STREAK: A Supervised Cell Surface Receptor Abundance Estimation Strategy for Single Cell RNA-Sequencing Data using Feature Selection and Thresholded Gene Set Scoring

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

[1]Department of Biomedical Data Science
Geisel School of Medicine
Dartmouth College
Hanover, NH 03755, USA
azka.javaid.gr@dartmouth.edu
rob.frost@dartmouth.edu

Abstract

The accurate estimation of cell surface receptor abundance is a pressing problem for the analysis of single cell transcriptomics given its centrality to tasks like cell type and phenotype categorization and cell-cell interaction quantification. We previously developed an unsupervised receptor abundance estimation technique named SPECK (Surface Protein abundance Estimation using CKmeans-based clustered thresholding) to address the challenges associated with accurate abundance estimation. In that paper, we concluded that SPECK results in improved correspondence with Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq) data relative to comparative unsupervised techniques and provides a framework for abundance estimation using only single-cell RNA-sequencing (scRNA-seq) data. In this paper, we outline a new supervised receptor abundance estimation method called STREAK (gene Set Testing-based Receptor abundance Estimation using Adjusted distances and cKmeans thresholding) that leverages gene expression associations learned from joint scRNA-seq/CITE-seq training data to estimate receptor abundance for scRNA-seq target data using a thresholded gene set scoring mechanism. We evaluate STREAK's performance relative to unsupervised and supervised techniques using two evaluation strategies on three joint scRNA-seq/CITE-seq datasets that represent two human tissue types. We conclude that STREAK outperforms other abundance estimation techniques using a more biologically interpretable and transparent statistical model.

Index terms— receptor abundance estimation, gene set scoring, single cell transcriptomics

1 Introduction

Single cell RNA-sequencing (scRNA-seq) technologies, such as the 10X Chromium system [1], can now cost effectively profile gene expression in tens-of-thousands of cells dissociated from a single tissue sample [2, 3]. The transcriptomic data captured via scRNA-seq gives researchers unprecedented insight into the cell types and phenotypes that comprise complex tissues such as the tumor microenvironment or brain. Single cell transcriptomics can also facilitate the characterization of cell-cell signaling [4] and the identification of co-regulated genetic modules and gene-regulatory networks [5]. A critical element of these cell type/phenotype identification and cell-cell interaction analysis tasks is the accurate estimation of receptor protein abundance. Although direct receptor protein measurements are sometimes available, via either fluorescence-activated cell sorting (FACS)-based [6] enrichment prior to scRNA-seq analysis or through joint scRNA-seq/CITE-seq (Cellular Indexing of Transcriptomes and Epitopes by Sequencing) [7] profiling, most single cell data sets contain only gene expression data. For such data sets, receptor abundance estimates can be generated using either a supervised method (i.e., a method that leverages a predictive model trained on joint transcriptomic/proteomic data) or an unsupervised method (i.e., a method that generates estimates from the target scRNA-seq data without reference to a trained model).

A common unsupervised approach for estimating receptor abundance uses the expression of the associated mRNA transcript as a proxy for the receptor protein. While this approach is plausible, it can result in low quality estimates

given the significant sparsity of scRNA-seq data [8]. This sparsity problem is well illustrated by scRNA-seq data for FAC sorted immune cells from Zheng et al. [1], which found that a large fraction of cells have no detectable expression of the transcript whose corresponding receptor is positively expressed in those cells, e.g., only 20% of CD19+ B cells expressed the CD19 transcript.

To overcome scRNA-seq sparsity, the standard bioinformatics workflow uses the average expression of the receptor transcript across large populations or cluster of cells to estimate receptor protein abundance. Although a cluster-based analysis mitigates sparsity, it has two major limitations. First, it assumes that receptor abundance is uniform across all cells in a given cluster, ignoring potentially significant within cluster heterogeneity. Second, it generates only a small number of independent receptor abundance estimates (one per cluster), which limits insight into the joint distribution of different receptors. A key benefit of single cell transcriptomics is the dramatic increase in sample size, with each cell providing a distinct expression profile. These large sample sizes can significantly improve estimates of the marginal and joint distribution of gene expression values. To support the cluster-free and unsupervised estimation of receptor abundance, we recently developed the Surface Protein abundance Estimation using CKmeans-based clustered thresholding (SPECK) method. SPECK uses a low-rank reconstruction of scRNA-seq data followed by a clustering-based thresholding of the reconstructed gene expression data to generate non-sparse estimates of receptor transcripts. While the SPECK approach significantly outperforms the naïve approach of directly using the receptor transcript, accuracy is poor for a number of biologically important receptors, e.g., receptors such as CD69 for which protein abundance is not closely correlated with transcript expression [9]. For such receptors, a supervised approach is needed that can leverage joint transcriptomic/proteomic data to identify the gene expression signature that is most closely associated with protein abundance.

Existing supervised methods for receptor abundance estimation include cTP-net (single cell Transcriptome to Protein prediction with deep neural network) [10] and PIKE-R2P (Protein–protein Interaction network-based Knowledge Embedding with graph neural network for single-cell RNA to Protein prediction) [11]. cTP-net uses a multiple branch deep neural network (MB-DNN) on joint scRNA-seq/CITE-seq training data to generate cell-level surface protein abundance estimates for target scRNA-seq data. While the abundance estimates generated using cTP-net are highly correlated with the corresponding protein measurements, only 24 immunophenotype markers/receptors are supported. Additionally since cTP-net uses a deep learning model trained on the same immune cell populations (i.e., peripheral blood mononuclear cells (PBMC), cord blood mononuclear cells (CBMC) and bone marrow mononuclear cells (BMMC)) to perform estimation via transfer learning, it may not be able to capture gene expression patterns that are specific to an individual dataset or that generalize to non-immune-related surface markers. PIKE-R2P uses a PPI (protein-protein interactions)-based graph neural network (GNN) integrated with prior knowledge embeddings to estimate receptor abundance values. This method is based on the assumption that since gene expression regulation mechanisms are likely shared between proteins, such similarities can be used to generate protein-protein interactions that can then be leveraged in a GNN. Like cTP-net, PIKE-R2P is limited to an analysis and assessment of model prediction results on 10 receptors. A second important limitation of neural network-based models such as PIKE-R2P and cTP-net is that the weights retrieved via the training process are treated as fixed entities and are not sufficiently transparent or biologically interpretable to be manually fine-tuned by medical practitioners.

Given the performance issues with unsupervised methods like SPECK and the limited support, generalizability and applicability of existing supervised estimation techniques such as PIKE-R2P and cTP-net, alternative receptor abundance estimation methods are needed. Here, we propose a novel supervised approach for cell-level receptor abundance estimation for scRNA-seq data named STREAK (gene Set Testing-based Receptor abundance Estimation using Adjusted distances and cKmeans thresholding) that leverages co-expression associations learned from joint scRNA-seq/CITE-seq training data to perform subsequent gene set scoring and thresholding on target scRNA-seq data. We validate STREAK on three datasets representing two different tissue types and compare its performance against both unsupervised receptor abundance estimation techniques, such as SPECK and normalized RNA transcript counts, and supervised techniques, such as cTP-net. In addition to supervised receptor abundance estimation techniques that leverage trained associations learned from separate training datasets like cTP-net, we also evaluate STREAK against the Random Forest (RF) algorithm, a comparison that is motivated by a recent study which analyzed performance of tree-based ensemble methods and neural networks and found RF to have a superior performance versus neural nets for the task of receptor abundance estimation [12]. We do not perform comparison against PIKE-R2P since implementations do not currently exist on CRAN or via code repository platforms like Github. We evaluate all comparisons between the estimated receptor abundance profiles and CITE-seq ADT data using the Spearman rank correlation coefficient. Overall, we observe that STREAK has superior performance relative to comparative methods on all analyzed datasets, which highlights the accuracy and generalizability of our proposed estimation approach. STREAK has additional

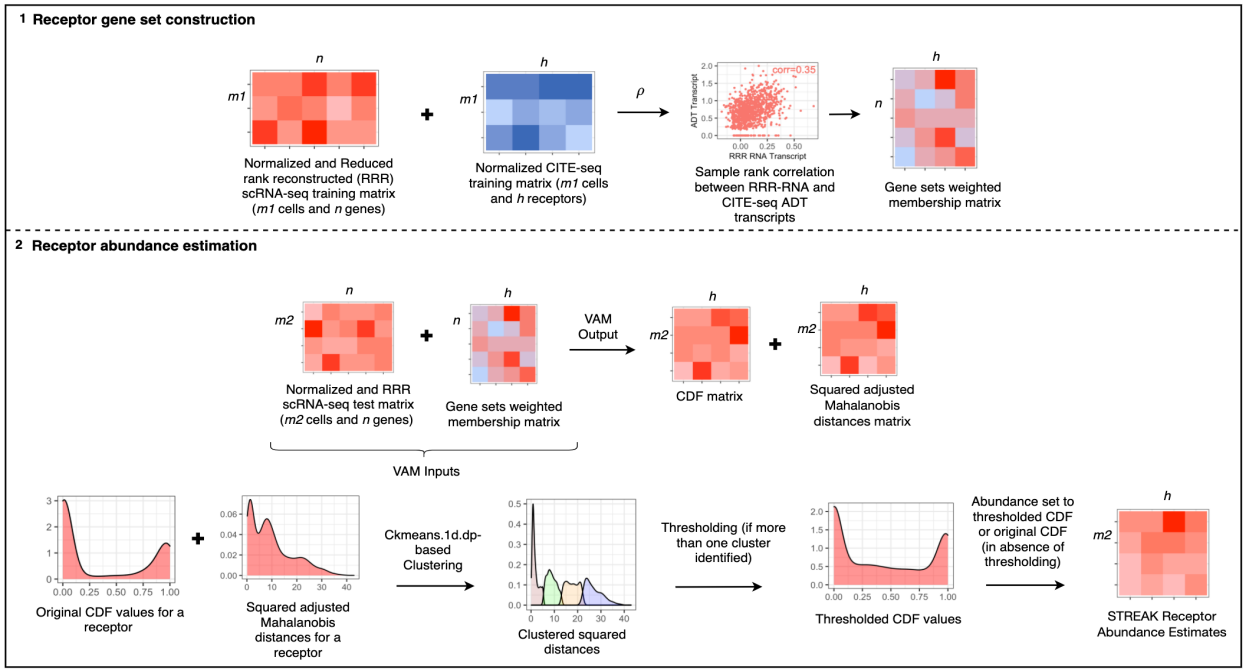


Figure 1: STREAK schematic with step (1) corresponding to the training co-expression analysis and (2) corresponding to gene set scoring and subsequent clustering and thresholding to achieve cell-specific estimated receptor abundance profiles.

clinical utility as co-expression-trained gene weights can be customized for subsequent gene set scoring to facilitate a more interpretable and adaptable receptor abundance estimation strategy.

2 Methods

2.1 STREAK method overview

STREAK estimates receptor abundance levels by leveraging associations between gene expression and protein abundance to enable receptor gene set scoring of scRNA-seq target data (see Figure 1). STREAK generates weighted receptor gene sets using joint scRNA-seq/CITE-seq training data with the gene set for each receptor containing the genes whose normalized and reconstructed scRNA-seq expression values are most strongly correlated with CITE-seq receptor protein abundance

2.2 Receptor gene set construction

Receptor abundance estimation using STREAK requires the specification of a weighted gene set for each supported receptor. To generate the results presented in this paper, we created the receptor gene sets using joint scRNA-seq/CITE-seq training data. It is important to note, however, that the weighted gene sets used by STREAK can be generated using alternative approaches, e.g., manually specified based on prior biological knowledge. The first step in the training process employed for this paper involved generating $\mathbf{X}^*_{tR}$, a rank-k reconstruction of the $m_1 \times n$ training scRNA-seq matrix $\mathbf{X}_{tR}$ holding the normalized gene expression counts for n genes in m_1 cells. Reduced rank reconstruction (RRR) of the scRNA-seq data was performed using the randomzied SVD algorithm from the rsvd R package [13] based on the rank selection procedure proposed in the SPECK paper [9], which leverages the rate of change in the standard deviation of the non-centered sample principal components to compute the estimated rank-k. The analogous $m_1 \times h$ CITE-seq counts matrix $\mathbf{X}_{tP}$ from the joint scRNA-seq/CITE-seq training data was normalized using the centered log-ratio (CLR) transformation supported by Seurat [14–17]. After normalization and RRR, the $\mathbf{X}^*_{tR}$ and $\mathbf{X}_{tP}$ matrices were utilized to define gene sets for each of the h supported receptors, with the set for each receptor containing the genes whose reconstructed expression values from $\mathbf{X}^*_{tR}$ have the largest positive Spearman rank correlation with the corresponding receptor's normalized CITE-seq values from $\mathbf{X}_{tP}$. The rank correlation values are also used to define positive weights for the genes in each set. For the results reported in this paper, a gene set size of 10 was used.

Algorithm 1 STREAK Algorithm (Receptor gene set construction)

Input: $X_{tR} \in \mathbb{R}^{m_1 \times n}$ - scRNA-seq training counts
Input: $X_{tP} \in \mathbb{R}^{m_1 \times h}$ - CITE-seq training counts
Output: $A \in \mathbb{R}^{n \times h}$ - Gene sets weighted membership matrix

Normalization and RRR of training data

$X_{tR} \xrightarrow{\text{Log-Normalization+RRR}} X^*_{tR}; X_{tP} \xrightarrow{\text{CLR-Normalization}} X_{tP}$

Co-expression analysis

for $i \leftarrow 1$ to h **do** ▹ Associations between scRNA-seq and CITE-seq training data
  for $j \leftarrow 1$ to n **do**
    $r_x \leftarrow rank(X^*_{tR}[,j])$
    $r_y \leftarrow rank(X_{tP}[,i])$
    $A \leftarrow \frac{cov(r_x, r_y)}{\sigma_{r_x}\sigma_{r_y}}$ ▹ Pearson correlation on rank data (i.e., Spearman rank correlation)
  end for
end for

2.3 Receptor abundance estimation

2.3.1 Single cell gene set scoring

Given weighted gene sets for h target receptors, the STREAK method uses single cell gene set scoring to generate receptor abundance estimates. For an $m_2 \times n$ matrix X_R that holds scRNA-seq data for m_2 cells and n genes, STREAK first performs normalization and RRR to produce the $m_2 \times n$ matrix $\mathbf{X}^*_R$. Using $\mathbf{X}^*_R$, STREAK next computes cell-level scores for each of the h target receptors using the Variance-adjusted Mahalanobis (VAM) [18] method.

Execution of the VAM method requires two input matrices:

1. $\mathbf{X}^*_R$: a $m_2 \times n$ target matrix containing the positive normalized and RRR counts for n genes in m_2 cells as measured by scRNA-seq.

2. $\mathbf{A}$: a $n \times h$ matrix that captures the weighted annotation of n genes to h gene sets, i.e., gene sets for each of the h receptor proteins. If gene i is included in gene set j, then element $a_{i,j}$ holds the gene weight, otherwise, $a_{i,j} = 0$.

VAM outputs a $m_2 \times h$ matrix, $\mathbf{M}$, that holds the cell-specific squared modified Mahalanobis distances for m_2 cells and h gene sets, and a $m_2 \times h$ matrix, $\mathbf{S}$, that holds the cell-level scores for m_2 cells and h gene sets. The computation for both matrices $\mathbf{M}$ and $\mathbf{S}$ is detailed below (see the VAM paper [18] for additional details).

1. **Technical variances estimation:** The length n vector σ^2_{tech} holding the technical variance of each gene in $\mathbf{X}^*_R$ is first computed using the Seurat variance decomposition approach for either the log-normalized or SCTransform-normalized data [19]. Alternatively, the elements of σ^2_{tech} can be set to the sample variance of each gene in $\mathbf{X}^*_R$ under the assumption that the observed marginal variance of each gene is entirely technical.

2. **Modified Mahalanobis distances computation:** Given $\mathbf{M}$, a $m_2 \times h$ matrix of squared values of a modified Mahalanobis distance, each column k of $\mathbf{M}$, which holds the cell-level squared distances for gene set k, is calculated as $M[,k] = X_k^T(I_g\sigma^2_{g,tech})^{-1}X_k$ where g corresponds to the size of the gene set k, X_k is a $m_2 \times g$ matrix containing the g columns of $\mathbf{X}^*_R$ relating to the members of set k, I_g is a $g \times g$ identity matrix, and $\sigma^2_{g,tech}$ contains the elements of σ^2_{tech} relating to the g genes in set k. Mahalanobis distances are additionally recomputed on a version of $\mathbf{X}^*_R$ where the row labels of each column are randomly permuted, which captures the distribution of the squared modified Mahalanobis distances under the H_0 that the normalized and RRR expression values in $\mathbf{X}^*_R$ are uncorrelated with only the technical variance. $\mathbf{X_p}$ thereby represents the row-permuted version $\mathbf{X}^*_R$ and $\mathbf{M_p}$ represents the $m_2 \times h$ matrix that contains the squared modified Mahalanobis distances computed on $\mathbf{X_p}$.

3. **Gamma distribution fit:** A gamma distribution is individually fit to the non-zero elements in each column of $\mathbf{M_p}$ using the method of maximum likelihood. Alternatively, gamma distributions can be fit directly on $\mathbf{M}$ to alleviate the computational costs of generating $\mathbf{X_p}$ and $\mathbf{M_p}$.

4. **Cell-specific scores computation:** Cell-level gene set scores are defined to be the gamma cumulative distribution function (CDF) value for each element of **M**.

2.3.2 Univariate clustering and thresholding

Following the generation of cell-specific squared distances, matrix **M**, and cell-specific gene set scores, matrix **S**, the $Ckmeans.1d.dp$ algorithm from Ckmeans.1d.dp v3.3.3 [20, 21] is used to perform univariate clustering on **M**. Each column k of **M**, which holds the cell-specific squared distances for gene set k, is clustered with the number of computed clusters bound between one and four. If more than one cluster is identified, then all the non-zero cell-specific gene set scores corresponding to the indices of the least-valued cluster for k from the analogous matrix **S** are set to zero. All zero values corresponding to the indices of the least-valued cluster and the remaining non-zero and zero values corresponding to the indices of the higher-valued clusters for k are retained. If only one cluster is identified for gene set k of **M**, then thresholding is not performed and the analogous cell-specific gene set scores for gene set k of **S** are preserved as estimated abundance estimates.

Algorithm 2 STREAK Algorithm (Receptor abundance estimation)

Input: $X_R \in \mathbb{R}^{m_2 \times n}$ - scRNA-seq target counts
Input: $A \in \mathbb{R}^{n \times h}$ - Gene sets weighted membership matrix
Output: $S_R \in \mathbb{R}^{m_2 \times h}$ - Estimated abundance profiles

Normalization and RRR of target data

$X_R \xrightarrow{\text{Log-Normalization+RRR}} X_R^*$

Gene set scoring and thresholding

```
for k ← 1 to h do
    topGenes ← names(A[1 : 10, k])
    topGenesWeights ← A[1 : 10, k]
    testRRR ← X*_R[, topGenes]                                   ▷ Top 10 most co-expressed genes
    vamOut ← vam(testRRR, gene.weights = topGenesWeights, gamma = T, center = F)   ▷ Gene set scoring
    vamCDF ← vamOut["cdf.value"]
    vamSqDist ← vamOut["distance.sq"]
    ckRes ← Ckmeans.1d.dp(vamSqDist, k = c(1 : 4))   ▷ Clustering of cell-specific squared distances
    numCluster ← ckRes["cluster"]
    valCenters ← ckRes["centers"]
    if length(valCenters) > 1 then
        minVal ← which(valCenters = min(valCenters))
        minNum ← which(numCenters = minVal)
        vamCDF[minNum] ← 0                                       ▷ Thresholding of cell-specific gene set scores
    end if
    S_R ← t(vamCDF)
end for
```

2.4 Evaluation

2.4.1 Datasets

Comparative evaluation of STREAK was performed on three publicly accessible joint scRNA-seq/CITE-seq datasets generated on two human tissue types: 1) the Hao et al. [14] human PBMC dataset (GEO [22] series GSE164378) contains 161,764 cells profiled using 10X Chromium 3' with 228 TotalSeq A antibodies, 2) the Unterman et al. [23] human PBMC dataset (GEO series GSE155224) contains 163,452 cells profiled using 10X Chromium 5' with 189 TotalSeq C antibodies, and 3) the 10X Genomics [24] human extranodal marginal zone B-cell tumor/mucosa-associated lymphoid tissue (MALT) dataset contains 8,412 cells profiled using 10X Chromium 3' with 17 TotalSeq B antibodies. The Hao et al. data (GSE164378) was generated on PBMC samples obtained from eight volunteers enrolled in a HIV vaccine trial [25, 26], with age ranging from 20 to 49 years. In comparison, the PBMC samples for the Unterman et al. data were obtained from 10 COVID-19 patients and 13 matched controls with a mean age of 71 years. The two PBMC

datasets thus consist of different underlying patient populations with varying age ranges. All three scRNA-seq/CITE-seq datasets were processed using Seurat v.4.1.0 [14–17] in R v.4.1.2 [27].

We performed method evaluation on each dataset using two approaches. First, we used 5-fold cross-validation where we utilized a subset of cells from each of the three joint scRNA-seq/CITE-seq datasets as training data and a second distinct subset of the same dataset as target data. This approach is suitable for analytical scenarios where joint scRNA-seq/CITE-seq data is only captured for a subset of cells from the original, larger population of cells for which only scRNA-seq data is available. Given this first scenario, receptor abundance levels can be estimated by mapping associations learned on the smaller subset of cells with joint scRNA-seq/CITE-seq training data to the larger population with only scRNA-seq data. Our second evaluation strategy used a cross-training-based approach to learn and evaluate associations on separate datasets. This latter approach, which was used to learn associations on the Hao data and evaluate on the Unterman data, is suitable for analytical cases where co-expression patterns learned from one joint scRNA-seq/CITE-seq data can be leveraged to perform receptor abundance estimation for another disjoint data that is of the same tissue type as the first data but only contains quantified scRNA-seq expression profiles. For both the cross-validation and the cross-training approaches, the relative correspondence between the estimated receptor abundance values for the target scRNA-seq data and the analogous ADT transcripts from the joint, target scRNA-seq/CITE-seq data was measured using the Spearman rank correlation, which was selected due to its ability to quantify relationships between the relative ranks, as opposed to the absolute values, of two variables [28].

2.4.2 Comparison methods

STREAK was evaluated against existing unsupervised and supervised-learning-based receptor abundance estimation techniques. SPECK was the only unsupervised receptor abundance estimation approach evaluated since it previously performed better than both comparative unsupervised RRR-based abundance estimation techniques such as the Markov Affinity-based Graph Imputation of Cells (MAGIC) [29] and RRR and thresholding-based methods such as the Adaptively thresholded Low-Rank Approximation (ALRA) [30]. All comparative approaches are detailed below.

- RNA transcript: RNA transcript associated with the log-normalized scRNA-seq target counts matrix was set as the receptor abundance estimate using Seurat's log-normalization procedure [17].
- SPECK: scRNA-seq target counts matrix was normalized, RRR and thresholded with the $speck$ function in the SPECK v0.1.1 R package [31]. Receptor abundance was set to this estimated output.
- cTP-net: Estimation was performed on the scRNA-seq target counts matrix with the $cTPnet$ function and default parameters using cTP-net v1.0.3 [10]. Receptor abundance was set to this estimated output. scRNA-seq matrix was not denoised with the SAVER-X package since cTP-net maintainers note that cTP-net can predict protein abundance relatively accurately without denoising. Denoising using SAVER-X was also not performed due to the associated high time complexity for large datasets and the pending Python package dependency updates required by the SAVER-X package maintainers.
- Random Forest (RF): A random forest model was trained using the $randomForest$ function with default parameters from the randomForest v4.7-1.1 R package [32]. Model training leveraged the 10 genes whose expression in the normalized scRNA-seq/CITE-seq training data was most highly correlated with the normalized CITE-seq counts for the target receptor. This trained RF model was applied to the target scRNA-seq data to generate estimated receptor abundance values.

2.4.3 Benchmark setup

Each of the cross-validation and cross-training evaluation approaches had a distinct benchmark setup. For the cross-validation approach, five individual subsets of different sizes were selected from each data to get a 20:80 train:test split where 20% of total cells were used for training and 80% were used for testing. The number of cells used in the training and target data subsets and number of total cells (i.e., the sum of cells from training and target data subsets) for each data are indicated in Table 1 for the cross-validation approach. Evaluation of the PBMC cross-training approach was performed on gene sets trained on a subset of 5,000 cells from the Hao data. For each trained subset, evaluation was performed on five subsets of 5,000, 7,000 and 10,000 cells and one subset of 50,000 cells from the Unterman data.

The lower and upper cell count limits for the cross-validation approach were individually determined for each dataset. For the Hao data, the upper limit of 60,000 total cells (i.e., 12,000 training cells and 48,000 target cells) was determined by the ability to perform RRR on 16 CPU cores without any virtual memory allocation errors. For the Unterman data, the upper limit of 50,000 total cells (i.e., 10,000 training cells and 40,000 target cells) was determined

by the total number of cells in the CITE-seq assay (50,438). For the MALT dataset, the upper limit of 8,410 total cells (i.e., 1,682 training cells and 6,728 target cells) was similarly determined by the total number of cells in the CITE-seq assay (8,412). The upper limit of 50,000 target cells for the cross-training approach was determined by the total number of cells in the Unterman data CITE-seq assay (50,438 cells).

Dataset	# Cells (Training Data)	# Cells (Target Data)	# Cells (Total Data)
PBMC Hao et al. (GSE164378)	1000	4,000	5,000
	2,500	10,000	12,500
	5,000	20,000	25,000
	10,000	40,000	50,000
	12,000	48,000	60,000
PBMC Unterman et al. (GSE155224)	1000	4,000	5,000
	2,500	10,000	12,500
	5,000	20,000	25,000
	10,000	40,000	50,000
10X Genomics MALT	1000	4,000	5,000
	1,250	5,000	6,250
	1,500	6,000	7,500
	1,682	6,728	8,410

Table 1: Number of cells in individual training and target data subsets and total number of cells for each of the three datasets for the 5-fold cross-validation approach.

From the initial 228 antibodies included in the Hao data, antibodies mapping to multiple HGNC (HUGO Gene Nomenclature Committee) [33] symbols or antibodies with their HGNC symbols not present in the feature/gene names corresponding to the scRNA-seq matrix were removed. Final cross-validation assessment was performed for 217 receptors for each subset of the Hao data. From the initial 200 antibodies included in the Unterman data, antibodies mapping to multiple HGNC symbols and mouse/rat specific antibodies were removed, resulting in 168 receptors. Since assessment was performed for a varying number of cells, receptors not expressed in smaller cell groups were dropped, resulting in assessment of either 167 or 168 receptors for the Unterman data. From the initial 17 antibodies included in the MALT data, three mouse/rat specific antibodies (IgG2a, IgG1 and IgG2b control) were removed. Final assessment was performed on 14 receptors for the MALT dataset data. For the PBMC cross-training approach, 124 receptors, overlapping between the 217 antibodies from the Hao data and the 168 antibodies from the Unterman data, were assessed.

3 Results

3.1 STREAK abundance estimates have better overall correspondence with CITE-seq data

We first quantified the proportion of abundance profiles that have high rank correlation with corresponding CITE-seq data when estimated using the STREAK method and estimated using SPECK and RNA normalization methods for the 5-fold cross-validation approach for each of the Hao, the Unterman and the MALT datasets. Similarly, we compared this percentage for abundance profiles estimated using STREAK and estimated with cTP-net and the RF algorithm for all three datasets. Each set of proportion estimates were separately visualized. As Figures 2a, 2b and 2c indicate, the percentage of receptors for which the STREAK estimates have the highest rank correlation with the CITE-seq data is larger than the corresponding percentages for SPECK or the normalized RNA approach for the Hao, Unterman and MALT datasets.

In addition to comparisons against existing unsupervised receptor abundance estimation strategies, we compared STREAK against cTP-net and the RF algorithm and similarly quantified the proportion of receptors that are estimated by each technique and have higher correspondence with CITE-seq data as measured using rank correlation. Figures 2d, 2e and 2f show comparisons between STREAK and cTP-net while Figures 2g, 2h and 2i indicate comparisons between STREAK and RF for the Hao, the Unterman and the MALT datasets, respectively. Both sets of figures show that STREAK overall performs better than cTP-net and RF across all datasets and training subset sizes as evident from the larger percentage of abundance profiles estimated using STREAK that have the highest rank correlations with analogous CITE-seq data.

We performed a similar relative comparison of STREAK for the cross-training evaluation strategy. We observed an equivalent pattern of improved correspondence for the STREAK receptor abundance estimates relative to the estimates generated using SPECK (Figure 2j), the normalized RNA approach (Figure 2j), cTP-net (Figure 2k), or RF (Figure 2l). While the relative performance benefit of STREAK for the cross-training evaluation is more incremental than for the cross-validation evaluation, the results still emphasize STREAK's ability to consistently generate more accurate abundance estimates than the comparison techniques. The fact that the relative performance of STREAK is insensitive to training and test dataset size across three datasets and two evaluation strategies further stresses the robustness of this approach for supervised receptor abundance estimation.

3.2 Individual comparisons of rank correlation estimates for a select train/test split

After performing overall comparisons of average rank correlations between the abundance profiles estimated using STREAK and comparative methods with the corresponding CITE-seq data across different training and test dataset size, we inspected trends in rank correlation estimates over a select train/test split group for the Hao, Unterman and MALT datasets. For the 5-fold cross-validation evaluation approach, we trained and evaluated the abundance values estimated using STREAK, SPECK, the normalized RNA approach, cTP-net and RF model using train/test splits of 12,000/48,000, 10,000/40,000 and 1,682/6,728 cells resulting in total cell sizes of 60,000, 50,000 and 8,410 for the Hao, Unterman and MALT datasets, respectively. We visualized rank correlations via a heatmap and used bold text format to distinguish receptors whose STREAK estimates have the highest rank correlation with analogous CITE-seq data. Figures 3a, 4a and 5a indicate that while there is some general agreement in rank correlation values between the CITE-seq data and estimated abundance profiles generated by different unsupervised estimation approaches, the abundance values estimated with STREAK have predominantly higher correlations with analogous CITE-seq data, as evident from the individual rank correlation estimates displayed in the heatmaps. An examination of individual receptor correlations reveals that with the exception of CD137L, CD15 and CD99 receptors from the Unterman data (see Figure 4a), abundance values estimated for all other receptors using STREAK from the Hao, Unterman and MALT data are positively correlated with the analogous CITE-seq data. The overall trend in positive correlation values between the STREAK-estimated abundance profiles and corresponding CITE-seq data is consistent even for receptors that are estimated using SPECK or the normalized RNA approach and are negatively correlated with analogous CITE-seq data (i.e., CD69 and CD83 from the Hao and Unterman datasets and CD15 and CD45RA from the MALT data).

In addition to comparisons against unsupervised abundance estimation methods, we visualized correspondence between CITE-seq data and abundance profiles generated using supervised techniques such as cTP-net and RF. Figures 3b, 4b and 5b indicate the rank correlation comparisons between STREAK and cTP-net while Figures 3c, 4c and 5c show the rank correlation comparisons between STREAK and RF for the Hao, Unterman and MALT datasets, respectively. While STREAK's performance lead over RF is less impressive than over SPECK and the normalized RNA approach, it is nonetheless marginally better than RF based on the quantified correspondence with CITE-seq data. STREAK-estimated abundance profiles are similarly more correlated with CITE-seq data compared to the abundance profiles estimated using cTP-net.

We similarly visualized rank correlations between CITE-seq data and abundance profiles estimated using STREAK and comparative unsupervised and supervised methods on a subset of 50,000 cells from the Unterman data for the cross-training evaluation approach. Figures 6a, 6c and 6b show comparisons of STREAK against unsupervised abundance estimation techniques, and the RF and cTP-net methods, respectively, and indicate that STREAK's performance, as learned from trained co-expression associations constructed using the Hao data, is marginally better than methods like SPECK and the normalized RNA approach as well as supervised abundance estimation techniques. STREAK's performance as evaluated using the 5-fold cross-validation strategy is, however, superior to the cross-training approach, which highlights the utility of learning from separately trained gene sets for each dataset for subsequent gene set scoring and thresholding.

In addition to the practical advantage of these correlation heatmaps as abundance estimation evaluation strategies, they also have broader scientific utility as they can be leveraged to identify the optimal abundance estimation strategy based on the documented associations between an estimated receptor and the corresponding CITE-seq data. As shown, while a majority of receptor abundance estimates quantified using STREAK have higher rank correlations with analogous CITE-seq data, some receptors still have higher rank correlations with CITE-seq data when estimated using alternative abundance estimation approaches. For example, the Spearman rank correlation computed between the estimated CD14 abundance profile and the corresponding CITE-seq values for the Hao data is higher for abundance values estimated using SPECK (ρ=0.754) as compared to estimates produced using STREAK (ρ=0.744). Similarly, rank correlation between the CD301 abundance profile and the corresponding CITE-seq values for the Hao data

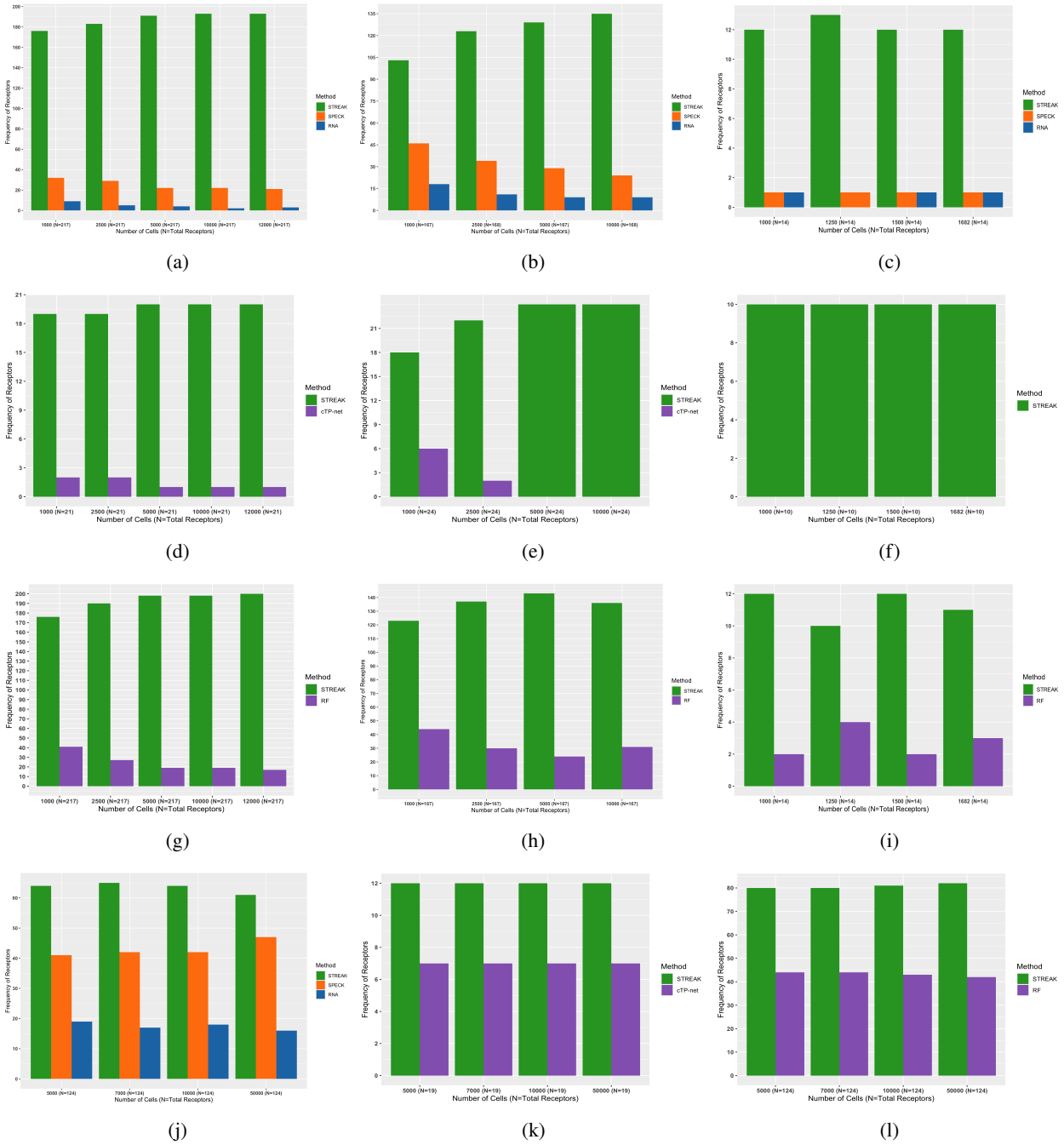
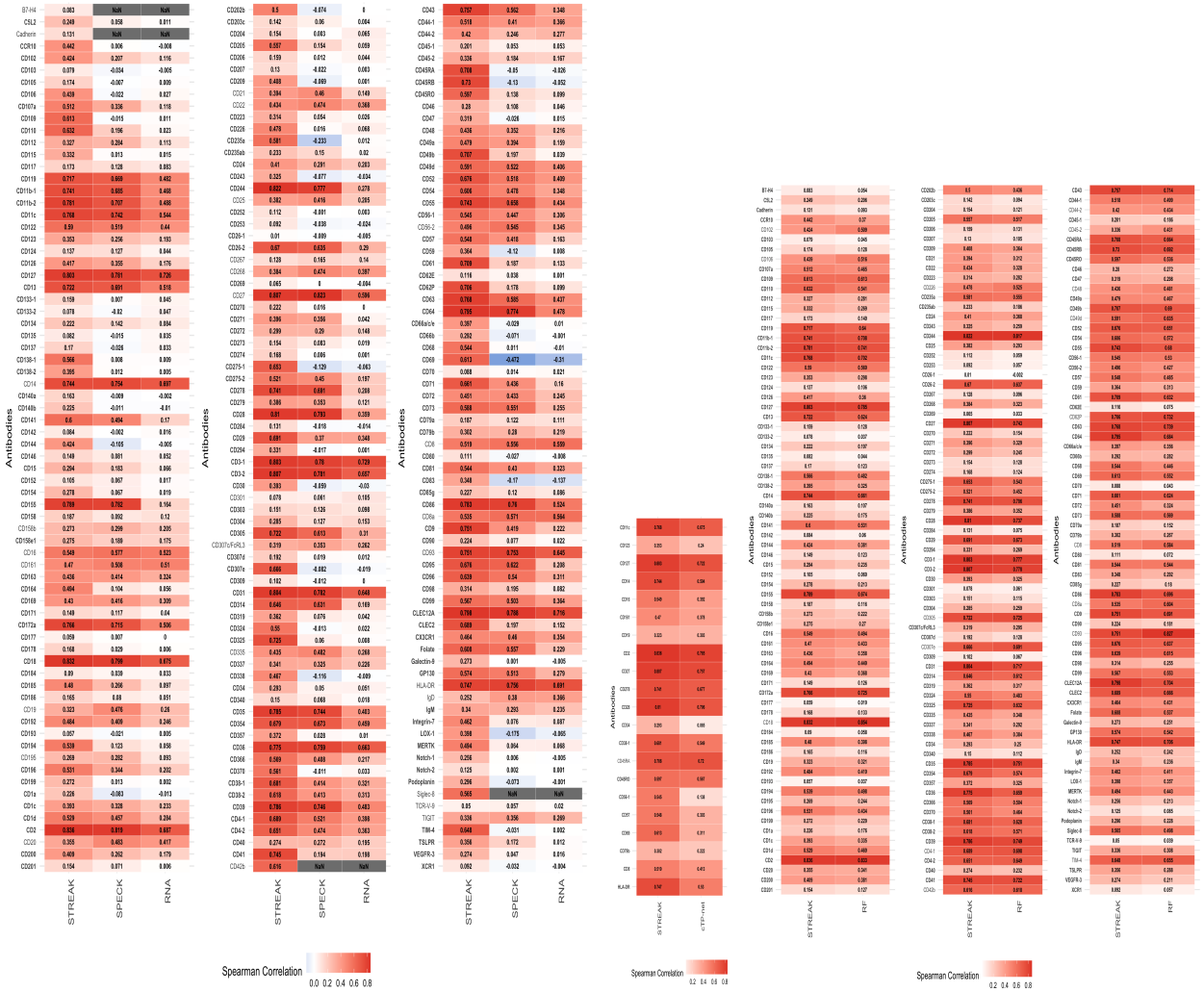
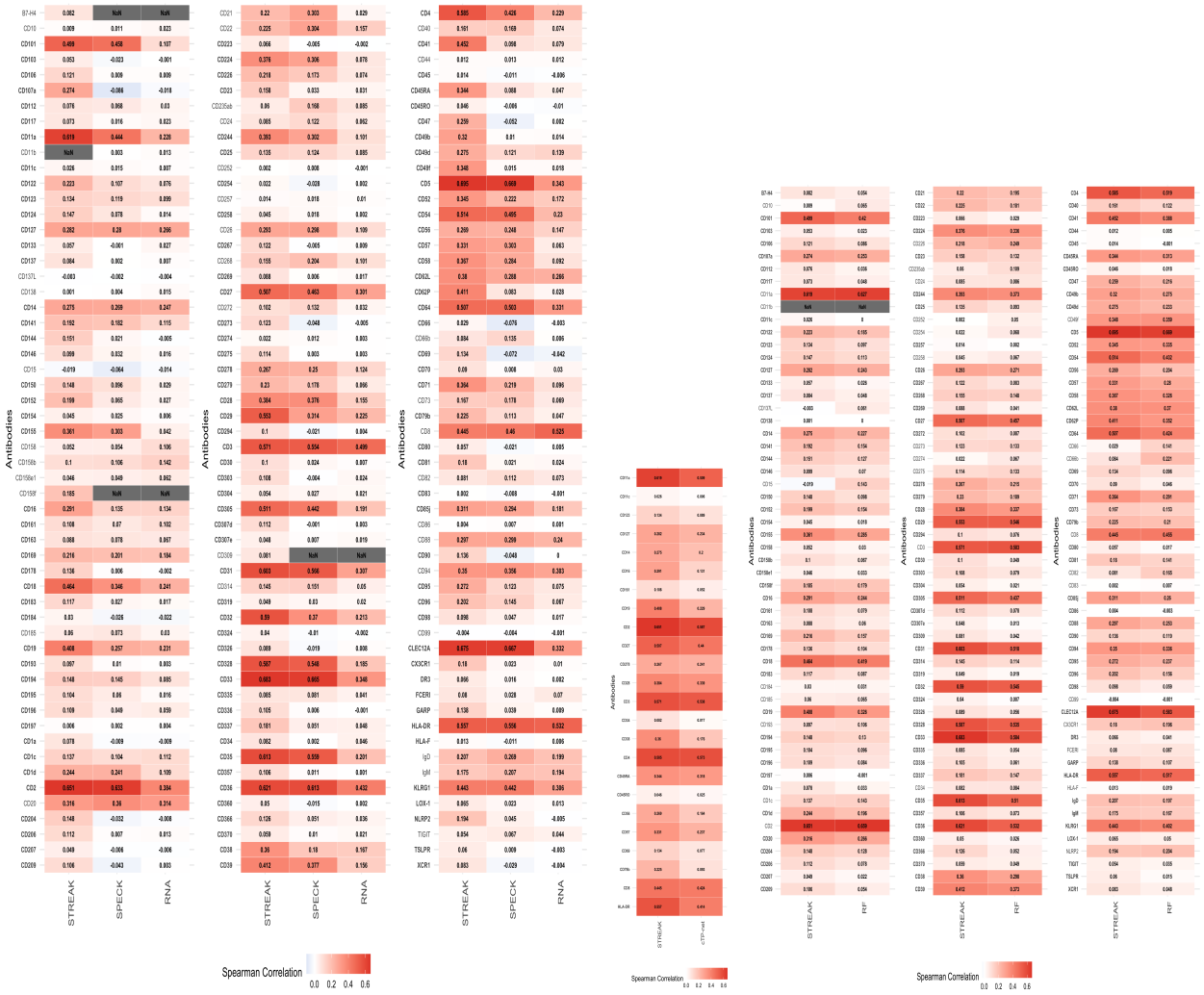


Figure 2: Frequency of receptors with highest average Spearman rank correlations between CITE-seq data and abundance profiles estimated using STREAK, SPECK or RNA transcript for the 5-fold cross-validation approach with training data ranging from 1000 to 12,000 cells for the Hao data (a), 1000 to 10,000 cells for the Unterman data (b) and 1000 to 1,682 cells for the MALT data (c) and 5,000 cells from the Hao data for the cross-training evaluation approach (j), receptors with highest average rank correlations between CITE-seq data and abundance profiles estimated using STREAK or estimated using cTP-net for the 5-fold cross-validation approach for the Hao data (d), Unterman data (e) and MALT data (f) and for the cross-training evaluation approach (k) and receptors with highest average rank correlations between CITE-seq data and abundance profiles estimated using STREAK or estimated using RF for the 5-fold cross-validation approach for the Hao data (g), Unterman data (h) and MALT data (i) and for the cross-training evaluation approach (l). The horizontal axis for the 5-fold cross-validation evaluation plots (a-i) indicates the number of cells used for training while the horizontal axis for the cross-training evaluation plots (j-l) indicates the number of cells from the Unterman data used for testing.



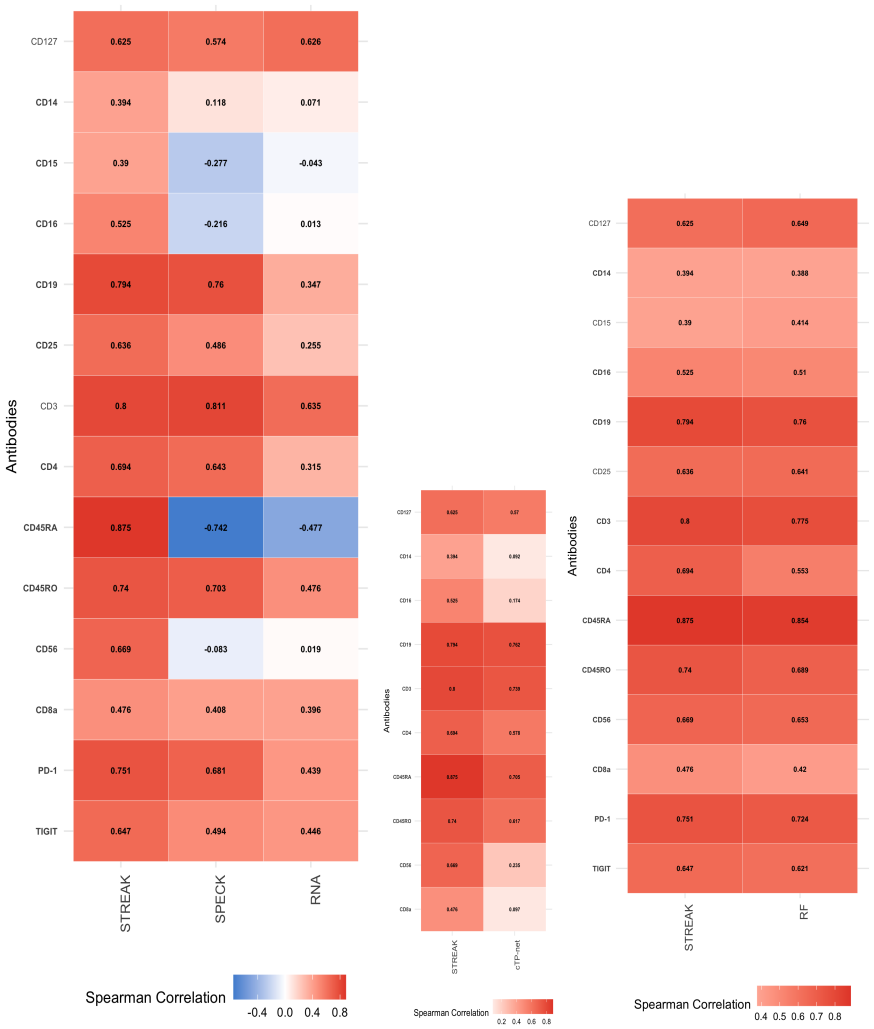
(a) (b) (c)

Figure 3: Average rank correlations between CITE-seq data and receptor abundance values estimated using STREAK, SPECK and RNA transcript (a), between CITE-seq data and abundance estimates generated using STREAK and estimated with RF (b) and between CITE-seq data and abundance generated using STREAK and estimated with cTP-net (c) evaluated using the 5-fold cross-validation approach for training data consisting of 12,000 cells from the Hao data. Bold text format is used to indicate receptors for which the STREAK estimate has highest correlation with corresponding ADT data compared to SPECK and RNA or RF or cTP-net. Limits of the gradient color scale are determined by the minimum and maximum average correlation values for all comparative methods combined.



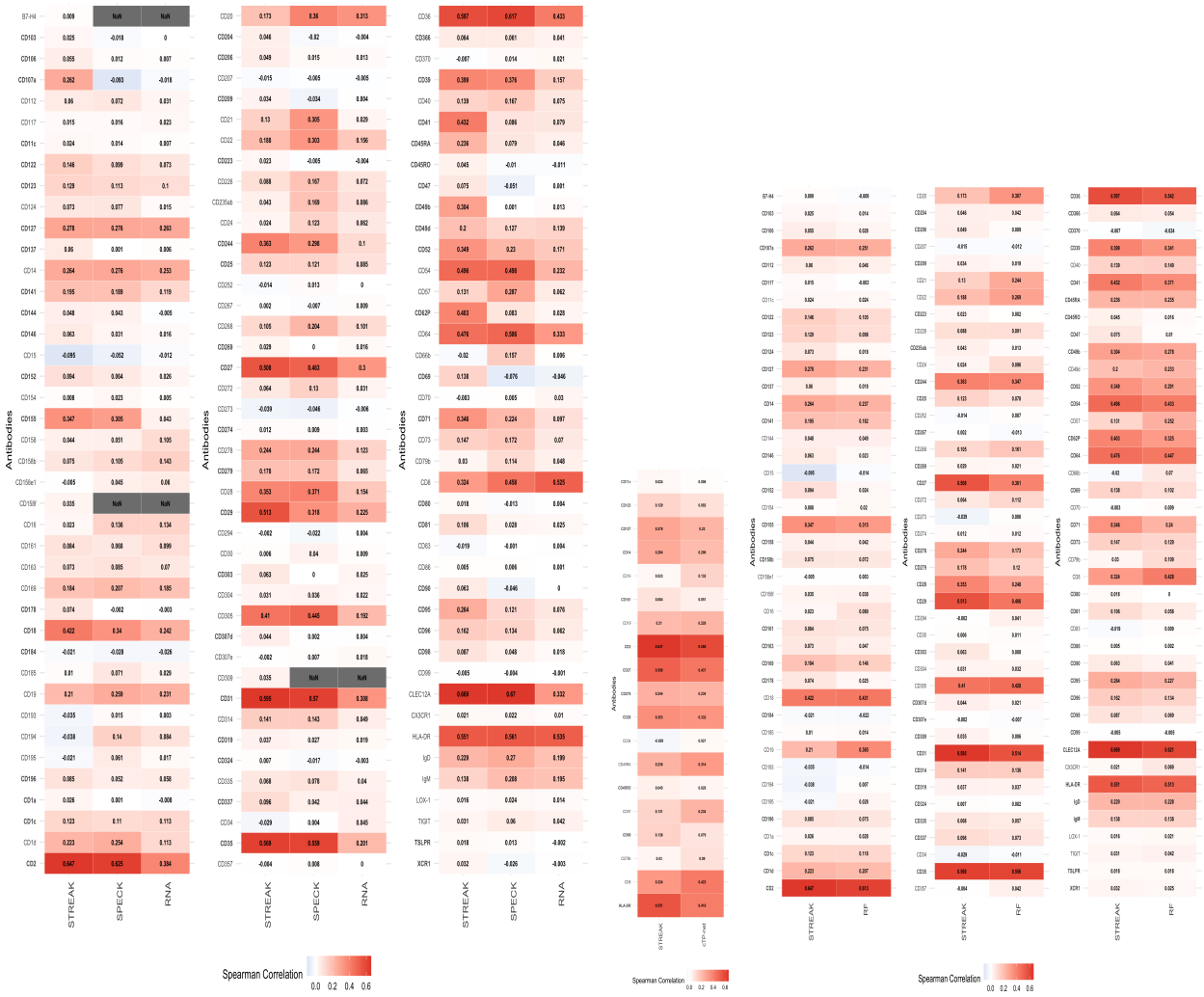
(a) (b) (c)

Figure 4: Average rank correlations between CITE-seq data and abundance values estimated using STREAK, SPECK and RNA transcript (a), between CITE-seq data and abundance estimates generated using STREAK and estimated with cTP-net (b) and between CITE-seq data and abundance values estimated using STREAK and estimated with RF (c) evaluated using the 5-fold cross-validation approach for training data consisting of 10,000 cells from the Unterman data.



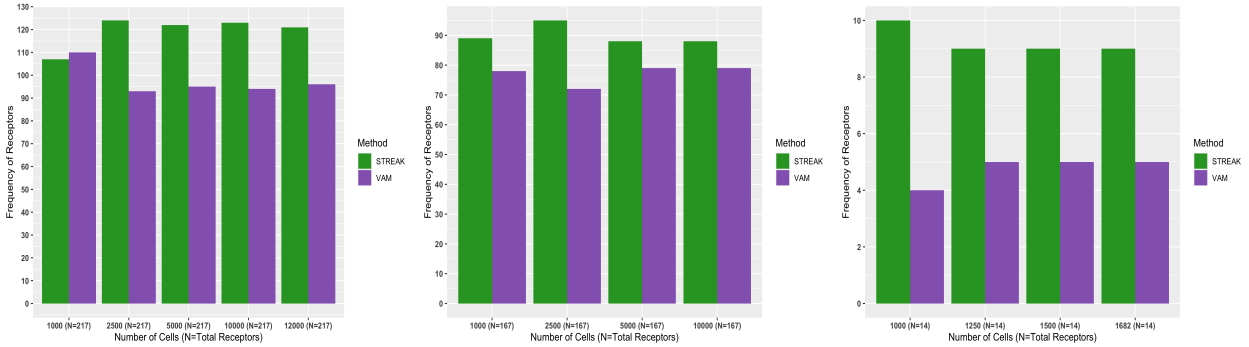
(a) (b) (c)

Figure 5: Average rank correlations between CITE-seq data and abundance values estimated using STREAK, SPECK and the normalized RNA approach (a), between data and abundance estimates generated using STREAK and estimated with cTP-net (b) and between CITE-seq data and abundance profiles estimated using STREAK and estimated with RF (c) evaluated using the 5-fold cross-validation approach for training data consisting of 1,682 cells from the MALT dataset.



(a) (b) (c)

Figure 6: Average rank correlations between CITE-seq data and abundance values estimated using STREAK, SPECK and the normalized RNA approach (a), between CITE-seq data and abundance profiles estimated using STREAK and estimated with cTP-net (b) and between CITE-seq data and abundance profiles estimated using STREAK and estimated with RF (c) evaluated using the cross-training approach trained on a subset of 5,000 cells from the Hao data and evaluated on a subset of 50,000 cells from the Unterman data.



(a) (b) (c)

Figure 7: Gene set scoring versus thresholding sensitivity analysis examining frequency of receptors with highest average rank correlations between CITE-seq data and abundance values estimated using STREAK (i.e., estimation via gene set scoring followed by thresholding) or VAM (i.e., estimation using just gene set scoring) evaluated using the 5-fold cross-validation approach with the indicated training data ranging from 1000 to 12,000 cells for the Hao data (a) 1000 to 10,000 cells for the Unterman data (b) and 1000 to 1,682 cells for the MALT data (c).

is higher for abundance profiles estimated using the normalized RNA approach (ρ=0.105) as compared to estimates produced using STREAK (ρ=0.078) (see Figure 3a). Similarly, when assessing STREAK against supervised abundance estimation techniques, the cTP-net-estimated profile for the CD45RA receptor from the Hao data is more correlated with the corresponding CITE-seq values (ρ=0.720) than the STREAK-estimated abundance profile (ρ=0.708) (see Figure 3b). These results suggest that SPECK may be a more suitable abundance estimation strategy for the CD301 receptor while cTP-net may be more optimal for the abundance estimation of CD45RA as compared to STREAK.

3.3 Gene set scoring versus thresholding and influence of gene set and training data size

Our next analysis evaluated the distinctive contributions of the gene set scoring and the thresholding steps on STREAK's performance. For this purpose, we compared the gene set scores (i.e., the CDF values) generated using VAM with the thresholded gene set scores generated using STREAK. Figures 7a, 7b and 7c indicate the proportion of receptors that have high rank correlations with analogous CITE-seq data that are estimated using STREAK and estimated using VAM for the 5-fold cross-validation evaluation strategy and the Hao, Unterman and MALT datasets, respectively. Overall, a greater percentage of receptors estimated with STREAK have higher rank correlations with CITE-seq data for all datasets and every training data size, aside from the 1,000 cells training category of the Hao data. This result emphasizes the important contribution of thresholding in further improving the correspondence of the estimated receptor abundance profiles with the analogous CITE-seq data over just VAM-generatd gene set scores.

We then performed sensitivity analysis to examine the influence of training data size on STREAK's performance for the cross-training evaluation approach. To accomplish this, we evaluated STREAK against SPECK, the normalized RNA approach, RF and cTP-net using co-expression associations built from joint transcriptomics and proteomics data as learned from separate individual subsets of the Hao dataset consisting of 5,000, 7,000, 10,000, 20,000 and 30,000 cells. Each set of trained associations were evaluated on five subsets of 5,000, 7,000 and 10,000 cells and a subset of 50,000 cells from the Unterman dataset. Figures 8a, 8b, 8c, 8d and 8e show proportion of receptors that have high rank correlations with CITE-seq data that are estimated using STREAK versus estimated using SPECK and the normalized RNA approach. Figures 8f, 8g, 8h, 8i and 8j indicate the analogous proportion of receptors that have higher rank correlations with CITE-seq data when estimated using STREAK versus cTP-net while Figures 8k, 8l, 8m, 8n and 8o visualize the equivalent results for the RF model trained on co-expression associations learned from subsets of 5,000, 7,000, 10,000, 20,000 and 30,000 cells, respectively. As observed, a higher proportion of receptors with abundance profiles estimated using STREAK are more correlated with corresponding CITE-seq data compared to abundance profiles estimated using comparative unsupervised estimation methods such as SPECK and the normalized RNA approach and supervised techniques such as cTP-net and RF. This result is moreover consistent across all training cell subsets, thereby underscoring STREAK's robustness to training data size.

Our subsequent analysis assessed sensitivity to gene set size for STREAK. We performed this analysis for the Hao dataset and the 5-fold cross-validation evaluation strategy using trained co-expression associations from the top 5, 10, 15, 20, 25 and 30 most correlated scRNA-seq transcripts with CITE-seq data. Figure 9a quantifies proportion

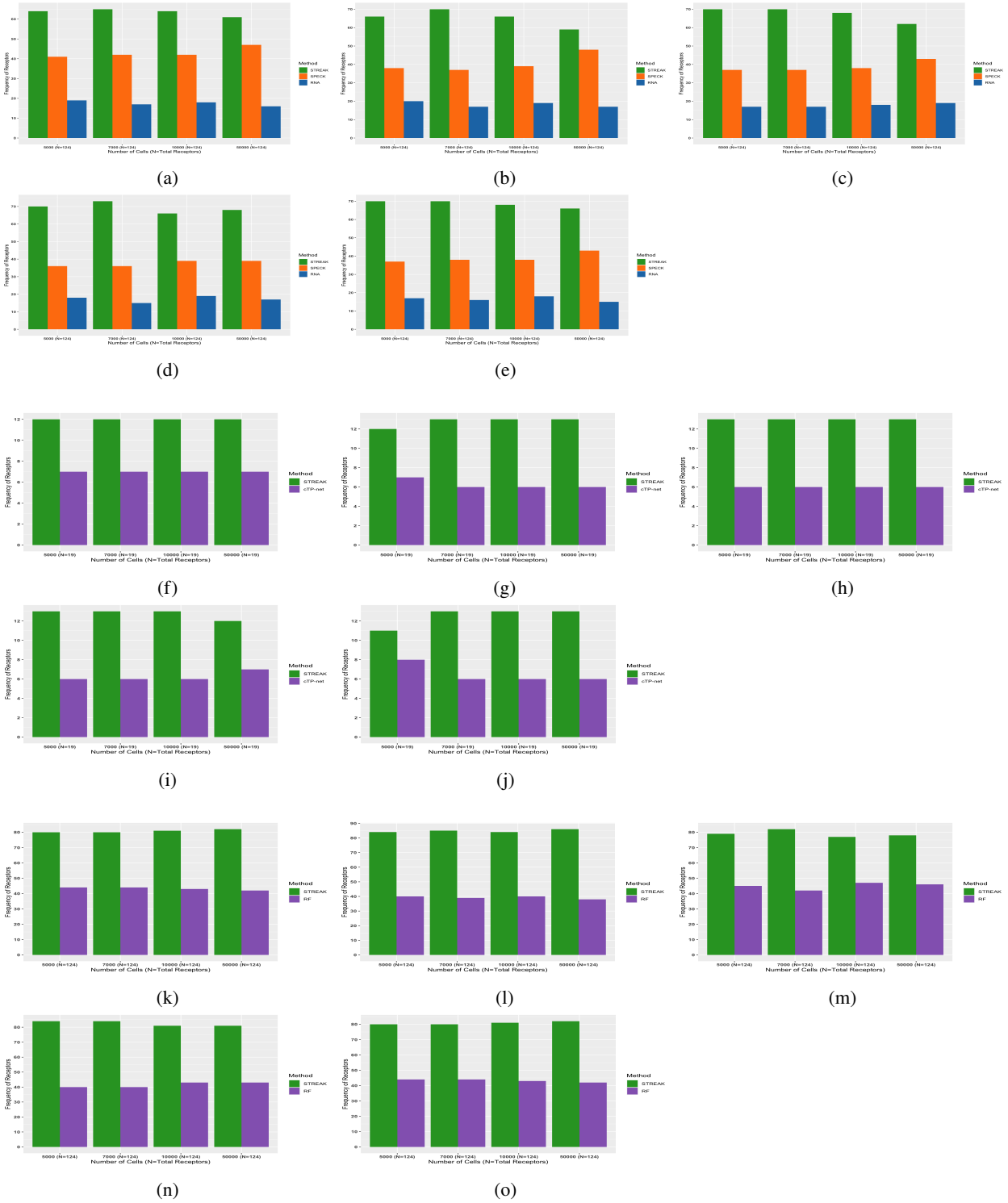


Figure 8: Training data sensitivity analysis examining frequency of receptors with highest average rank correlations between CITE-seq data and abundance values estimated using STREAK, SPECK or the normalized RNA approach evaluated using the cross-training approach with Hao training data consisting of 5,000 (a), 7,000 (b), 10,000 (c), 20,000 (d) and 30,000 (e) cells, between CITE-seq data and abundance profiles estimated using STREAK or estimated with cTP-net using the detailed cross-training approach with training data analogously consisting of 5,000 (f), 7,000 (g), 10,000 (h), 20,000 (i) and 30,000 (j) cells and between CITE-seq data and abundance profiles estimated using STREAK or estimated with RF using training data comprised of 5,000 (k), 7,000 (l), 10,000 (m), 20,000 (n) and 30,000 (o) cells. The horizontal axis for each plot indicates the number of test cells used from the Unterman data.

of receptors that have higher rank correlations with CITE-seq data and are estimated using STREAK, SPECK or the normalized RNA approach, Figure 9b compares proportion of receptors that are more correlated with corresponding CITE-seq data when estimated using STREAK or estimated with cTP-net while Figure 9c compares this proportion of receptors that are more highly correlated with corresponding CITE-seq data when estimated using STREAK or estimated with RF. The three plots emphasize that STREAK is mostly insensitive to gene set size and consistently performs better than comparative evaluation strategies using gene set size ranging from 5 to 30 genes.

4 Discussion

In this paper, we detail a novel supervised receptor abundance estimation method, STREAK, which functions by first learning associations between gene expression and protein abundance on joint scRNA-seq/CITE-seq training data and then performing gene set scoring and subsequent thresholding on target scRNA-seq data. We evaluated this method on three joint scRNA-seq/CITE-seq datasets representing two different tissue types via two evaluation strategies. The performance of STREAK was compared to unsupervised techniques such as SPECK and the normalized RNA approach and supervised methods such as cTP-net and the RF algorithm. This evaluation demonstrated that for the majority of the assessed receptors, the STREAK abundance estimates are more accurate than those produced by SPECK, the normalized RNA approach, cTP-net or RF.

A key strength of STREAK in comparison to neural network or ensemble-based supervised techniques is that the weighted gene sets used for cell-level estimation are simple to interpret and customize. Researchers can easily add or remove genes or adjust weights for empirically derived sets or define entirely new gene sets to reflect specific biological knowledge or better adapt to the expected pattern of gene expression in a given tissue type or environment.

One limitation of STREAK is that receptor gene set generation will typically require access to joint scRNA-seq/CITE-seq training data that is ideally measured on the same tissue type (and under similar biological conditions) as the target scRNA-seq data. While this limitation can be avoided if users feel comfortable manually defining the receptor gene sets, we anticipate that access to experimental training data will be necessary to generate accurate estimates for the majority of receptor proteins. A limitation of our proposed evaluation framework is that we do not compare estimated abundance values generated using STREAK against abundance values produced by RF and cTP-net on denoised target scRNA-seq data given the existing implementation challenges with the SAVER-X R package. We therefore could potentially be underestimating the performance of the RF and the cTP-net strategies in the context of receptor abundance estimation. In future, we plan to provide publicly trained gene sets and corresponding weights for the receptors estimated on the PBMC and the MALT datasets. An R package implementation of STREAK is currently in development and will be soon released on the Comprehensive R Archive Network (CRAN).

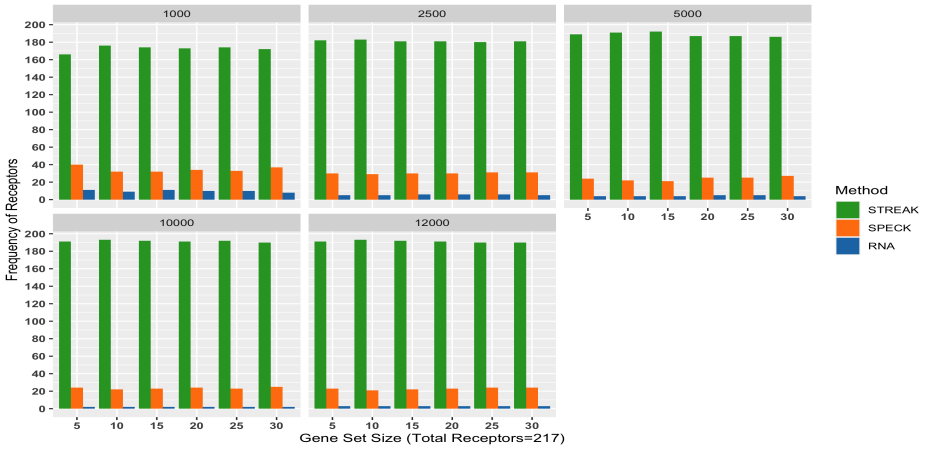
In short, we outline a new supervised receptor abundance estimation method that leverages joint associations between transcriptomics and proteomics data to generate abundance estimates via cell-level gene set scoring. The STREAK method produces more accurate abundance estimates relative to other unsupervised and supervised techniques with the potential to significantly improve the performance of single cell analysis tasks such as cell typing/phenotyping and cell-cell signaling estimation.

5 Competing interests

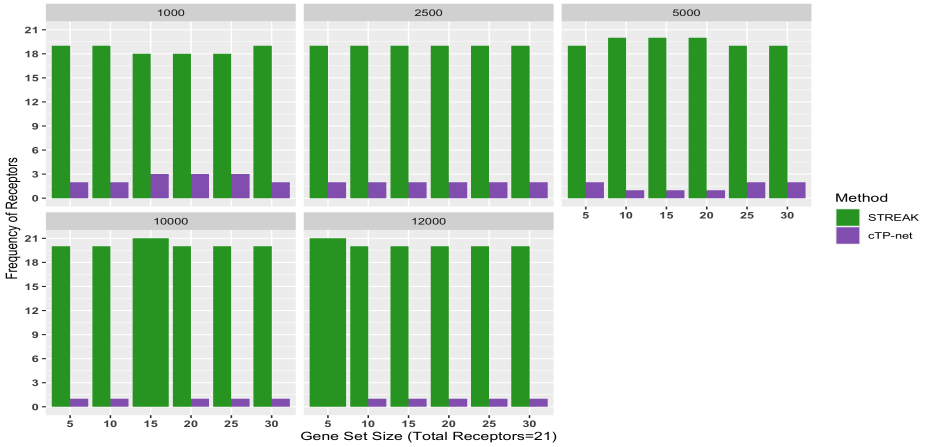
No competing interest is declared.

6 Acknowledgments

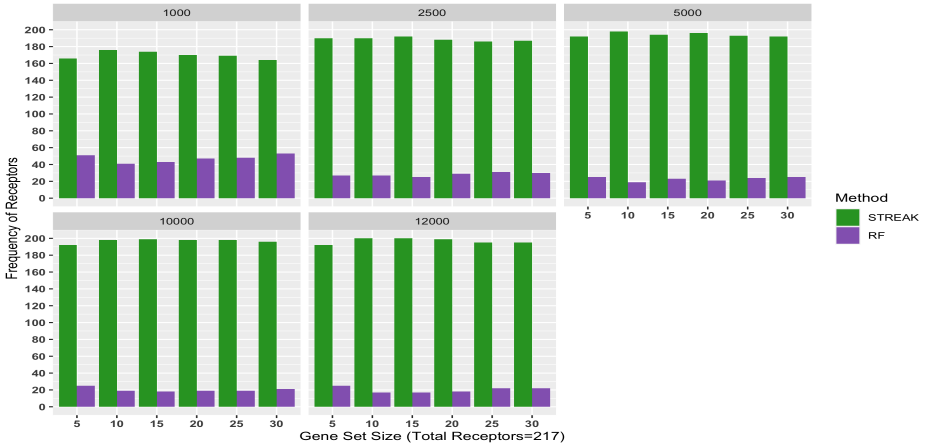
This work was funded by National Institutes of Health grants R35GM146586, R21CA253408, P20GM130454 and P30CA023108. We would like to acknowledge the supportive environment at the Geisel School of Medicine at Dartmouth where this research was performed.



(a)



(b)



(c)

Figure 9: Gene set size sensitivity analysis examining frequency of receptors with highest average rank correlations between CITE-seq data and abundance values estimated using STREAK, SPECK or the normalized RNA approach evaluated using the 5-fold cross-validation approach with the indicated training data ranging from 1000 to 12,000 cells for the Hao data and gene set consisting of 5, 10, 15, 20, 25 and 30 genes (a), between CITE-seq data and abundance profiles estimated with STREAK or estimated using cTP-net with analogous gene set size of 5, 10, 15, 20, 25 and 30 genes (b) and between CITE-seq data and abundance profiles estimated with STREAK or estimated using RF with gene set of 5, 10, 15, 20, 25 and 30 genes (c).

References

[1] G. X. Y. Zheng *et al.*, "Massively parallel digital transcriptional profiling of single cells", en, *Nature Communications*, vol. 8, no. 1, p. 14 049, Jan. 2017, Number: 1 Publisher: Nature Publishing Group. DOI: 10.1038/ncomms14049.

[2] F. Tang *et al.*, "mRNA-Seq whole-transcriptome analysis of a single cell", en, *Nature Methods*, vol. 6, no. 5, pp. 377–382, May 2009, Bandiera_abtest: a Cg_type: Nature Research Journals Number: 5 Primary_atype: Research Publisher: Nature Publishing Group. DOI: 10.1038/nmeth.1315.

[3] I. Tirosh *et al.*, "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq", en, *Science*, vol. 352, no. 6282, pp. 189–196, Apr. 2016, Publisher: American Association for the Advancement of Science Section: Research Article. DOI: 10.1126/science.aad0501.

[4] E. Armingol, A. Officer, O. Harismendy, and N. E. Lewis, "Deciphering cell–cell interactions and communication from gene expression", en, *Nature Reviews Genetics*, vol. 22, no. 2, pp. 71–88, Feb. 2021, Bandiera_abtest: a Cg_type: Nature Research Journals Number: 2 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Computational biology and bioinformatics;RNA sequencing;Systems biology Subject_term_id: computational-biology-and-bioinformatics;rna-sequencing;systems-biology. DOI: 10.1038/s41576-020-00292-x.

[5] A. Haque, J. Engel, S. A. Teichmann, and T. Lönnberg, "A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications", *Genome Medicine*, vol. 9, no. 1, p. 75, Aug. 2017. DOI: 10.1186/s13073-017-0467-4.

[6] W. A. Bonner, H. R. Hulett, R. G. Sweet, and L. A. Herzenberg, "Fluorescence activated cell sorting", *Review of Scientific Instruments*, vol. 43, no. 3, p. 404, 1972, https://doi.org/10.1063/1.1685647.

[7] M. Stoeckius *et al.*, "Simultaneous epitope and transcriptome measurement in single cells", en, *Nature Methods*, vol. 14, no. 9, pp. 865–868, Sep. 2017, Number: 9 Publisher: Nature Publishing Group. DOI: 10.1038/nmeth.4380.

[8] D. Lähnemann *et al.*, "Eleven grand challenges in single-cell data science", *Genome Biology*, vol. 21, no. 1, p. 31, Feb. 2020. DOI: 10.1186/s13059-020-1926-6.

[9] A. Javaid and H. R. Frost, "SPECK: An Unsupervised Learning Approach for Cell Surface Receptor Abundance Estimation for Single Cell RNA-Sequencing Data", en, Bioinformatics, preprint, Oct. 2022. DOI: 10.1101/2022.10.08.511197.

[10] Z. Zhou, C. Ye, J. Wang, and N. R. Zhang, "Surface protein imputation from single cell transcriptomes by deep neural networks", en, *Nature Communications*, vol. 11, no. 1, p. 651, Jan. 2020, Number: 1 Publisher: Nature Publishing Group. DOI: 10.1038/s41467-020-14391-0.

[11] X. Dai, F. Xu, S. Wang, P. A. Mundra, and J. Zheng, "PIKE-R2P: Protein–protein interaction network-based knowledge embedding with graph neural network for single-cell RNA to protein prediction", en, *BMC Bioinformatics*, vol. 22, no. Suppl 6, 2021, Publisher: BioMed Central. DOI: 10.1186/s12859-021-04022-w.

[12] F. Xu, S. Wang, X. Dai, P. A. Mundra, and J. Zheng, "Ensemble learning models that predict surface protein abundance from single-cell multimodal omics data", en, *Methods*, Machine learning for the analysis of multiomics data, vol. 189, pp. 65–73, May 2021. DOI: 10.1016/j.ymeth.2020.10.001.

[13] N. B. Erichson *et al.*, "Randomized matrix decompositions using r", *Journal of Statistical Software*, vol. 89, no. 11, pp. 1–48, 2019, https://www.jstatsoft.org/article/view/v089i11.

[14] Y. Hao *et al.*, "Integrated analysis of multimodal single-cell data", en, *Cell*, vol. 184, no. 13, 3573–3587.e29, Jun. 2021. DOI: 10.1016/j.cell.2021.04.048.

[15] T. Stuart *et al.*, "Comprehensive Integration of Single-Cell Data", en, *Cell*, vol. 177, no. 7, 1888–1902.e21, Jun. 2019. DOI: 10.1016/j.cell.2019.05.031.

[16] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, "Integrating single-cell transcriptomic data across different conditions, technologies, and species", *Nature Biotechnology*, vol. 36, pp. 411–420, 2018. DOI: 10.1038/nbt.4096.

[17] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev, "Spatial reconstruction of single-cell gene expression data", en, *Nature Biotechnology*, vol. 33, no. 5, pp. 495–502, May 2015, Number: 5 Publisher: Nature Publishing Group. DOI: 10.1038/nbt.3192.

[18] H. R. Frost, "Variance-adjusted Mahalanobis (VAM): A fast and accurate method for cell-specific gene set scoring", en, p. 20,

[19] C. Hafemeister and R. Satija, "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression", en, *Genome Biology*, vol. 20, no. 1, p. 296, Dec. 2019. DOI: 10.1186/s13059-019-1874-1.

[20] M. Song and H. Zhong, "Efficient weighted univariate clustering maps outstanding dysregulated genomic zones in human cancers", *Bioinformatics*, vol. 36, no. 20, pp. 5027–5036, Dec. 2020. DOI: 10.1093/bioinformatics/btaa613.

[21] H. Wang and M. Song, "Ckmeans.1d.dp: Optimal k-means Clustering in One Dimension by Dynamic Programming", en, *The R Journal*, vol. 3, no. 2, p. 29, 2011. DOI: 10.32614/RJ-2011-015.

[22] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository", *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, Jan. 2002. DOI: 10.1093/nar/30.1.207.

[23] A. Unterman *et al.*, "Single-cell multi-omics reveals dyssynchrony of the innate and adaptive immune system in progressive COVID-19", en, *Nature Communications*, vol. 13, no. 1, p. 440, Jan. 2022, Number: 1 Publisher: Nature Publishing Group. DOI: 10.1038/s41467-021-27716-4.

[24] *10k Cells from a MALT Tumor - Gene Expression with a Panel of TotalSeq™-B Antibodies*, en.

[25] M. L. Elizaga *et al.*, "Safety and tolerability of HIV-1 multiantigen pDNA vaccine given with IL-12 plasmid DNA via electroporation, boosted with a recombinant vesicular stomatitis virus HIV Gag vaccine in healthy volunteers in a randomized, controlled clinical trial", en, *PLOS ONE*, vol. 13, no. 9, e0202753, Sep. 2018, Publisher: Public Library of Science. DOI: 10.1371/journal.pone.0202753.

[26] S. S. Li *et al.*, "DNA Priming Increases Frequency of T-Cell Responses to a Vesicular Stomatitis Virus HIV Vaccine with Specific Enhancement of CD8+ T-Cell Responses by Interleukin-12 Plasmid DNA", *Clinical and Vaccine Immunology*, vol. 24, no. 11, e00263–17, Nov. 2017, Publisher: American Society for Microbiology. DOI: 10.1128/CVI.00263-17.

[27] *R: The R Project for Statistical Computing*.

[28] P. Schober, C. Boer, and L. A. Schwarte, "Correlation Coefficients: Appropriate Use and Interpretation", en-US, *Anesthesia & Analgesia*, vol. 126, no. 5, pp. 1763–1768, May 2018. DOI: 10.1213/ANE.0000000000002864.

[29] D. v. Dijk *et al.*, "Recovering Gene Interactions from Single-Cell Data Using Data Diffusion", English, *Cell*, vol. 174, no. 3, 716–729.e27, Jul. 2018, Publisher: Elsevier. DOI: 10.1016/j.cell.2018.05.061.

[30] G. C. Linderman *et al.*, "Zero-preserving imputation of single-cell RNA-seq data", en, *Nature Communications*, vol. 13, no. 1, p. 192, Jan. 2022, Number: 1 Publisher: Nature Publishing Group. DOI: 10.1038/s41467-021-27729-z.

[31] A. Javaid and H. R. Frost, *SPECK: Receptor Abundance Estimation using Reduced Rank Reconstruction and Clustered Thresholding*, https://CRAN.R-project.org/package=SPECK, 2022.

[32] A. Liaw and M. Wiener, "Classification and regression by randomforest", *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[33] S. Tweedie *et al.*, "Genenames.org: The HGNC and VGNC resources in 2021", *Nucleic Acids Research*, vol. 49, no. D1, pp. D939–D946, Jan. 2021. DOI: 10.1093/nar/gkaa980.