatable due to a dependence on the definition of the initial cluster centers. While, given the bimodal expression pattern of proteomics data, an upper bound of two seems reasonable for the Ckmeans.1d.dp-based parameter k, we specified a slightly higher upper bound to account for potential additional expression multimodality in ADT data beyond that accounted for by bimodality. We found that the upper limit of four provides a suitable selection choice to ensure that cells get mapped to a fairly granular and interpretable number of clusters. With the minimum cluster number set to one and the maximum cluster number set to four, thresholding was only performed if more than one cluster was identified. If so, then all the non-zero values in the RRR output for the gene corresponding to the indices of the least-valued cluster, as identified by the cluster mean, were set to zero. The zero values corresponding to the indices of the least-valued cluster and all other non-zero and zero values corresponding to higher-valued clusters in the RRR gene output were preserved. If, only one cluster was identified, thresholding was not performed and the RRR gene values were preserved. The final output consisted of a $X_{m,n}$ thresholded matrix.

Algorithm 1 SPECK Algorithm

$X_{m,n} \xrightarrow{\text{Normalization}} X_{m,n}$

Rank-k Selection and Singular Value Decomposition

rsvd.res $\leftarrow \sum_{i \leftarrow 1}^{100} \mu_i \sigma_i v_i^T$ ▷ Compute rank-100 SVD of $X_{m,n}$

rsvd.sdev $\leftarrow \sqrt{\frac{(\sigma_1, \sigma_2, \ldots, \sigma_{100})^2}{m-1}}$ ▷ Derive the eigenvalue standard deviation

rsvd.diff $\leftarrow list()$

for $q \leftarrow 1$ to $100-1$ **do** ▷ Compute the rate of change

  rsvd.diff $\leftarrow$ rsvd.sdev$[q+1]$ − rsvd.sdev$[q]$

end for

rsvdiff $\leftarrow list()$; jval $\leftarrow list()$

rlength $\leftarrow list()$; rvalue $\leftarrow list()$

$r \leftarrow 1$

rsvd.diff $\leftarrow$ |rsvd.diff| ▷ Compute absolute value of differences

while $r \leq 100-1$ **do** ▷ Compute the run length encoding

  count $\leftarrow 1$

  val $\leftarrow$ rsvd.diff$[r]$

  $s \leftarrow r$

  while $s < 100-1$ **do**

    if rsvd.diff$[s]$ = rsvd.diff$[s+1]$ **then**

      rsvdiff.$append$(rsvd.diff$[j+1]$)

      jval.$append(j+1)$

      count $\leftarrow$ count $+1$

      $s \leftarrow s+1$

    else

      break

    end if

  end while

  rlength.$append$(count); rvalue.$append$(val)

end while

min.val $\leftarrow \infty$

for $u \leftarrow 1$ to len(rvalue) **do** ▷ Compute the minimum index

  if rvalue$[u] \geq 0.01 \wedge$ rlength$[u] \geq 2 \wedge$ rvalue$[u] <$ min.val **then**

    min.val $\leftarrow$ rvalue$[u]$

  end if

end for

$k \leftarrow$ jval[rsvdiff.$index$(min.val)] ▷ Extract the estimated rank-k

$X_{m,n} \leftarrow \mu_{m,k} \times \sigma_{k,k} \times v_{k,n}^T$ ▷ Reconstruct the matrix

Clustered Thresholding

for $v \leftarrow 1$ to n **do** ▷ Threshold each gene of $X_{m,n}$

  ckmeans.res $\leftarrow Ckmeans.1d.dp(x = X_{m,n}[v],$
    $k = c(1:4))$

  clustnum $\leftarrow$ ckmeans.res["$centers$"]

  clustvalue $\leftarrow$ ckmeans.res["$cluster$"]

  if len(clustnum) > 1 **then**

    min.center $\leftarrow$(
      clustnum[clustvalue.$index(min$(clustvalue))]

    min.ind $\leftarrow$ [index for index,
      value in $enumerate$(clustnum)
      if value = min.center]

    $X_{m,n}[v][min.ind] \leftarrow 0$

  end if

end for

Evaluation

Comparison methods. Comparative evaluation of the SPECK method was performed against three other approaches described below.

- ALRA: Receptor abundance was estimated using the imputed value of the associated transcript generated by the ALRA method with v.1.0.0 (20). scRNA-seq data was normalized using ALRA's log-normalization followed by imputation with an automatically chosen rank-k.

- MAGIC: Receptor abundance was estimated using the imputed value of the associated transcript generated by the MAGIC method with Rmagic v.2.0.3 (19). scRNA-seq data was normalized using L1-normalization fol-

lowed by imputation with the number of nearest neighbors (knn) parameter set to 15.

- RNA transcript: Receptor abundance was estimated using the normalized value of the associated transcript with the log-normalization procedure from Seurat v.4.1.0 (21).

Public single cell data. SPECK was evaluated relative to ALRA, MAGIC and RNA on two publicly accessible CITE-seq datasets: 1) the Hao et al. (36) human peripheral blood mononuclear cell (PBMC) dataset (GEO (37) series GSE164378) that contains 161,764 cells profiled using 10X Chromium 3' with 228 TotalSeq A antibodies, and 2) the Stuart et al. (38) human bone marrow mononuclear cell (BMMC) dataset (GSE128639) that contains 33,454 cells similarly profiled with 25 TotalSeq A antibodies. For both datasets, the CITE-seq scRNA-seq data was used to generate receptor abundance estimates that were evaluated relative to corresponding CITE-seq ADT data. Each data was individually processed using Seurat v.4.1.0 (21) in R v.4.1.2 (39). ADT counts were normalized using the centered log-ratio transformation (40).

Setup and evaluation metrics. We evaluated the SPECK-estimated abundance profiles on five different subsets of the CITE-seq PBMC data with 33,538 genes and 13 cell categories ranging from 1000 to 60,000 with each interval consisting of 5000 cells. We determined the upper limit of 60,000 cells by the ability to perform RRR on 16 CPU cores without any virtual memory allocation errors. From the initial 228 antibodies included in the PBMC data, antibodies mapping to multiple HUGO Gene Nomenclature Committee (HGNC) (41) symbols were removed. Since assessment was performed for varying number of cells, receptors not expressed in smaller cell groups were dropped. Our final assessed receptors for the PBMC data ranged from 205 to 215. We also evaluated our method on CITE-seq BMMC data with 17,009 genes and seven cell categories ranging from 1000 to 30,000, with each interval consisting of 5000 cells, for 25 antibodies. We measured correspondence between the estimated receptor abundances and the ADT transcripts using the Spearman rank correlation. Spearman correlation was chosen since it can quantify relationships between the ranks of two variables, as opposed to their absolute values, which is appropriate for comparing the estimated relative abundance profiles produced by SPECK, ALRA, MAGIC and RNA against ADT data. Spearman correlation metric is moreover relatively robust to outliers and can be used to analyze nonlinear relationships (42). SPECK's performance, as quantified by the Spearman rank correlation coefficient, was compared with the performance of imputed expression profiles generated by ALRA and MAGIC and the normalized RNA counts.

Results

SPECK improves correspondence of abundance estimates with CITE-seq ADT data. We first quantified the number of receptors estimated by SPECK or imputed with ALRA, MAGIC or RNA that have the highest rank correlation values, as averaged over five subsets from each of the 13 cell categories, with PBMC ADT data. These frequencies were similarly quantified for each of the seven cell categories for the BMMC data. Statistical testing was used to quantify the pairwise differences within each cell category between the number of receptors that were estimated by SPECK and had the highest rank correlation with ADT data and receptors that were imputed by a comparative technique and had the second highest correlation using post-hoc tests following a significant chi-square homogeneity test. The Benjamini–Hochberg method (43) was used to control the False Discovery Rate (FDR) for multiple comparisons.

As Figures 2 and 3 show, SPECK-imputed abundance profiles consistently have a better correspondence with ADT data compared to imputed profiles obtained from comparative methods across all cell categories for the PBMC and the BMMC data, respectively. For example, for the 10,000 cell subset category from the PBMC data, 76 receptors (36%) estimated with SPECK have higher correlation with corresponding ADT data when compared to imputed abundance estimates produced by ALRA, MAGIC and RNA. This compares with 47 (22%) receptors that are imputed with ALRA and have higher correlation with ADT data compared to other techniques, including SPECK, 46 (22%) receptors imputed with MAGIC and 42 (20%) receptors substituted with the normalized RNA transcript that have higher rank correlation with ADT data versus alternative methods. The difference between the number of receptor abundance profiles that are estimated with SPECK and have the highest rank correlation with corresponding ADT data and the number of receptors imputed by a comparative method that have the second highest correlation with ADT data are all statistically significant for each cell category, except for the 55,000 cell category for the PBMC data and the 1000 cell category for the BMMC data, at an alpha level of 0.05.

To further examine the distinctive contributions of SPECK's RRR and thresholding procedures in individually affecting rank correlations between the estimated abundance profiles and the corresponding ADT data, we quantified and tested the frequency of receptors that have higher rank correlation with ADT data when estimated with the full SPECK method (i.e. including both the RRR and thresholding steps) versus when estimated by SPECK's RRR-only method (i.e. the RRR matrix) using a two-sample z-test of proportions. Figure 4 indicates that a greater number of receptors estimated by the full SPECK method have higher correlation with ADT data when compared to receptors estimated by SPECK'S RRR procedure for the PBMC data. We obtained comparable results for the BMMC data, as shown by Figure 5. All differences in receptor frequency were statistically significant at an alpha level of 0.05 for every cell category of the PBMC and the BMMC datasets. These results emphasize the central contribution of SPECK's thresholding procedure in considerably improving the accuracy of the estimated abundance profiles over SPECK's RRR-only approach.

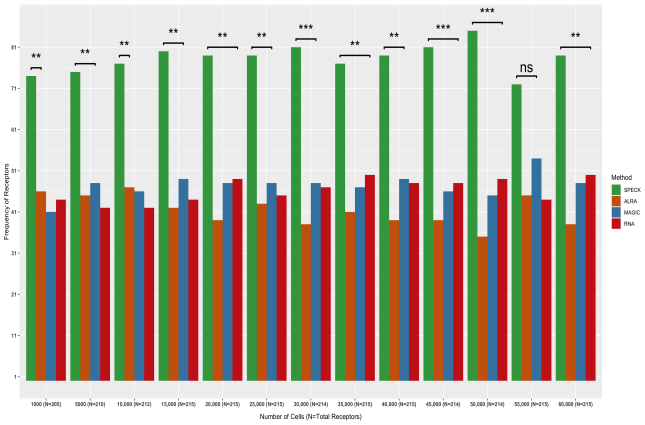

Fig. 2. Frequency of receptors with the highest averaged rank correlation values between PBMC ADT data and abundance estimates produced by SPECK or imputed profiles produced by ALRA, MAGIC or RNA for five subsets of cells ranging from 1000 to $60{,}000$ for the PBMC data are displayed. Asterisks indicate the adjusted p-value significance levels for differences between the top two, pairwise proportions for each cell category $(***, 0.0001 < P < 0.001; **, 0.001 < P < 0.01; *, 0.01 < P < 0.05;$ns, $0.05 < P < 1)$.

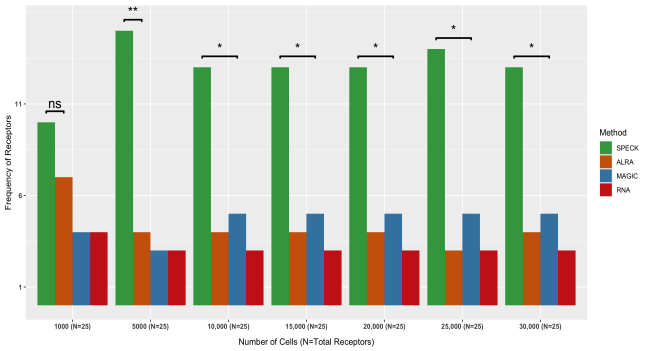

Fig. 3. Frequency of receptors with the highest rank correlation between BMMC ADT data and abundance profiles produced by a specified estimation/imputation technique, including SPECK, ALRA, MAGIC and RNA, averaged over five subsets of cells ranging from 1000 to $30{,}000$, are indicated. Asterisks indicate the adjusted p-value significance levels for differences between the top two proportion values $(**, 0.001 < P < 0.01; *, 0.01 < P < 0.05;$ns, $0.05 < P < 1)$.

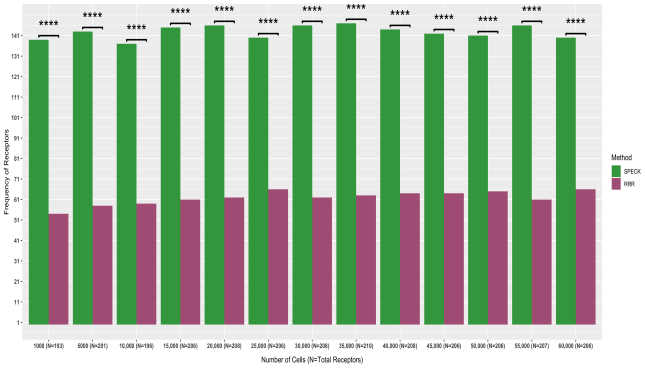

Fig. 4. SPECK-based RRR and thresholded estimates are compared with SPECK-based RRR-only values and frequency of receptors with the highest Spearman rank correlation between PBMC ADT data and estimated values produced by either estimates, averaged over five subsets of cells ranging from 1000 to $60{,}000$, are displayed. Receptors estimated with both techniques that have the same correlation value against corresponding ADT data are removed. Asterisks indicate the significance levels of p-values from the proportion tests $(****, 0 < P < 0.0001)$.

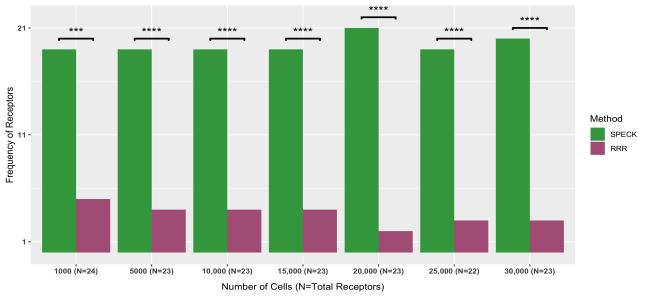

Fig. 5. Frequency of receptors with the highest rank correlation between BMMC ADT data and abundance profiles produced by SPECK (RRR and thresholding) and SPECK-based RRR-only, as averaged over five subsets of cells ranging from 1000 to $30{,}000$, are shown. Receptors with ties in rank correlation values are removed. Asterisks indicate the p-value significance levels for the proportion tests $(****, 0 < P < 0.0001; ***, 0.0001 < P < 0.001)$.

Examining individual abundance values using density estimation. We next visually assessed correspondence between the estimated abundance profiles produced by SPECK, ALRA, MAGIC and RNA with analogous ADT data using kernel density estimation-based smoothed scatterplots for individual subsets of $10{,}000$ cells from the PBMC and the BMMC datasets. The CD14, CD19 and CD79b receptors were selected for visualization since each receptor functions as an important marker for an immune cell population. While CD14 is a marker for monocytes, types of leukocytes that actively contribute to an adaptive immune response by differentiating as macrophages and dendritic cells, CD19 and CD79b are markers for B cells, types of lymphocytes that are primarily involved in producing antibodies against foreign antigens and serving as antigen-presenting cells. The lymphocyte to monocyte ratio has further clinical utility as it can serve as an independent predictor of overall survival in patients undergoing curative resection with colorectal cancer (44).

Figures 6 and 7 show these scatterplots for select receptors and, overall, indicate that compared to the imputed abundance profiles obtained by techniques such as ALRA, MAGIC and RNA, SPECK-based estimated abundance profiles for CD14, CD19 and CD79b show a distinctive separation of cells in high-density regions that are highly correlated with ADT data, thereby underscoring an important contribution of SPECK in producing relatively accurate abundance estimates compared to other unsupervised imputation techniques.

Visualization of individual estimates using a low-dimensional space. Next, we visually assessed correspondence between the estimated abundance profiles and the analogous ADT data for the same select subset of $10{,}000$ cells, as used for Figures 6 and 7, from each of the PBMC and the BMMC datasets via a projection of cells onto the first two Uniform Manifold Approximation and Projection (UMAP) (21) dimensions. This second visualization allowed us to capture the estimated abundance profiles on an underlying low-dimensional manifold, thereby enabling mapping of these values over both the local and global topological structure of the data. Figures 8 and 9 show a feature plot visualization of ADT data and the estimated abundance profiles pro-

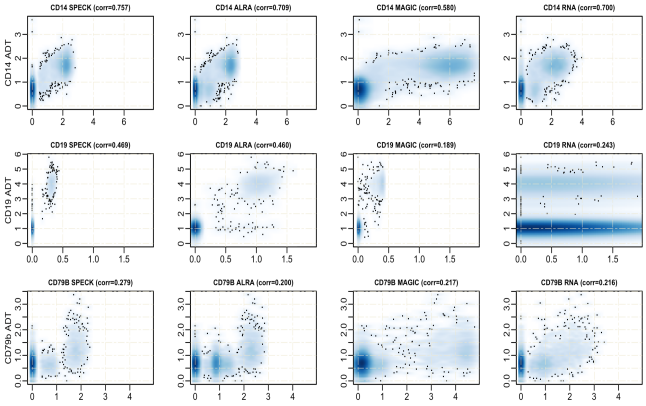

Fig. 6. Smoothed scatterplot representation of SPECK-estimated abundance profiles and imputed abundance values produced by ALRA, MAGIC and RNA against ADT data for CD14, CD19 and CD79b receptors. Plot is based on a two-dimensional kernel density estimate of estimated abundance profiles for a subset of $10{,}000$ cells from the PBMC data. The same lower and upper limits for the x-axis scale are set for every receptor based on the range of the estimated abundance profiles produced by SPECK and all other comparative methods.

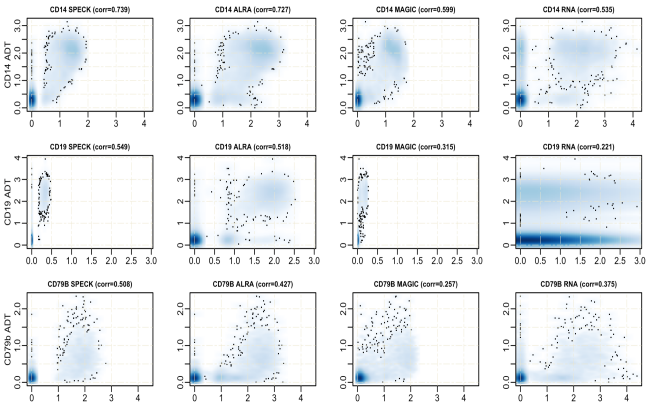

Fig. 7. Density-based smoothed scatterplot representation of SPECK-estimated values and ALRA, MAGIC and RNA-imputed abundance profiles against ADT data for CD14, CD19 and CD79b receptors for a subset of $10{,}000$ cells from the BMMC data. The same lower and upper limits of the x-axis are analogously determined based on the range of the abundance estimates produced by all methods combined.

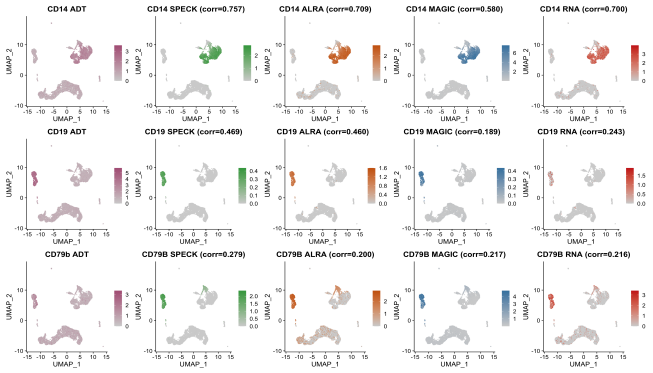

Fig. 8. Low-dimensional projection of abundance profiles for CD14, CD19 and CD79b receptors as estimated by SPECK and imputed by ALRA, MAGIC and RNA and corresponding ADT data for a subset of $10{,}000$ cells from the PBMC data.

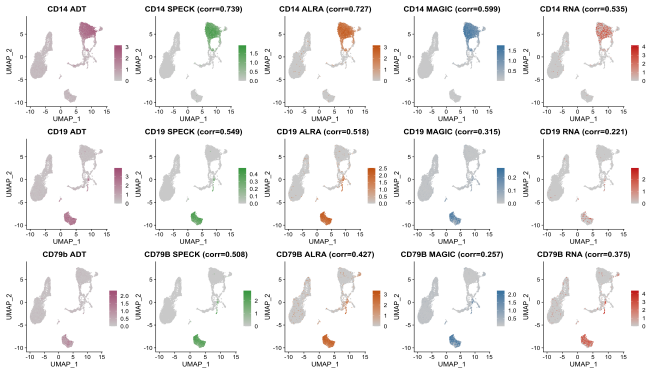

Fig. 9. Projection of estimated abundance profiles, produced by SPECK, ALRA, MAGIC and RNA, for CD14, CD19 and CD79b receptors for a subset of $10{,}000$ cells from the BMMC data.

duced by the SPECK method and imputed expression values produced by the ALRA, MAGIC and RNA techniques for the CD14, CD19 and CD79b receptors from the PBMC and the BMMC datasets, respectively. The abundance estimates produced by SPECK have a more visually defined expression profile that is primarily relegated to region of cells with high analogous ADT expression. This trend is especially visible for the CD79b receptor from both the PBMC and the BMMC datasets for which comparative techniques such as ALRA and RNA indicate a homogeneous and less distinctive expression pattern. Our results indicate a potential utility of the SPECK abundance estimation technique to the development of an improved cell-type composition identification strategy that can more precisely differentiate cells into separate, non-overlapping clusters.

Selection framework for an informed abundance estimation strategy. Following the overall evaluation of abundance profiles estimated by SPECK across all subsets and an examination of the distribution of estimated abundance values for select receptors for an individual subset, we quantified the magnitude and direction of the rank correlation values between ADT data and the SPECK-estimated and ALRA, MAGIC and RNA-imputed abundance profiles. Figure 10 displays a heatmap plot of these correlation values for 215 receptors averaged over five subsets of $60{,}000$ cells from the PBMC data while Figure 11 shows the heatmap for 25 re-

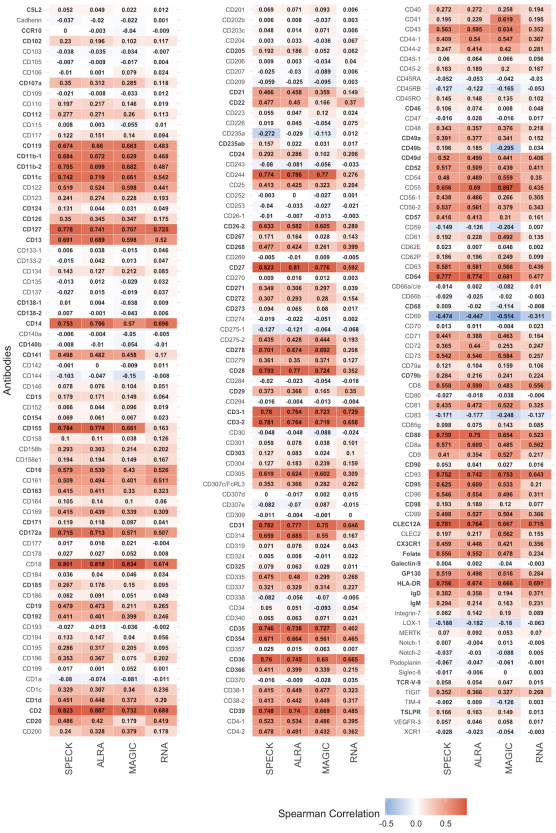

Fig. 10. Individual rank correlation values between the PBMC ADT data for 215 receptors and the SPECK-estimated or the ALRA, MAGIC and RNA-imputed abundance values, as averaged over five subsets of 60,000 cells, for 215 receptors are displayed. Bold text format is used to indicate receptors that are estimated with SPECK and are highly correlated with ADT data as compared to imputed profiles produced by ALRA, MAGIC and RNA.

ceptors averaged over five subsets of 30,000 cells from the BMMC data. In addition to confirming the results from Figures 2 and 3, which capture that SPECK overall outperforms ALRA, MAGIC and RNA in the task of abundance estimation, these figures have scientific utility as they can be referenced, when performing selective abundance estimation, to decide the most appropriate estimation/imputation strategy for a receptor based on the documented correlation with corresponding ADT data.

Discussion

In this work, we propose a new unsupervised learning method, SPECK, that uses RRR and thresholding to estimate cell surface receptor abundance for scRNA-seq data. We evaluated our approach on PBMC and BMMC CITE-seq datasets for a large set of human receptors. This comparative evaluation demonstrates that SPECK generates more accurate receptor abundance estimates over the RRR-based imputation methods ALRA and MAGIC or the direct use of the receptor RNA transcript.

Two key novel contributions of SPECK include an approach for rank-k selection and a strategy for cluster-based thresholding of the reconstructed gene expression values. Although our rank selection procedure is a heuristic method, motivated by the commonly used elbow method for determining the number of principal components, we find that it works well in practice relative to techniques that have a more formal statistical basis (e.g., methods that rely on the random matrix theory-based null distribution of eigenvalues (45), or methods that generate a null distribution via resampling (21, 46)). One strength of our rank selection procedure relative to the technique used by ALRA is that it does not depend on categorizing the smallest 80-100 singular values as noise, a rationale which needs additional justification. Our approach instead relies on ensuring that there exists marginal spacing between consecutive eigenvalue standard deviation estimates using a cutoff of 0.01 (47). The proposed thresholding mechanism also differs from existing thresholding schemes, e.g., the ALRA thresholding, in two ways. First, its performance is not contingent on assumptions corresponding to data where biological zeros are known. Second, since our strategy does not always threshold a gene and, moreover, does not necessarily apply the same quantile to threshold each gene, we are able to recover more accurate estimated profiles.

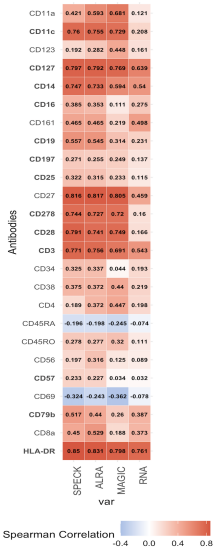

Fig. 11. Rank correlation values between the BMMC ADT data and SPECK-estimated abundance profiles or ALRA, MAGIC and RNA-imputed values, as averaged over five subsets of 30,000 cells, for 25 receptors are displayed. Receptors in bold text are estimated with SPECK and are most correlated with ADT data as compared to imputed profiles produced by ALRA, MAGIC and RNA.

A third important novel contribution of this proposed framework is that it outlines an extensive evaluation strategy for benchmarking the effectiveness of unsupervised methods for estimating receptor abundance using scRNA-seq data. With performance comparisons performed across multiple cell subsets for at least 205 and at most 215 receptors from the PBMC data and 25 receptors from the BMMC data, this evaluation provides information on the relative performance of unsupervised method across a wide range of human receptors.

One limitation of our proposed approach is that SPECK is not necessarily the best abundance estimation strategy for all receptors. For example, SPECK produces a negative correlation for CD69 for both the PBMC and the BMMC data as shown by Figures 10 and 11, respectively. While this correlation is in agreement with the negative correlations returned by

comparative methods, it still points to a deficiency in the use of transcriptomics data to infer protein abundance. We hope to address this limitation in future work by using pathway analysis to identify the biological characteristics of receptors whose abundance can be accurately estimated from scRNA-seq using unsupervised methods such as SPECK. Such insights could then be leveraged by researchers to help determine whether unsupervised estimation of receptor abundance is feasible for a given investigation or whether direct proteomic measurements are motivated. A second limitation of our approach stems from a key deficiency of scRNA-seq data, namely, that current scRNA-seq protocols can only capture static transcriptional states of cells at a specific point in time. This limitation prevents the applicability of SPECK to surface protein abundance estimation for dynamic biological processes such as tissue regeneration. We aim to address this limitation in future work by providing an implementation of SPECK that can incorporate the recently developed RNA velocity method (48) for estimating RNA abundance at future transcriptional cell states, thereby extending applicability to abundance estimation for dynamic processes such as cell differentiation. In addition to these expanded features, an R package implementation for SPECK is currently in development and will be soon released on the Comprehensive R Archive Network (CRAN).

In conclusion, SPECK is a promising approach for unsupervised estimation of surface receptor abundance for scRNA-seq data that addresses limitations of existing imputation methods such as ALRA and MAGIC. The cell surface receptor abundance profiles generated by SPECK have important scientific utility for the analysis of single cell data with specific relevance to the identification of the cell (sub)types, cell phenotypes, and cell-cell signaling present in a tissue. Improved support for these single cell analysis tasks will have a meaningful impact on the basic research supporting precision medicine applications, especially in the immunology domain.

ACKNOWLEDGEMENTS

This work was funded by National Institutes of Health grants R35GM146586, R21CA253408, P20GM130454 and P30CA023108. We would like to acknowledge the supportive environment at the Geisel School of Medicine at Dartmouth where this research was performed.

Bibliography

1. Scott J. Emrich, W. Brad Barbazuk, Li Li, and Patrick S. Schnable. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research*, 17(1):69–73, January 2007. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.5145806. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.
2. Nicole Cloonan, Alistair R. R. Forrest, Gabriel Kolle, Brooke B. A. Gardiner, Geoffrey J. Faulkner, Mellissa K. Brown, Darrin F. Taylor, Anita L. Steptoe, Shivangi Wani, Graeme Bethel, Alan J. Robertson, Andrew C. Perkins, Stephen J. Bruce, Clarence C. Lee, Swati S. Ranade, Heather E. Peckham, Jonathan M. Manning, Kevin J. McKernan, and Sean M. Grimmond. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, 5(7):613–619, July 2008. ISSN 1548-7105. doi: 10.1038/nmeth.1223. Number: 7 Publisher: Nature Publishing Group.
3. Ali Mortazavi, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, July 2008. ISSN 1548-7105. doi: 10.1038/nmeth.1226. Number: 7 Publisher: Nature Publishing Group.
4. Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B. Tuch, Asim Siddiqui, Kaiqin Lao, and M. Azim Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6 (5):377–382, May 2009. ISSN 1548-7105. doi: 10.1038/nmeth.1315. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 5 Primary_atype: Research Publisher: Nature Publishing Group.
5. Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, May 2015. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2015.05.002. Publisher: Elsevier.
6. Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049, January 2017. ISSN 2041-1723. doi: 10.1038/ncomms14049. Number: 1 Publisher: Nature Publishing Group.
7. Dvir Aran, Agnieszka P. Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P. Naikawadi, Paul J. Wolters, Adam R. Abate, Atul J. Butte, and Mallar Bhattacharya. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, 20(2):163–172, February 2019. ISSN 1529-2916. doi: 10.1038/s41590-018-0276-y. Number: 2 Publisher: Nature Publishing Group.
8. Alma Andersson, Joseph Bergenstråhle, Michaela Asp, Ludvig Bergenstråhle, Aleksandra Jurek, José Fernández Navarro, and Joakim Lundeberg. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Communications Biology*, 3(1):1–8, October 2020. ISSN 2399-3642. doi: 10.1038/s42003-020-01247-y. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational models;Data integration;Transcriptomics Subject_term_id: computational-models;data-integration;transcriptomics.
9. Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A. Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, 14(10):979–982, October 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4402. Number: 10 Publisher: Nature Publishing Group.
10. Simon Cabello-Aguilar, Mélissa Alame, Fabien Kon-Sun-Tack, Caroline Fau, Matthieu Lacroix, and Jacques Colinge. SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Research*, 48(10):e55–e55, June 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa183.
11. W. A. Bonner, H. R. Hulett, R. G. Sweet, and L. A. Herzenberg. Fluorescence Activated Cell Sorting. *Review of Scientific Instruments*, 43(3):404–409, March 1972. ISSN 0034-6748. doi: 10.1063/1.1685647. Publisher: American Institute of Physics.
12. Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K. Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, September 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4380. Number: 9 Publisher: Nature Publishing Group.
13. Zilu Zhou, Chengzhong Ye, Jingshu Wang, and Nancy R. Zhang. Surface protein imputation from single cell transcriptomes by deep neural networks. *Nature Communications*, 11 (1):651, January 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-14391-0. Number: 1 Publisher: Nature Publishing Group.
14. Yansheng Liu, Andreas Beyer, and Ruedi Aebersold. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*, 165(3):535–550, April 2016. ISSN 00928674. doi: 10.1016/j.cell.2016.03.014.
15. Jingyi Jessica Li, Peter J. Bickel, and Mark D. Biggin. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ*, 2: e270, February 2014. ISSN 2167-8359. doi: 10.7717/peerj.270. Publisher: PeerJ Inc.
16. Marko Jovanovic, Michael S. Rooney, Philipp Mertins, Dariusz Przybylski, Nicolas Chevrier, Rahul Satija, Edwin H. Rodriguez, Alexander P. Fields, Schraga Schwartz, Raktima Raychowdhury, Maxwell R. Mumbach, Thomas Eisenhaure, Michal Rabani, Dave Gennert, Diana Lu, Toni Delorey, Jonathan S. Weissman, Steven A. Carr, Nir Hacohen, and Aviv Regev. Dynamic profiling of the protein life cycle in response to pathogens. *Science*, 347(6226): 1259038, March 2015. doi: 10.1126/science.1259038. Publisher: American Association for the Advancement of Science.
17. Asfa Alli Shaik, Sheena Wee, Rachel Hai Xia Li, Zhen Li, Tom J. Carney, Sinnakaruppan Mathavan, and Jayantha Gunaratne. Functional Mapping of the Zebrafish Early Embryo Proteome and Transcriptome. *Journal of Proteome Research*, 13(12):5536–5550, December 2014. ISSN 1535-3893. doi: 10.1021/pr5005136. Publisher: American Chemical Society.
18. Wenpin Hou, Zhicheng Ji, Hongkai Ji, and Stephanie C. Hicks. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biology*, 21(1):218, August 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02132-x.
19. David van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J. Carr, Cassandra Burdziak, Kevin R. Moon, Christine L. Chaffer, Diwakar Pattabiraman, Brian Bierie, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe'er. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, 174(3):716–729.e27, July 2018. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2018.05.061. Publisher: Elsevier.
20. George C. Linderman, Jun Zhao, Manolis Roulis, Piotr Bielecki, Richard A. Flavell, Boaz Nadler, and Yuval Kluger. Zero-preserving imputation of single-cell RNA-seq data. *Nature Communications*, 13(1):192, January 2022. ISSN 2041-1723. doi: 10.1038/s41467-021-27729-z. Number: 1 Publisher: Nature Publishing Group.
21. Rahul Satija, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, May 2015. ISSN 1546-1696. doi: 10.1038/nbt.3192. Number: 5 Publisher: Nature Publishing Group.
22. Arnav Kapur, Kshitij Marwah, and Gil Alterovitz. Gene expression prediction using low-rank matrix completion. *BMC Bioinformatics*, 17(1):243, June 2016. ISSN 1471-2105. doi:

10.1186/s12859-016-1106-6.

23. Matt Silver, Peng Chen, Ruoying Li, Ching-Yu Cheng, Tien-Yin Wong, E.-Shyong Tai, Yik-Ying Teo, and Giovanni Montana. Pathways-Driven Sparse Regression Identifies Pathways and Genes Associated with High-Density Lipoprotein Cholesterol in Two Asian Cohorts. *PLOS Genetics*, 9(11):e1003939, November 2013. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003939. Publisher: Public Library of Science.
24. Zhenlong Sun, Jing Yang, and Xiaoye Li. Differentially Private Singular Value Decomposition for Training Support Vector Machines. *Computational Intelligence and Neuroscience*, 2022: 2935975, March 2022. ISSN 1687-5265. doi: 10.1155/2022/2935975.
25. Yuji Nakatsukasa and Nicholas J. Higham. Stable and Efficient Spectral Divide and Conquer Algorithms for the Symmetric Eigenvalue Decomposition and the SVD. *SIAM Journal on Scientific Computing*, 35(3):A1325–A1349, January 2013. ISSN 1064-8275. doi: 10.1137/120876605. Publisher: Society for Industrial and Applied Mathematics.
26. Koki Tsuyuzaki, Hiroyuki Sato, Kenta Sato, and Itoshi Nikaido. Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. *Genome Biology*, 21(1):9, January 2020. ISSN 1474-760X. doi: 10.1186/s13059-019-1900-3.
27. M. Gu. Subspace Iteration Randomization and Singular Value Problems. *SIAM Journal on Scientific Computing*, 37(3):A1139–A1173, January 2015. ISSN 1064-8275. doi: 10.1137/130938700. Publisher: Society for Industrial and Applied Mathematics.
28. N. Benjamin Erichson, Sergey Voronin, Steven L. Brunton, and J. Nathan Kutz. Randomized Matrix Decompositions Using R. *Journal of Statistical Software*, 89:1–48, June 2019. ISSN 1548-7660. doi: 10.18637/jss.v089.i11.
29. N. Halko, P. G. Martinsson, and J. A. Tropp. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review*, 53 (2):217–288, January 2011. ISSN 0036-1445. doi: 10.1137/090771806. Publisher: Society for Industrial and Applied Mathematics.
30. Maciej Dobrzyński, Dirk Fey, Lan K. Nguyen, and Boris N. Kholodenko. Bimodal Protein Distributions in Heterogeneous Oscillating Systems. In David Gilbert and Monika Heiner, editors, *Computational Methods in Systems Biology*, Lecture Notes in Computer Science, pages 17–28, Berlin, Heidelberg, 2012. Springer. ISBN 978-3-642-33636-2. doi: 10.1007/978-3-642-33636-2_3.
31. William J. Blake, Gábor Balázsi, Michael A. Kohanski, Farren J. Isaacs, Kevin F. Murphy, Yina Kuang, Charles R. Cantor, David R. Walt, and James J. Collins. Phenotypic Consequences of Promoter-Mediated Transcriptional Noise. *Molecular Cell*, 24(6):853–865, December 2006. ISSN 1097-2765. doi: 10.1016/j.molcel.2006.11.003. Publisher: Elsevier.
32. Sabrina L. Spencer, Suzanne Gaudet, John G. Albeck, John M. Burke, and Peter K. Sorger. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature*, 459(7245): 428–432, May 2009. ISSN 1476-4687. doi: 10.1038/nature08012. Number: 7245 Publisher: Nature Publishing Group.
33. Murat Acar, Jerome T. Mettetal, and Alexander van Oudenaarden. Stochastic switching as a survival strategy in fluctuating environments. *Nature Genetics*, 40(4):471–475, April 2008. ISSN 1546-1718. doi: 10.1038/ng.110. Number: 4 Publisher: Nature Publishing Group.
34. Haizhou Wang and Mingzhou Song. Ckmeans.1d.dp: Optimal k-means Clustering in One Dimension by Dynamic Programming. *The R Journal*, 3(2):29, 2011. ISSN 2073-4859. doi: 10.32614/RJ-2011-015.
35. S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, March 1982. ISSN 1557-9654. doi: 10.1109/TIT.1982.1056489. Conference Name: IEEE Transactions on Information Theory.
36. Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, June 2021. ISSN 0092-8674. doi: 10.1016/j.cell.2021.04.048.
37. Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, January 2002. ISSN 0305-1048. doi: 10.1093/nar/30.1.207.
38. Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902.e21, June 2019. ISSN 0092-8674. doi: 10.1016/j.cell.2019.05.031.
39. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. https://www.R-project.org/.
40. John Aitchison. Measures of location of compositional data sets. *Mathematical Geology*, 21(7):787–790, October 1989. ISSN 1573-8868. doi: 10.1007/BF00893322.
41. Susan Tweedie, Bryony Braschi, Kristian Gray, Tamsin E M Jones, Ruth L Seal, Bethan Yates, and Elspeth A Bruford. Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Research*, 49(D1):D939–D946, January 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa980.
42. Patrick Schober, Christa Boer, and Lothar A. Schwarte. Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768, May 2018. ISSN 0003-2999. doi: 10.1213/ANE.0000000000002864.
43. Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246.
44. Joseph C. Y. Chan, David L. Chan, Connie I. Diakos, Alexander Engel, Nick Pavlakis, Anthony Gill, and Stephen J. Clarke. The Lymphocyte-to-Monocyte Ratio is a Superior Predictor of Overall Survival in Comparison to Established Biomarkers of Resectable Colorectal Cancer. *Annals of Surgery*, 265(3):539–546, March 2017. ISSN 0003-4932. doi: 10.1097/SLA.0000000000001743.
45. Iain M. Johnstone. On the Distribution of the Largest Eigenvalue in Principal Components Analysis. *The Annals of Statistics*, 29(2):295–327, 2001. ISSN 0090-5364. Publisher: Institute of Mathematical Statistics.
46. Neo Christopher Chung and John D. Storey. Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics*, 31(4):545–554, February 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu674.
47. I.T. Jolliffe. *Principal Component Analysis*. Springer, New York, NY, 2nd edition, 2002. ISBN 9780387954424.
48. Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V. Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, August 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0414-6. Number: 7719 Publisher: Nature Publishing Group.